Volume 2, Número 1 - Abril 2010

linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 2, Número 1 – Abril 2010

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Editorial

Este é o terceiro número da ***Linguamática*** *e o primeiro de 2010, um número que termina o percurso da revista ao longo de um ano. Trata-se de uma edição especial com artigos seleccionados do* Sétimo Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana *(STIL'09), o que demonstra o interesse da nossa comunidade científica na* ***Linguamática****.*

Todos os artigos deste número especial são publicados na seccção dedicada aos Artigos de Investigação*. Agradecemos a colaboração dos autores seleccionados e dos organizadores do STIL na elaboração deste número da* ***Linguamática****.*

Finalmente, queremos marcar mais uma etapa na revista celebrando a indexação da ***Linguamática*** *em catálogos de bibliotecas digitais e em índices públicos de revistas electrónicas, entre os quais salientamos o* Latindex — Sistema Regional de información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal —*, o* DOAJ — Directory of Open Access Journals —*, o* Google Scholar *e o* The Linguist List*.*

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Prólogo

Uma visão geral dos avanços no Simpósio de Tecnologia da Informação e Linguagem Humana

Esta edição especial da Linguamática contém uma seleção dos artigos apresentados no 7º Simpósio de Tecnologia da Informação e Linguagem Humana (STIL 2009), que ocorreu de 8 a 11 de setembro de 2009 na Universidade de São Paulo (campus São Carlos), Brasil (`http://www.nilc.icmc.usp.br/til/stil2009_English`*). O STIL*[1] *é o evento anual de Tecnologia da Linguagem apoiado pela Sociedade Brasileira de Computação (SBC) e pela Comissão Especial de Processamento de Linguagem Natural. Este evento tem um caráter multidisciplinar, abrangendo um amplo espectro de disciplinas relacionadas à Tecnologia da Linguagem Humana, tais como Lingüística, Ciências da Computação, Psicologia, Ciência da Informação, entre outros, e tem por objetivo reunir participantes acadêmicos e da indústria que atuam nessas áreas.*

Os tópicos de interesse anunciados no Call for Papers estiveram centrados em torno dos trabalhos em tecnologia da linguagem humana em geral realizados a partir de perspectivas tão diversas como Ciências da Computação, Lingüística e Ciência da Informação, incluindo entre outros a mineração de texto, processamento da linguagem escrita e falada, a terminologia, lexicologia e lexicografia, modelagem e gestão de conhecimento e geração de linguagem natural. Foram submetidos 60 artigos longos e 26 curtos. Cada proposta foi analisada por três membros do Comitê de Programa, composto por 88 pesquisadores de 13 países e 45 instituições.

Após um rigoroso processo de revisão 18 artigos completos e 12 curtos foram selecionados, com taxas de aceitação de 30% e 42%, respectivamente. Os autores dos artigos completos foram convidados a submeter versões estendidas e revisadas dos seus trabalhos para esta edição especial, passando por um novo processo de revisão, desta vez pelos revisores da Linguamática, que selecionaram 7 dos artigos submetidos.

Estes artigos representam uma amostra do rico e variado trabalho apresentado no STIL e envolvem pesquisadores de instituições acadêmicas e industriais no Brasil, Portugal e França. Por exemplo, o primeiro artigo, Identificação de expressões multipalavra em domínios específicos *de Aline Villavicencio et al., propõe uma abordagem para a identificação de Expressões Multipalavra, tais como compostos nominais e verbos frasais, em corpora técnicos. A proposta apresentada combina medidas de associação com informações linguísticas e de alinhamentos lexicais, e o artigo examina a influência de diversos fatores sobre o seu desempenho.*

Os dois próximos artigos são relacionados a aplicações de PLN. Em Classificação

[1] Este evento era anteriormente conhecido como TIL (Workshop de Informação e Tecnologia da Linguagem Humana).

automática de textos por período literário utilizando compressão de dados através do PPM-C, *Bruno Barufaldi et al. propõem a aplicação do método Prediction by Partial Matching (PPM) para a tarefa de classificação de textos de acordo com períodos literários da literatura brasileira. Já Carolina Scarton e Sandra Aluísio, em* Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do Coh-Metrix para o Português, *investigam a adaptação de métricas da ferramenta Coh-Metrix para o português do Brasil (Coh-Metrix-Port), primeiramente avaliando as diferenças entre textos complexos para adultos e versões mais simples para crianças e também analisando o desempenho de classificadores para discriminar textos dedicados a adultos e a crianças, que podem ser usados para avaliar a simplicidade de textos disponíveis na Web.*

O quarto artigo Caracterização e processamento de expressões temporais em português *de Caroline Hagège, Jorge Baptista e Nuno Mamede também aborda a questão do tratamento de expressões, mas desta vez o foco é em expressões temporais tais como* de manhã *e* nesta semana. *Os autores propõem uma classificação para estas expressões do português e apresentam uma ferramenta de anotação delas em corpora.*

Quanto a construção de recursos linguísticos para o português, o artigo Extração de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação *de Hugo Oliveira, Diana Santos e Paulo Gomes apresenta o PAPEL, um recurso lexical que contém relações entre palavras, como sinonímia, automaticamente extraídas de um dicionário através de regras, discutindo ainda uma avaliação do mesmo.*

Outra tarefa abordada neste volume é a de sumarização, no artigo Estratégias de seleção de conteúdo com base na CST (Cross-document Structure Theory) para sumarização automática multidocumento *de Maria Jorge e Thiago Pardo. Os autores discutem a definição, formalização e avaliação de estratégias de seleção de conteúdo para sumarização automática multidocumento com base na teoria discursiva Cross-document Structure Theory.*

Por fim a tarefa de entendimento e linguagem natural é abordada no artigo Um analisador semântico inferencialista de sentenças em linguagem natural *de Vladia Pinheiro et al, onde é descrito o Analisador Semântico Inferencialista (SIA), um raciocinador semântico sobre o conteúdo inferencial de conceitos e padrões de sentenças, avaliado em um sistema de extração de informações sobre crimes.*

Nossos agradecimentos para os editores da Linguamática e revisores dos artigos tanto da Linguamática quanto do STIL 2009.

Aline Villavicencio
Horácio Saggion
Maria das Graças Volpe Nunes
Thiago Pardo

Comissão Científica

Alberto Álvarez Lugrís, Universidade de Vigo
Alberto Simões, Universidade do Minho
Aline Villavicencio, Universidade Federal do Rio Grande do Sul
Álvaro Iriarte Sanroman, Universidade do Minho
Ana Frankenberg-Garcia, ISLA e Universidade Nova de Lisboa
Anselmo Peñas, Universidad Nacional de Educación a Distancia
Antón Santamarina, Universidade de Santiago de Compostela
António Teixeira, Universidade de Aveiro
Belinda Maia, Universidade do Porto
Carmen García Mateo, Universidade de Vigo
Diana Santos, SINTEF ICT
Ferran Pla, Universitat Politècnica de València
Gael Harry Dias, Universidade Beira Interior
Gerardo Sierra, Universidad Nacional Autónoma de México
German Rigau, Euskal Herriko Unibertsitatea
Helena de Medeiros Caseli, Universidade Federal de São Carlos
Horacio Saggion, University of Sheffield
Iñaki Alegria, Euskal Herriko Unibertsitatea
Joaquim Llisterri, Universitat Autònoma de Barcelona
José Carlos Medeiros, Porto Editora
José João Almeida, Universidade do Minho
José Paulo Leal, Universidade do Porto
Joseba Abaitua, Universidad de Deusto
Lluís Padró, Universitat Politècnica de Catalunya
Maria Antònia Martí Antonín, Universitat de Barcelona
Maria das Graças Volpe Nunes, Universidade de São Paulo
Mercè Lorente Casafont, Universitat Pompeu Fabra
Mikel Forcada, Universitat d'Alacant
Nieves R. Brisaboa, Universidade da Coruña
Pablo Gamallo Otero, Universidade de Santiago de Compostela
Salvador Climent Roca, Universitat Oberta de Catalunya
Susana Afonso Cavadas, University of Sheffield
Tony Berber Sardinha, Pontifícia Universidade Católica de São Paulo
Xavier Gómez Guinovart, Universidade de Vigo

Artigos de Investigação

Identificação de Expressões Multipalavra em Domínios Específicos

Aline Villavicencio[1,2], Carlos Ramisch[1,3], André Machado[1],
Helena de Medeiros Caseli[4], Maria José Finatto[5]

[1]Instituto de Informática, Universidade Federal do Rio Grande do Sul (Brasil)
[2]Department of Computer Sciences, Bath University (Inglaterra)
[3]GETALP – Laboratoire d'Informatique de Grenoble, Université de Grenoble (França)
[4]Departmento de Ciência da Computação, Universidade Federal de São Carlos (Brasil)
[5]Instituto de Letras, Universidade Federal do Rio Grande do Sul (Brasil)
{avillavicencio,ceramisch,ammachado}@inf.ufrgs.br,
helenacaseli@dc.ufscar.br, mfinatto@terra.com.br

Resumo

Expressões Multipalavra (EM) são um dos grandes obstáculos para a obtenção de sistemas mais precisos de Processamento de Linguagem Natural (PLN). A cobertura limitada de EM em recursos linguísticos pode impactar negativamente o desempenho de tarefas e aplicações de PLN e pode levar à perda de informação ou a problemas de comunicação, especialmente em domínios técnicos, em que EM são particularmente frequentes. Este trabalho investiga algumas abordagens para a identificação de EM em corpora técnicos com base em medidas de associação, informações morfossintáticas e de alinhamento lexical. Primeiramente, examina-se a influência de alguns fatores sobre o seu desempenho, tais como fontes de informação para a identificação e avaliação. Se, por um lado, as medidas de associação enfatizam revocação, por outro, o método de alinhamento centra-se em precisão. Neste trabalho, propõe-se uma abordagem combinada que une os pontos fortes das diferentes abordagens e fontes de informação utilizando um algoritmo de aprendizado de máquina para produzir resultados mais robustos e precisos. A avaliação automática dos resultados mostra que o desempenho do método combinado é superior aos resultados individuais das abordagens associativa e baseada em alinhamento para a extração de EM de português e inglês. Além disso, é discutida a efetividade de cada um desses métodos para a identificação de EM específicas em comparação com EM de domínio genérico. O método proposto pode ser usado para auxiliar o trabalho lexicográfico, fornecendo uma lista de candidatos a EM.

1 Introdução

A cobertura dos recursos lexicais tem um impacto significativo sobre o desempenho de muitas tarefas e aplicações de Processamento de Linguagem Natural (PLN), e nesse sentido, muitas pesquisas têm se dedicado à proposição de métodos para automatizar a aquisição lexical. Nos últimos anos, alguns desses trabalhos têm se centrado em um conjunto de fenômenos para os quais recursos lexicais são particularmente carentes de cobertura, entre os quais destacam-se as *Expressões Multipalavra* (EM) (Baldwin, 2005; Villavicencio et al., 2007).

Essas expressões podem ser definidas como combinações de palavras que apresentam idiossincrasias lexicais, sintáticas, semânticas, pragmáticas ou estatísticas (Sag et al., 2002), e incluem, entre outros fenômenos, verbos frasais (*carry up, consist of*), verbos de suporte (*tomar um banho, dar uma caminhada*), compostos (*carro de polícia, bode expiatório*) e expressões idiomáticas (*engolir o sapo, nadar contra a corrente*). EM são muito numerosas dentro de uma língua e, segundo Biber et al. (1999), podem corresponder de 30% a 45% do inglês falado e 21% da linguagem acadêmica. De acordo com Jackendoff (1997), as EM têm a mesma ordem de magnitude, no léxico de um falante nativo, do número de palavras simples. No entanto, essas proporções são provavelmente subestimadas se considerarmos a linguagem de um domínio específico na qual: (i) o vocabulário especializado e a terminologia especializada vão ser compostos, na sua maior parte, por EM (*aquecimento global, sequenciamento de proteínas, litíase renal crônica*) e (ii) que novas EM estão sendo constantemente introduzidas na linguagem (*melhoramento genético, gripe suína).*

Os problemas causados pela cobertura limitada dos recursos lexicais podem ser ilustrados, por exemplo, no contexto de um analisador sintático. Em uma amostra aleatória de 20.000 sentenças do *British National Corpus* (Burnard, 2007), a baixa cobertura de EM no léxico utilizado resultou em 8% dos erros cometidos pelo analisador sintático (Baldwin et al., 2004), mesmo com uma gramática de ampla cober-

tura como a *English Resource Grammar* (Copestake e Flickinger, 2000).

Portanto, EM devem ser identificadas e tratadas adequadamente, pois, do contrário, a qualidade dos sistemas pode ser seriamente deteriorada, especialmente para tarefas de PLN que envolvam algum tipo de processamento semântico (Sag et al., 2002). Para tanto, acredita-se que métodos (semi-)automáticos robustos para a aquisição de informações lexicais sobre EM possam aumentar a cobertura dos recursos lexicais. Por exemplo, o número de construções verbo-partícula listadas em um dicionário, como o *Alvey Natural Language Tools* (Carroll e Grover, 1989), pode ser significativamente aumentado através da adição de construções verbo-partícula automaticamente extraídas de um corpus, como o *British National Corpus* (Baldwin, 2005).

Neste trabalho, são investigadas algumas abordagens para a identificação de EM a partir de *corpora técnicos*. Uma avaliação detalhada do desempenho destas abordagens é realizada, examinando-se o impacto das fontes de informação utilizadas. A mesma inclui uma comparação dos resultados obtidos para um domínio específico usando um corpus paralelo inglês–português (en–pt) composto por artigos científicos de uma revista brasileira bilíngue de Pediatria. O propóstito é verificar de que forma uma segunda língua pode fornecer pistas relevantes para a identificação de EM em português. São também discutidos alguns aspectos que influenciam uma avaliação mais profunda dos resultados, tais como a proporção de termos específicos e genéricos nas listas de referência, a filtragem dos candidatos e o número de palavras de cada *n*-grama.

Após avaliar as abordagens associativa e baseada em alinhamento separadamente em trabalhos anteriores (Caseli et al., 2009a; Villavicencio, Caseli e Machado, 2009), neste trabalho investiga-se sua combinação ponderada a fim de propor-se um método mais robusto que resulte em um conjunto mais preciso de EM candidatas do que as dos métodos individuais. A abordagem proposta pode ser utilizada para: a) auxiliar o trabalho de produção de dicionários especializados, quer sejam repertórios de termos ou de fraseologismos, fornecendo uma lista de EM candidatas para manter os recursos lexicais atualizados; e, b) também para a melhoria da qualidade dos sistemas de PLN, que poderiam vir a integrar listas de EM verificadas manualmente ou listas de candidatas a EM extraídas de forma totalmente automática.

O restante deste artigo está estruturado da seguinte forma. A seção 2 apresenta uma visão geral sobre EM e sobre alguns trabalhos relacionados que tratam da sua extração automática. A seção 3 descreve os recursos utilizados nos experimentos, enquanto a seção 4 descreve os métodos propostos para extrair EM. A seção 5 apresenta a metodologia de avaliação e de análise dos resultados. A seção 6 encerra este artigo com as conclusões e com algumas perspectivas de trabalhos futuros.

2 *Expressões Multipalavra: Problemas e Soluções para Identificação*

O termo Expressão Multipalavra vem sendo utilizado para descrever um grande número de construções distintas, mas fortemente relacionadas, tais como verbos de suporte (*fazer uma demonstração, dar uma palestra*), compostos nominais (*quartel general*), frases institucionalizadas (*pão e manteiga*), e muitos outros. Sag et al. (2002) definem EM como *interpretações idiossincráticas que cruzam os limites (ou espaços) entre as palavras*. Esses autores tratam da diferença que existe entre a interpretação de uma EM (por exemplo, *bode expiatório*) como um todo e os significados isolados das palavras individuais que a compõem (*bode* e *expiatório*). Os mesmos consideram que a definição de EM engloba um grande número de construções, tais como expressões fixas, compostos nominais e construções verbo-partícula. Ainda nessa linha, Calzolari et al. (2002) definem EM como *uma sequência de palavras que atua como uma única unidade, em algum nível de análise linguística*, a qual exibe algumas das seguintes características:

- transparência sintática e/ou semântica reduzida;
- composicionalidade reduzida;
- flexibilidade sintática reduzida;
- violação de regras sintáticas gerais;
- elevado grau de lexicalização;
- elevado grau de convencionalidade.

Para Moon (1998) *não há um fenômeno unificado que se possa descrever como EM, mas sim um complexo de atributos que interagem de formas diversas, muitas vezes desordenadas, e que representam um amplo contínuo entre o não-composicional (ou idiomático) e grupos composicionais de palavras*. Outros autores utilizam a noção de frequência e definem EM como sequências ou grupos de palavras que co-ocorrem com mais frequência do que seria esperado por acaso, e podem ultrapassar fronteiras sintagmáticas (Evert e Krenn, 2005). Isso incluiria também fórmulas de saudação como *Tudo bem? Como você vai?* e sequências lexicais, como *eu não sei se*. Santos (2008) aborda a questão de EM em relação a uma aplicação em particular, a tradução automática, e os desafios causados por expressões multipalavra ou expressões complexas, que envolvem tanto casos de traduções de uma palavra em muitas

(exemplo *miss* como *sentir a falta*), de muitas palavras traduzidas em uma (exemplo *get up early* como *madrugar*) e de sequências de palavras traduzidas como sequências também (exemplo *kick the bucket* como *bater as botas*). Por conseguinte, autores diferem nas definições que usam para EM em função dos aspectos particulares que estão sendo enfatizados e dos grupos de palavras e construções que consideram como EM.

As EM são muito frequentes na linguagem corrente e isso se reflete em várias gramáticas e recursos lexicais existentes, em que quase metade das entradas são dedicadas a EM. No entanto, devido às suas características heterogêneas, EM apresentam grandes desafios tanto sob o ponto de vista linguístico quanto computacional (Sag et al., 2002). Primeiramente, algumas EM são fixas, e não apresentam variação interna, como *ad hoc*, enquanto outras permitem diferentes graus de variabilidade interna e modificação, como *levar chumbo/ferro/pau* e *ir/descer/perder-se (por) água abaixo*. Em termos de semântica, algumas EM são mais opacas em seu significado como *lavar roupa suja* significando *discutir assunto particular geralmente conflituoso*, enquanto outras são mais transparentes, e seus significados podem ser inferidos a partir de seus componentes, tal como *carro de polícia*, em que o sintagma preposicional *de polícia* adiciona informação de função para a palavra *carro*.

No contexto de textos de um domínio específico, tem-se uma definição importante relacionada a EM, a de termo. Segundo Krieger e Finatto (2004), para especialistas de um domínio, termos são uma representação do conhecimento da área específica, ou seja, as terminologias contêm unidades lexicais que expressam conceitos abstratos ou mesmo elementos concretos de um domínio. Existem várias diferenças entre EM genéricas e termos. Primeiramente, termos podem ser compostos por uma única palavra ou por múltiplas, como locuções nominais, enquanto EM são inerentemente compostas por duas ou mais palavras. Em segundo lugar, EM são um fenômeno que integra tanto à linguagem técnica e científica quanto à linguagem cotidiana de propósito geral, enquanto termos são tipicamente relacionados com a primeira. Além, disso, é preciso considerar que uma mesma terminologia, quando ocorre simultaneamente em textos de linguagem cotidiana e em textos científicos, tende a adquirir trações semânticos mais e menos específicos conforme o tipo de comunicação envolvida. Essas são diferenças importantes, pois será necessário determinar até que ponto os métodos computacionais disponíveis para lidar com EM em textos genéricos podem ser aplicados para lidar com corpora de domínio especializados e vice-versa. Por outro lado, EM e termos têm também aspectos comuns: ambos têm idiosincrasia semântica e ambos são um desafio para os sistemas de PLN (Ramisch, 2009).

Uma classificação de EM que permite agrupá-las em classes de dificuldade para métodos de identificação automáticos, é a proposta por Sag et al. (2002). Eles classificam as EM divididas em dois grandes grupos: expressões institucionalizadas e expressões lexicalizadas. Expressões institucionalizadas se caracterizam por serem sintaticamente e semanticamente composicionais, mas estatisticamente idiossincráticas se comparadas a qualquer outra alternativa do mesmo conceito (*café forte* x ?*café possante*).[1] Dentre essas EM convencionalizadas, ou seja, observadas com uma frequência muito maior do que qualquer outra formulação equivalente, as mais representativas são as colocações (*sal e pimenta*, *bagagem emocional*, etc.). De acordo com Smadja (1993), as colocações podem variar muito quanto ao seu comprimento, porém, elas geralmente contêm uma média de duas a cinco palavras. Além disso, algumas vezes a colocação envolve palavras não adjacentes em uma frase, e nesses casos a distância entre as partes que a compõem depende da sintaxe da língua, sendo potencialmente tão longa quanto se queira. Essa característica acarreta em dificuldades para a identificação de EM, à medida que não se pode saber *a priori* quais os limites das EM a serem extraídas automaticamente de textos. Algumas características relevantes das colocações são:

- não composicionalidade: o seu significado é constituído pela composição dos significados de suas partes, somando-se a isto um componente semântico adicional não previsível a partir das palavras isoladas que a compõem.
- não substituibilidade: não é possível substituir cada uma das suas componentes por palavras que possuam o mesmo significado que estas (*sal e pimenta* x ?*sal e malagueta*).
- não modificação: existem restrições a quanto às possibilidades de modificação sintática de uma colocação, que variam em grau de rigidez de expressão para expressão (?*bagagem vermelha emocional*).

As expressões lexicalizadas, por outro lado, compreendem EM que *possuem pelo menos sintaxe ou semântica parcialmente idiossincráticas, ou contêm palavras que não ocorrem isoladamente*, ou seja, são expressões que apresentam certa rigidez formal. Como tipos de expressões lexicalizadas têm-se as expressões fixas, expressões semi-fixas e expressões sintaticamente flexíveis.

[1] A notação usada neste artigo marca sentenças não-gramaticais com "*"e sentenças gramaticais possíveis mas não usuais para um falante nativo com "?".

- As *expressões fixas*, tais como *ad hoc*, *Porto Alegre*[2] são consideradas as mais rígidas de todas e se caracterizam por não apresentarem variações morfossintáticas e não permitirem modificações internas.

- *Expressões semi-fixas*, diferentemente das expressões fixas, permitem certo nível de variação lexical. Esta variação pode ser referente à flexão, à forma reflexiva e à escolha de determinantes. Dentre estas, tem-se as expressões idiomáticas não decomponíveis, que permitem variações apenas quanto à flexão e quanto à forma reflexiva mas que não apresentam variabilidade sintática, tais como modificações internas ou até mesmo a transformação para a voz passiva. Um exemplo é a expressão em inglês *kick the bucket*, que apesar de permitir a conjugação do verbo *kick* (*kicked the bucket*), não admite modificações feitas internamente (**kick the big bucket*). Outros casos são os compostos nominais que não permitem variabilidade sintática e são caracterizados por permitir flexões de número (*wine glass* (taça de vinho), *orange juice* (suco de laranja) e *guarda-chuva*) e os nomes próprios que são altamente idiossincráticas do ponto de vista sintático.

- As *expressões sintaticamente flexíveis*, ao contrário das expressões semi-fixas, apresentam uma variabilidade sintática mais ampla. EM desse tipo incluem construções verbo-partícula do inglês (*look up* e *break up*), expressões idiomáticas decomponíveis (*ser barra pesada* e *a barra pesou*), verbos de suporte (*tomar um banho, dar uma caminhada*, etc). Os componentes de EM desse tipo podem estar separados uns dos outros por aceitarem constituintes variáveis ou devido a variação na ordem sintática causada por fenômenos como passivização, topicalização, entre outros. Por exemplo, em alguns verbos frasais do inglês, o verbo pode estar separado da partícula por complementos de tamanhos não previsíveis, como *eat up* em *eat up the delicious and very expensive Belgian chocolate* x ?*eat the delicious and very expensive Belgian chocolate up*. De acordo com Riehemann (2001), este grau de flexibilidade varia de expressão para expressão e é geralmente imprevisível. Por exemplo, *spill the beans* and *kick the bucket* são duas expressões idiomáticas formadas por verbo transitivos e sintagma nominais mas que têm comportamentos bem diversos em termos de flexibilidade, com a primeira sendo sintaticamente flexível e a segunda semi-fixa.

[2]Cabe aqui salientar que, embora *Porto Alegre* seja um nome próprio e que esse tipo de EM possa receber tratamentos específicos, neste trabalho traz-se um enfoque propositalmente mais geral de diferentes tipos de EM. Além disso, a inclusão de nomes próprios como EM é aceita por alguns autores, e o critério adotado neste trabalho irá considerar também nomes próprios.

Em termos da identificação de EM, o grau de dificuldade da tarefa aumenta com o grau de flexibilidade da expressão. Consequentemente, muitos dos métodos tendem a se concentrar em capturar expressões fixas e semi-fixas, em particular, como discutido a seguir, pois essas quase não aceitam modificação na sintaxe e ocorrem sob a forma de palavras adjacentes. O desafio está em decidir os seus limites, e se há elementos variáveis, como determinantes alternativos (por exemplo, *engolir [um/o] sapo*). Desta forma, métodos baseados em *n*-gramas contíguos podem ser empregados para a sua identificação com bons resultados. Porém para as expressões flexíveis há ainda a dificuldade adicional de que a ordem dos seus componentes pode variar de diversas maneiras, e eles podem estar separados por um número imprevisível de palavras. Para este tipo de EM, a abordagem para identificação deve ser capaz de reconhecer combinações de palavras recorrentes mesmo se a ordem das mesmas muda, e se há elementos opcionais ou variáveis. Para lidar com esses casos, neste artigo é investigada a utilização do método baseado em alinhamento.

Neste trabalho, adota-se a definição de EM como combinações de palavras que apresentam idiossincrasias lexicais, sintáticas, semânticas ou estatísticas, que inclui entre outras construções expressões idiomáticas, verbos de suporte, compostos nominais, e nomes próprios, seguindo Sag et al. (2002). Esta definição abrangente é compatível com o uso de medidas estatísticas para a identificação de EM, pois elas são independentes de tipo. Desta forma, para se restringir a extração de EM a um tipo particular de expressões, filtros morfosintáticos podem ser aplicados. De fato, neste trabalho, tais filtros são empregados dando-se ênfase a expressões nominais, dada a natureza dos recursos disponíveis para a avaliação dos métodos, como dicionários e glossários terminológicos. Porém, os métodos aqui apresentados podem teoricamente ser aplicados para qualquer EM englobada por esta definição.

2.1 Identificação de EM

Uma grande variedade de abordagens têm sido propostas para a identificação automática de EM em função de seus diferentes tipos e propósitos de identificação. As abordagens diferem teórica e metodologicamente entre si em função dos tipos de EM abrangidos, da língua a que se aplicam e das fontes de informação que utilizam.

Alguns desses trabalhos utilizam informações sobre uma língua, como Baldwin (2005) e Villavicencio et al. (2007) aplicadas para o inglês, (Silva et

al., 1999) e (Dias e Nunes, 2001) aplicadas para o português. Outros trabalhos se beneficiam ainda de informações de uma segunda língua para ajudar a identificar e a lidar com EM (Villada Moirón e Tiedemann, 2006; Caseli et al., 2009b). Como base para ajudar a determinar se uma dada sequência de palavras é realmente uma EM (por exemplo, *ad hoc* é uma EM porém *o menino pequeno* não), algumas dessas propostas empregam conhecimentos linguísticos, enquanto outras empregam métodos ditos fracos ou estatísticos (por exemplo, Evert e Krenn (2005) e Villavicencio et al. (2007)) ou combinam vários tipos de informações tanto linguísticas, como propriedades sintáticas e semânticas (Van de Cruys e Villada Moirón, 2007), quanto de frequência e estatísticas, resultantes de processos como por exemplo o alinhamento lexical automático em um par de línguas (Villada Moirón e Tiedemann, 2006). A combinação de diversos tipos de informação pode ser realizada através de classificadores aprendidos automaticamente a partir de conjuntos de dados anotados (Pecina, 2008). Este trabalho investiga a influência de diferentes fontes de informação na tarefa de identificação de EM.

Medidas estatísticas de associação têm sido amplamente empregadas na identificação de EM, visto que elas podem ser democraticamente aplicadas a qualquer tipo de EM e de língua. A idéia por trás de seu uso é que elas são um meio de baixo custo para a detecção de padrões recorrentes, dado que se espera que as palavras componentes de uma EM coocorram frequentemente. Dessa forma, essas medidas podem indicar a probabilidade de que um candidato seja uma EM verdadeira, independentemente do tipo de EM e da língua. No entanto, algumas medidas parecem fornecer previsões mais precisas que outras sobre a chance de um determinado candidato ser de fato uma EM, e não há ainda consenso sobre qual medida é mais adequada para identificar EM em geral. Uma comparação de algumas destas medidas para a detecção de EM independentes de tipo indicaram que informação mútua diferencia melhor EM de não-EM do que χ_2 (Villavicencio et al., 2007). Várias medidas comuns de associação, como a informação mútua e χ_2, têm sido amplamente usadas para a detecção de EM, além de outras que têm sido especialmente propostas para esta tarefa, por exemplo, por Silva et al. (1999).

Outra questão importante é a generalização de algumas destas medidas para aplicação a *n*-gramas com tamanho arbitrário, principalmente no que diz respeito às MA baseadas em tabela de contingência e sua conhecida aplicação a bigramas. Silva et al. (1999), por exemplo, propõem uma abordagem em que, para um dado candidato, todas as possíveis divisões dele em duas partes são geradas, onde cada uma das duas partes pode ser maior que um unigrama. Assim um *n*-grama formado por 4 palavras $(w_1 \dots w_4)$ gera 3 bigramas: $(w_1) + (w_2, w_3, w_4)$, $(w_1, w_2) + (w_3, w_4)$ e $(w_1, w_2, w_3) + ((w_4)$, que são analisados em termos da força de associação entre as suas partes.

Além disso, para a identificação de EM, a eficácia de uma determinada medida parece depender de fatores como o tipo de EM sendo identificada, o domínio e o tamanho dos corpora utilizados, e a quantidade de dados de baixa frequência excluídos através da adoção de um limiar (Evert e Krenn, 2005). No que se refere aos tipos de corpora utilizados, tanto têm sido utilizados corpora paralelos (Maia, 2003), que envolvem originais e traduções, quanto corpora comparáveis (Maia e Matos, 2008), que implicam pares de textos escritos sobre um mesmo tema ou tópico originalmente produzidos em línguas diferentes por pessoas diferentes. Esses corpora geralmente são tratados à medida que recebem algum tipo de etiquetamento ou anotação. Há, todavia, também estudos que se dedicaram a corpora monolíngues não tratados, tal como o de Dias e Lopes (2005).

Quanto a trabalhos que envolveram a extração de terminologias em corpora, pode-se dizer que têm sido muitos e diferentes os estudos publicados. Todos, entretanto, enfrentaram a dificuldade de distinguir, com apoio computacional, os limites entre o léxico especializado e o léxico da linguagem cotidiana. Ranchhod e Mota (1998), por exemplo, fizeram um estudo que justamente procurou qualificar a identificação de itens especializados em um analisador de texto integrado por uma ferramenta que pesquisa, arrola e traz informações sobre os termos nele contidos. O tratamento da informação, entretanto, não partiu de um bloco geral de EM de corpora previamente reunidos, mas consistiu em agregar ao sistema de busca informações trazidas de dicionários gerais e de dicionários específicos pré-existentes. Nesse trabalho, ainda que os textos a analisar tenham sido do tipo técnico, também foi enfrentado o problema da presença simultânea de uma mesma dada expressão em diferentes dicionários, fato que já reforçava o problema da distinção controversa entre o léxico comum e o léxico especializado.

Assim, considerada a dificuldade implicada nessa diferenciação, são investigadas aqui algumas abordagens para a identificação de EM de um domínio especializado e alguns aspectos que podem ter influência sobre os resultados obtidos, para uma avaliação mais precisa destes métodos. Para português, a combinação de algumas medidas baseadas em frequências e heurísticas para a identificação de termos para a construção de uma ontologia a partir de textos de domínio específico resultou em uma medida F de até 11,51% para bigramas e 8,41% para trigramas (Vieira et al., 2009).

Entre os métodos que utilizam informações adicionais para extrair EM, o proposto por Villada Moirón e Tiedemann (2006) parece ser o mais semelhante à abordagem baseada em alinhamento empregada neste trabalho. A principal diferença entre eles é a maneira com que o alinhamento lexical é usado no processo de extração de EM. Neste trabalho, o alinhamento de palavras é a base do processo de extração de EM, enquanto o método de Villada Moirón e Tiedemann usa o alinhamento apenas para a classificação dos candidatos a EM que foram extraídos com base em medidas de associação e em heurísticas de dependência de núcleo (em dados sintaticamente analisados). Outro trabalho relacionado reporta a detecção automática de compostos não-composicionais (Melamed, 1997) que são identificados através da análise de modelos estatísticos de tradução treinados com um corpus enorme em um processo demorado. Santos e Simões realizaram experimentos envolvendo alinhamento lexical em corpora paralelos (Santos e Simões, 2008), buscando, entre outros objetivos, mensurar a importância da combinação de dicionários e corpora, do uso de informações sintáticas neste processo e da direção de tradução entre os idiomas. Nesse sentido, pode-se considerar que metodologias para a extração de sintagmas nominais bilíngues a partir de corpora paralelos, por exemplo, através do uso da *Pattern Description Language* (Simões e Almeida, 2008), constituem em si formas de identificação de EM.

3 O Corpus e as Listas de Referência

Nos experimentos descritos a seguir, utilizou-se um conjunto de 283 artigos de Pediatria, o corpus JPED-Coulthard. Trata-se de um corpus paralelo português–inglês que contém 785.488 palavras em português e 729.923 palavras em inglês. A língua-fonte, isto é, a língua na qual os artigos foram originalmente escritos, é o português, enquanto a língua-alvo é o inglês. Os textos foram publicados no *Jornal de Pediatria* entre 2003 e 2004. Vale destacar que esse corpus foi inicialmente organizado e estudado quanto à adequação das traduções por Coulthard (2005). Não foram considerados os resumos/*abstracts* e as referências bibliográficas no cômputo do número de palavras dos textos.

A partir desses 283 artigos bilíngues, foram criados alguns recursos lexicais: o *Dicionário de Pediatria* e o *Catálogo de Pediatria*[3]. Ambos recursos contêm um levantamento de expressões conceitual e linguisticamente importantes do domínio. Esses dois recursos, entretanto, são produtos diferenciados, pois são voltados para o uso do aprendiz de tradução brasileiro que começa a trabalhar na área médica. Sua finalidade é auxiliar esses iniciantes, graduandos em tradução da área de Letras/Linguística, que têm pouca experiência com construções, noções e terminologias desse domínio. Enquanto o *Dicionário* apresenta expressões recorrentes nesse corpus que geralmente contêm algum termo de Medicina cuja definição é apresentada, o *Catálogo* apresenta um levantamento de construções frequentemente empregadas na linguagem do domínio em foco, representada pelo corpus, mas que não estão associadas a uma terminologia ou nódulo conceitual. Em síntese, o foco de um é para expressões associadas a definições e o de outro é dirigido para expressões com exemplos de uso e padrões, o que inclui colocações e fraseologias.[4] Assim:

- o *Dicionário de Pediatria* contém 747 itens em português e 746 itens em inglês
- o *Catálogo de Pediatria* possui 702 itens em português e 698 itens em inglês

O processo de seleção das entradas a serem adicionadas aos dois recursos passou pelas seguintes etapas:

1. geração de *n*-gramas (bi, tri e quadrigramas[5]);
2. seleção de *n*-gramas com frequência maior ou igual a 5, sendo de palavras técnicas ou não;
3. exclusão de *n*-gramas puramente gramaticais (*o leite*);
4. exclusão de *n*-gramas que contivessem preposições, pronomes, conjunções: (por exemplo, *n*-grama iniciado ou terminado por *de* [PREP]). Essa lista foi definida a partir da análise manual da lista de palavras mais frequentes nos *n*-gramas;
5. exclusão de *n*-gramas do tipo ([DET]+N+X[+Y]), ou seja, uma sequência de determinante (DET) seguido de um nome (N) e até dois elementos (X e Y) (por exemplo, *o leite o leite materno/ o leite materno*

[3] Produzidos e disponibilizados gratuitamente por TEXTQUIM/TERMISUL `http://www.ufrgs.br/textquim`

[4] Aqui cabe esclarecer que a divisão dos itens entre *Dicionário* e *Catálogo*, em sendo um julgamento que reproduz a distinção entre: a) o que é específico do domínio, associado a uma definição; b) o que é expressão da linguagem cotidiana; c) o que é uma expressão de natureza híbrida, associado a um padrão sintagmático e que inclua linguagem geral e especializada, tornou-se algo extremamente complexo. Nesse sentido, a divisão dos dados nesses três blocos, acomodados os dois últimos no *Catálogo*, foi feita por um grupo de pesquisadores com formação em Letras e Tradução que se ocupam de produtos dicionarísticos. A divisão, além de espelhar critérios objetivos, também reflete uma percepção subjetiva do fenômeno envolvido e sempre comportará críticas. O trabalho relatado neste artigo, que reúne pesquisadores de PLN e terminógrafos, pode justamente qualificar esses tipos de repertórios de EM à medida que o retoma e confronta o procedimentos que os geraram.

[5] Um esclarecimento sobre essa nomenclatura pode ser encontrado em Manning e Schütze (1999, p. 193).

ordenhado). O padrão DET foi preenchido com várias possibilidades de determinantes: *os/o/as/a/um/uma/alguma/cuja(o)*;

6. remoção de *n*-gramas começados ou terminadas por verbo; e
7. retirada de *n*-gramas que fossem subpalavras de *n*-gramas maiores.

Em resumo, os recursos apresentam como entradas apenas *n*-gramas do corpus com frequência superior a 5, os quais foram filtrados mediante o uso de informações morfossintáticas e manualmente verificados. O processo de filtragem gerou um total de 2.407 entradas. Desses itens, 1.421 são bigramas e 730 são trigramas. Ao se comparar as listas em português e em inglês, tanto no dicionário quanto no catálogo, há diferença na quantidade de bi, tri e quadrigramas de língua para língua. Isso ocorre porque nem todas as construções em português possuem equivalentes com construções idênticas em inglês, com o mesmo número e a mesma ordem de palavras. Por exemplo, *recém-nascidos de baixo peso* é um quadrigrama; no entanto, seu correspondente, *low birth weight*, é um trigrama.

Neste trabalho, para avaliar o reconhecimento de expressões candidatas a EM, foram utilizados para integrar as listas de referência tanto os *n*-gramas do *Dicionário* quanto os do *Catálogo de Pediatria*. Além disso, as candidatas a EM em inglês são avaliadas usando-se um dicionário geral de EM em inglês (Cambridge, 1994), que contém 24.160 entradas (dos quais 9.174 são bigramas e 2.946 trigramas). As listas de referência contem as EM selecionadas por lexicógrafos e terminólogos para cada uma das línguas. Para o português as listas de referência incluem candidatos com frequência maior ou igual a 5. Portanto, qualquer candidato identificado pelos métodos usados (em particular pelo método de alinhamento), que não atinja a frequência mínima de 5 ocorrências, não será considerado como verdadeiro positivo por não se encontrar nas referências, mesmo quando se tratar de uma EM.

A lista resultante do processo de enriquecimento foi produzida conforme descrito em Lopes et al. (2009)[6], adicionando-se todos os bigramas válidos contidos em trigramas da lista que haviam sido removidos durante a construção dos recursos. Esse processo foi feito para as listas de ambas as línguas, e as versões finais das listas de referência têm 2.150 *n*-gramas em português e 1.424 *n*-gramas em inglês.

Para verificar a proporção de EM genéricas e específicas de domínio que ocorreram no corpus, utilizou-se metodologias distintas para cada uma das línguas. Em português, além das informações sobre EMs nas listas de referência, usou-se também julgamentos humanos, para anotar as EM no que se refere à pertinência ou não de cada item ao domínio de Pediatria/Medicina. Desta forma cada lista foi anotada com informação de domínio, de comum acordo, por três pesquisadores de Terminologia e Tradução que estiveram envolvidos na produção do dicionário e do catálogo. O anotador humano seguiu as heurísticas listadas abaixo para decidir sobre cada EM.

E se a EM corresponde a um termo de Medicina ou Pediatria ou área afim, recebeu a etiqueta E (*teste tuberculínico, fase de indução*);

G se a EM corresponde a uma expressão da linguagem cotidiana, de fácil compreensão para qualquer falante medianamente escolarizado do português do Brasil, recebeu a etiqueta G (*falta de apetite, grupo de risco*);

H se a EM corresponde a uma expressão da linguagem cotidiana, mas com sentido específico em Pediatria/Medicina, sendo um híbrido entre linguagem cotidiana e linguagem especializada, constituindo casos de julgamento ambíguo, recebeu a etiqueta H (*nível de sódio, saturação de oxigênio*).

Em inglês a natureza das listas de referência foi utilizada como indicador de domínio, dada a indisponibilidade de anotações dos dados por falantes nativos. Desta forma considerou-se todas as entradas do dicionário e do catálogo como construções específicas ao domínio enquanto as construções genéricas provêm de um dicionário geral, o *Cambridge International Dictionary of Idioms* (Cambridge, 1994), com 1.270 *n*-gramas para o inglês com ao menos uma ocorrência no corpus. Essas duas fontes estão marcados na tabela 1 como especializada (E) e genérica (G), respectivamente.

A lista de referência em português anotada por domínio contém uma grande maioria (76,83%) de EM específicas de domínio. Dentre os 1.421 bigramas, 977 foram considerados específicos do domínio de Pediatria, 226 genéricos e 218 híbridos. No grupo de trigramas, há 730 trigramas, dos quais 419 específicos e 195 genéricos e 116 híbridos.

Entre os candidatos, há expressões recorrentes como *prevalence of elevate blood pressure*, que não serão extraídas por nenhum dos métodos, dado o foco em bigramas e trigramas. Consequentemente, nas avaliações reportadas, a revocação apresentada é subestimada em relação ao seu valor real.

4 Métodos

Neste trabalho, investiga-se o uso de duas abordagens independentes para identificação de EM, e propõe-

[6]Disponível em `www.inf.pucrs.br/~ontolp/downloads-ontolplista.php`

Tipo	português	inglês
Específico	1.396	1.424
Genérico	421	1.270
Híbrido	334	–
Total	2.151	2.694

Tabela 1: Número de EM nas referências.

se uma abordagem combinada. A primeira abordagem, doravante denominada *abordagem associativa*, aplica Medidas de Associação (MA) para todos os bigramas e trigramas gerados a partir de cada corpus. As candidatas a EM são avaliadas em termos dos valores obtidos para elas por cada uma das medidas de associação utilizadas.

A segunda abordagem, doravante denominada *abordagem baseada em alinhamento*, tem por princípio a extração de EM a partir dos alinhamentos automáticos lexicais de versões em português e em inglês do Corpus de Pediatria gerados pelo alinhador estatístico lexical `GIZA++` (Och e Ney, 2000b). O método combinado proposto, por sua vez, combina as duas abordagens usando redes bayesianas.

Nesta seção, são descritos os experimentos realizados usando-se cada uma das abordagens para extrair EM do corpus. A avaliação é realizada de maneira automática, comparando-se as EM identificadas pela abordagem com as listas de referência descritas na seção 3 para cada uma dos línguas, em termos da precisão (P), revocação (R) e medida F, calculadas, respectivamente, como:

$$P = \frac{(\text{\#candidatas corretas})}{(\text{\#candidatas propostas})}$$
$$R = \frac{(\text{\#candidatas corretas})}{(\text{\#candidatas na referência})}$$
$$F = \frac{(2 \times P \times R)}{(P+R)}$$

Primeiramente, uma lista prévia de candidatas a EM é extraída do corpus JPED-Coulthard com cada abordagem. Em seguida, as candidatas são analisadas morfosintaticamente pelas ferramentas do `Apertium`[7] (analisador morfossintático e desambiguador lexical) com base nos dicionários morfológicos originais aumentados conforme descrito em Caseli, Nunes e Forcada (2006) e em Caseli (2007). Vale ressaltar, aqui, que o desempenho do analisador morfossintático está relacionado à cobertura do mesmo, ou seja, utilizando-se os dicionários morfológicos aumentados citados a cobertura é de 1.136.536 formas superficiais em português e de 61.601, em inglês. Com relação ao desempenho do desambiguador lexical, não foi encontrado nenhum relato a esse respeito. Por fim, sobre as listas de candidatas propostas por cada um dos métodos, são aplicados os seguintes filtros:

- **f0** lista original, nenhum filtro é aplicado às candidatas;
- **f1** candidatas após a remoção de n-gramas contendo pontuação e números;
- **f2** candidatas após (a) **f1** e (b) cujo número de ocorrências no corpus[8] é no mínimo 5;
- **f3** candidatas após (a) **f1, f2** excluindo ainda aquelas que (b) se iniciam por uma palavra funcional[9] e algumas formas superficiais, como flexões do verbo *ser* (*são*, *é*, *era*, *eram*), pronomes relativos (*qual*, *quando*, *quem*, *por que*) e preposições (*para*, *de*)[10].

Para f3, padrões alternativos de filtros morfossintáticos têm sido propostos na literatura, como o aplicado para a construção das listas de referência, e podem ser otimizados com informações específicas para cada língua. No entanto, como neste trabalho o objetivo é investigar o desempenho de um conjunto de métodos, f3 foi definido seguindo a proposta de Caseli et al. (2009b) para o inglês com padrões equivalentes para o português. Além disto, os filtros foram aplicados independentemente para cada uma das línguas, gerando uma lista filtrada de candidatos para cada uma.

Uma vez obtida a lista filtrada de candidatas a EM, o objetivo é remover apenas e qualquer n-grama que não seja uma EM. Este processo difere para cada uma das abordagens, conforme descrito nas próximas seções, e o sucesso é avaliado de acordo com as EM contidas nas listas de referência. Isso significa que os n-gramas candidatos que sejam encontrados nas listas de referência são considerados EM, mas os não contidos não necessariamente são não-EM. Há limitações de cobertura das referências a se considerar, dada a natureza dinâmica das línguas, bem como questões das características de uma EM qualquer, como transparência e frequência, entre outras.

[7] `Apertium` (Armentano-Oller et al., 2006) é um sistema de tradução automática de código-fonte aberto disponível em `http://www.apertium.org`.

[8] O método de alinhamento considera as frequências de alinhamento de uma candidata (número de vezes em que as palavras da candidata foram alinhadas juntas). No entanto, o filtro **f1** é aplicado sobre o número de ocorrências dessa candidata no corpus independentemente dos alinhamentos.

[9] Nesse trabalho, considera-se que uma palavra funcional é um artigo, verbo auxiliar, pronome, advérbio ou conjunção.

[10] E analogamente para o inglês, considerando-se a tradução literal dos termos de filtragem.

4.1 Abordagem Associativa

Na abordagem associativa, a filtragem é feita com base na força de associação de uma candidata medida de acordo com a probabilidade de co-ocorrência das palavras que a compõem. Evidências estatísticas de associação forte têm sido bastante empregadas em trabalhos recentes da área (Evert e Krenn, 2005; Ramisch et al., 2008; Pearce, 2002; Pecina, 2008; Ramisch, 2009). Uma visão geral destes trabalhos é apresentada em Ramisch (2009).

A validação de uma EM candidata é feita utilizando-se um conjunto de medidas de associação (MA): informação mútua pontual (PMI, do inglês *pointwise mutual information*), informação mútua (MI, do inglês *mutual information*), estatística *t* de student, estatística χ^2 de Pearson, coeficiente de Dice, teste exato de Fisher, medida de Poisson-Stirling (PS) e razão de chances (OR, do inglês *odds ratio*), implementadas no Ngram Statistics Package (Banerjee e Pedersen, 2003). Estas medidas típicas de associação são resumidas na tabela 2 (adaptada de Ramisch (2009)) e as fórmulas são calculadas com base nas frequências obtidas no corpus de cada língua. Globalmente, as medidas assumem que a frequência de co-ocorrência das palavras em uma EM é superior à frequência esperada para uma sequência randômica de *n* palavras. A segunda coluna da tabela mostra quais os valores de *n* (ou seja, o comprimento do *n*-grama) para os quais a MA pode ser aplicada, onde "*" representa a ausência de limitação de tamanho.

Formalmente, considera-se a candidata a EM como um *n*-grama composto de *n* palavras adjacentes w_1 a w_n. A contagem do número de ocorrências (frequência) de um *n*-grama em um corpus é denotada $c(w_1 \dots w_n)$. A medida da força de associação entre as palavras do *n*-grama $w_1 \dots w_n$ é feita através da comparação da frequência relativa *observada* $c(w_1 \dots w_n)$ com a frequência relativa *esperada* E. A última é calculada supondo-se como hipótese nula que palavras em um corpus são eventos independentes, ou seja, que a frequência de um *n*-grama é igual ao número de palavras N do corpus ponderado pelo produto das probabilidades de cada uma das palavras que o compõem:

$$E(w_1 \dots w_n) = \frac{c(w_1) \dots c(w_n)}{N^{n-1}}$$

Algumas das MA usadas na identificação associativa de EM são baseadas em tabelas de contingência. Isto significa que, além de considerar as frequências individuais das palavras, elas também levam em consideração a frequência de não-ocorrência dessas palavras, construindo uma tabela com as combinações possíveis. Nesses casos, usa-se a_i para representar ambas possibilidades, w_i e $\bar{w}_i$, em que a notação $\bar{w}_1$ corresponde à ocorrência de qualquer palavra exceto w_i. Em um dado *n*-grama $w_1 \dots w_n$, cada célula da tabela de contingência corresponde a uma combinação possível de $a_1 \dots a_n$. Essas medidas são muito robustas para eventos raros e são particularmente adequadas para os *n*-gramas onde $n = 2$ mas não são facilmente estendidas para candidatos com comprimento arbitrário, como mostra a coluna intermediária da tabela 2. As três últimas MA apresentadas na tabela são baseadas em tabelas de contingência e possuem limitação do valor de *n*. Portanto, para todos os trigramas candidatos, os valores dessas medidas não puderam ser calculados.

4.2 Abordagem Baseada em Alinhamento Lexical

Nos últimos anos, a utilização de textos paralelos e textos paralelos alinhados tem se tornado cada vez mais frequente em inúmeras aplicações de PLN. Os textos paralelos, segundo a terminologia estabelecida pela comunidade de linguística computacional, são textos acompanhados de sua tradução em uma ou várias línguas. Se esses textos possuírem marcas que identificam os pontos de correspondência entre o texto original (texto fonte) e sua tradução (texto alvo) eles são considerados alinhados.

Métodos automáticos de alinhamento de textos paralelos podem ser usados para encontrar os pontos de correspondências entre os textos fonte e alvo. O processo automático de alinhamento de textos paralelos, resumidamente, pode ser entendido como a "busca", no texto alvo, de uma ou mais sentenças (ou unidades lexicais) que correspondam à tradução de uma dada sentença (ou unidade lexical) no texto fonte. Quando a correspondência se dá entre sentenças dizemos que o alinhamento é sentencial, quando a mesma ocorre entre unidades lexicais, dizemos que o alinhamento é lexical.

O corpus paralelo utilizado nos experimentos apresentados neste artigo passou por ambos os processos de alinhamento. O alinhamento sentencial foi realizado automaticamente por uma versão do *Translation Corpus Aligner* (TCA) (Hofland, 1996), descrita em detalhes em Caseli (2003) e Caseli, Silva e Nunes (2004). Após o processamento automático, os casos potencialmente alinhados de maneira incorreta (alinhamentos diferentes de 1 : 1) foram verificados manualmente. O alinhamento lexical, por sua vez, foi desempenhado automaticamente por meio da ferramenta `GIZA++`[11] (Och e Ney, 2000b), porém sem a verificação manual uma vez que esta seria uma tarefa extremamente árdua já que os alinhamentos diferentes de 1 : 1 são muito mais frequentes no alinhamento

[11] `http://www.fjoch.com/GIZA++.html`

Medida	n	**Fórmula**
PMI	*	$\log_2 \frac{c(w_1 \dots w_n)}{E(w_1 \dots w_n)}$
t	*	$\frac{c(w_1 \dots w_n) - E(w_1 \dots w_n)}{\sqrt{c(w_1 \dots w_n)}}$
Dice	*	$\frac{n \times c(w_1 \dots w_n)}{\sum_{i=1}^{n} c(w_i)}$
MI	2,3	$\sum_{a_1 \dots a_n} \frac{c(a_1 \dots a_n)}{N} \log_2 \left[\frac{c(a_1 \dots a_n)}{E(a_1 \dots a_n)} \right]$
PS	2,3	$c(w_1 \dots w_n) \times \left[\log \frac{c(w_1 \dots w_n)}{E(w_1 \dots w_n)} - 1 \right]$
χ^2	2	$\sum_{a_1 a_2} \frac{[c(a_1 a_2) - E(a_1 a_2)]^2}{E(a_1 a_2)}$
Fisher	2	$\sum_{k=c(w_1 w_2)}^{\min\{c(w_1),c(w_2)\}} \frac{c(\bar{w}_1)!c(w_1)!c(\bar{w}_2)!c(w_2)!}{N!k!(c(w_1)-k)!(c(w_2)-k)!(c(\bar{w}_2)-c(w_1)+k)!}$
OR	2	$\frac{c(w_1 w_2)c(\bar{w}_1 \bar{w}_2)}{c(w_1 \bar{w}_2)c(\bar{w}_1 w_2)}$

Tabela 2: Medidas de associação utilizadas pelo método associativo

lexical do que no sentencial.

GIZA++ utiliza os modelos estatísticos da IBM (Brown et al., 1993) e o modelo oculto de Markov (HMM) (Vogel, Ney e Tillmann, 1996; Och e Ney, 2000b; Och e Ney, 2000a) para determinar as melhores correspondências entre palavras fonte e palavras alvo. Para os experimentos apresentados neste artigo, utilizou-se a versão 2.0 em sua configuração padrão na qual estão incluídas iterações dos modelos IBM-1, IBM-3, IBM-4 e HMM.

Os modelos utilizados por GIZA++ variam no modo como é calculada a probabilidade do alinhamento $\Pr({f_1}^S, {a_1}^S | {e_1}^T)$, na qual ${a_1}^S$ é um alinhamento que descreve o mapeamento da palavra fonte f_j na palavra alvo e_{a_j} considerando-se que ${f_1}^S$ é uma cadeia de caracteres fonte e ${e_1}^T$, uma cadeia de caracteres alvo. Por exemplo, no modelo IBM-1, todos os alinhamentos têm a mesma probabilidade. O modelo HMM, por sua vez, usa um modelo de primeira ordem $p(a_j|a_{j-1})$ no qual a posição do alinhamento a_j depende da posição do alinhamento anterior a_{j-1}. A partir do modelo IBM-3, um modelo de fertilidade $p(\phi|e)$ é adicionado ao cálculo da probabilidade. Esse modelo descreve o número de palavras ϕ alinhadas com a palavra alvo e. O modelo IBM-4, por sua vez, busca modelar o efeito de mudança de posição das palavras fonte na tradução e inclui, portanto, um modelo de distorção para simular o fato de que a tradução de uma palavra fonte é deslocada na frase alvo.

O alinhamento foi realizado por GIZA++ no sentido pt–en e no sentido en–pt e a combinação (união) dos alinhamentos foi gerada resultando no alinhamento final. A união foi selecionada como método de simetrização dos alinhamentos gerados nos dois sentidos de tradução por se tratar do método que apresentou melhor revocação em experimentos prévios (Caseli, 2007). O desempenho no alinhamento de lemas, configuração utilizada nos experimentos apresentados neste artigo, não foi avaliado especificamente para o corpus de Pediatria, porém em avaliação prévia realizada em outro corpus pt–en o desempenho relatado foi de 8,94% AER (*Alignment-Error Rate*), o que está de acordo com os valores relatados para outros pares de línguas. Detalhes sobre a avaliação do alinhamento de lemas produzido por GIZA++ podem ser obtidos em Caseli (2007).

Além dos alinhamentos lexical e sentencial, o corpus pt–en também foi etiquetado morfossintaticamente usando os dicionários morfológicos e as ferramentas do Apertium (Armentano-Oller et al., 2006). Em particular, o corpus foi analisado morfossintaticamente com base nos dicionários morfológicos originais aumentados conforme descrito em Caseli, Nunes e Forcada (2006) e em Caseli (2007). A partir desse processo de etiquetação morfossintática é que foi possível aplicar filtros de categorias gramaticais na lista inicial de candidatas a EM.

Um exemplo de um par de sentenças paralelas pt–en alinhado lexicalmente por GIZA++ é apresentado

na Figura 1. Nesse exemplo, cada palavra é apresentada em uma linha separada na ordem em que ocorrem na sentença, sua posição na sentença é indicada na primeira coluna e os alinhamentos lexicais podem ser recuperados pelo número que segue o ":" ao final de cada palavra. Alinhamentos de omissão estão representados pelo "0". Além disso, cada forma superficial da palavra nesta figura é seguida por seu lema, categoria gramatical e traços morfológicos retornados pelo etiquetador morfossintático que, quando não reconhece uma determinada palavra, indica que a mesma é desconhecida inserindo um "*" em seu início, como ocorre com as palavras em português *helicobacter* e *pylori*. Por fim, é possível notar um alinhamento envolvendo a candidata a EM "*precisa para*" com sua correspondente tradução em inglês "*needs to*".

	Sentença em português
1	o/o<det><def><m><sg>:1
2	único/único<adj><m><sg>:2
3	fato/fato<n><m><sg>:3
4	aceito/aceitar<vblex><pri><p1><sg>:3
5	é/ser<vbser><pri><p3><sg>:4
6	o/o<detnt>:0
7	de/de<pr>:0
8	que/que<cnjsub>:5
9	o/o<det><def><m><sg>:0
10	*helicobacter/helicobacter:6
11	*pylori/pylori:7
12	precisa/precisar<vblex><pri><p3><sg>:8_9
13	entrar/entrar<vblex><inf>:10
14	para/para<pr>:8_9
15	o/o<det><def><m><sg>:11
16	estômago/estômago<n><m><sg>:12
17	através/através<adv>:13
18	da/de<pr>+o<det><def><f><sg>:0
19	boca/boca<n><f><sg>:15

	Sentença em inglês
1	the/the<det><def><sp>:1
2	only/only<adj>:2
3	certainty/certainty<n><sg>:3_4
4	is/be<vbser><pri><p3><sg>:5
5	that/that<cnjsub>:8
6	*helicobacter/helicobacter:10
7	pylori/pylorus<n><pl>:11
8	needs/need<vblex><pri><p3><sg>:12_14
9	to/to<pr>:12_14
10	enter/enter<vblex><inf>:13
11	the/the<det><def><sp>:15
12	stomach/stomach<n><sg>:16
13	through/through<pr>:17
14	the/the<det><def><sp>:0
15	mouth/mouth<n><sg>:19

Figura 1: Exemplo de um par de sentenças paralelas alinhadas lexicalmente por GIZA++

Diferentemente da abordagem associativa, na abordagem baseada em alinhamento, as candidatas a EM são identificadas a partir das correspondências entre palavras e sequências de palavras da língua fonte e alvo definidas pelo alinhador. Mais especificamente, usando o alinhamento lexical entre uma sequência de palavras origem S ($S = s_1 \dots s_n$ com $n \geq 2$) e uma sequência de palavras destino T ($T = t_1 \dots t_m$ com $m \geq 1$), o método de extração baseado em alinhamento assume que a sequência S será uma candidata a EM. Por exemplo, a sequência de duas palavras em português *aleitamento materno* — que ocorre 202 vezes no corpus utilizado nos experimentos — é uma candidata a EM porque essas duas palavras foram alinhadas em conjunto 184 vezes com a palavra *breastfeeding* (um alinhamento 2 : 1), 8 vezes com a palavra *breastfed* (um alinhamento 2 : 1), 2 vezes com *breastfeeding practice* (um alinhamento 2 : 2) e assim por diante. É essa frequência de alinhamento, ou seja, o número de vezes em que a sequência de palavras da língua fonte ocorre em um alinhamento $n : m$ com $n \geq 2$, que será usada como atributo na combinação das abordagens. Por procurar sequências de palavras-origem que são frequentemente unidas durante o alinhamento, independentemente do número de palavras-alvo envolvidas, o método baseado em alinhamento prioriza precisão sobre revocação.

Algumas observações podem ser feitas a respeito de como o produto do alinhamento lexical influencia as candidatas de EM geradas. Por exemplo, na Figura 1, pode-se notar que duas palavras em português não consecutivas (*precisa* e *para*) foram alinhadas com duas palavras consecutivas do inglês (*needs to*). Essa característica traz um diferencial para o método de alinhamento quando comparado às medidas de associação uma vez que estas últimas recuperam apenas n-gramas e, sendo assim, a abordagem associativa nunca gera EM compostas por itens não consecutivos, diferentemente do método de alinhamento, que é capaz de gerá-las. Como consequência, a avaliação realizada com base nas listas de referência subestima a revocação do método baseado em alinhamento, uma vez que o processo de construção das listas levou em conta apenas sequências de palavras consecutivas.

4.3 Abordagem Combinada

Dado que as abordagens associativa e baseada em alinhamento têm características diferentes, que podem fazer com que se capture diferentes conjuntos de EM, a proposta deste trabalho é desenvolver um método combinado que maximize as vantagens de cada uma. Para isto, as diferentes MA e as frequências de alinhamento obtidas para as candidatas podem ser con-

n-grama (α)	n	$c(\alpha)$	**Abordagem alinhamento**	**Abordagem associativa**								**classe**
				Dice	OR	PMI	PS	t	MI	χ^2	Fisher	
abnormal findings	2	11	9	,03	114,1	6,74	25,70	2,62	0	734,73	0	não
renal insufficiency	2	26	0	,13	767,7	9,10	138	5,09	,0003	14249,6	0	sim
ato cirúrgico	2	7	3	,08	989,1	9,64	39,79	2,64	,0001	5584,2	0	sim
academia americana	2	24	0	,52	74302	13,3	197,4	4,9	,0004	244244	0	não

Tabela 3: Exemplos de entradas dos conjuntos de treinamento contendo todos os atributos usados em cada uma das estratégias de combinação.

sideradas como atributos para algoritmos de aprendizado de máquina, em uma abordagem semelhante à adotada por Pecina (2008) e Ramisch (2009). Para a abordagem combinada, foram utilizados os algoritmos implementados pelo pacote Weka (Witten e Frank, 2005).

O classificador para cada língua foi construído a partir do conjunto de *n*-gramas filtrados e anotados com os valores das medidas associativas e com o diagnóstico do alinhamento lexical sobre se o *n*-grama é uma possível EM, isto é, a frequência com que ele foi alinhado conjuntamente com uma palavra ou sequência de palavras na língua alvo. Na próxima seção, avalia-se duas possibilidades de combinação dos métodos: a primeira consiste em enriquecer os candidatos extraídos pelo método de alinhamento com as MA do método associativo; a segunda consiste em enriquecer os candidatos extraídos pelo método associativo (ou seja, todos os *n*-gramas do corpus que passaram pelos filtros) com a frequência de alinhamento. Em ambos os casos, os atributos usados para treinar o classificador são idênticos, e estão exaustivamente enumerados nas colunas da tabela 3.

Para adicionar a informação de classe para cada candidata foi feita uma avaliação das mesmas em relação as listas de referência: se o *n*-grama está contido nas listas, ele tem a classe *sim* (correspondendo a uma EM), caso contrário ele pertence à classe *não* (não-EM). A tabela 3 mostra alguns exemplos de entradas de inglês e português do conjunto de treinamento.

Como discutido na próxima seção, os conjuntos de dados disponíveis para treinar o classificador são desbalanceados, com uma proporção muito maior de não-EM do que de EM. Desta forma, optou-se por utilizar um algoritmo de construção de redes bayesianas com pesquisa de solução ótima através da árvore de cobertura. Este algoritmo, além de ser especialmente adequado para dados numéricos como os da tabela 3, tem se mostrado robusto e pouco sensível ao uso de classes com tamanhos muito diferentes.[12] Experimentos realizados em outros conjuntos de dados demonstraram que o algoritmo de máquina de vetor de suporte produz classificadores de boa qualidade. Neste trabalho, no entanto, optou-se por empregar um classificador do tipo rede bayesiana porque ele é menos oneroso em termos de recursos computacionais e de tempo de treinamento do que o algoritmo de máquina de vetor de suporte, além de produzir resultados comparáveis ao mesmo (Ramisch, 2009).

[12] Utilizando, por exemplo, árvores de decisão sobre os dados em inglês obteve-se um modelo com uma única classe com um mesmo diagnóstico para todos os candidatos (*não*).

5 Resultados

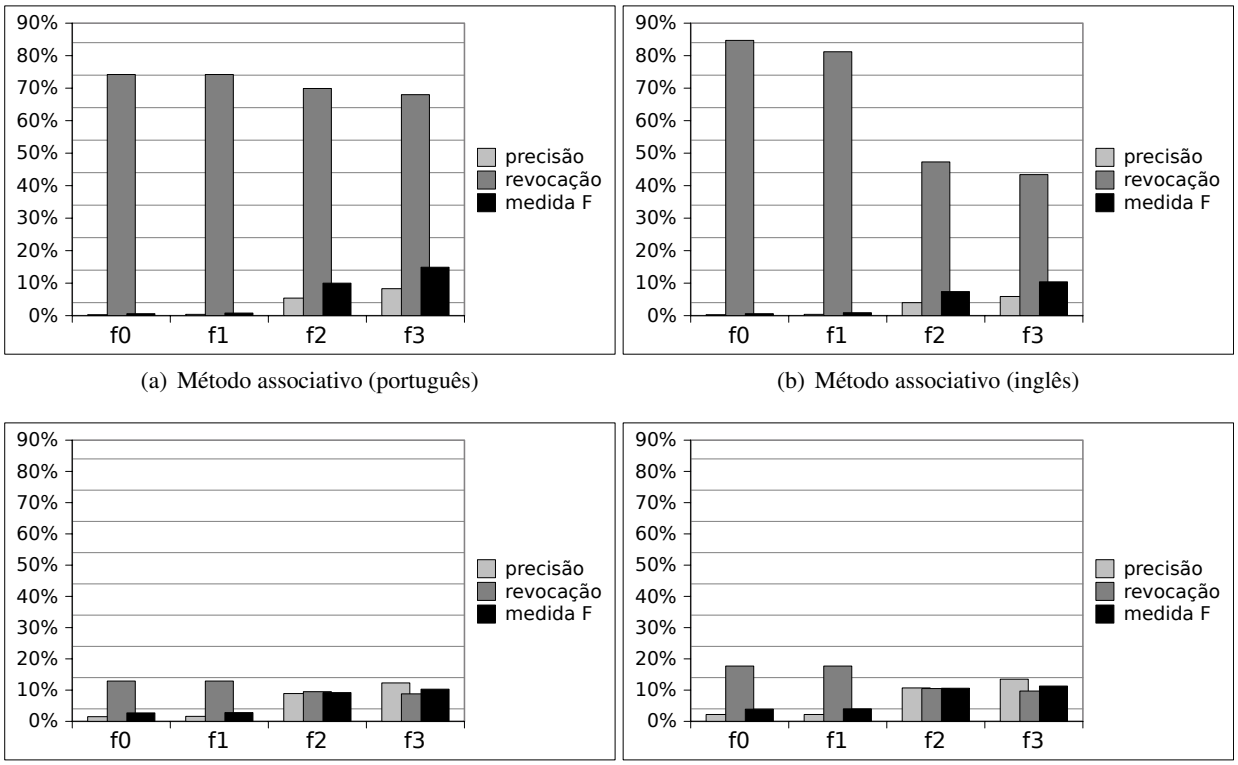
O desempenho obtido por cada um dos métodos na tarefa de identificação de EM será discutido a seguir. Após, discutir-se-á a taxa de acerto de cada método para EM de acordo com sua especificidade de domínio.

5.1 Identificação de EM

Primeiramente, descreve-se os resultados da avaliação da lista inicial de candidatas propostas por cada método e da aplicação dos vários filtros para remoção de ruído, como mostrado nas tabelas 4 e 5. Os resultados para português e inglês são descritos em termos de número de candidatos resultantes de cada processo e número de verdadeiros positivos (VP).

Para ambas as línguas e ambos os métodos a aplicação dos filtros melhorou os resultados em termos da precisão e da medida *F* (figura 2). Em particular o filtro f2 resultou em uma grande melhora da precisão, e mesmo nos casos onde houve uma redução na revocação, a medida *F* ainda refletiu a contribuição positiva do filtro. Por exemplo, para a abordagem associativa para o inglês, a revocação baixou em 33,9% mas ainda assim a medida *F* aumentou em 6,5%.

A diferença entre os métodos se refletiu em um número muito menor de candidatas a EM propostas pelo método baseado em alinhamento do que pelo método associativo: para o português 18.132 contra 572.893 respectivamente. Apesar desta grande diferença no número de candidatos propostos pelo alinhador (97% menos candidatos que a abordagem associativa), os resultados têm maior precisão para



(a) Método associativo (português)

(b) Método associativo (inglês)

(c) Método baseado em alinhamento (português)

(d) Método baseado em alinhamento (inglês)

Figura 2: Avaliação do efeito dos filtros aplicados a cada um dos métodos independentemente.

Abordagem associativa (pt)

	f0	**f1**	**f2**	**f3**
n-gramas	572.893	384.742	27.874	17.242
VPs	1.595	1.595	1.504	1.455

Abordagem baseada em alinhamento (pt)

	f0	**f1**	**f2**	**f3**
n-gramas	18.132	17.444	2.284	1.464
VP	277	277	204	189

Tabela 4: Desempenho dos filtros aplicados a ambos os métodos para os candidatos em *português*.

Abordagem associativa (en)

	f0	**f1**	**f2**	**f3**
n-gramas	586.431	391.850	25.478	15.399
VP	1.822	1.746	1.017	873

Abordagem baseada em alinhamento (en)

	f0	**f1**	**f2**	**f3**
n-gramas	17.516	16.972	2.108	1.527
VP	380	380	225	201

Tabela 5: Desempenho dos filtros aplicados a ambos os métodos para os candidatos em *inglês*.

ambas as línguas e maior medida F em todos os casos, exceto f2 e f3 para o português.

A tabela 6 mostra os resultados obtidos com f3 aplicado à intersecção entre os candidatos propostos por ambos os métodos, ou seja, considerando-se apenas candidatos extraídos simultaneamente pelo método baseado em alinhamento e pelo método associativo. Uma grande proporção dos candidatos propostos pelo alinhador estão contidos nos candidatos propostos pelas MA. Além disto, a precisão obtida melhora para ambas as línguas, com um número menor de candidatos que para o alinhador, mas o mesmo número de VP.

5.2 Combinação dos Métodos

A linha de base para a comparação do desempenho do método combinado é a obtida pelas abordagens associativa e baseada em alinhamento de forma independente.

São testadas duas alternativas diferentes para o método combinado, conforme mostrado na tabela 7. Em ambos os casos, foi realizada uma avaliação por validação cruzada fornecida pelo pacote Weka, usando-se 10 subconjuntos de dados. O tamanho do conjunto de dados fornecido é, respectivamente, de 1.464 e 17.242 candidatos para o português e de 1.527 e 15.399 para o inglês. A primeira estratégia,

	português	inglês
n-gramas	1.431	1.368
VP	190	208
Precisão	13,28%	15,20%
Revocação	8,83%	7,62%
Medida *F*	10,61%	10,16%

Tabela 6: Desempenho do filtro f3 aplicado à intersecção dos candidatos propostos pelas abordagens associativa e baseada em alinhamento.

alinhador → associativo, utiliza como base as candidatas propostas pelo método de alinhamento e usa os métodos associativos para subsequente validação. A segunda alternativa, **associativo → alinhador**, consiste em utilizar a lista de candidatas geradas pelas MA, adicionando às mesmas uma coluna que corresponde à informação fornecida pelo alinhador sobre a frequência de alinhamento da candidata. Vale relembrar que, em ambos os casos, os atributos usados pelo classificador são idênticos e foram descritos na tabela 3. Isso significa que as estratégias de combinação correspondem a duas maneiras diferentes de *escolher* quais candidatas serão consideradas pelo método combinado, mas *não têm influência no número ou tipo de atributos* usados pelo classificador. Por conseguinte, o conjunto de dados derivado da primeira estratégia contém algumas candidatas que não possuem nenhum valor de MA por se tratarem de *n*-gramas que não foram detectados pelo método associativo. Inversamente, o conjunto de dados derivado da segunda estratégia possui diversas candidatas contendo zero como informação de frequência de alinhamento, que correspondem a *n*-gramas identificados pelas MA mas que não fazem parte de nenhum alinhamento múltiplo.

A segunda alternativa parte de um número bem maior de candidatos do que a primeira, e para ambas as línguas tem-se um resultado muito superior em termos de medida *F* (mais de 30% superior para o português) do que o resultado obtido pela combinação dos métodos na direção contrária. Mesmo em relação ao desempenho máximo obtido por cada um dos métodos individuais, se pode ver uma melhora significativa nos resultados, em particular para o método associativo no português, onde a combinação resulta em uma aumento de quase 30% na medida *F*.

5.3 Especificidade dos Candidatos

Destes resultados gerais, o uso de uma lista de referência maior para o inglês, com EM genéricas, do que para o português, não parece ter contribuído para diferenças em resultado. Porém, a fim de determinar mais precisamente o desempenho de cada um dos métodos na identificação de termos específicos de domínio ou genéricos, os candidatos propostos por eles foram avaliados também em termos de tipo de EM, tabela 8. A maior proporção de candidatas específicas de domínio nas listas de referência foi também refletida nas candidatas VP retornadas por cada método. Para ambas as línguas, todas as abordagens têm melhor taxa de acerto para EM de domínio específico (E), com a identificação de um número maior destas candidatas. Estes resultados sugerem que a identificação de EM genéricas pode ser uma tarefa mais difícil do que a de EM específicas. O uso mais preciso e mais convencionalizado de EM dentro de um domínio pode contribuir para isto, com um menor grau de variabilidade léxica, sintática e semântica. Comparando-se as abordagens associativa e baseada em alinhamento, para EM específicas a abordagem que obteve uma melhor taxa de acerto foi a primeira, e para candidatas genéricas, foi a segunda. Isto pode ser devido à capacidade da abordagem baseada em alinhamento de identificar candidatos não-contíguos, sendo mais robusta à possível modificação ou variabilidade sintática. Porém, como as listas de referência possuem EM contíguas, não se pode calcular automaticamente qual o ganho trazido por esta capacidade. Investigações futuras serão feitas para avaliar este impacto. Além disso, a taxa de acerto para cada tipo obtida para o inglês é consideravelmente superior à obtida para o português.

6 Conclusões e Trabalhos Futuros

EM representam um conjunto complexo e heterogêneo de fenômenos que desafiam tentativas linguísticas e computacionais de capturá-los totalmente. Paradoxalmente, as EM têm um papel fundamental na comunicação oral e escrita e precisam ser levadas em conta quando da concepção de aplicações de processamento de linguagem que precisem de alguma interpretação semântica. Neste contexto, o tratamento de EM nos sistemas de PLN atuais é um grande desafio, dada a essência heterogênea e extremamente flexível dessas construções. Em decorrência do seu caráter simultaneamente complexo e essencial, as EM têm sido o foco de diversos trabalhos na comunidade científica, principalmente no que diz respeito à sua aquisição automática a partir de grandes bases textuais.

Neste trabalho, diferentemente de outros estudos com EM, lidou-se com um corpus especializado de originais e traduções e com listas de EM dele derivadas, as quais foram previamente identificadas por analisadores humanos como relevantes para a aprendizagem de tradução em Pediatria — tanto do ponto de vista conceitual quanto lingüístico — e incluídas em dois produtos de caráter dicionarístico diferentes. O conjunto geral dessas expressões, reunido em

português

	alinhador → associativo	**associativo → alinhador**
n-gramas	260	1.576
VP	137	787
Precisão	52,7%	49,9%
Revocação	6,4%	36,6%
Medida *F*	11,4%	42,2%

inglês

	alinhador → associativo	**associativo → alinhador**
n-gramas	97	1.130
VP	53	372
Precisão	54,6%	32,9%
Revocação	2,5%	17,3%
Medida *F*	4,7%	22,7%

Tabela 7: Desempenho do método combinado usando um classificador do tipo rede bayesiana.

português

	alinhamento	**associativa**	**alinhamento ∩ associativa**	**referências**
VP	190	1.463	190	2.151
E	55,79%	58,85%	55,79%	64,90%
G	24,21%	21,60%	24,21%	19,57%
H	20,00%	19,55%	20,00%	15,53%

inglês

	alinhamento	**associativa**	**alinhamento ∩ associativa**	**referências**
VP	208	934	208	2.694
E	63,46%	73,13%	63,46%	53,45%
G	36,54%	26,76%	36,54%	46,55%
H	0%	0%	0%	0%

Tabela 8: Proporção de EM por tipo em candidatos propostos por abordagens e na lista de referência

uma única lista de EM, com itens que integram tanto o léxico geral quanto o especializado, foi reavaliado pela mesma equipe e então dividido em três tipos: itens do léxico especializado, do léxico geral e itens de um "léxico híbrido", que constituiria, em tese, confluência entre linguagem cotidiana e linguagem especializada. O desafio aqui colocado foi o de encontrar metodologias de identificação para os itens associados ao léxico especializado combinando os diferentes fatores envolvidos nos materiais sob exame.

Nesse intuito, procurou-se investigar em que medida é possível utilizar e combinar recursos heterogêneos para automatizar a extração de EM, em específico no caso de textos técnicos em que grande parte das expressões possui simultaneamente um estatuto terminológico. Em primeiro lugar, analisou-se separadamente dois métodos de extração de EM: o método associativo, cuja lista de candidatos resultantes é gerada com base nas frequências de coocorrência das palavras que o formam; e o método baseado em alinhamentos, que por sua vez supõe que, em um corpus paralelo bilíngue, as expressões serão alinhadas de forma múltipla, extraindo-se assim a partir dos alinhamentos $n : m$ uma lista de candidatos a EM.

A fim de avaliar o desempenho individual e as possíveis estratégias de combinação de ambos os métodos, gerou-se uma lista de candidatos para cada

método e para cada língua a partir do corpus paralelo em português e em inglês do Jornal de Pediatria. Em um primeiro momento, foi investigado o impacto de diferentes fontes de informação na identificação de EM de domínios técnicos, através da aplicação de filtros sobre essas listas de candidatos. Os três filtros testados se mostraram bastante eficazes na remoção de ruídos e resultaram em melhoras significativas na medida F. Dentre esses, o que apresentou melhor compromisso entre um aumento na precisão e uma queda na revocação foi o filtro de frequência (**f2**); porém, o filtro morfossintático (**f3**) foi uma maneira simples e eficaz de eliminar o ruído com poucos efeitos colaterais.

Em termos das abordagens utilizadas (associativa e de alinhamento), uma avaliação dos desempenhos individuais indicou a natureza complementar de cada uma delas: a primeira identifica um maior número de candidatas, porém a segunda propõe um conjunto mais focado de candidatas com maior precisão. Ambos os métodos demonstraram maior sucesso na identificação de EM específicas de domínio, associadas ao léxico especializado, o que sugere que as EM genéricas apresentam um maior desafio para estes métodos. Está prevista uma investigação mais detalhada dos fatores que podem estar causando isso, como flexibilidade de uso e frequência, com comparação dos resultados obtidos em corpora genéricos. Pretende-se também verificar a portabilidade dos métodos para outros domínios.

Comparando as abordagens individuais, a associativa teve uma maior taxa de acerto nas EM específicas. Porém para as candidatas genéricas, a abordagem de alinhamento teve maior taxa de acerto. Dadas as diferenças dos candidatos propostos pelas duas abordagens, a combinação delas, proposta neste trabalho, trouxe um aumento significativo de desempenho na tarefa de identificação de EM. Foram avaliados dois modos para combinação dos resultados, e o que apresentou melhor desempenho foi que adotou o enriquecimento dos candidatos propostos pelos métodos associativos com informação de alinhamento. Neste caso a medida F aumentou de 14% para 42%.

Métodos como os apresentados neste artigo podem acelerar significativamente o trabalho de produção de repertórios de expressões recorrentes em corpora de textos científicos. Os resultados obtidos mostram que a adoção de abordagens simples, de baixo custo computacional e de conhecimento, pode trazer melhoras consideráveis de desempenho.

Para trabalhos futuros está prevista a investigação de maneiras alternativas para se obter a combinação ponderada das abordagens associativa e baseada em alinhamento, para produzir um conjunto de EM candidatas que é ainda mais precisa do que a fornecida pela primeira, mas que tem mais cobertura que a segunda. Além disso, seguindo a tendência de alguns trabalhos da área que exploram a extração de conhecimento de corpus comparável ao invés de corpus paralelo, como Fung (1998) e Haghighi et al. (2008), pretende-se, também, avaliar como as técnicas apresentadas neste artigo se comportam na extração de EM a partir de textos comparáveis. Por fim, a utilização dos resultados obtidos por este trabalho na construção semi-automática de ontologias também será investigada.

Agradecimentos

Este trabalho contou com a colaboração do grupos TERMISUL/TEXTECC da UFRGS, que disponibilizou o corpus de Pediatria JPED-Coutlhard e as listas de referência. Esses grupos têm apoio apoio financeiro do CNPq, FINEP e SEBRAE, e a pesquisa foi parcialmente realizada no projeto COMUNICA (FINEP/SEBRAE 1194/07).

Referências

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, e Miriam A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. Em R. Vieira, P. Quaresma, M.d.G.V. Nunes, N.J. Mamede, C. Oliveira, e M.C. Dias, editores, *Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006*, volume 3960 of *Lecture Notes in Computer Science*. Springer-Verlag, pp. 50–59, May, 2006.

Baldwin, T. 2005. The deep lexical acquisition of english verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Baldwin, Timothy, Emily M. Bender, Dan Flickinger, Ara Kim, e Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. Em *of the Fourth (LREC 2004)*, Lisbon, Portugal, May, 2004.

Banerjee, S. e T. Pedersen. 2003. The Design, Implementation and Use of the Ngram Statistics Package. Em *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 370–381.

Biber, D., S. Johansson, G. Leech, S. Conrad, e E. Finegan. 1999. *Grammar of Spoken and Written English*. Longman, Harlow.

Brown, P., V. Della-Pietra, S. Della-Pietra, e R. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–312.

Burnard, Lou. 2007. User Reference Guide for the British National Corpus. Relatório técnico, Oxford University Computing Services, February, 2007.

Calzolari, Nicoletta, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, e Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. Em *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1934–1940, Las Palmas, Canary Islands.

Cambridge. 1994. *Cambridge International Dictionary of English*. Cambridge University Press.

Carroll, J. e C. Grover. 1989. The derivation of a large computational lexicon of English from LDOCE. Em B. Boguraev e E. Briscoe, editores, *Computational Lexicography for Natural Language Processing*. Longman.

Caseli, H. M. 2003. Alinhamento sentencial de textos paralelos português-inglês. Tese de Mestrado, Instituto de Ciencias Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP). 101 p.

Caseli, H. M. 2007. *Indução de léxicos bilíngües e regras para a tradução automática*. Tese de doutoramento, Instituto de Ciencias Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP). 158 p.

Caseli, H. M., M. G. V. Nunes, e M. L. Forcada. 2006. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20:227–245.

Caseli, H. M., A. M. P. Silva, e M. G. V. Nunes. 2004. Evaluation of Methods for Sentence and Lexical Alignment of Brazilian Portuguese and English Parallel Texts. Em *Proceedings of the SBIA 2004 (LNAI)*, number 3171, pp. 184–193, Berlin Heidelberg. Springer-Verlag.

Caseli, H. M., A. Villavicencio, A. Machado, e M. J. Finatto. 2009a. Statistically-driven alignment-based multiword expression identification for technical domains. Em *Proceedings of the 2009 Workshop on Multiword Expressions (ACL-IJCNLP 2009)*, pp. 1–8.

Caseli, Helena Medeiros, Carlos Ramisch, Maria das Graças Volpe Nunes, e Aline Villavicencio. 2009b. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 1:1–20.

Copestake, Ann e Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. Em *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*.

Coulthard, R. J. 2005. The application of corpus methodology to translation: the jped parallel corpus and the pediatrics comparable corpus. Tese de Mestrado, Universidade Federal de Santa Catarina.

Dias, Gaël e Gabriel Pereira Lopes, 2005. *Extração Automática de Unidades Polilexicais para o Português*, pp. 155–184. Mercado de Letras / FAPESP, Campinas, SP, Brasil.

Dias, Gaël e Sergio Nunes. 2001. Combining evolutionary computing and similarity measures to extract collocations from unrestricted texts. Em *Proceedings of RANLP 2001 (Recent Advances in NLP)*, September, 2001.

Evert, S. e B. Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Fung, Pascale. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. Em David Farwell, Laurie Gerber, e Eduard Hovy, editores, *Machine Translation and the Information Soup: Proceedings of the Third Conference for Machine Translation in the Americas, AMTA'98*, volume 1529, pp. 1–17. Springer-Verlag, October, 1998.

Haghighi, Aria, Percy Liang, Taylor Berg-Kirkpatrick, e Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. Em *of the 46th : (ACL-08: HLT)*, Columbus, OH, USA, June, 2008.

Hofland, K. 1996. A program for aligning English and Norwegian sentences. Em S. Hockey, N. Ide, e G. Perissinotto, editores, *Research in Humanities Computing*, pp. 165–178, Oxford. Oxford University Press.

Jackendoff, R. 1997. Twistin' the night away. *Language*, 73:534–59.

Krieger, M. G. e M. J. B. Finatto. 2004. *Introdução à Terminologia: teoria & prática*. Editora Contexto.

Lopes, Lucelene, Renata Vieira, Maria José Finatto, Daniel Martins, Adriano Zanette, e Luiz Carlos

Ribeiro Jr. 2009. Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. *RECIIS. Electronic journal of communication information and innovation in health (English edition.Online)*, 3:72–84.

Maia, Belinda. 2003. Using corpora for terminology extraction: Pedagogical and computational approaches. Em B. Lewandowska-Tomasczczyk, editor, *PALC 2001 – Practical Applications of Language Corpora*, pp. 147–164.

Maia, Belinda e Sérgio Matos. 2008. Corpografo v4 - tools for researchers and teachers using comparable corpora. Em *LREC 2008 Workshop on Comparable Corpora (LREC 2008)*, pp. 79–82, May, 2008.

Manning, Christopher D. e Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, USA.

Melamed, I. Dan. 1997. Automatic discovery of non-compositional compounds in parallel data. Em *of the 2nd (EMNLP-2)*, Brown University, RI, USA, August, 1997.

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English. A Corpus-based Approach.* Oxford: Clarendon Press.

Och, F. J. e H. Ney. 2000a. A comparison of alignment models for statistical machine translation. Em *Proceedings of the 18th International Conference on Computational Linguistics (COLING'00)*, pp. 1086–1090, Saarbrücken, Germany, August, 2000.

Och, F. J. e H. Ney. 2000b. Improved statistical alignment models. Em *Proceedings of the 38th Annual Meeting of the ACL*, pp. 440–447, Hong Kong, China, October, 2000.

Pearce, Darren. 2002. A comparative evaluation of collocation extraction techniques. Em *of the Third (LREC 2002)*, Las Palmas, Canary Islands, Spain, May, 2002.

Pecina, Pavel. 2008. A machine learning approach to multiword expression extraction. Em *Proceedings of the LREC Workshop Towards a Shared Task for MWE 2008*, Marrakech, Morocco, June, 2008.

Ramisch, Carlos. 2009. Multiword terminology extraction for domainspecific documents. Tese de Mestrado, École Nationale Supérieure d'Informatiques et de Mathématiques Appliquées, Grenoble, França.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, e Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. Em *of the LREC Workshop Towards a Shared Task for (MWE 2008)*, pp. 50–53, Marrakech, Morocco, June, 2008.

Ranchhod, Elisabete Marques e Cristina Mota. 1998. Dicionários electrónicos de léxicos terminológicos. "Seguros". Em *Actas do Workshop sobre Linguística Computacional da APL*. APL.

Riehemann, Susanne. 2001. *A Constructional Approach to Idioms and Word Formation*. Tese de doutoramento, Stanford University.

Sag, I. A., T. Baldwin, F. Bond, A. Copestake, e D. Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. Em *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002)*, volume 2276 of *(Lecture Notes in Computer Science)*, pp. 1–15, London, UK. Springer-Verlag.

Santos, Diana e Alberto Simões. 2008. Portuguese-english word alignment: some experiments. Em *of the Sixth (LREC 2008)*, Marrakech, Morocco, May, 2008.

Silva, Joaquim Ferreira da, Gaël Dias, Sylvie Guilloré, e José Gabriel Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. Em *Proceedings of the 9th Portuguese Conference on Artificial Intelligence (EPIA '99)*, volume 1695, pp. 113–132, London, UK. Springer-Verlag.

Simões, Alberto e José J. Almeida. 2008. Bilingual terminology extraction based on translation patterns. *Procesamiento del Lenguaje Natural*, 41:281–288, September, 2008.

Smadja, Frank A. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Van de Cruys, T. e B. Villada Moirón. 2007. Semantics-based Multiword Expression Extraction. Em *Proceedings of the Workshop on A Broader Prespective on Multiword Expressions*, pp. 25–32, Prague, June, 2007.

Vieira, Renata, Maria José Finatto, Daniel Martins, Adriano Zanette, e Luiz Carlos Ribeiro Jr. 2009. Extração automática de termos compostos para construção de ontologias: Um experimento na área da saúde. *Reciis - Revista Eletronica de Comunicação Informação e Inovação em Saúde*, 3:76–88.

Villada Moirón, B. e J. Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. Em *Proceedings of the*

Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006), pp. 33–40, Trento, Italy.

Villavicencio, A., H. M. Caseli, e A. Machado. 2009. Identification of multiword expressions in technical domains: Investigating statistical and alignment-based approaches. Em *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology*, pp. 1–9.

Villavicencio, A., V. Kordoni, Y. Zhang, M. Idiart, e C. Ramisch. 2007. Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. Em *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1034–1043, Prague, June, 2007.

Vogel, S., H. Ney, e C. Tillmann. 1996. HMM-based word alignment in statistical translation. Em *COLING'96: The 16th International Conference on Computational Linguistics*, pp. 836–841, Copenhagen, August, 1996.

Witten, Ian H. e Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Classificação Automática de Textos por Período Literário Utilizando Compressão de Dados Através do PPM-C

Bruno Barufaldi
Departamento de Informática,
Universidade Federal da Paraíba (UFPB)
bruno.barufaldi@gmail.com

Milton Marques Junior
Departamento de Letras Clássicas Vernáculas, Universidade Federal da Paraíba (UFPB)
marquesjr45@hotmail.com

Eduardo Freire Santana
Departamento de Informática,
Universidade Federal da Paraíba (UFPB)
eduardo.freire.87@gmail.com

JanKees van der Poel
Programa de Pós-graduação em Engenharia Mecânica, Universidade Federal da Paraíba (UFPB)
jkvdpoel@yahoo.com.br

José Rogério Bezerra Barbosa Filho
Departamento de Informática,
Universidade Federal da Paraíba (UFPB)
jose.rogerio.filho@gmail.com

Leonardo Vidal Batista
Programa de Pós-graduação em Informática, Universidade Federal da Paraíba (UFPB)
leonardo@di.ufpb.br

Resumo

Métodos e técnicas para compressão de dados têm sido utilizados para o reconhecimento de padrões, incluindo a classificação automática de textos. A eficiência do método Prediction by Partial Matching (PPM) como classificador textual já foi comprovada em diversos trabalhos, entre eles a atribuição de autoria para textos em português. As classes utilizadas no processo de classificação não precisam ficar restringidas a apenas um autor. Ao incluir dois ou mais autores numa mesma classe pode-se definir um estilo literário. Esse trabalho objetiva a aplicação do modelo estatístico PPM-C para a classificação de textos dos períodos literários da literatura brasileira.

1. Introdução

O aumento da popularidade da Internet nos últimos anos fez com que o número de dados circulando na rede crescesse abruptamente. Imagens digitais, textos e arquivos de áudio são armazenados e compartilhados entre usuários, muitas vezes com seu conteúdo marcado incorretamente ou de forma não confiável. A maioria das ferramentas de busca na *World Wide Web* utiliza algoritmos para filtrar e detectar parâmetros textuais passados pelo usuário a fim de recuperar informação de forma automática, sem levar em consideração o conteúdo daquilo que se procura. Isso acarreta em um excesso de informações circulando atualmente na rede mundial que não conta com mecanismos inteligentes de busca ou classificação de conteúdo.

O Reconhecimento de Padrões é a disciplina que tem como objetivo a classificação de objetos em um determinado número de categorias ou classes [Theodoris e Koutroumbas, 2006]. Assim como os sinais da natureza estão sujeitos a regras e geram padrões, um texto – que pode ser entendido como um sinal – está sujeito a regras de linguagem e também gera padrões. Por esse motivo, o reconhecimento de padrões pode ser utilizado na Classificação Automática de Textos (CAT). As utilizações da CAT não se limitam em apenas melhorar mecanismos de busca, mas também pode ser utilizada em diversas outras aplicações. Dentre essas aplicações podem ser citadas a filtragem de *spam*, a identificação de conteúdo adulto, a organização de documentos em bibliotecas digitais e quaisquer outras aplicações que necessitem de seleção e organização de documentos.

O método de compressão de dados sem perdas *Prediction by Partial Match* (PPM) constrói um modelo estatístico a partir de uma determinada fonte de informação [Cleary e Witten, 1984]. Esse modelo é usado para diminuir a entropia dos símbolos da fonte e,

assim, obter uma compressão sobre o sinal. Isso significa que quanto mais se conhece sobre a fonte, menor é a surpresa que seus símbolos causam ao aparecer e menor a quantidade de dados necessária para representá-los. Este método pode ser utilizado para o reconhecimento de padrões mapeando sinais (objetos) para modelos (classes) que obtiverem maior compressão sobre a fonte de informação [Coutinho et al., 2005].

A eficiência do PPM na classificação de textos já foi provada, superando inclusive classificadores Naïve Bayes, cujos modelos são baseados em palavras [Teahan e Harper, 2001]. A utilização de técnicas de compressão para classificação possui a vantagem de não necessitar extrair características dos textos além de registros de seqüências de caracteres [Coutinho et al., 2005]. Características como tamanho médio das palavras, tamanho do dicionário ou número de palavras repetidas não são utilizadas tornando o método mais simples. Atualmente, o PPM está consolidado como um meio efetivo de atribuição de autoria para textos [Stamatatos, 2009][Coutinho et al., 2005].

Este trabalho tem como objetivo utilizar o método de compressão de dados PPM-C para classificação automática de textos de escolas literárias brasileira. As escolas literárias Barroco, Arcadismo, Romantismo e Realismo foram contempladas no escopo deste trabalho.

2. Fundamentação Teórica

2.1 *Prediction by Partial Matching* (PPM) e Codificação Aritmética

A predição por emparelhamento parcial (*Prediction by Partial Matching*) é um dos mais eficientes métodos utilizados para compressão de dados sem perdas, sendo atualmente considerado o estado da arte nesta área. O PPM é um método para compressão de dados que mantém atualizado um modelo estatístico contextual adaptativo de uma fonte de informação [Salomon, 2007]. O modelo armazena a ocorrência de seqüências de símbolos e procura associar novas seqüências com aquelas anteriormente armazenadas. A cada símbolo lido, novas seqüências são armazenadas. O PPM realiza a predição levando em consideração os últimos símbolos lidos ao invés de trabalhar com as freqüências de cada símbolo de forma isolada. Neste trabalho foi utilizado o PPM-C [Moffat, 1990], uma das variantes do PPM.

O modelo PPM utiliza um conjunto de no máximo k símbolos precedentes como contexto para estimar a distribuição de probabilidades condicionais para o próximo símbolo da mensagem. Este modelo alimenta um codificador aritmético [Witten et al., 1987], que atribui a cada símbolo um número de bits inversamente proporcional à sua probabilidade.

Dado um novo símbolo S a ser comprimido em um contexto C_k de tamanho k, o PPM utiliza seu modelo estatístico para calcular a probabilidade condicional da ocorrência do símbolo S e passa essa probabilidade para o codificador aritmético. Caso não haja ocorrência do símbolo S no contexto C_k, um símbolo especial de ESCAPE é codificado e é realizada uma nova busca no contexto C_{k-1}, que é a seqüência de símbolos C_k reduzida de um símbolo. Caso o símbolo não seja encontrado em nenhum dos contextos, ele é codificado utilizando um modelo que considera equiprováveis todos os símbolos possíveis de ocorrer. Após a codificação do símbolo, o modelo atualiza as probabilidades condicionais do símbolo S. Este processo é repetido para cada novo símbolo a ser comprimido.

No final do processo, o codificador aritmético gera uma seqüência de símbolos codificados. Quanto menor for o tamanho dessa seqüência em relação ao tamanho do texto de entrada, maior será a compressão obtida.

A Tabela 1 mostra o modelo gerado pelo PPM após comprimir a cadeia de caracteres “hocuspocus”, utilizando contexto com tamanho máximo de $k = 2$. Na tabela a seguir c indica o contador do símbolo em questão (número de vezes que o símbolo apareceu num determinado contexto) e p sua probabilidade, derivada do seu contador.

Contexto k = 2				**Contexto k = 1**				**Contexto k = 0**		
Contexto	Símbolo	c	P	Predição	Símbolo	c	p	Predição	c	P
ho	c Esc	1 1	$^1/_2$ ½	h	o Esc	1 1	$^1/_2$ $^1/_2$	h o c u s p	1 2 2 2 2 1	$^1/_{10}$ $^2/_{10}$ $^2/_{10}$ $^2/_{10}$ $^2/_{10}$ $^1/_{10}$
oc	u Esc	2 1	$^2/_3$ $^1/_3$	o	c Esc	2 1	$^2/_3$ $^1/_3$			
cu	s Esc	2 1	$^2/_3$ $^1/_3$	c	u Esc	2 1	$^2/_3$ $^1/_3$			
us	p Esc	1 1	$^1/_2$ ½	u	s Esc	2 1	$^2/_3$ $^1/_3$			
sp	o Esc	1 1	$^1/_2$ ½	s	p Esc	1 1	$^1/_2$ $^1/_2$			
po	c Esc	1 1	$^1/_2$ ½	p	o Esc	1 1	$^1/_2$ $^1/_2$			

Tabela 1: Modelo PPM depois do processamento da cadeia de caracteres hocuspocus.

O PPM-C é uma variante do PPM que utiliza o mecanismo de exclusão. Esse mecanismo remove temporariamente os símbolos cuja ocorrência é impossível em um determinado contexto no momento da codificação. Isso aumenta a probabilidade dos símbolos que de fato serão codificados, melhorando o modelo de compressão. Quando um símbolo não é encontrado em um determinado contexto *k* e o ESCAPE é codificado, todos os símbolos deste contexto são removidos temporariamente do contexto *k-1*, onde a nova busca será realizada. Isto acontece porque a probabilidade destes símbolos tornam-se nula, visto que já apareceram em um contexto superior e não eram o objeto de procura.

Em geral, o PPM utiliza a codificação aritmética. Nela, a mensagem é representada inicialmente dentro do intervalo real [0,1). Este intervalo é alterado à medida que os símbolos e suas probabilidades são inseridos no codificador. Quanto maior o tamanho da mensagem, menor o intervalo e mais casas decimais são necessárias para sua representação [Witten et al., 1987].

2.2 Literatura Brasileira

Literatura é a arte da palavra que atua como instrumento de comunicação e de interação social. Suas primeiras manifestações no Brasil ocorreram durante o período colonial (de 1500 a 1822), fortemente influenciada pela cultura portuguesa, tendo principalmente o propósito informativo. Atualmente, os poetas e prosadores se expressam de maneira diversificada, contribuindo com a arte mesmo sem que haja um projeto literário em comum [Cereja e Magalhães, 2002]. Apesar da origem da literatura brasileira ser bastante recente, comparada a outros países, a produção de textos literários no Brasil merece destaque e reconhecimento.

Um estilo literário pode ser entendido como um conjunto de textos com diversas características em comum. Apesar de não serem classificados como um mesmo estilo literário, o Barroco e o Arcadismo no Brasil são encontrados numa época, conhecida como fase luso-brasileira. Houve ecos do Barroco europeu entre os séculos XVII e XVIII, e sua transição para o Arcadismo buscou por esquemas rítmicos mais graciosos de forma específica e de menor beleza [Bosi, 2007]. Algumas características podem ser ressaltadas, tais como o cultismo e o conceptismo no Barroco, e o bucolismo e a simplicidade no conteúdo do Arcadismo.

Os períodos literários do Realismo e do Romantismo consolidaram-se no país e tiveram a contribuição de textos de diversos autores consagrados. Um dos traços essenciais do Romantismo brasileiro é o nacionalismo, que explora características como o indianismo, o regionalismo e a pesquisa histórica. Já os escritores realistas são motivados pelas teorias científicas e filosóficas da época, desejando retratar o homem e a sociedade em sua totalidade [Cereja e Magalhães, 2002].

3. Materiais e Métodos

Para classificar os textos, foram utilizadas quatro classes, as quais correspondem aos períodos literários Barroco, Arcadismo, Romantismo e Realismo. Os textos escolhidos estão listados a seguir, juntamente com seus respectivos autores e períodos literários.

• **Barroco**: Antonio Vieira (Sermão *da Primeira Dominga do Advento, Sermão da Sexagésima, Sermão do Espírito Santo e Sermão do Bom Ladrão*) e Gregório de Matos (*Coletânea de Obras Líricas, Coletânea de Obras Satíricas e Coletânea de Obras Religiosas*);

• **Arcadismo**: Alvarenga Peixoto (*Coletânea de Obras*), Cláudio Manoel da Costa (*Poemas Escolhidos*), Basílio da Gama (*O Uraguai*) e Tomás Antônio Gonzaga (*Cartas Chilenas, Marília de Dirceu*);

• **Romantismo**: Joaquim Manuel de Macedo (*O Moço Loiro, A Moreninha, Os Dois Amores*), José de Alencar (*O Guarani, Senhora, Ubirajara, Iracema*), Machado de Assis (*A Mão e a Luva, Helena, Iaiá Garcia*), Manuel Antônio de Almeida (*Memórias de um Sargento de Milícias*) e Bernardo Guimarães (*A Escrava Isaura*);

• **Realismo**: Adolfo Caminha (*O Bom Crioulo, A Normalista*), Aluízio Azevedo (*O Mulato, O Homem, O Coruja*), Franklin Távora (*O Cabeleira*), Júlio Ribeiro (*A Carne*), Machado de Assis (*Memórias Póstumas de Brás Cubas, Dom Casmurro*) e Raul Pompéia (*O Ateneu, 14 de Julho na Roça, As Jóias da Coroa, Uma Tragédia no Amazonas*).

A coletânea de textos de Alvarenga Peixoto foi feita a partir dos poemas presentes no livro "A poesia dos inconfidentes: poesias completas de Cláudio Manuel da Costa, Tomás Antônio Gonzaga e Alvarenga Peixoto" de Domício Proença Filho. As coletâneas de Gregório de Matos foram obtidas do livro "Poemas escolhidos: Gregório de Mattos" de José Miguel Wisnik [Mattos, 1999]. O restante dos textos foi obtido através do sítio Domínio Público [Portal Domínio Público, 2009] e do sítio Biblioteca Digital de Literatura do NUPILL [Biblioteca Digital de Literatura, 2009].

O processo de classificação pode ser divido em três etapas: formatação dos textos, construção dos modelos e comparação da razão de compressão.

3.1 Formatação dos Textos

Antes da elaboração dos modelos e classificação, os textos passam por uma fase de padronização. São eliminados acentuação, grande parte das pontuações, tabulação e quebras de linha, restando apenas as 26 letras do alfabeto (minúsculas) e os caracteres de espaçamento e ponto. Esta etapa tem por finalidade descartar símbolos pouco relevantes ou mesmo que dificultem a classificação correta, enquanto preserva a essência do texto, as palavras e frases.

3.2 Construção dos modelos

Os modelos criados são compostos por informações estatísticas sobre a ocorrência de símbolos dentro de determinados contextos que serão utilizadas para compressão. Uma vez criado, o modelo usado para classificação não será alterado.

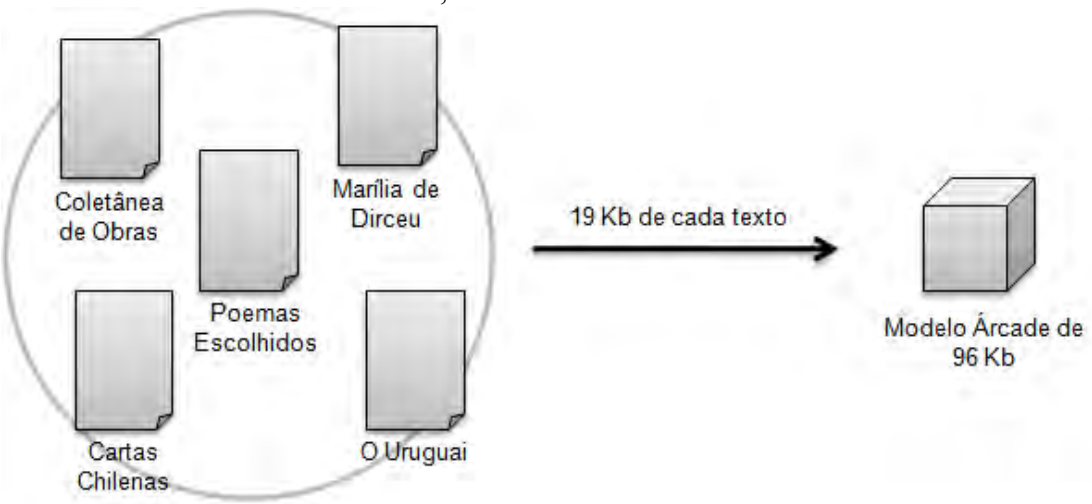


Figura 1: Construção de um modelo de 96kb a partir de cinco textos.

Para cada texto a ser classificado, são gerados quatro modelos, sendo um para cada classe. Os modelos são construídos utilizando todos os textos presentes em uma classe. São lidos os *n* primeiros símbolos de cada texto, onde *n* é determinado pela razão entre o tamanho do treinamento e o número de textos por classe. O tamanho do treinamento é a quantidade de informação que será lida para a construção do modelo, independentemente de quantos textos existam em uma determinada classe. Por exemplo, para um treinamento de tamanho 96kb e uma classe com cinco textos, os 19kb iniciais de cada texto serão utilizados para a construção do modelo.

O texto que se deseja classificar não deve ser utilizado como parte do treinamento para a construção do modelo. Isto é feito para que os modelos não possuam nenhuma informação sobre o texto desconhecido, garantindo assim, uma classificação apenas por afinidade com os demais textos.

Os testes realizados utilizaram tamanhos de treinamento de 8kb, 16kb, 48kb, 96kb e 128kb.

3.3 Comparação da Razão de Compressão

O texto a ser classificado é comprimido utilizando-se cada um dos quatro modelos PPM gerados. O texto será classificado como pertencente à classe cujo modelo obtiver maior compressão. Para a classificação, foram realizados testes variando os tamanhos máximos de contexto entre 0 e 10.

4. Resultados

Os testes realizados variaram a quantidade de informação para treinamento e o tamanho máximo de contextos utilizados pelo PPM. A Figura 3 mostra que a maior taxa média de acerto foi de 85%, encontrada ao se utilizar 48kb para o treinamento dos modelos. Este índice médio de acertos representa a média de acertos encontrada em cada contexto testado, do contexto $k = 0$ até o contexto $k = 10$.

A Figura 4 mostra os resultados obtidos separados por tamanho máximo de contexto e utilizando 48kb para treinamento dos modelos. Pode-se observar que o melhor resultado foi encontrado com tamanho máximo de contexto $k = 4$. Uma pequena queda no desempenho do classificador ocorreu quando utilizados contextos com tamanhos maiores que 4. Tal fato acontece devido à natureza do PPM, cuja curva de aprendizado tem característica assintótica e pára de crescer a partir de certo contexto. [Salomon, 2007].

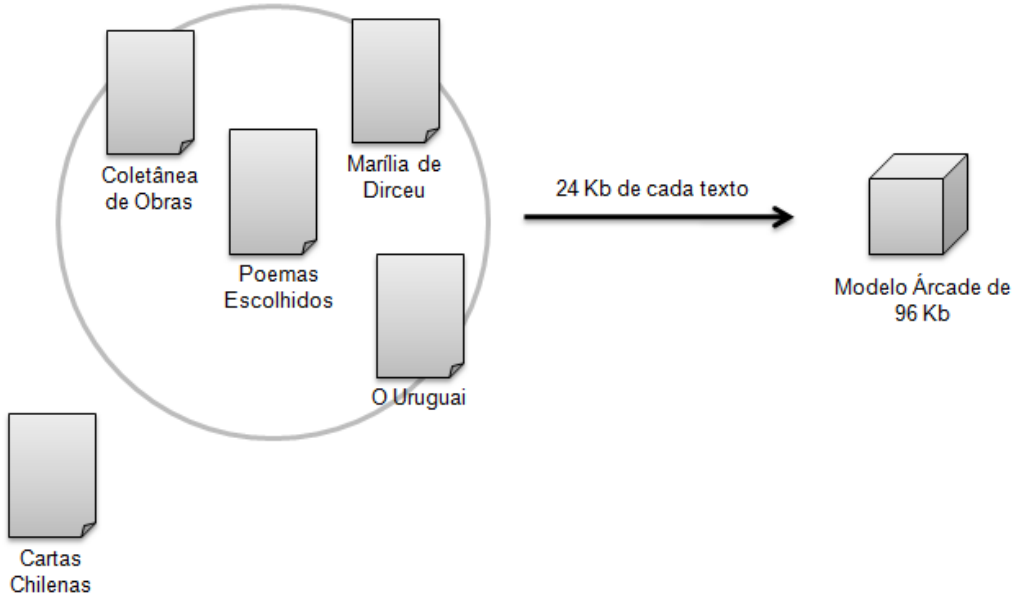


Figura 2: O texto árcade "Cartas Chilenas" não participa da criação do modelo árcade durante sua classificação.

A Tabela 2 é a tabela de confusão obtida na classificação quando utilizados 48kb de informação para treinamento e um contexto k=4. Desta tabela pode-se inferir que apenas três textos foram classificados erroneamente: a coletânea de obras líricas de Gregório de Matos, Helena de Machado de Assis e Memórias de Um Sargento de Milícias de Manuel Antônio Bandeira. Possíveis razões para esse erro na classificação são discutidas na próxima seção deste trabalho.

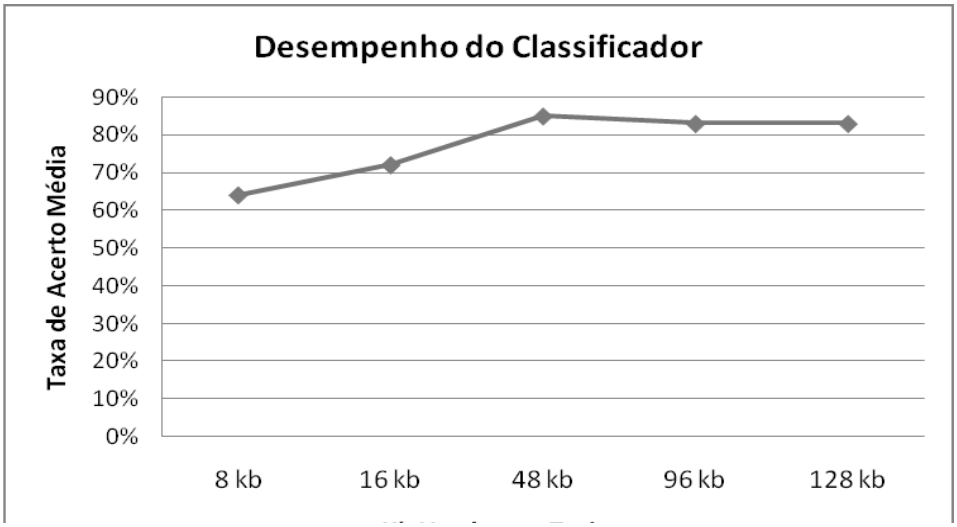


Figura 3: Gráfico de acerto médio por tamanho de treinamento.

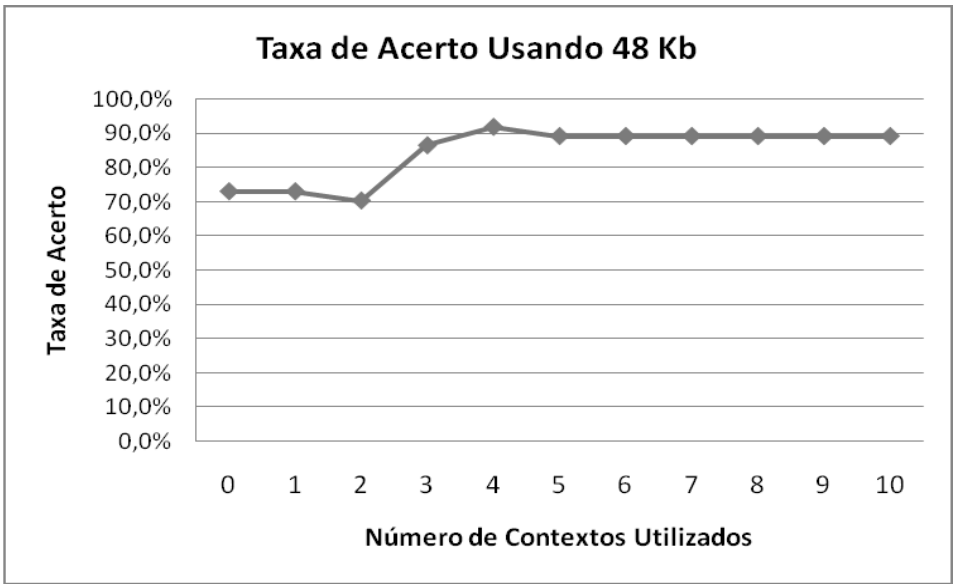


Figura 4: Gráfico de acerto obtido na classificação usando 48kb e diferentes contextos.

Estilos Literários/Obras \ Classificadas como		Barroco	Arcadismo	Romantismo	Realismo
Barroco	Sermão da Primeira Dominga do Advento	X			
	Sermão da Sexagésima	X			
	Sermão do Espírito Santo	X			
	Sermão do Bom Ladrão	X			
	Coletânea de obras líricas		X		
	Coletânea de obras satíricas	X			
	Coletânea de obras religiosas	X			
Arcadismo	Coletânea		X		
	Poemas Escolhidos		X		
	O Uraguai		X		
	Cartas Chilenas		X		
	Marília de Dirceu		X		
Romantismo	O Moço Loiro			X	
	A Moreninha			X	
	Os Dois Amores			X	
	O Guarani			X	
	Senhora			X	
	Ubirajara			X	
	Iracema			X	
	A Mão e a Luva			X	
	Helena				X
	Iaiá Garcia			X	
	Memórias de um Sargento de Milícias				X
	A Escrava Isaura			X	
Realismo	O Bom Crioulo				X
	A Normalista				X
	O Mulato				X
	O Homem				X
	O Coruja				X
	O Cabeleira				X
	A Carne				X
	Memórias Póstumas de Brás Cubas				X
	Dom Casmurro				X
	O Ateneu				X
	14 de Julho na Roça				X
	As Jóias da Coroa				X
	Uma Tragédia no Amazonas				X

Figura 2: Tabela de confusão da classificação usando 48 Kb e contexto k = 4.

5. Conclusões e Discussões

Obteve-se uma taxa de acerto máxima de 91,89% utilizando 48kb de informação para o treinamento e modelos PPM com tamanho máximo de contexto $k = 4$. Com esses parâmetros de treinamento e compressão, ocorreram apenas três classificações incorretas: *Helena* (Machado de Assis), Gregório de Matos no estilo lírico e *Memórias de Um Sargento de Milícias* (Manuel Antônio de Almeida). Estes erros podem ser atribuídos às particularidades presentes nessas obras.

Machado de Assis, romancista consagrado entre especialistas da área, tem características marcantes que iniciaram o movimento realista no país. Apesar de a classificação pelo PPM obter resultados satisfatórios em suas obras, o marcante "estilo machadiano" pode influenciar nos resultados, considerando que textos do autor foram utilizados tanto na construção do modelo romântico quanto na construção do modelo realista. Como exemplo disto, o romance *Helena* foi classificado como realista, um equívoco que não se repetiu em outros textos de sua autoria.

Gregório de Matos, um dos autores barrocos utilizados na pesquisa, possui características distintas dos outros autores. Suas obras foram selecionadas e associadas a estilos satíricos, líricos e religiosos. Contudo, Gregório de Matos no estilo lírico persiste na classificação árcade com uma diferença de compressão em torno de 3% para a compressão obtida pelo modelo barroco. A utilização de referentes clássicos e algumas metáforas com elementos da natureza nos textos líricos podem ter influenciado sua classificação como árcade.

A obra *Memórias de um Sargento de Milícias*, romance de Manuel Antônio de Almeida, foi classificada como sendo de estilo realista. Isso pode se justificar por esta possuir características dos estilos romântico e realista. Apesar de esta sua obra ser do início do Romantismo, possui características que antecipam o Realismo e assim foi na maioria dos testes classificada como realista. Contudo, em todas as classificações incorretas o modelo romântico conseguiu obter a segunda melhor compressão, com uma diferença de 1% para o modelo realista.

Uma das maiores dificuldades encontradas durante a pesquisa foi a pouca disponibilidade de textos originais no formato digital. Sendo assim, atualmente a pesquisa está focada em obras barrocas, árcades, românticas e realistas. Apesar disso, tem-se a perspectiva de refinar o modelo criado através da inserção de novos textos e estilos literários.

Trabalhos futuros irão estudar a utilização de atributos textuais para auxiliar a classificação automática de textos em conjunto com o PPM. Esta abordagem investigaria uma possível melhora na classificação dos textos levando em consideração atributos como tamanho médio das palavras, riqueza vocabular e entropia dos bigramas.

Cabe aqui salientar que não existem na literatura pesquisas utilizando o PPM (ou quaisquer outros métodos) para classificar textos da literatura brasileira por **período literário**. Por esta razão, não foram realizadas comparações entre este trabalho e outras abordagens para classificação.

O Professor Milton Marques Junior, doutor em Letras pela Universidade Federal da Paraíba, auxiliou na pesquisa que culminou com o presente artigo, colaborando com seus conhecimentos na área. Por ser um especialista, o professor orientou os alunos através da disponibilização de textos e discussões relacionadas à literatura brasileira.

Referências

Biblioteca Digital de Literatura. Núcleo de Pesquisas em Informática, Literatura e Lingüística da UFSC (NUPILL). Disponível em <http://www.literaturabrasileira.ufsc.br/>. Acessado em 24 de maio de 2009.

Bosi, A. (2007). "História concisa da Literatura Brasileira", Editora Cultrix, 44ª Edição.

Cereja, W. R.; Magalhães, T. C. (2002). "Literatura Brasileira", Editora: Atual Editora, 2ª Edição.

Cleary, J.G.; Witten, I. H. (1984). "Data compression using adaptive coding and partial string matching", IEEE Transactions on Communications, v. 32, n. 4, pp. 396-402.

Coutinho, B. C.; Macedo, J. L. de M.; Júnior, A. R.; Batista, L. V. (2005). "Atribuição de Autoria usando PPM". In: III Workshop em Tecnologia

da Informação e da Linguagem Humana, 2005, São Leopoldo. Anais do XXV Congresso da Sociedade Brasileira de Computação, 2005. v. 1. p. 2208-2217.

Mattos, G. (1999). "Poemas escolhidos: Gregório de Mattos"; seleção, introdução e notas de José Miguel Wisnik. 7ª Edição. São Paulo: Cultrix.

Moffat, A. (1990). "Implementing the PPM data compression scheme". IEEE Transactions on Communications, v. 38, n.11, pp. 1917-1921.

Peixoto, A. (1996). "Poesias". In: "A poesia dos inconfidentes: poesias completas de Cláudio Manuel da Costa, Tomás Antônio Gonzaga e Alvarenga Peixoto"; organização de Domício Proença Filho; artigos, ensaios e notas de Eliana S. Muzzi, João Ribeiro, Letícia Malard, Lúcia Helena, Luciano Figueiredo, Manuel Bandeira, Manuel Rodrigues Lapa, Melânia Silva de Aguiar e Paulo Roberto Dias Pereira. Rio de Janeiro: Nova Aguilar.

Portal Domínio Público. Disponível em <http://www.dominiopublico.gov.br/>. Acessado em 24 de maio de 2009.

Salomon, D. (2007). Data Compression, Springer-Verlag, 4th Edition.

Stamatatos, E. (2009). "A survey of modern authorship attribution methods". Journal of the American Society for Information Science and Technology, v. 60, n. 3, pp. 538-556.

Teahan, W. J.; Harper, D. J. (2003). "Using compression-based language models for text categorization". In: W. B. Croft and J. Lafferty (Eds.), Language Modeling for Information Retrieval, pp. 141-166. Kluwer Academic Publishers, 2003.

Theodoris, S.; Koutroumbas, K. (2006), "Pattern Recognition", 3rd Edition.

Witten, I. H.; Neal, R. M.; Cleary, J. G. (1987). "Arithmetic Coding For Data Compression". In Journal of the ACM, v. 30, n. 6.

Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português

Carolina Evaristo Scarton, Sandra Maria Aluísio
NILC – ICMC – Universidade de São Paulo
São Carlos – SP, Brasil
{carolina@grad.,sandra@}icmc.usp.br

Resumo

Este artigo apresenta o projeto de adaptação de métricas da ferramenta Coh-Metrix para o português do Brasil (Coh-Metrix-Port). Descreve as ferramentas de processamento de língua natural para o português que foram utilizadas, juntamente com as decisões tomadas para a criação da Coh-Metrix-Port. O artigo traz duas aplicações da ferramenta Coh-Metrix-Port: (i) a avaliação de textos jornalísticos e sua versão para crianças, mostrando as diferenças entre os textos supostamente complexos e textos simples, isto é, os textos reescritos; (ii) a criação de classificadores binários (com córpus de textos dedicados a adultos e crianças), analisando a influência do gênero no desempenho destes classificadores (gêneros jornalístico e de divulgação científica) e de textos de outras fontes. A precisão do melhor classificador treinado foi conseguida com a implementação de *Support Vector Machines* (SMO) do WEKA e foi de 97%. Como as métricas desta ferramenta ajudam a discriminar com boa precisão textos dedicados a adultos e a crianças, acreditamos que elas possam também ajudar a avaliar se textos disponíveis na Web são simples o suficiente para serem inteligíveis por analfabetos funcionais e pessoas com outras deficiências cognitivas, como afasia e dislexia, e também para crianças e adultos em fase de letramento e assim permitir o acesso dos textos da Web para uma gama maior de usuários.

1. Introdução

Leffa (1996) apresenta os aspectos essenciais no processo de compreensão de leitura de um texto: o texto, o leitor e as circunstâncias em que se dá o encontro. Ele destaca que o levantamento feito em estudos publicados até a data de seu trabalho mostra que a compreensão da leitura envolve diversos fatores que podem ser divididos em três grandes grupos: i) relativos ao texto, ii) relativos ao leitor e, iii) relativos à intervenção pedagógica. Entre os fatores relativos ao texto, destacam-se, tradicionalmente, a legibilidade (apresentação gráfica do texto) e a inteligibilidade (uso de palavras freqüentes e estruturas sintáticas menos complexas). É bem sabido que sentenças longas, com vários níveis de subordinação, cláusulas embutidas (relativas), sentenças na voz passiva, uso da ordem não canônica para os componentes de uma sentença, além do uso de palavras de baixa frequência aumentam a complexidade de um texto para leitores com problemas de compreensão como, por exemplo, analfabetos funcionais, afásicos e dislexos (Siddharthan, 2002). Atualmente, há também, uma preocupação com a macroestrutura do texto além da microestrutura, em que outros fatores são visto como facilitadores da compreensão como a organização do texto, coesão, coerência, o conceito do texto sensível ao leitor. Este último apresenta características que podem facilitar a compreensão como proximidade na anáfora, o uso de marcadores discursivos entre as orações, a preferência por definições explícitas ou a apresentação de informações completas (Leffa, 1996).

Neste artigo, nosso foco é principalmente no texto e como suas características podem ser utilizadas para se avaliar a dificuldade ou facilidade de compreensão de leitura. Segundo DuBay (2004), até 1980 já existiam por volta de 200 fórmulas superficiais de inteligibilidade, para a língua inglesa. As fórmulas mais divulgadas no Brasil são o *Flesch Reading Ease* e o *Flesch-Kincaid Grade Level*, pois se encontram disponíveis em processadores de texto como o MSWord. Entretanto, as fórmulas de inteligibilidade superficiais são limitadas. Estas duas acima se baseiam somente no número de palavras das

a) *Sometimes you did not pick the right letter. You did not click on the letter 'd'.*
b) *Sometimes you did not pick the right letter. For example, you did not click on the letter 'd'.*
c) *Sometimes you did not pick the right letter. You did not, for example, click on the letter 'd'.*
d) *Sometimes you did not pick the right letter – you did not click on the letter 'd', for example.*
e) *You did not click on the letter 'd'. Sometimes you did not pick the right letter.*
f) *Sometimes you did not pick the right letter. For instance, you did not click on the letter 'd'.*

Figura 1: Exemplo dos problemas do índice Flesch (Willians, 2004)

sentenças e no número de sílabas por palavra para avaliar o grau de dificuldade/facilidade de um texto. Para exemplificar nossa afirmação, considere os exemplos em inglês de (a) – (f) apresentados na Figura 1, retirados de Williams (2004). De acordo com o índice Flesch, os itens (a) e (e) são os mais inteligíveis, com (b) e (c) em segundo lugar, seguidos por (f) e, em último, (d).

Porém, (a) e (e) são os exemplos menos compreensíveis, pois eles não contêm marcadores de discurso para explicar que a relação entre as duas sentenças é de exemplificação, isto é, uma é um exemplo para outra.

As fórmulas de inteligibilidade superficiais não conseguem capturar a coesão e dificuldade de um texto (McNamara et al., 2002) nem avaliar mais profundamente as razões e correlações de fatores que tornam um texto difícil de ser entendido. Para o inglês, a ferramenta Coh-Metrix[1] (Graesser et al., 2004; McNamara et al., 2002; Crossley et al., 2007) foi desenvolvida com a finalidade de capturar a coesão e a dificuldade de um texto, em vários níveis (léxico, sintático, discursivo e conceitual). Ela integra vários recursos e ferramentas, utilizados na área de Processamento de Língua Natural (PLN): léxicos, *taggers*, *parsers*, lista de marcadores discursivos, entre outros. Para o português do Brasil, a única ferramenta de análise da inteligibilidade de textos adaptada foi o índice Flesch (Martins et al., 1996), que, como dito acima, é um índice superficial. A língua portuguesa já dispõe de várias ferramentas e recursos de PLN que poderiam ser utilizados para a criação de uma ferramenta que analisasse vários níveis da língua e fosse calibrada com textos de vários gêneros, por exemplo, jornalísticos e científicos, tanto os adaptados para crianças como os dedicados a adultos.

Neste artigo, apresentamos uma análise das fórmulas de inteligibilidade e das ferramentas que utilizam métodos de PLN para a tarefa, como é o caso do Coh-Metrix (Seção 2); o processo de adaptação de um conjunto das métricas do Coh-Metrix para o português (Seção 3); e um estudo das aplicações do Coh-Metrix-Port (Seção 4). Este estudo é dividido em quatro partes: apresentação dos córpus[2] utilizados (Seção 4.1), avaliação de textos jornalísticos e sua versão reescrita para crianças (Seção 4.2) e a criação de classificadores de textos "simples" (para crianças) e "complexos" (para adultos) (Seção 4.3). O trabalho descrito neste artigo faz parte de um projeto maior que envolve a Simplificação Textual do Português para Inclusão e Acessibilidade Digital – o PorSimples (Aluisio et al., 2008a, 2008b; Caseli et al., 2009, Candido Jr. et al., 2009) que propõe o desenvolvimento de tecnologias para facilitar o acesso à informação dos analfabetos funcionais e, potencialmente, de pessoas com outras deficiências cognitivas, como afasia e dislexia.

2. Análise da Inteligibilidade: as métricas do Coh-Metrix e de trabalhos relacionados

2.1 Índice Flesch

Os índices *Flesch Reading Ease* e o *Flesch-Kincaid Grade Level* são fórmulas que avaliam, superficialmente, a inteligibilidade de um texto. Apesar de serem superficiais, elas merecem destaque, pois a primeira é a única métrica de inteligibilidade já adaptada para o português (Martins et al., 1996) e incorpora o conceito de séries escolares da segunda. Estas

[1] http://cohmetrix.memphis.edu/cohmetrixpr/index.html

[2] Neste trabalho escolhemos o aportuguesamento da palavra corpus/corpora para córpus/córpus.

métricas são consideradas superficiais, pois medem características superficiais do texto, como o número de palavras em sentenças e o número de letras ou sílabas por palavra:

Flesch reading Ease
A saída desta fórmula é um número entre 0 e 100, com um índice alto indicando leitura mais fácil:

$$206.835 – (1.015 \text{ x ASL}) – (84.6 \text{ x ASW})$$

em que ASL = tamanho médio de sentenças (o número de palavras dividido pelo número de sentenças) e ASW = número médio de sílabas por palavra (o número de sílabas dividido pelo número de palavras)

Flesch-Kincaid Grade Level
Esta fórmula converte o índice *Reading Ease Score* para uma série dos Estados Unidos:

$$(0.39 \text{ x ASL}) + (11.8 \text{ x ASW}) – 15.59$$

Para o português, a adaptação do *Flesch Reading Ease* resultou na fórmula:

$$248.835 – (1.015 \text{ x ASL}) – (84.6 \text{ x ASW})$$

que corresponde à fórmula do *Flesch Reading Ease* somada com o número 42 que, de acordo com Martins et al. (1996), é, na média, o número que diferencia textos em inglês de textos em português. Os valores desse índice variam entre 100-75 (muito fácil), 75-50 (fácil), 50-25 (difícil) e 25-0 (muito difícil), que correspondem, respectivamente, às duas séries da educação primária (1-4 e 5-8), secundária (9-11) e ensino superior.

2.2 As métricas do Lexile

O *framework* Lexile[3] (Burdick e Lennon, 2004) é uma abordagem científica para leitura e tamanho de textos. Ele consiste de dois principais componentes: a medida Lexile e a escala Lexile. O primeiro é a representação numérica de uma habilidade do leitor ou de uma dificuldade do texto, ambos seguidos de "L" (Lexile). Já o segundo é uma escala para o domínio da leitura variando de 200L (leitores iniciantes) até 1700L (leitores avançados). As medidas Lexile são baseadas em dois fatores: frequência de palavras e tamanho da sentença, mais formalmente chamadas de dificuldade semântica e complexidade sintática. No framework Lexile há um programa de software (Lexile *Analyzer*) desenvolvido para avaliar a inteligibilidade de textos. Este programa avalia um texto dividindo-o em pedaços e estudando suas características de dificuldade semântica e sintática (frequência de palavras e tamanho da sentença). Sentenças longas e com palavras de baixa frequência possuem um alto valor Lexile, enquanto que sentenças curtas e com palavras de alta frequência possuem baixo valor Lexile. Já para avaliar os leitores é necessário utilizar algum método padronizado de teste de leitura reportando os resultados em Lexiles. Um exemplo é o *Scholastic Reading Inventory* (SRI[4]), que é uma avaliação padronizada desenvolvida para medir quão bem os estudantes leem textos explicativos e da literatura de várias dificuldades. Cada item deste teste consiste de uma passagem do texto de onde é retirada uma palavra ou frase e são dadas opções ao leitor para completar a parte que falta na passagem, de forma similar como fazem os testes de Cloze (Santos et al., 2002). Como um exemplo de aplicações das medidas Lexiles, podemos citar professores que podem utilizar as medidas para selecionar os textos que melhor se enquadrem no grau de inteligibilidade de seus alunos.

2.3 Coh-Metrix

A ferramenta Coh-Metrix, desenvolvida por pesquisadores da Universidade de Memphis, calcula índices que avaliam a coesão, a coerência e a dificuldade de compreensão de um texto (em inglês), usando vários níveis de análise lingüística: léxico, sintático, discursivo e conceitual. A definição de coesão utilizada é que esta consiste de características de um texto que, de alguma forma, ajudam o leitor a conectar mentalmente as idéias do texto (Graesser et al., 2003). Já coerência é definida como características do texto (ou seja, aspectos de coesão) que provavelmente contribuem para a coerência da representação mental. O Coh-Metrix 2.0 é a versão livre desta ferramenta

[3] http://www.lexile.com

[4] http://www2.scholastic.com/

que possui 60 índices que vão desde métricas simples (como contagem de palavras) até medidas mais complexas envolvendo algoritmos de resolução anafórica. Vale comentar que a ferramenta Coh-Metrix possui cerca de 500 métricas que estão disponíveis somente para os pesquisadores da Universidade de Memphis (Graesser et al., 2008).

Os 60 índices estão divididos em seis classes que são: Identificação Geral e Informação de Referência, Índices de Inteligibilidade, Palavras Gerais e Informação do Texto, Índices Sintáticos, Índices Referenciais e Semânticos e Dimensões do Modelo de Situações. A primeira classe corresponde às informações que referenciam o texto, como título, gênero entre outros. A segunda contém os índices de inteligibilidade calculados com as fórmulas *Flesch Reading Ease* e *Flesch Kincaid Grade Level*. A terceira classe possui quatro subclasses: Contagens Básicas, Frequências, Concretude, Hiperônimos. A quarta possui cinco subclasses: Constituintes, Pronomes, Tipos e Tokens, Conectivos, Operadores Lógicos e Similaridade sintática de sentenças. A quinta classe está subdividida em três subclasses: Anáfora, Co-referência e *Latent Semantic Analysis (LSA)* (Deerwester et al., 1990). Por fim, a sexta classe possui quatro subclasses: Dimensão Causal, Dimensão Intencional, Dimensão Temporal e Dimensão Espacial.

Para todas essas métricas, vários recursos de PLN são utilizados. Para as métricas de freqüências, os pesquisadores utilizaram o CELEX, uma base de dados do *Dutch Centre for Lexical Information* (Baayen et al., 1995), que consiste nas frequências da versão de 17,9 milhões de palavras do córpus COBUILD. Para as métricas de concretude, o Coh-Metrix 2.0 utiliza o *MRC Psycholinguistics Database* (Coltheart, 1981), que possui 150.837 palavras com 26 propriedades psicolinguísticas diferentes para essas palavras. O cálculo de hiperônimos é realizado utilizando a WordNet (Fellbaum, 1998), sistema de referência lexical, que também é utilizado para calcular as métricas de dimensão causal, dimensão intencional e dimensão espacial. Para os índices sintáticos, foi utilizado o *parser* sintático de Charniak (Charniak, 2000). Os conectivos foram identificados utilizando listas com os conectivos classificados em várias classes. Por fim, a Análise Semântica Latente *(LSA)* recupera a relação entre documentos de texto e significado de palavras, ou semântica, o conhecimento base que deve ser acessado para avaliar a qualidade do conteúdo.

3. Adaptando o Coh-Metrix para o Português

Para a adaptação do Coh-Metrix para o português, chamada aqui de Coh-Metrix-Port, é necessário o estudo dos recursos e ferramentas de PLN existentes para o português. Infelizmente, o português não possui a vasta quantidade e variedade de recursos que existem para o inglês, porém, pretendemos integrar as ferramentas com os melhores desempenhos.

3.1 Ferramentas e Recursos de PLN Selecionados

Primeiramente, foi necessário o estudo e a escolha de um *tagger* e *parser*. Para o português do Brasil, um dos melhores *parsers* desenvolvidos é o PALAVRAS, criado durante o doutorado de Eckard Bick, e que está sendo constantemente melhorado (Bick, 2000). Embora use um conjunto de etiquetas bastante amplo, o *parser* alcança – com textos desconhecidos – a precisão de 99% em termos de morfossintaxe (classe de palavras e flexão), e 97-98% em termos de sintaxe (Bick, 2005). Porém, vale comentar que, dependendo de como se faz a avaliação e qual a versão do PALAVRAS utilizada estes valores poderão variar. No entanto, como no projeto Coh-Metrix-Port buscamos utilizar soluções livres sempre que possível, decidimos restringir o uso do PALAVRAS somente quando extremamente necessário.

As 34 métricas do Coh-Metrix que inicialmente decidimos implementar não utilizam a análise sintática total, somente a parcial (identificação de sintagmas), então não utilizamos o PALAVRAS.

Para a extração de sintagmas, utilizamos a ferramenta de Identificação de Sintagmas Nominais Reduzidos (Oliveira et al., 2006), que classifica cada palavra de acordo com o

tagset {I, O, B} (*In Noum Phrase, Out Noum Phrase, Border with Noum Phrase*). Para seu funcionamento, é necessário um *tagger* que pré-processa os textos. Foram disponibilizados pelo NILC[5] vários *taggers* treinados com vários córpus e *tagsets*. Dentre eles, escolhemos o MXPOST (Ratnaparkhi, 1996) que, em estudos anteriores, apresentou os melhores resultados. Submetemos o *tagger* MXPOST, treinado com o córpus e *tagset* do projeto Lácio-Web[6] (MacMorpho), a um teste comparativo com o *parser* PALAVRAS, usando 10 textos originais do jornal ZeroHora[7]. Após a conversão entre *tagsets*, construímos tabelas comparando as etiquetas palavra-a-palavra. Verificamos que o MXPOST erra em casos que a classificação da palavra é única (por exemplo, a palavra *daquele* é sempre uma contração da preposição *de* mais o pronome *aquele*, cuja etiqueta no MXPOST é sempre PREP|+). Por isso, construímos uma lista com as palavras de classificação única e sua respectiva etiqueta correta, para um pós-processamento. Porém, ainda tínhamos o problema dos erros que não podiam ser tratados, ou seja, erros em palavras de classes abertas. Por isso, decidimos utilizar um modelo para o MXPOST treinado com um t*agset* menor, chamado NILC *tagset*[8] que, mesmo tendo sido treinado com um córpus menor (10% do Mac-Morpho), apresentou melhor precisão. Entretanto, para o uso da ferramenta de Identificação de Sintagmas Nominais, é necessário utilizar o *tagset* do Lácio-Web e, portanto, neste caso, utilizaremos o tagger MXPOST com o *tagset* do Lácio-Web após o pós-processamento.

Outro recurso que precisou ser avaliado foi uma lista de palavras com suas repectivas frequências, vindas de um grande córpus do português. Decidimos utilizar a lista de frequências do córpus Banco do Português (BP)[9], compilada por Tony Sardinha da PUC-SP, com cerca de 700 milhões de unidades. Outros córpus como o córpus NILC e o de referência do Lácio-Web também foram cogitados, porém o BP é o córpus maior e mais balanceado existente para o português do Brasil, o que justifica nossa escolha. Um recurso necessário para o cálculo das métricas de concretude é uma lista de palavras com seu grau de concretude. Para o português, encontramos o trabalho de Janczura et al. (2007) que compilou uma lista com 909 palavras e seus respectivos valores de concretude. Vale ressaltar que este recurso é muito limitado, porém, até o momento, é o único que possuímos e, assim, decidimos não implementar a métrica de avaliação da concretude[10]. Outras listas de frequências que poderão ser utilizadas neste trabalho são as da Linguateca[11], que são de domínio público. O estudo comparativo destas listas será reservado para trabalhos futuros.

Estamos analisando também a MultiWordNet[12] (Pianta et al., 2002), que possui relações de hiperonímia para substantivos. O NILC[13] (Núcleo Interinstitucional de Linguística Computacional), ao qual os autores estão vinculados, irá adquirir a MultWordNet, o que torna possível a extração da métrica de hiperônimos de substantivos. Além da MultiWordNet, pretendemos analisar o PAPEL (Gonçalo Oliveira et al., 2008 e Santos et al., 2009), que é um recurso lexical baseado no Dicionário PRO da Língua Portuguesa[14]. O PAPEL também possui relações de hiperonímia para substantivos, nos permitindo, então, escolher entre os dois recursos (MultiWordNet ou PAPEL).

Outro recurso utilizado foi a WordNet.Br (Dias-da-Silva et al., 2002, Dias-da-Silva e Moraes, 2003 e Dias-da-Silva et al., 2008), desenvolvida nos moldes da WordNet de Princeton (WordNet.Pr[15]) (Fellbaum, 1998). A construção da base de relações da WordNet.Br é feita por meio de um alinhamento com a WordNet.Pr. Um linguísta começa o procedimento selecionando um verbo na lista do WordNet.Br; após a escolha

[5] http://www.nilc.icmc.usp.br/nilc/index.html
[6] http://www.nilc.icmc.usp.br/lacioweb/ConjEtiquetas.htm
[7] http://www.zh.com.br/
[8] http://www.nilc.icmc.usp.br/nilc/TagSet/ManualEtiquetagem.htm
[9] http://www2.lael.pucsp.br/corpora/bp/index.htm

[10] Mais detalhes podem ser encontrados em http://caravelas.icmc.usp.br/wiki/index.php/Carolina_Scarton
[11] http://www.linguateca.pt/lex_esp.html
[12] http://multiwordnet.itc.it/english/home.php
[13] http://www.nilc.icmc.usp.br
[14] Dicionáio PRO da Língua Portuguesa. Porto Editora, Porto (2005)
[15] http://wordnet.princeton.edu/

é realizada uma busca em um dicionário bilíngue *online* Português do Brasil - Inglês e o verbo selecionado é relacionado com sua versão em inglês. Assim, relações de hiperonímia podem ser herdadas automaticamente. Por exemplo: na WordNet.Pr consta que *risk* é hipônimo de *try*, no procedimento descrito anteriormente, *risk* é relacionado com *arriscar* e *try* com *tentar*, de modo que na WordNet.Br constará *arriscar* como hipônimo de *tentar* (Dias-da-Silva et al., 2008). O trabalho de Scarton e Aluísio (2009) implementou a herança automática das relações de hiperonímia da Wordnet.Br, assim foi possível a implementação da métrica que conta hiperônimos de verbos.

Para as métricas que contam Conectivos, elaboramos listas em que os marcadores são classificados em duas dimensões (seguindo a classificação do Coh-Metrix). Na primeira dimensão, a extensão da situação descrita pelo texto é determinada. Conectivos positivos ampliam eventos, enquanto que conectivos negativos param a ampliação de eventos (Louwerse, 2002; Sanders et al., 1992). Na segunda dimensão, os marcadores são classificados de acordo com o tipo de coesão: aditivos, causais, lógicos ou temporais. Nossa lista de marcadores foi construída utilizando listas já compiladas por outros pesquisadores (Pardo e Nunes, 2004; Moura Neves, 2000) e traduzindo alguns marcadores das listas em inglês.

Outro recurso que utilizamos é o Separador Silábico desenvolvido no projeto ReGra (Nunes et al., 1999).

Estendemos o trabalho de Scarton et al. (2009) criando mais sete métricas para o Coh-Metrix-Port. Para isso, além da Wordnet.Br com as relações de hiperonímia, foi necessário o uso de um outro recurso, o TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil (Maziero et al., 2008), que já disponibiliza as opções de consulta de sinonímia e de antonímia da WordNet.Br. Seu conjunto completo de dados – que conta com cerca de 20.000 entradas, distribuídas em 6.000 verbos, 2.000 substantivos e 12.000 adjetivos – está disponível para download e pode ser incorporado em diversas aplicações. Este recurso foi necessário para identificar o grau de ambiguidade das palavras.

3.2 Métricas Selecionadas

Para o Coh-Metrix-Port, contamos com o Índice Flesch (Martins et al., 1996), além das 40 seguintes métricas:

- Contagens Básicas: número de palavras, número de sentenças, número de parágrafos, sentenças por parágrafos, palavras por sentenças, sílabas por palavras, número de verbos, número de substantivos, número de advérbios, número de adjetivos, número de pronomes, incidência de palavras de conteúdo (substantivos, adjetivos, advérbios e verbos) e incidência de palavras funcionais (artigos, preposições, pronomes, conjunções e interjeições).
- Constituintes: incidência de sintagmas nominais, modificadores por sintagmas nominais e palavras antes de verbos principais.
- Frequências: frequência de palavras de conteúdo e mínimo das frequências de palavras de conteúdo.
- Conectivos: incidência de todos os conectivos, incidência de conectivos aditivos positivos, incidência de conectivos temporais positivos, incidência de conectivos causais positivos, incidência de conectivos lógicos positivos, incidência de conectivos aditivos negativos, incidência de conectivos causais negativos, incidência de conectivos temporais negativos e incidência de conectivos lógicos negativos.
- Operadores Lógicos: incidência de operadores lógicos, número de *e*, número de *ou*, número de *se* e número de negações.
- Pronomes, Tipos e Tokens: incidência de pronomes pessoais, pronomes por sintagmas e relação tipo/token.
- Hiperônimos: hiperônimos de verbos.
- Ambiguidades: ambiguidade de verbos, de substantivos, de adjetivos e de advérbios.

Entretanto, as métricas relacionadas com anáforas também poderão ser implementadas, dado que já existem métodos de resolução anafórica para pronomes (Cuevas e Paraboni, 2008) e descrições definidas (Souza et al., 2008). O Coh-Metrix-Port está sendo

desenvolvido em Ruby com o framework Rails. Tomamos esta decisão, pois esta linguagem possibilita um desenvolvimento ágil e bem estruturado. Para o banco de dados, decidimos utilizar o MySQL que, em projetos anteriores, mostrou-se muito bom para tecnologias Web.

4. Aplicações do Coh-Metrix-Port

Na Seção 4.2, ilustramos uma das utilidades de nossa ferramenta em desenvolvimento via um experimento com dois córpus, para mostrar as diferenças entre textos supostamente complexos e textos simples, isto é, textos reescritos para crianças, amparados pela abordagem de Crossley et al. (2007). Um dos córpus é composto de textos originais de notícias do jornal ZeroHora (ZH), dos anos 2006 e 2007, e outro de textos reescritos para crianças da seção *Para o seu filho ler* (PSFL), destinada a crianças entre 7 e 11 anos, dos correspondentes textos complexos do jornal ZeroHora. Na Seção 4.3, analisamos as métricas do Coh-Metrix-Port para verificar quais são mais significativas para o treinamento de classificadores binários (textos complexos e simples). Além disso, analisamos a influência (i) do gênero no desempenho destes classificadores, trabalhando com textos simples e complexos em dois gêneros: jornalístico e de divulgação científica e (ii) de textos de outras fontes. Na Seção 4.1 descrevemos todos os córpus utilizados nas duas aplicações apresentadas neste artigo, com exceção dos córpus para avaliação do desempenho com outras fontes que são descritos na Seção 4.3.

4.1 Descrição dos córpus de trabalho

Na Tabela 1, apresentamos algumas estatísticas dos quatro córpus principais utilizados neste artigo, provindos das seguintes fontes: ZH[16], PSFL, Ciência Hoje[17] (CH) e Ciência Hoje das Crianças[18] (CHC). Os córpus utilizados para a avaliação do Coh-Metrix-Port estão disponíveis na wiki do projeto PorSimples[19].

Córpus	Número de textos	Número de palavras	Média de palavras por textos
ZH	166	63996	385,518
CH	130	81139	624,146
PSFL	166	19257	116,006
CHC	127	56096	441,701

Tabela 1: Descrição dos córpus utilizados nas aplicações do Coh-Metrix-Port

Na Figura 2 apresentamos um gráfico com a distribuição dos córpus de análise e treinamento em relação ao número médio de palavras por textos e a Figura 3 mostra trechos dos córpus disponíveis para crianças.

Na Figura 3a, vemos o uso do pronome "você" que tem a função de aproximar o leitor do texto (esta característica é freqüente nestes textos) e na Figura 3b vemos o uso de uma definição via reformulação de um conceito (a reação do corpo face à aplicação de vacinas). A reformulação é geralmente antecedida por determinadas expressões lingüísticas como: "ou seja", "isto é" e "em outras palavras" e é muito comum neste córpus.

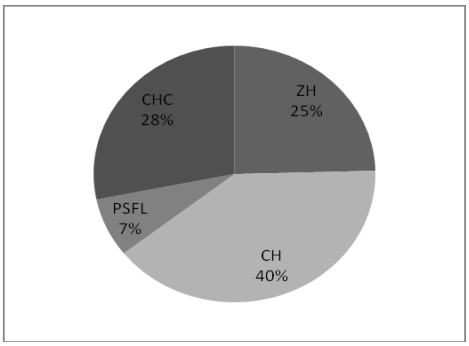


Figura 2: Distribuição dos córpus de treinamento em relação ao número médio de palavras por texto

O córpus ZH é composto por 166 textos jornalísticos, dos anos de 2006 e 2007. Neste trabalho, consideramos o córpus ZH como "complexo", pois este é escrito para adultos. Para o seu filho ler é uma seção do jornal ZeroHora destinada a crianças entre 7 e 10. Neste caso, o córpus PSFL é considerado com "simples".

[16] http://zerohora.clicrbs.com.br/
[17] http://cienciahoje.uol.com.br/revista-ch
[18] http://www.chc.org.br/
[19] http://www.nilc.icmc.usp.br/coh-metrix-port/avaliacao/

(a) *Os Estados Unidos acham que o Irã quer construir bombas atômicas, que podem matar milhares de pessoas. O Irã diz que não é verdade e que só pretende produzir eletricidade. Os americanos ameaçam aprovar medidas contra os iranianos na Organização das Nações Unidas, como proibir que o Irã compre produtos de outros países. Os Estados Unidos também podem começar uma guerra contra o Irã para impedir a fabricação das bombas. Ontem, o presidente iraniano desafiou seus inimigos e disse que não acredita em guerra ou castigos contra seu país. O presidente do Irã não gosta de Israel e também fez críticas aos israelenses.*

(b) *Tudo funciona da seguinte maneira: quando nós e nossos animais domésticos tomamos vacina, uma pequena dose de vírus, bactérias, protozoários etc. é dada ao corpo na medida certa, de tal maneira que não causa doença, mas é suficiente para ativar o sistema imunológico. Assim, a partir da aplicação da vacina, o corpo reage, ou seja, cria anticorpos que nos protegem, caso algum invasor igual ao que nos foi inoculado tente entrar em nosso organismo para atacar nossa saúde.*

Figura 3: (a) Trecho do córpos Para ser Filho ler (notícia do dia 25/04/2006); (b) Trecho do córpus Ciência Hoje das Crianças (artigo da edição 186 de dezembro de 2007)

O córpus CH é composto por 130 textos científicos, extraídos da revista Ciência Hoje (CH) dos anos de 2006, 2007 e 2008. Este córpus também é considerado como "complexo". O córpus CHC é composto por 127 textos do gênero científico, extraídos da revista Ciência Hoje das Crianças (CHC) (dos anos 2006, 2007 e 2008,) que tem como público alvo crianças entre 8 e 14 anos. Este córpus é considerado como "simples".

4.2 Avaliação de textos jornalísticos e sua versão reescrita para crianças

Em Crossley et al. (2007), é apresentada uma análise de dois córpus, utilizando o Coh-Metrix: um com textos reescritos e outros com textos originais. No final, os resultados obtidos são comparados e relacionados com hipóteses de pesquisadores da área de psicolinguística. Para ilustrar uma das utilidades de nossa ferramenta em desenvolvimento, resolvemos realizar um experimento também com dois córpus, o ZH e o PSFL, apresentados na Seção 4.1. Esse estudo de caso serve para comparar resultados e inferir conclusões sobre as diferenças e semelhanças entre os córpus. A Tabela 2 apresenta esta análise.

Para validar as métricas que citaremos a seguir, utilizamos o teste t-student, considerando $p < 0,05$. Na tabela 2, temos as métricas que foram aplicadas a ambos os córpus (originais e reescritos).

O número de palavras e o número de sentenças foi maior no texto original, o que era esperado, pois os textos originais são bem maiores do que os textos reescritos para crianças, os quais apenas apresentam a idéia do assunto. O número de pronomes (7,09% reescritos; 3,71% originais com p = 1,06E-14) e o número de pronomes por sintagmas (0,275 reescritos; 0,130 originais com p = 2,27E-13) foi maior nos textos reescritos. De acordo com a documentação do Coh-Metrix, deveríamos esperar o contrário, pois um maior número de pronomes por sintagmas dificulta ao leitor identificar a quem ou a que o pronome se refere.

Para entender este número elevado, fizemos uma análise em 50 textos à procura dos pronomes. Há um número elevado de pronome pessoal "você" em orações como "Quando viajar de carro com seus pais, você pode aproveitar o tempo livre para brincar.", que são usadas para aproximar o leitor do texto. O uso de pronomes como "ele(s)"/"ela(s)", que são os principais responsáveis por dificultar a leitura, acontece na maioria das vezes na sentença seguinte ou na mesma sentença (37 vezes vs. 4 numa sentença longe da definição da entidade) e o uso de cadeias de "ele(s)"/"ela(s)" é mínimo (6). Desta forma, os perigos do uso de pronomes são minimizados nos textos reescritos para crianças.

		Originais	**Reescritos**
Contagens Básicas	Número de palavras	63996	19257
	Número de sentenças	3293	1165
	Palavras por Sentença	19,258	16,319
	Número de parágrafos	1750	405
	Sentenças por Parágrafos	1,882	2,876
	Sílabas por Palavras de Conteúdo	2,862	2,530
	Número de Verbos	9016 (14,09%)	3661 (19,01%)
	Número de Substantivos	21749 (33,98%)	5349 (27,78%)
	Número de Adjetivos	4179 (6,53%)	1226 (6,37%)
	Número de Advérbios	2148 (3,36%)	980 (5,09%)
Frequências	Frequências de palavras de conteúdo	210075,48	267622,22
	Mínimo de freqüências de palavras de conteúdo	401,37	832,45
Constituintes	Palavras antes de verbo principal / Sentenças	4,096	2,900
	Sintagmas Nominais por palavras (x 1000)	283,72	257,26
Pronomes, Tipos e Tokens	Número de Pronomes	2372 (3,71%)	1365 (7,09%)
	Pronomes pessoais	298 (0,47%)	224 (1,16%)
	Proporção Type-Token	0,310	0,345
	Pronomes por Sintagmas Nominais	0,130	0,275
Operadores Lógicos	Número de *e*	1480 (2,31%)	476 (2,47%)
	Número de *ou*	116 (0,18%)	84 (0,44%)
	Número de *se*	352 (0,55%)	177 (0,92%)
	Número de negações	516 (0,81%)	247 (1,28%)
Conectivos	Todos os conectivos	8660 (13,57%)	3266 (17,03%)
	Aditivos Positivos	3529 (5,53%)	1356 (7,07%)
	Temporais Positivos	832 (1,30%)	311 (1,62%)
	Causais Negativos	4156 (6,51%)	1548 (8,07%)
	Lógicos Positivos	3083 (4,83%)	1192 (6,21%)
	Aditivos Negativos	559 (0,88%)	201 (1,05%)
	Temporais Negativos	7 (0,01%)	5 (0,03%)
	Causais Negativos	38 (0,06%)	4 (0,02%)
	Lógicos Negativos	170 (0,27%)	47 (0,24%)

Tabela 2 – Análise de 2 córpus utilizando algumas métricas do Coh-Metrix

Já a métrica de palavras antes do verbo principal merece um destaque especial. Na documentação do Coh-Metrix, afirma-se que este índice é muito bom para medir a carga da memória de trabalho, ou seja, sentenças com muitas palavras antes do verbo principal são muito mais complexas, pois sobrecarregam a memória de trabalho dos leitores. Em nosso experimento, obtivemos uma marca de 4,096 para córpus de textos originais e 2,900 para o córpus de textos reescritos, o que é um bom resultado, pois espera-se que os textos reescritos para crianças facilitem a leitura (com p = 1,19E-17). Outros resultados que merecem ser citados são a porcentagem de partículas "ou", a porcentagem de partículas "se" e a porcentagem de negações ("não", "jamais", "nunca", "nem", "nada", "nenhum", "nenhuma") que foram consideravelmente superiores nos textos reescritos (0,44%, 0,92% e 1,28%, respectivamente) em relação aos textos originais (0,18%, 0,55% e 0,81%, respectivamente). Porém, para estes últimos resultados não obtivemos um p significativo: 0,154; 0,173 e 0,176, respectivamente, o que não nos permite afirmar que textos reescritos possuem mais dessas partículas.

As métricas que calculam freqüência também merecem destaque. Os textos reescritos obtiveram um índice maior de freqüências de palavras de conteúdo 267622,22, contra 210075,48 dos textos originais (com p = 2,37E-28). Com isso,

concluímos que textos reescritos apresentam mais palavras freqüentes do que textos originais, o que já era esperado. Já a métrica de mínimo de freqüências de palavras de conteúdo, merece destaque pois, segundo a documentação do Coh-Metrix, essa métrica avalia, sentença a sentença, as palavras mais raras.

Como os textos simplificados apresentaram um número maior para esta métrica 832,45, contra 401,37 dos textos originais (com p = 1,23E-41), podemos inferir que os textos originais possuem mais palavras raras do que os textos reescritos. Quanto à métrica que conta conectivos, podemos dizer que os textos reescritos possuem mais conectivos (17,3%) do que os textos originais (13,57%) com p = 5,80E-05. Para ilustrar a utilidade desta métrica, voltemos as quatro sentenças em inglês citadas na introdução. Com essas métricas que contam marcadores conseguimos identificar que as sentenças (a) e (e) não possuem marcadores, enquanto que (b), (c), (d) e (f) possuem. Como estamos avaliando sentenças semelhantes, poderíamos concluir que as sentenças (a) e (e) são menos inteligíveis. Calculamos também as métricas de conectivos divididas em duas dimensões de acordo com a documentação do Coh-Metrix (descrevemos estas dimensões na Seção 3.1). Os resultados dessas métricas para os dois grupos de textos também são apresentados na Tabela 2.

4.3 Aprendizado de Máquina aplicado à avaliação da inteligibilidade

Na Seção 4.2 ilustramos uma das utilidades de nossa ferramenta em desenvolvimento via um experimento com dois córpus (ZH e PSFL), para mostrar as diferenças entre textos supostamente complexos e textos simples, amparados pela abordagem de Crossley et al. (2007). Esse estudo de caso serviu para comparar resultados entre corpus e a inferir conclusões sobre as diferenças e semelhanças entre eles.

Nesta seção, analisaremos as métricas do Coh-Metrix-Port para verificar quais são mais significativas para uma classificação entre textos complexos e simples. Além disso, propomos um classificador binário para textos "simples" e "complexos" em dois gêneros: jornalístico e de divulgação cientifica. Para isso, utilizamos quatro córpus para o treinamento, descritos na Seção 4.1.

4.3.1 Análise da contribuição das métricas do Coh-Metrix-Port na classificação de textos simples e complexos

Utilizando a ferramenta WEKA (Witten e Frank, 2005) com o algoritmo de seleção de atributos InfoGainAttributeEval, avaliamos as métricas do Coh-Metrix-Port em três cenários. O primeiro com os córpus ZH e Para o seu filho ler. O segundo com os córpus CHC e CH. Por fim, o último, com todos os quatro córpus. Na Figura 4 apresentamos um gráfico com as métricas ordenadas de acordo com o primeiro cenário. A ordem das métricas do segundo cenário é apresentada na Figura 5 e a do terceiro cenário na Figura 6.

Nos três casos podemos observar que as métricas mais distintivas são as métricas básicas (contagens e índice Flesch). Porém, métricas como incidência de pronomes por sintagmas, incidência de substantivos e incidência de verbos são bem classificadas e, por isso, podemos dizer que elas têm grande contribuição na classificação dos textos. Outra observação interessante é que há uma intersecção considerável entre as métricas que aparecem nos três casos.

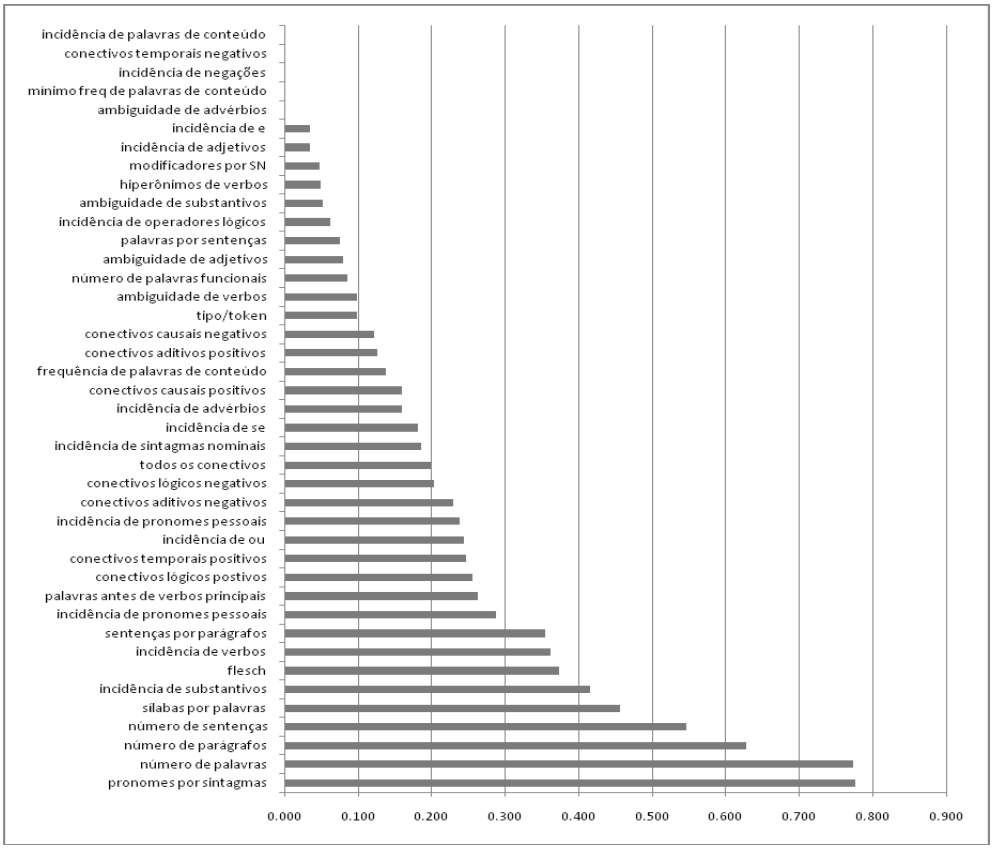


Figura 4: Ordem de importância das métricas do Coh-Metrix-Port utilizando os córpus ZH e PSFL

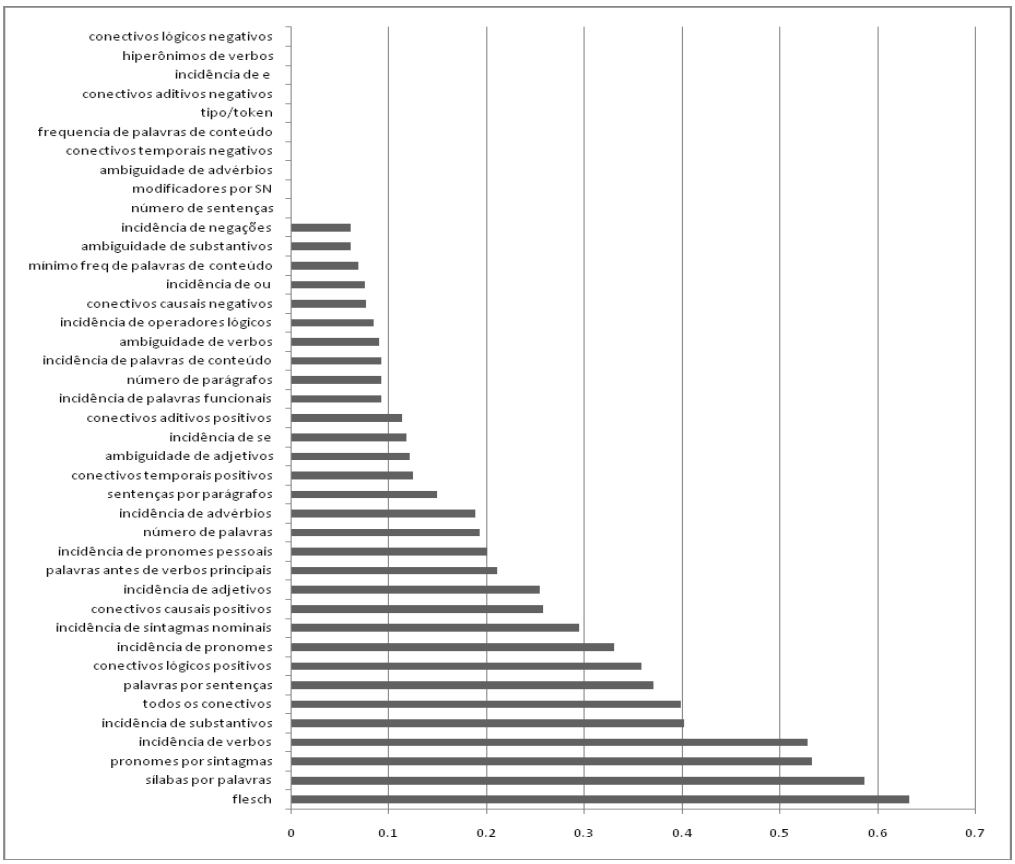


Figura 5: Ordem de importância das métricas do Coh-Metrix-Port utilizando os córpus CH e CHC

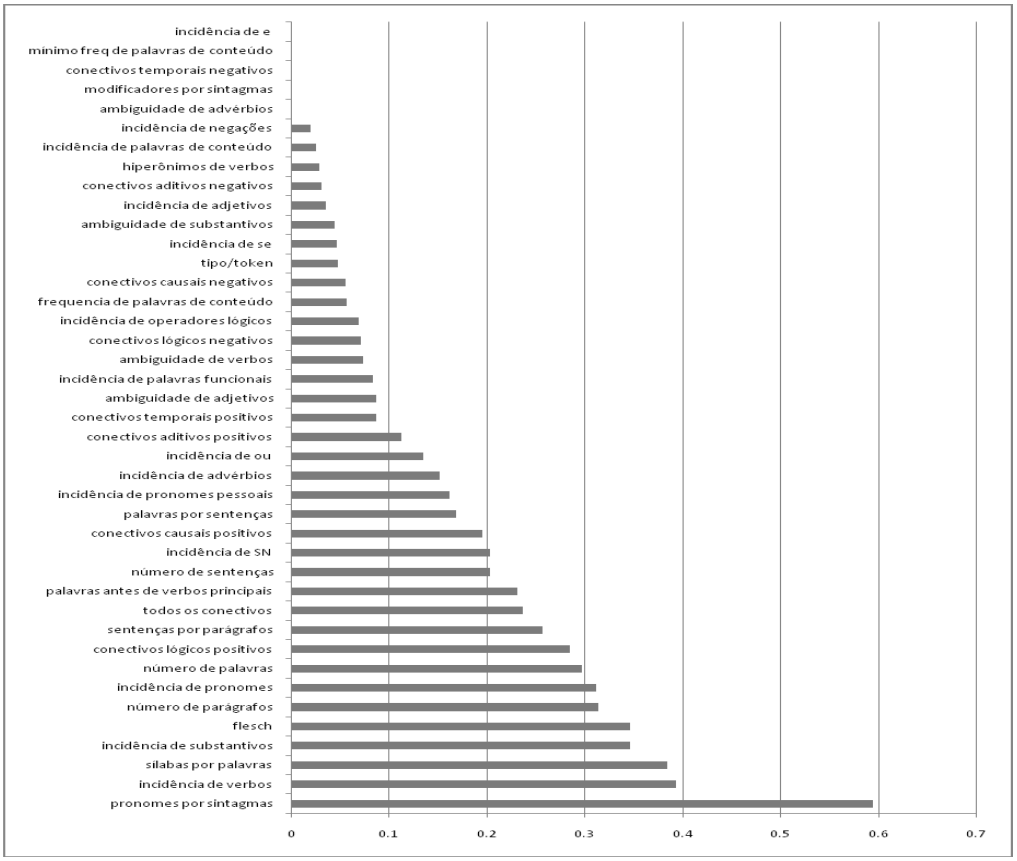


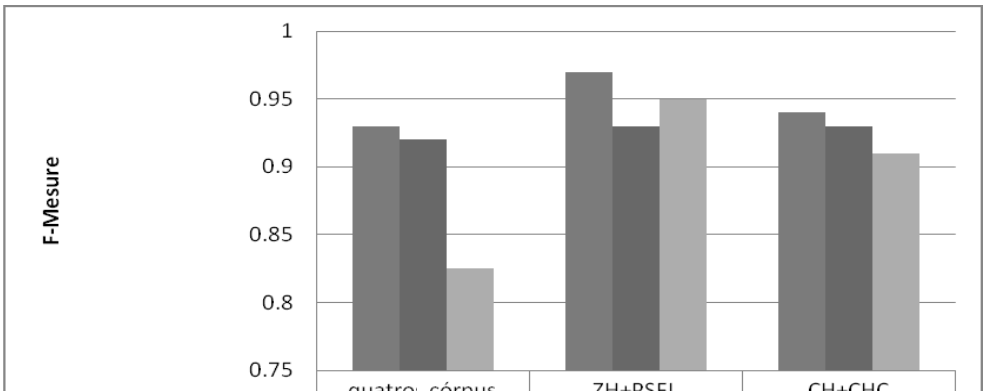
Figura 6: Ordem de importância das métricas do Coh-Metrix-Port utilizando os córpus ZH, PSFL, CH e CHC

4.3.2 Criação de classificadores de textos simples e complexos

Utilizando o algoritmo de classificação SMO da ferramenta WEKA, realizamos nove experimentos, considerando duas classes: simples ou complexos. Os textos classificados como "simples" são os do córpus PSFL e os do córpus CHC. Já os textos classificados como "complexos" estão nos córpus ZH e CH. Os nove experimentos são descritos a seguir:

- Utilizando os quatro córpus
  - Classificação somente com o índice Flesch e suas componentes (número de palavras, número de sentenças, palavras por sentenças e sílabas por palavras)
  - Classificação com as métricas do Coh-Metrix-Port sem o Flesch
  - Classificação com todas as métricas (Coh-Metrix-Port+Flesch)
- Utilizando ZH+PSFL
  - Classificação somente com o índice Flesch e suas componentes (número de palavras, número de sentenças, palavras por sentenças e sílabas por palavras)
  - Classificação com as métricas do Coh-Metrix-Port sem o Flesch
  - Classificação com todas as métricas (Coh-Metrix-Port+Flesch)
- Utilizando CH+CHC
  - Classificação somente com o índice Flesch e suas componentes (número de palavras, número de sentenças, palavras por sentenças e sílabas por palavras)
  - Classificação com as métricas do Coh-Metrix-Port sem o Flesch
  - Classificação com todas as métricas (Coh-Metrix-Port+Flesch)

Os valores de *F-Mesure* para todos os casos são mostrados na Figura 7. Como podemos observar na Figura 7, a precisão de uma classificação feita utilizando somente o índice Flesch e suas componentes é de 82,5% para os quatro córpus, 95% para ZH+PSFL e 91% para CH+CHC. O único caso em que o índice Flesch obteve resultado superior que os demais é para o córpus ZH+PSFL, o que



	quatro córpus	ZH+PSFL	CH+CHC
Coh-Metrix-Port+Flesch	0.93	0.97	0.94
Coh-Metrix-Port	0.92	0.93	0.93
Flesch	0.825	0.95	0.91

Figura 7: Valores de *F-Mesure* para os experimentos realizados

pode ser justificado pela grande diferença de tamanho de sentenças (palavras por sentenças) e tamanho de palavras (sílabas por palavras) entre os dois córpus, uma vez quem os textos da seção *Para o seu filho ler* consistem, geralmente, de somente um parágrafo com um média de, aproximadamente, 116 palavras por textos. Já os textos do córpus ZH possuem uma média bem maior, aproximadamente, 385 palavras por texto. Quando excluímos o índice Flesch do conjunto de métricas do Coh-Metrix-Port, o valor de *F-Mesure* aumenta em dois casos (com os quatro córpus e com os córpus CH e CHC). Por fim, se considerarmos todas as métricas do Coh-Metrix-Port mais o índice Flesch os resultados não só aumentam como são satisfatórios. Portanto, podemos concluir que as métricas presentes no Coh-Metrix-Port são autossuficientes em alguns casos. Porém, a melhor maneira de utilizá-las é como um completo ao índice Flesch.

4.3.3 Avaliação do desempenho dos classificadores binários em textos de novas fontes

Na Seção 4.3.2, observamos que as métricas do Coh-Metrix-Port, junto com o índice Flesch, apresentam uma boa precisão na classificação de textos como simples ou complexos. Por isso, resolvemos fazer um experimento utilizando o classificador (com todos os córpus: ZH, PSFL, CH e CHC) construído na Seção 4.3.2 visando avaliar textos que não pertencem a estes córpus de treinamento. Escolhemos, então, seis córpus. O primeiro córpus (conjunto_PSFL) contém 222 textos da seção Para o seu filho ler que não pertencem ao córpus de treinamento. O segundo córpus (conjunto_JCC) contém 80 textos do suplemento semanal JC Criança do Jornal da Cidade de Bauru[20] que são textos destinados a crianças de 8 a 14 anos. O terceiro (conjunto_ FSP) contém 50 textos do Caderno Ciência do Jornal Folha de São Paulo[21]. O quarto córpus (conjunto_ZH) contém 513 textos do jornal Zero Hora que não pertencem ao córpus de treinamento. O quinto (conjunto_CHC) contém 40 textos da revista Ciência Hoje das Crianças do ano de 2009. Por fim, o sexto córpus (conjunto_CH) contém 54 textos da revista Ciência Hoje do ano de 2009. Uma descrição dos seis conjuntos é apresentada na Tabela 6. Na Tabela 7, apresentamos a classificação esperada do classificador para cada conjunto.

Na Tabela 8 apresentamos a porcentagem de acerto para cada conjunto e os números de textos classificados erroneamente.

Pelos resultados apresentados na Tabela 3, observamos que possuímos um bom classificador para distinguir textos "simples" (para crianças) e "complexos" (para adultos). O pior resultado obtido foi para o conjunto JCC o que pode ser justificado pela grande

[20] http://www.jcnet.com.br/
[21] http://www1.folha.uol.com.br/fsp/

diferença deste córpus em relação ao córpus de treinamento (com o mesmo gênero) considerado simples (a média do número de palavras por texto do corpus PSFL é de 116,006 enquanto que a do JCC é de 442,413). O córpus JCC possui textos jornalísticos, como o PSFL, porém os textos são consideravelmente maiores.

Córpus	Número de textos	Número de palavras	Média de palavras por textos
ZH	513	147923	288,349
CH	54	25197	466,611
FSP	50	16530	330,600
PSFL	222	26548	119,586
CHC	40	14271	356,775
JCC	80	35393	442,413

Tabela 6: Descrição dos córpus

Conjunto	Classe
conjunto_PSFL	simples
conjunto_JCC	simples
conjunto_CHC	simples
conjunto_ZH	complexo
conjunto_CH	complexo
conjunto_FSP	complexo

Tabela 7: Classificação esperada

Conjunto	Porcentagem de acerto	Número de textos classificados errados
conjunto_PSFL	95%	11
conjunto_JCC	61,3%	31
conjunto_CHC	90%	4
conjunto_ZH	85,2%	76
conjunto_CH	87%	7
conjunto_FSP	94%	3

Tabela 8: Resultado do classificador

Quanto aos conjuntos ZH, CH e FSP, observamos que poucos textos foram classificados como "simples" o que garante que haverá poucos problemas em uma classificação que se deseja classificar um texto de acordo com seu público alvo: infantil ou adulto. Além disso, o conjunto FSP composto de textos completamente diferentes dos córpus de treinamento apresentou um bom resultado.

5.Conclusão

O projeto Coh-Metrix-Port é um início de uma pesquisa para satisfazer uma carência muito grande na área de inteligibilidade para a língua portuguesa. Buscamos com a construção da ferramenta o suporte necessário para o estudo detalhado dos fatores que tornam um texto complexo, para termos as diretrizes para simplificá-lo. A literatura sobre simplificação textual nos ajuda a compreender o que é considerado um texto difícil de ser lido. Como comentado na introdução, sentenças longas, com vários níveis de subordinação, cláusulas embutidas (relativas), sentenças na voz passiva, uso da ordem não canônica para os componentes de uma sentença, além do uso de palavras de baixa frequência aumentam a complexidade de um texto para leitores com problemas de leitura. Dessas características, todas as relacionadas com o uso de um *parser* (sentenças com vários níveis de subordinação, cláusulas embutidas – relativas –, sentenças na voz passiva, uso da ordem não canônica para os componentes de uma sentença) não foram ainda computadas e estão reservadas para trabalhos futuros. Um dos resultados desta pesquisa é a criação de métodos que contribuem com a inclusão social no âmbito do direito ao acesso à informação. Estes dão suporte à reescrita de textos apropriados para que pessoas com alfabetização em níveis básicos, as crianças em processo de alfabetização ou pessoas com alguma deficiência cognitiva possam assimilar melhor as informações lidas.

Vale comentar que a ferramenta Coh-Metrix-Port é de domínio público e seu código fonte será disponibilizado ao fim da pesquisa, em julho de 2010, para que outros pesquisadores possam utilizá-lo.

Agradecimentos

Os autores agradecem o apoio da agência de fomento à pesquisa Fapesp para o desenvolvimento desta pesquisa.

Referências

Aluísio, Sandra Maria, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick G. Maziero e Renata P. M. Fortes (2008b). Towards

Brazilian Portuguese Automatic Text Simplification Systems. Em Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), páginas 240-248, São Paulo, Brasil.

Aluísio, Sandra Maria, Lúcia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena de Medeiros Caseli e Renata P. M. Fortes (2008a) "A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems " In: Proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008), pp. 15-22.

Baayen, Harald R., Richard Piepenbrock e Leon Gulikers (1995). The CELEX lexical database (CD-ROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bick, Eckhard (2000). The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Tese de Doutorado. Aarhus University.

Bick, Eckhard (2005), Gramática Constritiva na Análise Automática de Sintaxe Portuguesa. In: Berber Sardinha, Tony (ed.), A Língua Portuguesa no Computador. Campinas: Mercado de Letras, São Paulo: FAPESP. ISBN: 85-7591-044-2

Burdick, Hal e Colleen Lennon (2004). The Lexile Framework as an approach for reading measurement and success. A white paper from The Lexile Framework for Reading. Disponível em: http://www.paseriesmathematics.org/downloads/Lexile-Reading-Measurement-and-Success-0504.pdf

Candido Jr., Arnaldo, Erick G. Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lúcia Specia e Sandra Maria Aluísio (2009). Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In: Proceedings of NAACL 2009 Workshop of Innovative Use of NLP for Building Educational Applications, pp. 34-42.

Caseli, Helena de Medeiros, Tiago de Freitas Pereira, Lúcia Specia, Thiago A. S. Pardo, Caroline Gasperin e Sandra Maria Aluísio (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In Alexander Gelbukh (ed), Advances in Computational Linguistics, Research in Computer Science, vol 41, pp. 59-70. 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009), March 01–07, Mexico City.

Charniak, Eugene (2000). A Maximum-Entropy-Inspired Parser. Em Proceedings of NAACL'00, páginas 132-139, Seattle, Washington.

Coltheart, Max (1981). The MRC psycholinguistic database. Em Quartely Jounal of Experimental Psychology, 33A, páginas 497-505.

Crossley, Scott A., Max M. Louwerse, Philip M. McCarthy e Danielle S. McNamara (2007). A linguistic analysis of simplified and authentic texts. Em Modern Language Journal, 91, (2), páginas 15-30.

Cuevas, Ramon Ré Moya e Ivandré Paraboni (2008). A Machine Learning Approach to Portuguese Pronoun Resolution. Em Proceedings of the 11th Ibero-American Conference on Ai: Advances in Artificial intelligence, Lisboa, Portugal.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer e Richard Harshman (1990). Indexing By Latent Semantic Analysis. Em Journal of the American Society For Information Science, 41, páginas 391-407.

Dias-da-Silva, Bento Carlos, Mirna F. De Oliveira e Helio Roberto de Moraes (2002). Groundwork for the development of the brazilian portuguese wordnet. In PorTAL'02: Proceedings of the Third International Conference on Advances in Natural Language Processing, pages 189–196, London, UK. Springer-Verlag.

Dias-da-Silva, Bento Carlos e Helio Roberto de Moraes (2003). A construção de um thesaurus eletrônico para o português do Brasil. ALFA, Vol. 47, N. 2, pp. 101-115.

Dias-da-Silva, Bento Carlos, Ariani Di Felippo e Maria das Graças Volpe Nunes (2008). The automatic mapping of Princeton WordNet lexicalconceptual relations onto the Brazilian Portuguese WordNet database. Em *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Dubay, Willian H. (2004) The Principles of Readability A brief introduction to readability research. http://www.eric.ed.gov/ERICDocs/data/ericdo

cs2sql/content_storage_01/0000019b/80/1b/bf/46.pdf

Fellbaum, Christiane (1998). WordNet: An electronic lexical database. MIT Press, Cambridge, Massachusetts.

Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes e Nuno Seco (2008). "PAPEL: a dictionary-based lexical ontology for Portuguese". Em *Proceedings do VII Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada,* (PROPOR 2008), páginas 31-40. Aveiro, Portugal.

Graesser, Arthur C., Danielle S. McNamara e Max M. Louwerse (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? Em A. P. Sweet e C. E. Snow, editores, *Rethinking reading comprehension,* páginas 82-98. Guilford Publications Press, New York, Estados Unidos.

Graesser, Arthur C., Moongee Jeon, Zhiqiang Cai and Danielle S. McNamara (2008). Automatic analyses of language, discourse, and situation models. In J. Auracher and W. van Peer (Eds.), New beginnings in literary studies, pp. 72–88, Cambridge, UK: Cambridge Scholars Publishing.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse e Zhiqiang Cai (2004). Coh-Metrix: Analysis of text on cohesion and language. Em *Behavioral Research Methods, Instruments, and Computers*, 36, páginas 193-202.

Janczura, Gerson Américo, Goiara de Mendonça Castilho, Nelson Oliveira Rocha, Terezinha de Jesus Cordeiro van Erven e Tin Po Huang (2007). Normas de concretude para 909 palavras da língua portuguesa. Em *Psic.: Teor. e Pesq*. [online], vol. 23, páginas 195-204.

Witten, Ian H. e Eibe Frank (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition.

Leffa, Vilson José (1996) Fatores da compreensão na leitura. Em *Cadernos no IL*, v.15, n.15, páginas 143-159, Porto Alegre. < http://www.leffa.pro.br/textos/trabalhos/fatores.pdf)>. Acesso em julho de 2009.

Louwerse, Max M. (2002). An analytic and cognitive parameterization of coherence relations. Em *Cognitive Linguistics*, páginas 291-315.

Martins, Teresa B. F., Claudete M. Ghiraldelo, Maria das Graças Volpe Nunes e Osvaldo Novais de Oliveira Junior (1996). Readability formulas applied to textbooks in brazilian portuguese. Notas do ICMC, N. 28, 11p.

McNamara, Danielle S., Max M. Louwerse e Arthur C. Graesser (2002) Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Grant proposal. Disponível em: http://csep.psyc.memphis.edu/mcnamara/pdf/IESproposal.pdf

Maziero, Erick G., Thiago Alexandre Salgueiro Pardo, Ariani Di Felipo e Bento Carlos Dias-da-Silva (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. Em *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana TIL, 2008*, Vila Velha, ES.

Moura Neves, Maria Helena de (2000). Gramática de Usos do Português. Editora Unesp, 2000, 1040 p.

Nunes, Maria das Graças Volpe, Denise Campos e Silva Kuhn, Ana Raquel Marchi, Ana Cláudia Nascimento, Sandra Maria Aluísio e Osvaldo Novais de Oliveira Júnior (1999). Novos Rumos para o ReGra: extensão do revisor gramatical do português do Brasil para uma ferramenta de auxílio à escrita. Em *Proceedings do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada,* (PROPOR 1999), páginas 167-182. Évora, Portugal.

Oliveira, Cláudia, Maria Cláudia Freitas, Violeta Quental, Cícero Nogueira dos Santos, Renato Paes Leme e Lucas Souza (2006). A Set of NP-extraction rules for Portuguese: defining and learning. Em *7th Workshop on Computational Processing of Written and Spoken Portuguese*, Itatiaia.

Pardo, Thiago Alexandre Salgueiro e Maria das Graças Volpe Nunes (2004). Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil. Relatório Técnico NILC.

Pianta, Emanuele, Luisa Bentivogli e Christian Girardi (2002). MultiWordNet: developing an aligned multilingual database. Em *Proceedings of the First International*

Conference on Global WordNet, páginas 293-302, Mysore, India.

Ratnaparkhi, Adwait (1996). A Maximum Entropy Part-of-Speech Tagger. Em *Proceedings of the First Empirical Methods in Natural Language Processing Conference*, páginas133-142.

Sanders, Ted J. M., Wilbert P. M. Spooren e Leo G. M. Noordman (1992). Toward a taxonomy of coherence relations. Em *Discourse Processes*, 15, páginas 1-35.

Santos, Acácia A. Angeli dos, Ricardo Primi, Fernanda de O. S. Taxa e Claudette M. M. Vendramini (2002). O teste de Cloze na avaliação da compreensão em leitura. Em *Psicol. Reflex. Crit.* [online]., v. 15, n. 3, páginas 549-560.

Santos, Diana, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo Gonçalo Oliveira, José Carlos Medeiros e Rosário Silva (2009). "O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o PAPEL". Em *XXV Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, Portugal.

Scarton, Carolina Evaristo e Sandra Maria Aluísio (2009). Herança Automática das Relações de Hiperonímia para a Wordnet.Br. Série de Relatórios do NILC. NILC-TR-09-10, Dezembro, 48p.

Siddharthan, Advaith (2002). An Architecture for a Text Simplification System. Em *Proceedings of the Language Engineering Conference (LEC)*, páginas 64-71.

Souza, José Guilherme, Patrícia Gonçalves e Renata Vieira (2008). Learning Coreference Resolution for Portuguese Texts. In *Proceedings of the 8th international Conference on Computational Processing of the Portuguese Language*, Aveiro, Portugal.

Williams, Sandra (2004). Natural Language Generation (NLG) of discourse relations for different reading levels. Tese de Doutorado, University of Aberdeen.

Caracterização e Processamento de Expressões Temporais em Português

Caroline Hagège
Xerox Research Centre Europe – XRCE
6 Chemin de Maupertuis – Meylan – France
Caroline.Hagege@xrce.xerox.com

Jorge Baptista
Universidade do Algarve, FCHS
L2F, INESC-ID Lisboa
Campus de Gambelas – Faro – Portugal
jbaptis@ualg.pt

Nuno Mamede
Instituto Superior Técnico
L2F, INESC-ID Lisboa
Rua Alves Redol, 9 – Lisboa – Portugal
Nuno.Mamede@inesc-id.pt

Resumo

A dimensão temporal é um elemento estruturante fundamental para a informação veiculada em textos e constitui um desafio para o processamento de língua natural, sendo igualmente importante para muitas aplicações do processamento das línguas. Este artigo constitui mais um passo para o ambicioso objectivo de tratamento da informação temporal. Para tal, apresenta-se uma proposta de classificação das expressões temporais do Português que permita esclarecer algumas incertezas relativas ao estatuto de diferentes expressões temporais e constitui uma base para a anotação destas expressões. Utilizando esta classificação, foi desenvolvida uma ferramenta de anotação automática das expressões temporais do Português, cujo desempenho foi avaliado.

1 Introdução

A descrição do tempo, assim como os processos de inferência que levam em conta a informação temporal, são assuntos que há muito tempo despertaram interesse em áreas tão diversas como a lógica, a filosofia e a linguística. Reichenbach em (Reichenbach, 1947) propõe um sistema explicativo dos tempos verbais utilizando três pontos de referência temporal: o tempo do evento (E), o tempo de referência (R) e o tempo do discurso (S). Nos anos 50, Prior em (Prior, 1957) propõe um teoria de lógica temporal, onde introduz a representação formal dos tempos usando operadores temporais.

Mais recentemente têm aparecido novos trabalhos relacionados com processos de inferência temporal. Um dos mais conhecidos em Inteligência Atificial (IA) e Processamento de Linguagem Natural (PLN) é o trabalho de Allen descrito em (Allen, 1991). Só muito recentemente, porém, apareceram os primeiros sistemas que fazem efectivamente algum tipo de processamento da informação temporal. Esta nova tendência foi impulsionada pelo facto de um tratamento adequado da componente temporal em textos permitir um melhor desempenho numa ampla gama de tarefas, tais como a resposta a perguntas, a sumarização (uni- e multidocumento) e, de um modo geral, a extracção de informação a partir de documentos. Um dos factores que ajudaram a desenvolver este renovado interesse pelo processamento de expressões temporais (ET) foi a criação do projecto TimeML (Saurí et al., 2006). Este projecto fornece um conjunto de directrizes para a anotação de expressões temporais e de eventos para o Inglês. Estas orientações foram adaptadas para o Francês (ver (Bittar, 2008)), para o Italiano e para o Romeno. Outras abordagens para a descrição e normalização de expressões temporais são apresentadas por Battistelli, Minel e Schwer em (Battistelli, Minel e Schwer, 2006). Nesta última abordagem, o tratamento temporal tem como finalidade ser usado por um sistema de navegação temporal nos textos. Para obter uma representação adequada da informação temporal, um subconjunto de expressões (designadas expressões temporais) são descritas como termos e sua composição é feita através de operadores pré-definidos.

Têm vindo a ser desenvolvidos alguns sistemas automáticos dedicados à anotação temporal e, recentemente, foi organizado um concurso para avaliar a precisão dos processadores tem-

porais automáticos para Inglês (Verhagen et al., 2007). As anotações baseiam-se nas directrizes TimeML mencionadas anteriormente e a maioria dos sistemas utiliza técnicas de aprendizagem automática e *corpora* anotados para o treino supervisionado. Infelizmente, este tipo de *corpora* com anotações temporais só estão disponíveis para o Inglês (TimeBank). O facto de um recurso como este exigir um grande esforço em termos de recursos humanos para a sua construção explica, naturalmente, a falta de recursos equivalentes noutras línguas. Mais recentemente, Parent e colegas (Parent, Gagnon e Muller, 2008) e Hagège e Tannier ((Hagège e Tannier, 2008) apresentaram sistemas baseados em regras para a anotação e normalização de expressões temporais em Francês e Inglês. Para Português, um primeiro passo para a anotação temporal foi realizado no âmbito do Segundo HAREM (Mota e Santos, 2008).

O nosso objectivo a longo prazo é o desenvolvimento de um sistema capaz de ancorar e de ordenar temporalmente os processos expressos nos textos. Para alcançar este objectivo, é necessário dar os seguintes passos:

- identificação e etiquetagem das expressões temporais que ocorrem nos textos;
- resolução das expressões temporais referenciais para que se possa proceder à sua normalização;
- identificação dos eventos associados às expressões temporais;
- caracterização das relações entre eventos e expressões temporais, o que inclui normalmente considerar o tempo, o aspecto e a modalidade;
- realização de inferência temporal.

O significado das ET referenciais não pode ser obtido directamente a partir dos elementos da expressão, requerendo algum tipo de cálculo quanto à sua referência temporal. A normalização das ET consiste, justamente, em representar esse valor de uma forma que permita esse cálculo.

Estes passos são, contudo, estreitamente interdependentes, visando um tratamento adequado da temporalidade. Este artigo aborda algumas das questões acima mencionadas. É proposto um conjunto de orientações para lidar com a identificação e etiquetagem de expressões temporais que aparecem em textos em Português. Nessa caracterização, as diferenças de estatuto referencial dessas expressões são levadas em consideração, pois levam à utilização de diferentes métodos para a normalização das expressões temporais. Desenvolvemos uma ferramenta para etiquetar automaticamente as expressões temporais de acordo com essas orientações e para realizar uma primeira etapa de normalização temporal.

O artigo começa por explicar as motivações para este trabalho, que teve um forte impulso a partir da participação na campanha de avaliação conjunta do Segundo HAREM (Mota e Santos, 2008), mostrando que uma caracterização adequada das expressões temporais não é uma tarefa trivial e que precisa de levar em consideração não apenas os elementos lexicais por que as ET são formadas mas, e de forma fundamental, também o contexto mais amplo em que se elas se encontram inseridas, por forma a que esta tarefa possa ser adequadamente executada. Serão, então, sucintamente apresentadas as directrizes para a anotação de entidades temporais e explicitaremos em que aspectos nos demarcamos das directrizes do projecto TimeML. Finalmente, apresenta-se o anotador temporal por nós desenvolvido e os resultados obtidos com o sistema naquela campanha de avaliação.

2 Motivação

A fim de motivar a dificuldade da tarefa de anotação temporal apresentam-se os seguintes exemplos:

(1) *Banana* ***de manhã*** *emagrece. Será?*
(2) *Partiu* ***esta manhã***
(3) ***A manhã*** *é um momento mágico do dia*
(4) ***Numa bela manhã****, resolveu partir*

Todas estas expressões são sintagmas nominais (SN) ou preposicionais (SP) que têm a mesma cabeça lexical (*manhã*). Mas, cada expressão tem de ser interpretada de forma diferente. Como é afirmado por Ehrmann e Hagège em (Ehrmann e Hagège, 2009), não se pode realizar de forma adequada a interpretação de uma ET sem levar em conta as suas relações com os outros constituintes da frase.

No primeiro caso, a expressão *de manhã* tem de ser interpretada como um agregado temporal (isto é, uma expressão temporal que vai ancorar o processo associado mais do que uma vez na linha do tempo). Além do mais, este agregado temporal tem um período regular (a expressão é aproximadamente equivalente a *todas as manhãs*)[1].

[1] A análise deste tipo de situação é, porém, bastante complexa, já que se trata de uma construção elíptica, correspondendo à frase: (*alguém*) *comer banana de manhã emagrece* (*alguém*); o valor genérico de *banana* é dado pela sua determinação (determinante zero ou ausência de determinante), e o valor frequentativo de *comer banana*

No segundo caso, trata-se de uma ET referencial cujo antecedente é o momento da enunciação (ou seja, partiu na manhã do dia em que a frase foi produzida). No terceiro caso, trata-se de uma expressão genérica temporal. Isto significa que a expressão não fornece qualquer ancoragem temporal para o predicado associado. Finalmente, a última expressão é uma expressão temporal vaga: existe de facto uma ancoragem do processo associado na linha temporal, mas não se pode especificar de maneira precisa onde este se situa na linha do tempo.

Estes exemplos mostram claramente que um simples esquema de emparelhamento de padrões não é suficiente para realizar uma caracterização adequada de entidades temporais. Por esta razão, parece perfeitamente defensável o ponto de vista que rejeita a inclusão da tarefa de reconhecimento de ET como uma mera sub-tarefa da tarefa mais geral de Reconhecimento de Entidades Mencionadas (como tem sido feito, por exemplo, em (Mota e Santos, 2008)).

3 Directivas para a identificação e Classificação de ET em Português

Um dos pontos-chave nas directivas que aqui se apresentam é justamente a ideia de que as ET só podem ser devidamente classificadas e anotadas quando consideradas em relação aos processos que modificam. Apesar de tal observação poder parecer óbvia, mesmo nas orientações do projecto TimeML (Saurí et al., 2006) permanece alguma incerteza quanto ao estatuto das relações entre as ET e os processos que modificam enquanto factor determinante para a sua interpretação, especialmente quando as ET são citadas, sem qualquer contexto.

3.1 A nossa proposta vs. estado da arte

O nosso trabalho inscreve-se na linha geral do projecto TimeML, embora com algumas diferenças que explicitaremos e exemplificaremos já a seguir.

O projecto TimeML constitui sem nenhuma dúvida valioso contributo e incontornável referência no quadro do processamento da informação temporal. O TimeML propõe não só uma classificação e uma normalização das ET, mas também uma proposta de anotação da informação acerca dos processos (aspecto, tempo, modalidade), informação que deve ser tomada em consideração para o tratamento adequado da temporalidade.

A nossa proposta é mais modesta, pois, neste momento, preocupámos-nos exclusivamente com expressões temporais e não propusemos ainda qualquer anotação específica para representar a informação relevante associada aos processos modificados pelas ET. Assim, por exemplo, o problema do estatuto das ET associadas a predicados modificados por diferentes modalidades não foi sequer considerado nesta altura. Por outras palavras, não tentamos dar resposta à pergunta sobre como interpretar temporalmente a frase seguinte: *É possível que venham na próxima quarta-feira*, na qual não se sabe se o processo *venham* vai ocorrer ou não. Na nossa proposta, vamos circunscrever-nos ao problema da identificação e classificação das ET, procurando desde já avançar no sentido de uma normalização da informação temporal por elas veiculada. Nesta proposta, apresentamos critérios formais operativos e reprodutíveis para identificação, segmentação e classificação das ET. Neste sentido, salientam-se desde já os aspectos em que nos distinguimos das directivas do projecto TimeML, não deixando, no entanto, de o considerar como uma referência fundamental neste domínio.

Os pontos onde nos distanciamos do TimeML são os seguintes:

- Integração sistemática da preposição que introduz uma ET;
- Proposta de critérios formais, claros e reprodutíveis, para segmentação de ET complexas;
- Clara distinção entre a anotação e os passos intermédios necessários para a realizar.

3.1.1 Integração da preposição

Consideramos que a preposição que introduz o grupo preposicional (SP) que contém uma expressão temporal deve fazer parte integrante da ET. Esta posição distingue-se da solução apresentada pelo TimeML, que anota as preposições introdutoras de ET com a categoria *SIGNAL* e as separa da expressão temporal propriamente dita. As razões desta escolha são as seguintes:
A preposição é um elemento formal que muitas vezes permite caracterizar ou classificar de forma inequívoca a expressão temporal. Assim, por exemplo, em (*a partir de/até/desde*)*segunda-feira*, o significado das ET seguintes está estreitamente ligado à escolha da preposição que introduz o nome de tempo *segunda-feira*.

está muito provavelmente relacionado com o infinitivo e a redução de um sujeito genérico (*alguém*); do mesmo modo, o valor genérico deste emprego de *emagrecer* parece resultar do uso do presente do indicativo e da redução de um complemento directo indefinido (*alguém*).

De facto, as propriedades sintácticas da construção destes adjuntos adverbiais de tempo estão directamente relacionadas com a preposição. Assim, por exemplo, enquanto que com as preposições acima o nome de tempo não obriga à presença de um artigo, se se tiver a preposição *em*, o artigo torna-se obrigatório: *na segunda-feira/*em segunda-feira*. Também a inserção de um advérbio quantificador indefinido como *aproximadamente* não se verifica com todas as preposições: (*a partir de/até/desde/*em*)*aproximadamente segunda-feira*. Em segundo lugar, a mesma ET, quando traduzida para outra língua, pode ou não ser introduzida por preposição. Por exemplo, a expressão em Português *na segunda-feira* poderá ser traduzida simplesmente em Inglês por *Monday* (sem preposição) ou por *on Monday* (com preposição). Já em Francês não se admite qualquer preposição: (*le*) *lundi*.

3.1.2 Segmentação de ET complexas

No que diz respeito à segmentação de ET complexas, a norma TimeML não fornece elementos suficientes para decidir sem ambiguidade se uma expressão complexa deve ser considerada como uma única ET ou se deverá ser segmentada várias ET independentes. Neste sentido, iremos propor, como veremos, um conjunto de critérios sintácticos e semânticos que permitem tomar esta decisão de forma clara e reprodutível.

3.1.3 Distinção entre resultado da anotação e etapas de processamento para a anotação

Finalmente, as directrizes do TimeML obrigam em certos casos a indicar as etapas intermédias de anotações (que correspondem possivelmente a diferentes etapas de processamento automático das expressões temporais). Assim, para a expressão *two days before yesterday* em *John left two days before yesterday.*, o guia de anotação TimeML preconiza a anotação de *two days* com o tipo *DURATION*, a anotação de *before yesterday* com o tipo *DATE*, e finalmente uma anotação global da expressão *two days before yesterday*. Este forma de anotar parece-nos indesejável, já que inclui etapas intermédias antes de fornecer a anotação final. Efectivamente, consideramos que as directivas para anotação não devem pressupor os meios que poderão ser utilizados para alcançar a anotação preconizada. O facto de se introduzir possíveis passos intermédios para se chegar à anotação final obriga, de certa forma, os anotadores automáticos a seguir um certo algoritmo de anotação, o que ultrapassa claramente a função de directivas.

Estando feitas estas clarificações relativamente à nossa posição perante o estado da arte, apresentamos, nas secções seguintes, a nossa proposta de identificação e classificação das ET.

3.2 Identificação

Para identificar de forma objectiva expressões temporais, apresentam-se vários critérios. Uma expressão é uma ET quando satisfaz simultaneamente os critérios 1 e 2 ou, então, é considerada uma ET genérica, definida pelo critério 3:

1. **Critério 1** - uma expressão temporal em contexto pode responder adequadamente a uma das interrogativas *quando?*, *quanto tempo?*, eventualmente precedido de uma preposição, ou *com que frequência?*;
2. **Critério 2** - uma expressão temporal contém pelo menos uma unidade lexical que corresponda a um dos seguintes tipos:
   (a) uma data numérica ou alfanumérica (por exemplo, 21-Mar-2008), tanto para expressar as datas do calendário como os diferentes formatos de hora (12:30), incluindo as abreviaturas dos meses, e certas expressões adverbiais (por exemplo, *AM*, *GMT* e *a.C.*);
   (b) uma unidade de tempo (*segundo*); este conjunto inclui também unidades de tempo que não pertençam ao sistema internacional e que são de emprego "informal" como *fim-de-semana*;
   (c) os substantivos correspondentes à designação de algumas destas unidades de tempo, como os nomes dos meses (*Janeiro*), os dias da semana (*segunda-feira*) e advérbios derivados de unidades de tempo (*diariamente*);
   (d) os substantivos que designam festividades e efemérides de natureza religiosa, política, histórica ou cultural; o nome das estações do ano e o de dias festivos, que podem ou não incluir o substantivo *dia*;
   (e) advérbios de tempo simples e não ambíguos (por exemplo *ontem*) ou advérbios compostos (*depois de amanhã*), juntamente com advérbios tempo derivados, formados com o sufixo *-mente* (*futuramente*);
   (f) um grupo preposicional (SP), cuja cabeça é um substantivo de tempo genérico (por exemplo, *altura*, *data*, *instante*, *momento* e *vez*); estes substantivos são geralmente acompanhados

de diversos determinantes como, por exemplo, quantificadores de tipos diferentes (*por duas/várias/diversas vezes*), os pronomes demonstrativos (*nessa altura*), outros pronomes com função determinativa, inclusive pronomes possessivos (*no meu tempo*); podem também ser modificados por diferentes adjectivos e até por orações relativas (*na altura em que ela vivia em Lisboa*); n.b.: a ET não inclui a oração relativa; inclui-se ainda no conjunto dos modificadores os adjectivos (normalmente em maiúsculas), que designam um período histórico (*durante o período Barroco*); n.b. o adjectivo deverá ser considerado como estando incluído na ET;

(g) os chamados complementos determinativos envolvendo numerais e unidades de tempo que quantificam temporalmente um nome (predicativo) designativo de evento, estado ou processo (*uma viagem de 5 dias*); n.b.: a preposição *de* deve ser incluída na ET;

(h) os grupos preposicionais com unidades de tempo modificadas por vários adjectivos com valor referencial (*no ano passado*, *no próximo mês*, *durante o corrente ano* e *nos séculos vindouros*); ou em que as unidades de tempo se encontram modificadas por orações relativas envolvendo os verbos *passar*, *vir*, ou outros : *no ano que passou*, *para o mês que vem*;n.b.: estes verbos constituem um conjunto fechado;

(i) expressões com os verbos *fazer* ou *haver* e unidades de tempo: *há três anos*, *faz duas semanas*; n.b.: as expressões com *fazer* ou *ter* que indicam a idade (*O Pedro já fez/tem 18 anos*) não deverão ser classificadas como ET.

3. **Critério 3** - A expressão envolve um ou mais dos itens lexicais (ou formatos numéricos) descritos no critério 2, mas não cumpre o critério 1: *A Primavera é a mais bela estação do ano.*

3.3 Segmentação

As entidades temporais incluem a preposição, quando a ET é um grupo (ou sintagma) preposicional (SP: *no ano passado*), ou o determinante se a expressão é um grupo nominal (SN: *dois dias depois*). No caso das ET complexas, que podem eventualmente constituir sequências ambíguas, adoptam-se os critérios de segmentação definidos em (Hagège e Tannier, 2008):

Uma expressão temporal complexa deve ser dividida em unidades menores se e só se as seguintes condições forem ambas verdadeiras:

1. Cada componente da expressão é sintacticamente válida, quando combinada com o processo que modifica;
2. Cada componente da expressão é logicamente implicada na expressão complexa, ou, por outras palavras, se a expressão complexa for verdadeira, então cada expressão componente deve também ser verdadeira.

Por exemplo, na frase *Visitei o Pedro* ***dois dias nesta semana***, a expressão de tempo complexa deve ser dividida em duas ET, pois cada uma das expressões componentes se pode combinar com o evento: *Visitei o Pedro* ***dois dias*** */ Visitei o Pedro* ***nesta semana***, e cada expressão componente é tão verdadeira quanto o valor de verdade da expressão temporal complexa. Pelo contrário, na frase seguinte: *Visitei o Pedro* ***dois dias depois*** (= *dois dias mais tarde*), apenas uma ET deverá ser considerada, uma vez que, apesar de cada uma das expressões menores poder combinar-se sintacticamente com o evento: *Visitei o Pedro* ***dois dias*** */ Visitei o Pedro* ***depois***, o significado de cada combinação individual torna-se diferente do significado global da expressão complexa.

3.4 Classificação

A classificação é proposta juntamente com um conjunto de critérios. Esta classificação é inspirada em trabalhos anteriores (Saurí et al., 2006) mas também é influenciada pelo resultado da experiência de anotação temporal do Segundo HAREM (Baptista, Hagège e Mamede, 2008). Por último, está também intimamente relacionada com a classificação feita em (Ehrmann e Hagège, 2009).

O principal critério utilizado para classificar entidades temporais consiste no tipo de ancoragem (ou localização) dos processos temporais que operam. Quatro tipos principais são assim considerados:

1. DATA – a ET corresponde a uma ancoragem única do processo na linha do tempo;
2. DURAÇÃO – a ET não ancora o processo na linha do tempo, exprimindo porém uma quantificação de ordem temporal;
3. FREQUÊNCIA – a ET relaciona o processo com a linha do tempo através de várias instâncias de ancoragem;

4. GENÉRICO – a expressão não ancora qualquer processo na linha do tempo; não é realmente uma expressão temporal no sentido de que nenhuma informação temporal está associada a qualquer processo, mas mantém um significado temporal que pode ser importante para a resolução de referências temporais.

Enquanto que os três primeiros e principais tipos de expressões temporais podem constituir uma resposta adequada às interrogativas [2] com *(Prep) quando?*, *(Prep) quanto tempo?*, ou *com que frequência?*, respectivamente, o tipo genérico não pode.

A subclassificação destes tipos principais depende, sobretudo, da estrutura simples ou complexa da ET. Assim, o tipo DATA pode ser estruturado nos seguintes subtipos:

- DATAs simples, inclui não só as datas do calendário, mas também expressões temporais com horas (e.g. *20/05/2009 11:45 TMG*);
- INTERVALOs, expressões temporais envolvendo duas DATAs (*de 5 a 15 de Maio*); e
- o subtipo COMPLEXO, que corresponde a ET envolvendo expressões de DATA e de DURAÇÃO (*de hoje a quinze dias*).

Na mesma maneira, o tipo DURAÇÃO inclui um subtipo simples (por exemplo, *A reunião durará **2 horas***) e um subtipo de intervalo; o último envolve duas expressões quantificadas (*A reunião durará **entre 1 e 2 horas***).

Além disso, o tipo DATA também é classificado com base na referência temporal da ET e/ou na sua indeterminação quanto à ancoragem do processo na linha do tempo. Neste sentido, distinguem-se os seguinte subtipos:

- data ABSOLUTA, directamente computável a partir da ET (e.g. *em Maio de 2009*);
- data RELATIVA, envolvendo o cálculo de uma referência temporal; estas ET são ainda subdivididas, consoante se refiram ao momento de ENUNCIAÇÃO (e.g. *ontem*) ou a outro elemento TEXTUAL, algures no texto (e.g. *no dia seguinte*).

Uma propriedade especial, denominada *INDET*[3] em (Baptista, Mamede e Hagège, 2009) é usada em diferentes tipos de ET. Por exemplo, certas ET do tipo DATA fornecem, para o processo que modificam, um ponto de ancoragem na linha de tempo, no entanto, este ponto de ancoragem não é especificado. Assim, em: ***Numa bela manhã**, resolveu partir* o evento está ancorado no tempo, mas nada na expressão nem mesmo num contexto mais alargado permite indicar o ponto de ancoragem preciso. O mesmo tipo de indeterminação pode ser encontrado em ET dos tipos DURAÇÃO (*durante algum tempo*) e FREQUÊNCIA (*de tempos a tempos*).

Considera-se ainda outra propriedade, a que se chamou *FUZZY*, para as expressões temporais que, embora apresentem os elementos necessários para a sua normalização, se encontram modificadas por diferentes tipos de expressões que tornam *imprecisa* essa DATA(*por volta do dia 10*), DURAÇÃO (*durante cerca de 2 horas*) ou FREQUÊNCIA (*praticamente dois dias por semana*).

4 Desenvolvimento de um Sistema de Análise Temporal

Foi desenvolvido um sistema de reconhecimento de expressões temporais baseado nas directivas de identificação e de classificação acima apresentadas. Este módulo pretende ser o ponto de partida de uma cadeia de processamento das expressões temporais mais ambiciosa, isto é, que não se limite à mera identificação das ET mas que seja capaz de as classificar adequadamente, tendo como objectivo final a capacidade de ancorar temporalmente os processos expressos nos textos, assim como estabelecer relações de ordem temporal parciais entre estes mesmos processos.

Ora, como já o demonstrámos, processar a informação temporal desta maneira mais complexa obriga a ter em conta o contexto, por vezes bastante alargado, em que a ET se encontra: é necessário ter em consideração a natureza do evento, estado ou processo associado à ET; é necessário, também, que o sistema seja capaz de resolver anáforas; finalmente, é necessário ter em consideração os fenómenos de tempo e de aspecto verbal. Por estas razões, parece-nos que um sistema de reconhecimento de ET deve poder contar com informação linguística rica, a qual inclui desde a informação morfológica (para o processamento dos tempos e modos verbais) até à informação sobre cadeias anafóricas. Naturalmente, o sistema deverá poder contar com a ligação correcta entre ET e os processos modificados por estas ET. A nossa estratégia, tendo em conta estes requisitos, consistiu em integrar o processamento temporal num sistema mais geral

[2] A fim de capturar todos os tipos relevantes, outras formas interrogativas também são utilizadas, mas a sua lista completa é dada por Baptista e colegas em (Baptista, Mamede e Hagège, 2009).

[3] Este tipo de expressões é também chamado de *data indeterminada* em (Ehrmann e Hagège, 2009) e (Gosselin, 1996).

de análise linguística, considerando que o tratamento em paralelo da informação temporal e da informação sintáctica clássica deveria beneficiar tanto a análise sintáctica como o processamento temporal.

4.1 Apresentação do XIP

Uma característica importante do nosso módulo de anotação temporal é o facto de ele estar integrado numa ferramenta mais geral de processamento linguístico: O XIP-PT.

XIP (Xerox Incremental Parser) (Aït-Mokhtar, Chanod e Roux, 2002) é uma ferramenta de análise linguística cujo objectivo é a extracção de dependências sintácticas. A ferramenta oferece um formalismo de análise linguística que permite expressar um leque importante de regras, que vão da desambiguação das categorias das palavras até à construção de dependências, passando pela delimitação de sintagmas nucleares. Foram desenvolvidas gramáticas para diferente línguas no XIP. Para a gramática do Português, o sistema foi desenvolvido em conjunto no L2F, INESC-ID Lisboa e na Xerox. Este sistema é designado XIP-PT.

As várias etapas do processamento são as seguintes:

- uma fase de pré-processamento que inclui a segmentação, análise morfológica;
- a desambiguação das categorias de palavras;
- a análise sintáctica superficial;
- a análise sintáctica em dependências.

4.1.1 Pré-Processamento Linguístico

A etapa de pré-processamento inclui a segmentação e a análise morfológica das unidades textuais. O XIP-PT integra dois sistemas de pré-processamento desenvolvidos independentemente por cada uma das instituições que colaboram neste trabalho. A entrada do pré-processamento é um texto bruto ou XML. A saída do pré-processamento consiste num lista de unidades às quais é associada informação morfo-sintáctica (possivelmente ambígua). Nota-se que no XIP, uma entrada lexical é representada por um conjunto de traços (atributos:valores) que explicitam a informação linguística associada a esta entrada; trata-se, naturalmente, de informação morfos-sintáctica, mas também de natureza sintáctica e semântica.

4.1.2 Desambiguação

A desambiguação das categorias das palavras é feita de maneira híbrida: É utilizado um modelo escondido de Markov (HMM, *hidden Markov model*) em conjunto com regras construídas manualmente. Com efeito, o XIP oferece um formalismo de desambiguação que permite, considerando um contexto à esquerda e à direita de uma dada forma ambígua, escolher de entre um conjunto de categorias a categoria mais adequada ou preferencial.

4.1.3 Análise Sintáctica de Superfície

A análise sintáctica de superfície permite agrupar sequências de palavras, construindo sintagmas nucleares (sintagmas não recursivos no sentido dos *chunks* de Abney ((Abney, 1991)). Para o fazer, o XIP oferece um formalismo de regras de reescrita contextuais. É também graças a este formalismo que são elaboradas as regras do sistema de reconhecimento de entidades mencionadas (REM) integrado no XIP (Hagège, Baptista e Mamede, 2008a).

4.1.4 Análise Sintáctica em Dependências

A partir dos sintagmas nucleares delimitados na etapa anterior, e considerando a organização destes sintagmas e as propriedades lexico-sintácticas das unidades lexicais que os integram, é então possível estabelecer ligações (relações de dependência) entre os diversos elementos das frases. Estas ligações constituem relações orientadas, que são etiquetadas com o nome de uma função sintáctica. Assim, por exemplo, em Português, se um grupo nominal (NP) estiver à direita de um verbo e se um outro grupo nominal já tiver associado a esse mesmo verbo através da função sintáctica de sujeito, então, tipicamente o NP deverá desempenhar a função sintáctica de complemento directo; estas relações de dependência são construídas ligando a cabeça sintáctica dos sintagmas nucleares (chunks); no exemplo acima, a cabeça do NP estará ligada ao verbo com uma ligação de tipo complemento directo.

4.1.5 Ilustração

Para ilustrar as diferentes etapas de processamento, apresentamos a análise da sequência *Eis um exemplo que ilustra o funcionamento do XIP*.

Um excerto da saída do pré-processamento da frase inicial (apenas a cadeia *um exemplo que ilustra*) tem a seguinte forma (ver a figura 1): a sequência é inicialmente segmentada em entradas lexicais. O primeiro campo da saída corresponde à forma, o segundo ao lema da palavra e o terceiro à informação associada à palavra; os números correspondem ao *offset* da palavra; o resto da informação é apresentado sob a forma de traços booleanos. Note-se que as palavras *um*, *que* e *ilustra* são ambíguas, pelo que cada leitura corresponde a uma linha diferente.

```
um      um      +4+6+Pron+Indef+Masc+Sg+#lex+hmm_QUANTSG
um      um      +4+6+Art+Indef+Masc+Sg+#lex+hmm_QUANTSG
um      um      +4+6+Symbol+Meas+Abbr+#lex+hmm_SYM

exemplo exemplo +7+14+Noun+Masc+Sg+#lex+hmm_NSG

que     que     +15+18+Pron+Rel+MF+SP+#lex+hmm_PRONREL
que     que     +15+18+Pron+Interrog+MF+Sg+#lex+hmm_PRONSG
que     que     +15+18+Det+Interrog+MF+SP+#lex+hmm_DETINT
que     que     +15+18+Conj+#lex+hmm_CONJSUB

ilustra ilustrar +19+26+Verb+PresInd+3P+Sg+#lex+hmm_VERBF
ilustra ilustrar +19+26+Verb+Impv+2P+Sg+#lex+hmm_VERBF
```

Figura 1: A saída do pré-processamento para a sequência *um exemplo que ilustra*

À saída da fase de desambiguacão, a mesma sequência (ver a figura 2) já só apresenta para a palavra *um* uma única leitura, a de artigo indefinido.

Da mesma maneira, para a forma de entrada *que* já só subsiste a leitura de pronome relativo. No que diz respeito ao verbo, *ilustra*, na medida em que a ambiguidade se estabelece entre apenas duas formas verbais conjugadas (indicativo presente, terceira pessoa do singular e imperativo na segunda pessoa do singular) do mesmo lema *ilustrar*, não é ainda feita a sua desambiguação.

Na fase seguinte, o sistema procede a uma análise sintáctica de superfície que permite construir, a partir da frase inicial, uma sequência de *chunks*.

```
ADVP{Eis<ADV>}
NP{um<ART> exemplo<NOUN>}
SC{que<PRON>
   VF{ilustra<VERB>}
  }
NP{o<ART> funcionamento<NOUN>}
PP{de<PREP> o<ART> XIP<NOUN>}
```

Finalmente, são construídas, com base nesta primeira organização em *chunks*, uma série de relações de dependência entre os constituintes da frase:

```
MAIN(exemplo)
QUANTD(exemplo,um)
DETD(funcionamento,o)
DETD(XIP,o)
PREPD(XIP,do)
VDOMAIN(ilustra,ilustra)
MOD_POST(funcionamento,XIP)
MOD_POST(ilustra,XIP)
SUBJ(ilustra,que)
CDIR_POST(ilustra,funcionamento)
SUBORD(que,ilustra)
```

Assim, a relação `CDIR_POST` indica que o complemento directo de *ilustra* é *funcionamento*. A relação `DETD` liga a cabeça nominal *XIP* com o artigo *o*. Note-se que, nesta fase, não se desambigua a dependência do complemento preposicional *do XIP* (trata-se do conhecido problema do *PP-attachment*). Por esta razão, temos duas relações concorrentes de modificador envolvendo a palavra *XIP* e que exprime a ambiguidade de o complemento preposicional *do XIP* poder *a priori* encontrar-se ligado tanto a *funcionamento* como a *ilustra*.

4.2 Módulo XIP para a Anotação de Expressões Temporais

O desenvolvimento deste módulo foi iniciado em 2007 (Loureiro, 2007) e profundamente revisto para a campanha do HAREM em 2008 (Hagège, Baptista e Mamede, 2008b) na qual se propôs uma tarefa especial para anotação temporal.

Como se disse atrás, este módulo de processamento temporal está integrado num sistema mais geral de análise linguística, o XIP e executa as seguintes tarefas:

1. Reconhecimento e delimitação das ET nos textos;
2. Classificação destas ET;
3. Normalização (de um subconjunto) das ET;
4. Ligação entre as ET e os processos.

A realização destas tarefas é feita em paralelo às diferentes etapas de processamento do XIP descritas acima.

4.2.1 Pré-Processamento

Ao nível do pré-processamento, a implementação do módulo de análise temporal obriga à introdução de nova informação lexical. Com efeito, para o processamento temporal é necessário especificar mais a informação linguística de base associada a certos elementos lexicais (nome de

```
um       um       +4+6+Art+Indef+Masc+Sg+#lex+hmm_QUANTSG

exemplo exemplo +7+14+Noun+Masc+Sg+#lex+hmm_NSG

que      que      +15+18+Pron+Rel+MF+SP+#lex+hmm_PRONREL

ilustra ilustrar         +19+26+Verb+PresInd+3P+Sg+#lex+hmm_VERBF
ilustra ilustrar         +19+26+Verb+Impv+2P+Sg+#lex+hmm_VERBF
```

Figura 2: A saída da fase de desambiguacão para a sequência *um exemplo que ilustra*

meses, nome de dias), bem como a certas cadeias numéricas (números de 4 dígitos que possam corresponder a anos ou número de 1 a 2 dígitos potencialmente correspondentes à indicação de meses, etc.). Tecnicamente, esta especificação do léxico faz-se com introdução de novos traços, que serão depois utilizados nas gramáticas locais para reconhecimento de expressões temporais.

Por exemplo, à palavra *semana*, que, para o sistema geral de processamento do Português, é apenas considerada como um nome feminino, acrescenta-se o traço booleano `time_meas:+`, que indica tratar-se uma medida de tempo. De forma similar, ao lema nominal *primavera* acrescenta-se o traço *season:2*, que o especifica como um nome de estação do ano e o identifica com o número 2 (que será depois usado para cálculos).

4.2.2 Desambiguação de Categorias

O processamento temporal obrigou à introdução no sistema de novas regras de desambiguação. Por exemplo, o sistema de processamento do Português inicial considerava a palavra *Natal* como tendo apenas uma leitura, como nome próprio. É evidente, no entanto, que, num contexto de processamento do tempo, a distinção entre *Natal*, estado no Brasil, ou *Natal*, dia ou altura do ano, tem ser estabelecida. A regra seguinte determina que, quando a palavra *Natal* está precedida da preposição *durante*, a qual, por sua vez, pode ser seguida por um determinante e, eventualmente, por um adjectivo, esta palavra deverá ter apenas a leitura correspondente à expressão de tempo, passando, por esta razão a apresentar um traço específico `one_day` com valor `+`.

```
20> noun[maj:+,surface:Natal] %=
   |prep[lemma:durante],(art;?[dem:+]),(adj)|
   noun[one_day=+,maj=+,proper=+].
```

A primeira linha corresponde à parte esquerda da regra de desambiguação e significa que ela só será despoletada quando encontrar o nome *Natal*. A segunda linha corresponde ao padrão que deve seguir o contexto esquerdo da palavra para que a regra possa ser aplicada. Este contexto é uma expressão regular que descreve a seguinte sequência: a preposição *durante* seguida opcionalmente por um artigo ou um determinante demonstrativo, seguido ainda por um adjectivo opcional. Finalmente, a terceira linha indica os traços que devem ser acrescentadas à palavra *Natal* para que passe a ter apenas a leitura que corresponde à expressão temporal.

4.2.3 Gramáticas Locais

As gramáticas locais agrupam elementos lexicais, geralmente enriquecidos por nova informação relevante relativa ao tempo, para assim formar expressões temporais. Simultaneamente a este agrupamento, pode-se, em certos casos, proceder a uma primeira classificação de algumas expressões temporais. Por exemplo, no caso de datas completas (i.e., que incluem o número de dia, o nome de mês e o número de ano), como se trata de uma data absoluta não é necessário qualquer contexto para uma correcta classificação destas expressões. No mesmo sentido, o exemplo que se segue mostra a regra que permite construir uma ET a partir de o nome de uma estação do ano, seguido da preposição *de* e por uma sequência de dígitos correspondentes a um número que represente um ano, tal como *Primavera de 2002*.

```
18> noun[time=+,date=+,tipo_tempref=absolut]
    @=
    ?[season], prep[lemma:de],
    (?[lemma:o]), num[dig,year=+].
```

A parte esquerda da regra corresponde à expressão, constituída pelo elemento *Primavera*, que emparelha com o traço `?[season]` na parte direita; este elemento aparece então seguido pela preposição *de* e, por sua vez, por uma sequência numérica à que se vai acrescentar o traço `year:+`.

4.2.4 Dependências Sintácticas

As dependências vão permitir:

- Determinar a que predicado está ligado a ET; isto é feito graças à gramática geral

do Português, que calcula relações de modificação entre um predicado e um modificador;

- Caracterizar de forma mais pormenorizada certos tipos de ET que não podem ser classificados com um simples contexto local. Esta classificação pode ser feita graças às relações já calculadas, por exemplo entre o predicado e a ET-alvo que o modifica, mediante informação adicional obtida a partir desse predicado.

Considerem-se, por exemplo, as duas frases:

São **duas horas**
Ficou **duas horas** *em casa*

No primeiro caso, a expressão temporal constitui uma data relativa, com uma granularidade correspondente à unidade de medida hora. Trata-se, de facto, de uma expressão formular para indicar as horas, que apresenta alguma fixidez sintáctica. No segundo caso, estamos perante uma construção locativa, em que o sujeito está omisso, e que é facultativamente modificada por uma expressão de tempo do tipo DURAÇÃO e igualmente expressa em horas; nesta expressão, o verbo é tradicionalmente analisado como um verbo copulativo.

Para que um sistema automático seja capaz de fazer esta distinção, é necessário considerar o tipo de predicado expresso em cada frase e a forma como este se encontra associado à sequência *duas horas* (v.g. a construção formular de *ser*, no primeiro caso, e a construção locativa *ficar em casa*, no segundo). No léxico do sistema, está disponível a informação de que, entre outros traços, *ficar* pode ter um valor aspectual permansivo (anotado `permanency`). Com base nesta informação lexical, a regra seguinte determina que, perante um verbo copulativo com o valor permansivo, um complemento que possua o traço `time:+` deve ser classificado como uma duração.

```
// ficou 2 horas
if (^PREDSUBJ(#1[permanency],#2[time]))
        MOD[post=+](#1,#2),
        NE[tempo=+,duration=+](#2).
```

Na primeira linha, verifica-se a existência de uma relação `PREDSUBJ` entre um verbo copulativo (excluindo o verbo *ser*) e o complemento, que o verbo tem o traço `permanency`) e que o seu complemento constitui uma expressão de tempo, isto é, apresenta o traço `time`). As segundas e terceiras linhas correspondem às novas relações que são criadas se a condição expressa na primeira linha for verdadeira. Nesse caso, constrói-se uma expressão temporal `NE` do tipo DURAÇÃO e é estabelecida uma relação de modificação entre o verbo e a expressão temporal. Finalmente, é destruída a relação `PREDSUBJ` existente entre o verbo e a ET.

4.2.5 Cálculos Externos

Além destas tarefas que estão simplesmente integradas no analisador linguístico, há necessidade de realizar cálculos numéricos para proceder à normalização de expressões temporais dos tipos data absoluta, horas e durações. Assim, associam-se acções às regras para permitir realizar a normalização. Essas acções são chamadas a funções Python que podem ser executadas directamente a partir do analisador (Roux, 2006).

As expressões de subtipo DATA são normalizadas e o valor da normalização guardado no atributo `VAL_NORM` com o seguinte formato:

```
<Era><Ano><Mes><Dia>T<Hora><Minuto>
  E<ESTACAO>LM<limite\_aberto>
```

em que:

- `<Era>` corresponde a 2 caracteres: '+' ou '-', conforme a data seja depois ou antes da era de referência; e uma das seguintes letras maiúsculas, que representa a era de referência: C para a era cristã ocidental (valor por defeito), H para a era muçulmana (de Hijra, Hégira); M (anno Mundi) para o calendário judaico; P para a cronologia arqueológica (Presente = 1950); etc.
- `<Milenio>` corresponde a 2 caracteres de tipo dígito que representam o valor do milénio;
- `<Seculo>` corresponde a 2 caracteres de tipo dígito que representam o valor do século;
- `<Decada>` corresponde a 2 caracteres de tipo dígito que representam o valor da década;
- `<Ano>` corresponde a 4 dígitos que representam o valor do ano;
- `<Mes>` corresponde a 2 dígitos que representam o valor do mês;
- `<Dia>` corresponde a 2 dígitos que representam o valor do dia;
- `<Hora>` corresponde a 2 dígitos que representam o valor da hora;
- `<Minuto>` corresponde a 2 dígitos que representam o valor dos minutos;
- `<Segundo>` corresponde a 2 dígitos que representam o valor dos segundos;
- `<Milissegundo>` corresponde a 2 dígitos que representam o valor dos milissegundos;

- `<ESTACAO>` corresponde a 2 letras capitalizadas correspondente às estações do ano: `PR` para Primavera, `VE` para Verão, `OU` para Outono e `IN` para Inverno;
- `<limite_aberto>` indica se a expressão normalizada de data absoluta representa um intervalo de tempo com limite anterior ou limite posterior não determinado (em aberto). Os valores respectivos são: `A` no caso de limite anterior em aberto (neste caso a expressão temporal apresenta um limite posterior, e.g. *até 2009*); `P` no caso de limite posterior em aberto (neste caso, a expressão temporal tem um limite anterior, e.g. *desde 2009*); e, finalmente, `-`, quando a data absoluta corresponde a um intervalo sem limites abertos, e.g. *em 2009*.

No caso da data absoluta não ser expressa em termos de algum destes campos, o campo omiso é substituído por um ou mais "-". Por exemplo, a expressão *a 3 de Janeiro de 1986* é normalizada através de "`VAL_NORM=+19860103T----E--LM-`", a expressão *na Primavera de 1996* através de "`VAL_NORM=+1996----T----EPRLM-`"e a a expressão *antes das 3:00 da tarde* através de "`VAL_NORM="+--------T15--E--LMA`".

Para as expressões de tipo DURAÇÃO ou as DATAs relativas, o valor da normalização também é registada no atributo `VAL_DELTA` usando os seguintes campos:

```
A<digitos>D<digitos>H<digitos>
  M<digitos>S<digitos>M<digitos>
```

onde:

- as letras `A`, `D`, `H`, `M`, `S`, `M` são constantes que devem aparecer nesta ordem e que correspondem, respectivamente, ao valores de Anos, Dias, Horas, Minutos, Segundos e Milissegundos;
- os `<digitos>` à direita das letras correspondem ao valor dos campos respectivos; no caso das expressões de DURAÇÃO, este são simplesmente o valor temporal do intervalo de tempo; no caso das DATAs relativas, estes valores correspondem ao intervalo de tempo que se deve adicionar ou diminuir à data de referência para obter o valor temporal da expressão anotada.

Usa-se a tabela 1 para converter as unidades de tempo durante a normalização de durações. Como princípio geral, os valores de VAL_DELTA devem ser convertidos para indicar de forma precisa a duração temporal referida. Para essa consideram-se três intervalos de valores:

- valores superiores a 1 ano;
- valores inferiores a 1 ano e superiores a 1 dia;
- valores inferiores a 1 dia;

Para efectuar a conversão usam-se as seguintes regras:

- todos valores de VAL_DELTA superiores ao ano são convertidos em anos;
- todos valores de VAL_DELTA inferiores a 1 ano e superiores a 1 dia são convertidos em dias;
- para valores de VAL_DELTA inferiores a 1 dia utilizam-se as unidades imediatamente inferiores (horas, minutos, segundos e milissegundos)

As expressões com valores inteiros inferiores a 1 dia não sofrem qualquer conversão: as expressões temporais são normalizadas pela transposição dos valores referidos nas correspondentes unidades temporais (horas, minutos, segundos e milissegundos).

No caso das expressões fraccionárias:

- fracções das unidades temporais são convertidas para a unidade imediatamente inferior;
- se a conversão não corresponder a um inteiro, arredonda-se para o inteiro mais próximo;
- para durações que combinam várias unidades temporais, a conversão faz-se para cada um das quantidades individuais.

Por exemplo, as seguintes expressões devem ser normalizadas como indicado:

- *durante um ano e dez dias* ("`VAL_DELTA=A1D10H0M0S0M0`")
- *durante meio ano* ("`VAL_DELTA=A0D183H0M0S0M0`")
- *2/3 da semana* ("`VAL_DELTA=A0D5H0M0S0M0`")
- *meio dia* ("`VAL_DELTA=A0D0H12M0S0M0`")
- *por hora e meia* ("`VAL_DELTA=A0D0H1M30S0M0`")

Das conversões apresentadas há uma que merece uma chamada de atenção por ser uma aproximação, já que aceitando que um ano tem 365 dias (ignorando os anos bissextos), um mês tem na realidade 30,41(6) dias e não 30 dias.

Por outro lado, as unidades de tempo que efectivamente ocorrem na ET são registadas noutro

Unidade 1	Unidade 2
1 milénio	1000 anos
1 século	100 anos
1 década	10 anos
1 ano	365 dias
1 mês	30 dias
1 quinzena	14 dias
1 semana	7 dias
1 dia	24 horas
1 hora	60 minutos
1 minuto	60 segundos

Tabela 1: Conversão entre unidades para o cálculo do VAL_DELTA.

campo, *UMED*. Por exemplo, a expressão temporal na frase *Fiquei* **dois meses** *em Lisboa* é normalizada através de "`VAL_NORM=A0D60H0M0S0MO UMED=meses`".

A tarefa de normalização é obtida através da análise dos pares atributo:valor associados aos elementos constituintes de cada entidade temporal normalizável. Para simplificar a tarefa de normalização, atribuem-se alguns traços específicos aos diferentes elementos que constituem a expressão temporal, nomeadamente aos dígitos que podem representar anos, meses, dias, horas, minutos e segundos, assim como as sequências alfabéticas para os meses e respectivas abreviaturas e os nomes das estações do ano. Assim, na expressão *25/Dez/2009*, o número *25* é associado à propriedade `day:+`, *Dez* à propriedade `month:+` e *2009* recebe a propriedade `year:+`.

A normalização "resume-se", então, a percorrer todos os nós das entidades temporais e a converter para o formato adequado todos os nós que contiverem um dos traços relevantes para a normalização. Contudo, é ainda necessário tratar de forma especial todas as ET que:

- contém elementos em numeração romana (*século XVI*);
- envolvem unidades de tempo não representadas na normalização final (*fim-de-semana*, *semana*, *quinzena* ou *semestre*);
- exprimem fracções de unidades de tempo(*meio ano*, *um mês e meio*);
- constituem maneiras informais de indicação das horas, como por exemplo *meia-noite*, *3 horas da tarde*, *2 menos um quarto* e *5 para as 3*;
- incluem expressões não numéricas referentes a durações (*uma quinzena de dias*) ou advérbios de tempo (*amanhã*, *anteontem* e *antes de ontem*);
- incluem diferentes tipos de modificador com valor referencial particular (*no dia seguinte*, *na semana que vem*, *no mês passado* e *no ano que há-de vir*).

4.3 Resultados

A avaliação do Módulo de Anotação Temporal teve lugar na campanha do Segundo HAREM (tarefa de anotação de expressões temporais). Sete sistemas participaram nesta tarefa, embora nem todos pretendessem tratar as ET ao mesmo nível de granularidade. Ainda assim, esta forte participação mostra o interesse da comunidade de processamento computacional do Português por este tema. Apenas um sistema se apresentou com o objectivo de realizar todas as dimensões da tarefa, (incluindo a de normalização).

Os resultados obtidos pelo sistema apresentado neste artigo são bastante animadores. Considerando-se as tarefas de identificação e de classificação de expressões temporais[4], o sistema atingiu uma precisão de 0,85 e uma abrangência de 0,76. Alguns erros ficaram a dever-se ao facto de, por enquanto, a tarefa de identificação e o processo de classificação terem sido feitos apenas ao nível local e, por essa razão, a semântica particular do processo associado às ET não ter podido ainda ser levado em linha de conta. Outros erros deveram-se a uma ainda incompleta codificação no léxico dos elementos que funcionam como índices (*triggers*) lexicais temporais. A normalização das datas absolutas e normalização parcial de datas referenciais também produziu resultados promissores, tendo o sistema conseguido um resultado de 0,74 de medida-f. No entanto, é opinião dos autores de que apenas a consideração do contexto mais alargado que envolve as expressões temporais poderá vir a melhorar de forma significativa estes resultados.

5 Conclusão

O processamento temporal é uma tarefa ambiciosa mas importante no domínio mais amplo de extracção de informação a partir de textos. Esta linha de pesquisa tem vindo a ser desenvolvida há já algum tempo e para diversos idiomas. Para Português, no entanto, a investigação neste domínio está ainda no seu início.

Em primeiro lugar, uma das dificuldades, consiste justamente em caracterizar de forma adequada o que se entende por expressões temporais,

[4] As directrizes para a classificação das ET propostas no Segundo HAREM são ligeiramente diferente das que aqui apresentámos, no entanto, elas são compatíveis.

tendo em conta as suas propriedades referenciais e sem perder de vista o objetivo principal da tarefa, isto é, a ordenação parcial dos processos, estados ou eventos expressos nos textos ao longo do eixo temporal. Têm sido desenvolvidas diferentes orientações e critérios para identificar, segmentar e classificar expressões temporais. Este artigo apresenta um conjunto de directrizes, inspiradas nas normas das campanhas de avaliação internacionais, com o objectivo de dar, assim, um passo firme mas significativo no sentido de um processamento temporal eficiente de textos em Português.

Foi igualmente desenvolvido um módulo temporal para reconhecer e classificar automaticamente as expressões temporais que aparecem nos textos de acordo com essas orientações. Nesta fase, o módulo temporal só opera ao nível sintáctico, mas os resultados obtidos são já bastante encorajadores. É, no entanto, bastante óbvio que o contexto imediato da frase não é suficiente para ancorar temporalmente os eventos, efectuar uma ordenação (parcial) temporal precisa dos eventos ou resolver todas as relações de referência temporal. De facto, um cálculo adequado da dimensão temporal veiculada nos textos tem de ter em conta muito mais elementos, como o tempo e o aspecto verbal, e diferentes processos de modelização do discurso, que alteram as condições de ancoragem temporal dos eventos. Também deverá ser necessário ter em consideração fenómenos que relevam da organização do discurso como, por exemplo, os referidos por Lascarides e Asher ((Lascarides e Asher, 1993)).

Acreditamos que, perante a quantidade e a diversidade de parâmetros que devem ser considerados para o tratamento da dimensão temporal e dado o facto de a anotação manual da informação temporal ser um trabalho extremamente difícil e custoso, uma abordagem baseada em regras e que explora informação linguística construída manualmente é a estratégia mais adaptada para esta tarefa. O nosso trabalho constitui apenas um primeiro passo nessa direcção. Esperamos poder avançar a pouco e pouco neste caminho, integrando progressivamente na nossa ferramenta de processamento do Português uma caracterização cada vez mais precisa dos diferentes tipos de eventos, estados e processos e desenvolvendo um módulo de cálculo referencial.

Referncias

Abney, S. P. 1991. Parsing by chunks. Em R. C. Berwick, S. P. Abney, e C. Tenny, editores, *Principle-Based Parsing: Computation and Psycholinguistics*. Kluwer, Dordrecht, pp. 257–278.

Aït-Mokhtar, Salah, Jean-Pierre Chanod, e Claude Roux. 2002. Robustness beyond shallowness: Incremental deep parsing. Em *Natural Language Engineering, 8*. Cambridge University Press, New York, NY, USA, pp. 121–144.

Allen, James F. 1991. Time and time again: The many ways to represent time. Wiley and Sons, pp. 341–355.

Baptista, Jorge, Caroline Hagège, e Nuno Mamede. 2008. Capítulo 2: Identificação, classificação e normalização de expressões temporais do português: A experiência do segundo HAREM e o futuro. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Baptista, Jorge, Nuno Mamede, e Caroline Hagège, 2009. *Time Expressions in Portuguese. Guidelines for Indentification, Classification and Normalization (Internal Report L2F–INESC-ID)*, May, 2009.

Battistelli, Delphine, Jean-Luc Minel, e Sylviane Schwer. 2006. Représentation des expressions calendaires dans les textes: vers une application à la lecture assistée de biographies. *Traitement Automatique des Langues*, pp. 11–37.

Bittar, André. 2008. Annotation des informations temporelles dans des textes en français. Em *Actes de RECITAL 2008*.

Ehrmann, Maud e Caroline Hagège. 2009. Proposition de caractérisation et de typage des expressions temporelles en contexte. Em *Actes de TALN 2009*, Senlis, France.

Gosselin, Laurent. 1996. *Sémantique de la temporalité en français. Un modèle calculatoire et cognitif du temps et de l'aspect*. Duculot.

Hagège, Caroline e Xavier Tannier. 2008. Xtm: A robust temporal text processor. Em *Proceedings of CICLing 2008*, Haïfa, Israël.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede. 2008a. Capítulo 15: Reconhecimento de entidades mencionadas com o xip: Uma colaboração entre o INESC-L2F e a Xerox. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede, 2008b. *Proposta de anotação*

e normalização de expressões temporais da categoria TEMPO para o HAREM-MII. `http://www.linguateca.pt/aval_conjunta/HAREM/2008_04_13_Tempo.pdf`.

Lascarides, Alex e Nicholas Asher. 1993. Temporal interpretation, discourse relations, and commonsense entailment. Springer, http://www.springerlink.com, pp. 437–493.

Loureiro, João Miguel. 2007. Reconhecimento de Entidades Mencionadas (Obra, Valor, Relações de Parentesco e Tempo) e Normalização de Expressões Temporais. Tese de Mestrado, Universidade Técnica de Lisboa, Instituto Superior Técnico, Lisboa, Portugal, November, 2007.

Mota, Cristina e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.* Linguateca, Aveiro, Portugal.

Parent, Gabriel, Michel Gagnon, e Philippe Muller. 2008. Annotation d'expressions temporelles et d'événements en français. Em *Actes de TALN 2008*, Avignon, France.

Prior, Arthur N. 1957. *Time and Modality.* Oxford University Press.

Reichenbach, Hans. 1947. *Elements of Symbolic Logic.* Reprinted, 1980, Dover Publications, New York.

Roux, Claude. 2006. Coupling a linguistic formalism and a script language. Em *Proceedings of CSLP-06 - Coling-ACL*, Sydney, Australia.

Saurí, Roser, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, e James Pustejovsky, 2006. *TimeML Annotation Guidelines Version 1.2.1*, January, 2006. `http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf`.

Verhagen, M., R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, e J. Pustejovsky. 2007. Xrce-t: Xip temporal module. Em *SemEval-2007 - Task 15 TempEval Temporal Relation Identification.*

Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação

Hugo Gonçalo Oliveira
CISUC, Universidade de Coimbra, Portugal
hroliv@dei.uc.pt

Diana Santos
Linguateca, SINTEF ICT, Noruega
Diana.Santos@sintef.no

Paulo Gomes
CISUC, Universidade de Coimbra, Portugal
pgomes@dei.uc.pt

Resumo

Neste artigo apresentamos o PAPEL, um recurso lexical para o português, constituído por relações entre palavras, extraídas de forma automática de um dicionário da língua geral através da escrita manual de gramáticas para esse efeito. Depois de contextualizarmos o tipo de recurso e as opções tomadas, fornecemos uma visão do processo da sua construção, apresentando as relações incluídas e a sua quantidade. Apresentamos também uma primeira avaliação, que tomou duas formas: para as relações de sinonímia, a comparação com o TeP 2.0, um recurso publicamente acessível e de cobertura vasta; para as outras relações, interrogando corpos em português. Esta segunda forma pode ser efectuada automaticamente, ou recorrendo a avaliadores. Nesta última vertente, integrado no projecto AC/DC, é oferecido mais um serviço de validação de relações à comunidade do processamento computacional da língua portuguesa, onde qualquer utilizador pode actuar como avaliador.

1 Introdução

Cada vez mais os estudos do processamento da língua exigem que haja acesso computacional a informação semântica, e é cada vez mais frequente o recurso a redes ou ontologias lexicais que tentam cobrir o panorama lexical de uma língua toda, ao invés, ou como complemento, de terminologias, cujo objectivo é descrever uma área específica do conhecimento. A ontologia lexical paradigmática é a WordNet (Fellbaum, 1998), também chamada WordNet de Princeton (WordNet.Pr), embora uma ontologia mais relacionada com o nosso trabalho seja a MindNet (Richardson, Dolan e Vanderwende, 1998).

Neste artigo apresentamos o PAPEL, Palavras Associadas Porto Editora - Linguateca, `http://www.linguateca.pt/PAPEL` (desde 17 de Agosto de 2009 livre e publicamente acessível), que é pioneiro para o português, ao tentar obter uma ontologia lexical semi-automaticamente a partir de um dicionário, o *Dicionário PRO da Língua Portuguesa da Porto Editora* (DLP, 2005).

Como é notado por Sampson (2000) na sua apreciação da WordNet.Pr, é curioso que tenha sido uma abordagem manual a preferida pela comunidade do processamento de linguagem natural (PLN), mas o que é certo é que a maior parte dos projectos associados ou inspirados pela WordNet seguem uma metodologia que usa peritos para criar o recurso manualmente. Pensamos que uma das razões para isto se deve à questão dos direitos de autor, e nesse aspecto pode ser que o PAPEL seja o primeiro recurso totalmente público baseado num dicionário comercial, visto que a MindNet é propriedade de uma empresa.

Visto que não existe ainda uma terminologia completamente consensual, cumpre indicar aqui, na senda de Veale (2007), o que designamos por *ontologia lexical* de uma dada língua:

- uma estrutura de conhecimento que relaciona itens lexicais (vulgo, palavras) de uma língua entre si, por relações que têm a ver com o significado desses mesmos itens;
- uma estrutura que pretende abranger a língua toda e não conhecimento de um domínio em particular, ou seja, que não se encontre restrita a campos específicos.

Deixamos desde já bem claro que, dentro desta descrição razoavelmente abrangente, existem muitas perguntas específicas a que cada criador de recurso terá de dar uma resposta, assim como não há respostas precisas para o que é uma "palavra" (e de facto a maior parte das

ontologias lexicais de que temos conhecimento usam também expressões) ou o que é a "língua toda".

De um ponto de vista operacional, é mais natural desde já afirmarmos que o PAPEL não pretende ser uma resposta definitiva a estas questões, mas sim uma abordagem concreta que se apoiou no trabalho de lexicógrafos. Quanto à noção de palavra/entrada e ao conjunto de itens que fazem parte da língua geral, mas que, sabendo que a língua é uma entidade claramente dinâmica, a nossa intenção é vir a expandir o PAPEL tendo em conta esse facto.

Há no entanto duas questões, completamente ortogonais, que nos parecem estabelecer uma delimitação clara na paisagem das ontologias lexicais, e sobre as quais posicionamos de imediato aqui o PAPEL:

- o carácter público ou privado de um recurso: em que o PAPEL alinha com a WordNet, ou seja, é público;
- a construção manual ou automática a partir de um dicionário com definições (ou seja, um recurso já existente), em que o PAPEL alinha com a MindNet, ou seja, é construído a partir de um recurso.

Outras opções tomadas, e que nos separam de outros recursos ou abordagens, serão mencionadas à medida que as formos apresentando.

Por interessar mais à audiência deste texto, e também a nós, vamos centrar a discussão nos recursos que existem para o português de que temos conhecimento, nomeadamente a WordNet.PT (Marrafa, 2002), o TeP (Dias-Da-Silva e Moraes, 2003) (Maziero et al., 2008) e a WordNet.BR (Dias da Silva, Oliveira e Moraes, 2002), e ainda a MultiWordNet.PT (`http://mwnpt.di.fc.ul.pt`).

É importante contudo referir que não vemos nem desenvolvemos o PAPEL[1] como sendo um competidor em relação ao trabalho já existente, mas sim como mais uma contribuição para obter informação semântica de cobertura vasta para o português.

Consideramos, de facto, que a situação ideal seria a de ter uma ontologia lexical pública para todo o português, embora naturalmente entrando em conta com as diferenças entre as variedades ou variantes da língua (Barreiro, Wittmann e Pereira, 1996). Em Santos et al. (2009), seguido de Santos et al. (2010), apresentámos uma primeira comparação entre vários recursos que sublinha a sua complementaridade.

Nessa linha, tentaremos convencer os leitores de que as formas de avaliação que descrevemos na secção 4 constituem um bom início para uma ligação e consequente actualização de ambos os recursos envolvidos (o TeP e o PAPEL), além de apresentarmos também uma oferta de validação para outros recursos existentes ou que venham a ser desenvolvidos para o português em conjunto com a interrogação de corpos em português.

2 Contexto

Desde muito cedo que foi reconhecido que, para realizar o processamento computacional de uma língua, seria necessário o acesso a recursos de grande cobertura, como o são as ontologias lexicais, ou antes de esse termo ser cunhado, a dicionários em forma electrónica ou bases de dados lexicais, por um lado, ou bases de conhecimento sobre o mundo, por outro. Para uma excelente discussão da diferença e relação entre ontologias e bases de dados lexicais, veja-se Hirst (2004). Outras abordagens interessantes em relação a essa questão são Dahlgren (1995) e Marcellino e Dias da Silva (2009).

2.1 Modelos de ontologia lexical

2.1.1 A escola da WordNet

A WordNet.Pr é uma ontologia lexical para o inglês, construída manualmente, que procura representar a forma como o ser humano processa o vocabulário. Está disponível gratuitamente e ao longo dos anos tem sido amplamente utilizada pela comunidade do PLN. A sua estrutura mais básica é um grupo de sinónimos (do inglês, *synset*), ou seja, um conjunto de palavras que, em determinado contexto, podem ter o mesmo significado e ser utilizadas para representar o mesmo conceito. Uma rede semântica estabelece-se na WordNet.Pr através de ligações, correspondentes a relações semânticas, entre os nós, que correspondem aos grupos de sinónimos. Entre as relações cobertas encontram-se a hiponímia e a meronímia (entre substantivos) e a troponímia e a implicação (entre verbos). Há ainda a dizer que no léxico da WordNet.Pr há uma clara distinção entre nós que são substantivos, verbos, adjectivos, advérbios ou palavras gramaticais. Além de ser possível levantar gratuitamente várias versões da WordNet.Pr, através da sua página, em `http://wordnet.princeton.edu` é também possível interrogar a sua versão mais

[1] Quando o projecto de construção do PAPEL foi iniciado pela Linguateca em colaboração com a Porto Editora, após assinatura de um protocolo em Maio de 2006, não havia nenhum recurso publicamente disponível para o português. Congratulamo-nos muitíssimo pelo facto de existirem agora vários.

recente, a 3.0, através de uma interface na rede.

Dado o enorme sucesso da WordNet.Pr, o seu modelo foi seguido para representar ontologias lexicais noutras línguas. Dessas destacam-se as wordnets criadas para as línguas presentes no projecto EuroWordNet (Vossen, 1997; Vossen, 1998), mais propriamente o holandês, castelhano, italiano, francês, alemão e estónio. A ideia do EuroWordNet foi alinhar várias wordnets com a WordNet.Pr.

Destacamos ainda as wordnets para a língua portuguesa, a WordNet.PT, para a variante europeia, e a WordNet.BR, para a variante brasileira[2]. Há no entanto a lamentar que ambos os projectos tardem a tornar os seus conteúdos acessíveis para o público. Por exemplo, apesar da existência de uma interface na rede para interrogar a WordNet.PT (ou parte dela), a partir de `http://cvc.instituto-camoes.pt:8080/wordnet/index.jsp`, não é costume ser possível realizar pesquisas porque o sistema se encontra permanentemente em manutenção. Contudo, os grupos de sinónimos da WordNet.BR, bem como as relações de antonímia, encontram-se disponíveis no Thesaurus Electrônico do Português, o TeP (Maziero et al., 2008), também ele construído de acordo com os princípios da WordNet.Pr.

Inspirado pelo EuroWordNet, o projecto MultiWordNet (Pianta, Bentivogli e Girardi, 2002) procurou também alinhar várias wordnets com a WordNet.Pr, mas desta vez, ao invés de se procurar as correspondências possíveis entre as wordnets existentes nas diferentes línguas e a WordNet.Pr, a ideia foi criar novas wordnets onde fosse mantida a maior parte dos nós e relações presentes na WordNet.Pr. Desta forma, na MultiWordNet, wordnets para o italiano, o espanhol, o romeno, o hebraico, o latim e, mais recentemente, o português (`http://mwnpt.di.fc.ul.pt`) estão alinhadas com a WordNet.Pr.

2.1.2 A MindNet

Além do modelo da WordNet, outro tipo de recurso que pode ser visto como uma ontologia lexical é a base de conhecimento MindNet.

A MindNet é mais do que um recurso estático e pode ser visto como uma metodologia que envolve um conjunto de ferramentas para adquirir, estruturar, aceder e explorar, de forma automática, informação léxico-semântica contida em texto. Como, numa fase inicial, o recurso foi construído a partir de um dicionário para a língua inglesa, a sua estrutura é baseada em entradas de dicionário. Desta forma, para cada palavra definida, além de informação típica num dicionário (e.g. informações gramaticais) existe um conjunto de registos associados aos sentidos que a palavra pode ter. Por sua vez, para cada sentido, além da definição, encontram-se ligações a outras entradas, sendo que cada ligação tem um tipo correspondente a uma relação gramatical (e.g. sujeito típico, predicado típico) ou semântica (e.g. sinónimo, hiperónimo, parte, causa, finalidade, maneira). Estas relações são extraídas com base na aplicação de regras sobre árvores sintáctico-semânticas, produzidas por um analisador sintáctico de vasta cobertura. Cada relação estabelecida tem um peso atribuído de acordo com a sua saliência.

A MindNet pode ser interrogada através do MindNet Explorer (MNEX) (Vanderwende et al., 2005), a partir do `http://stratus.research.microsoft.com/mnex/Main.aspx`, onde é possível procurar caminhos (de relações semânticas) entre duas palavras.

2.1.3 Outros recursos semânticos

As bases de senso comum são outro tipo de recurso semântico, sendo o recurso mais conhecido o Cyc (Lenat, 1995), uma base de conhecimento baseada em lógica de predicados de primeira ordem, que vem sendo criada de forma manual.

Outro recurso deste tipo é a ConceptNet (Liu e Singh, 2004), construído de forma automática a partir do preenchimento de frases matriz, tal como *The effect of eating food is ...*, ou o *A knife is used for ....* A ConceptNet utiliza uma representação semelhante à do WordNet.Pr, mas inclui conhecimento mais informal, de uma natureza mais prática e além disso tem um maior elenco de relações (tais como propriedade de, subevento de, efeito de, utilizado para).

Tanto o Cyc como a ConceptNet têm associadas capacidades de raciocínio, de forma a ser possível inferir novas relações. No entanto, enquanto no Cyc o raciocínio é realizado sob representações em lógica de predicados, na ConceptNet o raciocínio é feito sobre representações em linguagem natural.

A FrameNet (Baker, Fillmore e Lowe, 1998), por seu lado, é uma rede semântica baseada no conceito de enquadramentos (em inglês, *frames*) (Fillmore, 1982). Nesta representação, cada enquadramento descreve um objecto, um evento ou um estado, que corresponde a um conceito e se pode relacionar com outros

[2]Para uma comparação entre as ontologias lexicais existentes para o português, mais propriamente a WordNet.PT, a WordNet.BR e o TeP, a MultiWordNet.PT, e ainda o PAPEL, recomenda-se a leitura de Santos et al. (2010).

enquadramentos, através de um conjunto de relações semânticas (e.g. herança, sub-frame, causador, utiliza). Para o português existe já um projecto seguidor deste modelo de recurso, o FrameNet Brasil (Salomão, 2009), `http://www.framenetbr.ufjf.br/`, veja-se também Afonso (2009).

Devemos também citar o Port4NooJ (Barreiro, No prelo), um conjunto de recursos linguísticos construídos no ambiente de desenvolvimento linguístico do NooJ (Silberztein e Varadi, No prelo), que tem em vista o processamento automático do português. Estes recursos encontram-se publicamente disponíveis em `http://www.linguateca.pt/Repositorio/Port4Nooj/` e são usados em várias ferramentas públicas para o português e outras línguas. Os recursos correspondem a léxicos e a gramáticas com finalidades diversas: análise morfológica, sintáctico-semântica, desambiguação, identificação de unidades lexicais multipalavra, parafraseamento e tradução. O Port4NooJ inclui além disso uma extensão bilingue, permitindo a sua utilização em aplicações como a tradução automática do português para o inglês. As diferentes propriedades associadas aos itens lexicais contidas nos recursos provêm do OpenLogos, um sistema de tradução automática em código aberto derivado do sistema Logos (Scott, 2003), mas novas propriedades têm sido adicionadas através do NooJ e encontram-se em fase de validação, entre as quais relações semânticas, como apresentado em Santos et al. (2010).

2.1.4 Sentidos numa ontologia lexical

Enquanto que, pela escola da WordNet, cada nó da rede representa um sentido e uma "mesma" palavra pode pertencer a vários nós, que são sim as unidades básicas, no PAPEL a única distinção de sentidos feita tem a ver com a categoria gramatical, ou seja, um nó do PAPEL é uma palavra gráfica (com uma dada categoria: substantivo, adjectivo, etc.). Esta opção tem duas razões de ser: uma filosófica e outra prática. A primeira prende-se com a concepção de que a língua é soberana (Santos, 2006) e distinções de sentido são sempre imprecisas (Kilgarriff, 1996) e artificiais; veja-se Saussure (1916) para a descrição de uma língua como sistema sincrónico, e Edmonds e Hirst (2002) sobre o problema dos quase-sinónimos. A segunda razão tem que ver com o facto de, nas definições de um dicionário, as palavras que ocorrem nas definições não aparecem indexadas pelos sentidos, tornando por isso quase impossível fazer essa identificação automaticamente.

Aliás, confrontado com o mesmo problema, no âmbito da MindNet, Dolan (1994) propôs fazer a "ambiguação" de sentidos relacionados. Desta forma, numa primeira fase de construção, a MindNet é uma rede entre palavras, tal e qual se encontram no dicionário, e os seus registos são relativos a palavras. Apenas numa segunda fase se procura atribuir um sentido a cada uma destas palavras, tirando partido dos campos de domínio ou de co-ocorrências nas definições.

Também a partir de uma rede onde a unidade básica é a palavra, sem qualquer distinção de sentidos, e onde as ligações, pesadas, apenas indicam a co-ocorrência em corpos, Dorow (2006) aplica algoritmos estatísticos sobre grafos para extrair informação semântica interessante. Por exemplo, quando dois nós não estão ligados ou têm uma ligação muito fraca (isto é, as palavras não co-ocorrem frequentemente), mas têm uma vizinhança semelhante, é provável que sejam sinónimos. Por outro lado, quando um nó é a única ligação entre duas sub-redes, é provável que se esteja perante uma palavra com dois sentidos.

Ainda relativamente à representação dos sentidos numa ontologia lexical, os recursos que resultam de uma tradução cega de um recurso deste tipo feito para uma língua diferente, como as MultiWordNets, têm de lidar, adicionalmente às questões decorrentes da imprecisão existente na identificação de sentidos, com problemas mais específicos relacionados com a tradução. Como línguas diferentes representam diferentes realidades sociais e culturais, estas não cobrem exactamente a mesma parte do léxico e, mesmo nas partes que lhes são comuns, os vários conceitos são normalmente lexicalizados de forma diferente (Hirst, 2004). Isto leva a que, por exemplo, na MultiWordNet.PT faltem palavras para identificar alguns conceitos importados da WordNet.Pr (Santos et al., 2009), assim como muito provavelmente faltarão conceitos específicos das realidades portuguesa e brasileira.

2.2 Abordagens para a construção de uma ontologia lexical

Há basicamente três formas consagradas de construção de um recurso semântico de cobertura larga: (i) trabalho manual; (ii) processamento de corpos; e (iii) processamento de dicionários; apesar de novas ideias terem surgido nos últimos tempos, como por exemplo através da análise de logs (Costa e Seco, 2008) ou jogos colaborativos.

O PAPEL (Gonçalo Oliveira et al., 2008) seguiu a terceira via: foi construído a partir da análise automática das definições constantes numa versão electrónica do *Dicionário PRO da Língua Portuguesa* . A utilização de dicionários

em formato electrónico com vista à construção de recursos lexicais iniciou-se há cerca de quarenta anos, com os estudos de Calzolari, Pecchia e Zampolli (1973) para o italiano e de Amsler (1981) para o inglês. Os autores que utilizaram dicionários apontam várias razões para a sua escolha como ponto de partida para a construção automática de uma ontologia lexical: além de serem uma enorme fonte de conhecimento lexical (Briscoe, 1991) e serem vistos como autoridades no que diz respeito ao sentido das palavras (Kilgarriff, 1997), a sua estrutura e a previsibilidade e simplicidade do vocabulário utilizado nas definições facilitam a sua utilização para a extracção e organização de informação léxico-semântica. Com base no trabalho de Amsler (1981), Chodorow, Byrd e Heidorn (1985) criaram procedimentos semi-automáticos para a extracção da relação de hiperonímia a partir de um dicionário. Alshawi (1989) desenvolveu uma gramática que tinha como único objectivo a derivação das definições de um dicionário específico, de forma a facilitar a extracção de relações que eram depois organizadas em estruturas semânticas. Montemagni e Vanderwende (1992), por outro lado, defenderam a utilização de um analisador sintáctico de grande cobertura, com o argumento de que este seria melhor para extrair informação mais específica dentro de uma definição.

Apesar de vários trabalhos com este objectivo, a MindNet terá sido a primeira base de dados lexical independente, criada de forma automática a partir de dicionários, mas não houve muitos continuadores nesta senda, talvez devido à análise sobre a inconsistência dos dicionários feita por Ide e Veronis (1995). Ainda assim, alguns trabalhos recentes nesta área são O'Hara (2005), Nichols, Bond e Flickinger (2005) e Zesch, Müller e Gurevych (2008), este último usando o Wikcionário[3].

Por outro lado, vários investigadores apontaram o facto de que algum conhecimento importante para o PLN não se encontrava presente em dicionários: algumas aplicações necessitam de conhecimento específico sobre determinados domínios, que é mais fácil de obter em corpos (Hearst, 1992; Riloff e Shepherd, 1997; Caraballo, 1999).

Para a extracção de conhecimento que não se consegue encontrar nem em dicionário, nem em outros recursos de vasta cobertura, como qualquer WordNet já existente, iniciou-se o processamento de recursos não estruturados.

No que diz respeito à utilização de recursos estruturados (ou semi-estruturados) para extrair conhecimento léxico-semântico, nos últimos anos tem também sido dada especial atenção à utilização de recursos colaborativos, como a Wikipédia[4] ou o já referido Wikcionário, veja-se por exemplo Medelyan et al. (2009), Navarro et al. (2009) ou Herbelot e Copestake (2006).

A referência mais conhecida no que diz respeito à extracção de conhecimento léxico-semântico a partir de corpos é o trabalho de Hearst (1992), que propõe um método para identificar padrões textuais indicadores da relação de hiponímia e que aplica um conjunto de padrões para extrair automaticamente relações deste tipo. Vários trabalhos tiveram como principal inspiração a abordagem de Hearst para descobrir padrões e para extrair relações, não só de hiponímia (Caraballo, 1999; Freitas e Quental, 2007), mas também outros tipos de relações, como por exemplo causais (Girju e Moldovan, 2002), ou de meronímia (ou parte de) (Berland e Charniak, 1999), e mais especificamente para relações geográficas em português (Chaves, 2009).

2.3 Abordagens para a avaliação de ontologias

Brank, Grobelnik e Mladenić (2005) apresentam quatro formas que têm sido utilizadas para avaliar ontologias de domínio: (i) avaliação manual; (ii) comparação com um recurso dourado; (iii) realização de uma tarefa independente, definida para avaliar uma ontologia; (iv) comparação com um conjunto de dados sobre o mesmo domínio.

Apesar de, regra geral, estas formas de avaliação se adaptarem a qualquer tipo de ontologia, é preciso notar que temos de distinguir entre as ontologias propriamente ditas (Gruber, 1993), que cobrem uma área específica e são baseadas numa conceptualização de um domínio, e as ontologias lexicais que, como já referimos, tentam descrever o sistema conceptual de uma língua inteira. Isto leva naturalmente a que nem todos os métodos possam ser adaptados cegamente a ontologias lexicais.

A avaliação manual é uma forma habitualmente escolhida para avaliar a qualidade de um recurso. Muitos trabalhos efectuam este tipo de avaliação — por exemplo Riloff e Shepherd (1997), Caraballo (1999), ou mesmo Richardson, Vanderwende e Dolan (1993), no âmbito do que viria a ser a MindNet — por ser provavelmente a forma mais fiável. No entanto, está sempre dependente de trabalho por parte dos indivíduos que realizam a avaliação. De forma a minimizar o

[3] `http://wiktionary.org/`

[4] `http://wikipedia.org`

esforço necessário para avaliar manualmente uma ontologia obtida automaticamente, Navigli et al. (2004) geraram definições em linguagem natural a partir do conteúdo dessa ontologia.

Para utilizar um recurso dourado, que pode eventualmente ser outra ontologia, é necessário que exista um elevado nível de confiança na sua correcção, possivelmente por ter sido criado manualmente por peritos. A qualidade de uma ontologia pode ser assim medida através da sua comparação com um recurso dourado, de acordo com determinados critérios. Neste tipo de avaliação, Santos (2007) refere que as medidas de precisão e abrangência, tradicionalmente utilizadas em recolha de informação (Salton e McGill, 1983), têm sido extremamente populares em PLN, sendo muitas vezes propostas sem uma total compreensão das suas limitações e adequação.

Outro problema desta abordagem de avaliação é que, sendo a criação de ontologias um assunto bastante recente, nem sempre existe um recurso dourado que se adeque aos critérios da avaliação. Para o inglês, no âmbito das ontologias lexicais, muitos autores utilizam a própria WordNet.Pr como recurso dourado na avaliação da sua ontologia (Hearst, 1992; Nichols, Bond e Flickinger, 2005).

Partindo do princípio de que uma ontologia serve para ser integrada noutras aplicações, com o objectivo de realizar determinadas tarefas, alguns autores propõem avaliar uma ontologia de forma indirecta. Desta forma a ontologia é utilizada numa aplicação para realizar uma tarefa específica, cujos resultados serão alvo de avaliação. No entanto, é necessário ter algum cuidado com as ilações tiradas deste tipo de avaliação, já que há muitas variáveis envolvidas e a qualidade dos resultados não está apenas dependente da qualidade da ontologia, mas também do resto da aplicação. Cuadros e Rigau (2006) realizaram uma avaliação indirecta de várias ontologias lexicais, incluindo a WordNet.Pr, no âmbito da desambiguação do sentido das palavras. Curiosamente, os recursos criados de forma automática obtiveram melhores resultados ao nível tanto da precisão como da abrangência. Outra conclusão a que chegaram foi a de que a qualidade dos resultados obtidos, ao combinar o conhecimento de todos os recursos utilizados no estudo, é muito próxima daquela que apenas selecciona o sentido mais frequente para cada palavra.

Quanto à última forma de avaliação, a comparação com outros dados referentes ao mesmo domínio, Brewster et al. (2004) propõem que a adequação de uma ontologia de domínio a um dado corpo seja avaliada através do número dos termos salientes do corpo, que será sobre o domínio em questão, que também constam na ontologia. Contudo, repare-se que, para obter os termos salientes num dado domínio, é preciso precisamente compará-lo com a linguagem geral e outros domínios, e obter os termos salientes na linguagem geral é algo que não faz muito sentido. Ainda assim, será possível medir a cobertura de um determinado corpo por um léxico, tal como Demetriou e Atwell (2001) propõem. A cobertura será medida através do número de palavras do corpo que se encontrarem no léxico.

A verdade, contudo, é que, tal como Raman e Bhattacharyya (2008) referem, a avaliação explícita de ontologias lexicais não é uma prática comum. A principal razão para esta situação será o facto de haver bastante confiança nestes recursos, que são na sua maioria criados manualmente por peritos, o que minimiza a possibilidade de erros. De forma a verificar se a confiança é justificada, Raman e Bhattacharyya (2008) levaram a cabo uma validação automática dos grupos de sinónimos (*synsets*) da WordNet.Pr, utilizando um dicionário. Nesse trabalho consideraram que uma palavra estava correctamente incluída num nó da WordNet se na sua definição fossem referidas palavras dos nós hiperónimos desse nó, ou outras palavras pertencentes ao mesmo nó (sinónimos). Como esperado, não foram encontrados muitos problemas.

Há ainda a referir um outro método de avaliação que tira partido da quantidade de texto que se consegue encontrar hoje em dia na Web, como fizeram, por exemplo, Etzioni et al. (2005) para calcular o nível de confiança de relações de hiperonímia entre classes e entidades mencionadas. Para o efeito, as relações foram primeiro transformadas em padrões textuais discriminadores, semelhantes aos de Hearst (1992). Em seguida, procuraram esses padrões na rede e calcularam o PMI-IR (Turney, 2001) entre os padrões envolvendo a entidade e as ocorrências da própria entidade.

3 Breve apresentação do PAPEL

Nesta secção descrevemos primeiro o procedimento semi-automático utilizado para construir o PAPEL e de seguida apresentamos os conteúdos da sua versão actual, incluindo a contabilização de itens lexicais, a contabilização de relações, e ainda exemplos destas últimas.

```
PARTE{
  nome:nome * PARTE_DE:INCLUI;
  nome:adj * PARTE_DE_ALGO_COM_PROPRIEDADE:PROPRIEDADE_DE_ALGO_QUE_INCLUI;
  adj:nome * PROPRIEDADE_DE_ALGO_PARTE_DE:INCLUI_ALGO_COM_PROPRIEDADE;
}
```

Figura 1: Exemplo da descrição do grupo de relações relativas à meronímia.

3.1 Construção

De forma resumida, visto que já foi detalhado noutras publicações (Gonçalo Oliveira et al., 2008; Gonçalo Oliveira e Gomes, 2008b; Gonçalo Oliveira e Gomes, 2008a), o processo de construção do PAPEL segue um ciclo de três passos até considerarmos ter chegado a um nível de desempenho suficiente, entrando depois no quarto e último passo.

1. **Criação de gramáticas semânticas**: foram criadas gramáticas para cada tipo de relação que se pretende extrair, por categoria gramatical (fornecida pelo dicionário). Na tabela 1 mostramos alguns dos padrões e as relações que pretendem descobrir e na figura 1 mostramos de que forma as relações que pretendemos extrair são descritas, de acordo com o grupo e especificando ainda a categoria dos argumentos e a sua relação inversa.

Padrão	Relação associada
`tipo|género|classe|forma de`	Hiperonímia
`parte|membro de`	Meronímia
`que causa|provoca|origina`	Causa
`usado|utilizado para`	Finalidade
`natural|originário de`	Local
uma palavra ou lista de palavras	Sinonímia

Tabela 1: Exemplos de padrões usados nas gramáticas.

2. **O processo de extracção**: usando um analisador automático, é feita a análise superficial das definições, a partir da qual são automaticamente extraídas relações (descritas no passo anterior) entre palavras na definição e a palavra definida, também chamada "verbete" (ver figura 2).

3. **Inspecção dos resultados**: usando um sistema de regressão para identificar mais facilmente as diferenças entre resultados anteriores, procede-se à inspecção manual dos resultados obtidos, com o eventual regresso ao primeiro passo para corrigir problemas detectados ou melhorar as gramáticas.

4. **Ajuste das relações**: aqui procura-se corrigir (ou eliminar) de forma automática relações com argumentos inválidos.

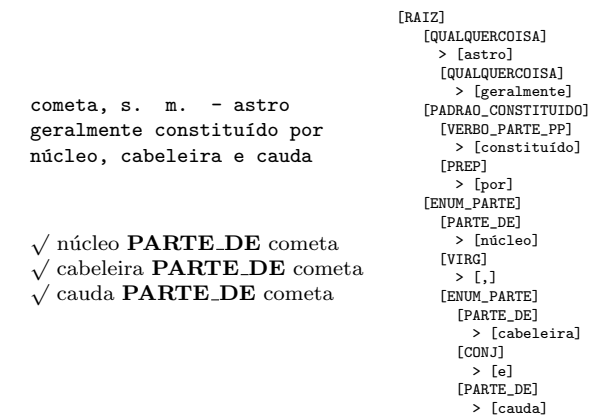


Figura 2: O resultado da análise da definição de *cometa*.

O último passo é realizado em dois tempos. Inicialmente, todas as relações são transformadas no tipo directo[5]. Por exemplo, *manga* INCLUI *punho* é convertida para *punho* PARTE_DE *manga*, e *dor* RESULTADO_DE *distensão* é transformada em *distensão* CAUSADOR_DE *dor*.

Visto que as gramáticas não fazem uma análise sintáctica das definições, não atribuindo por exemplo a classe gramatical, e que as definições do dicionário apenas incluem a classificação da vedeta, em alguns casos o processo de construção automática do PAPEL resulta em relações entre palavras de categorias erradas. É por isso preciso verificar, também de uma forma automática, esses casos, usando primeiro a própria lista de palavras/vedetas do dicionário e em seguida o analisador morfológico Jspell (Simões e Almeida, 2002). Se conseguirmos apurar que há um desajuste nas categorias mas que pode ser corrigido através da escolha de outra relação pertencente ao mesmo grupo, substituímos, senão removemos esse triplo. Por exemplo, a relação *loucura* ACCAO_QUE_CAUSA *desvario* – que pressupõe um verbo como primeiro argumento – é transformada automaticamente em *loucura* CAUSADOR_DE *desvario*, visto que ambos os argumentos são

[5] A escolha de um tipo directo e outro inverso foi arbitrariamente efectuada pelos criadores das gramáticas por um critério de naturalidade, e não de frequência, no dicionário ou em texto, e não tem qualquer consequência excepto a de facilitar a arrumação e depuração do recurso.

Categoria	Simples	Multipalavra	Total
Substantivo	52.599	3.334	55.933
Verbo	10.195	13.866	24.061
Adjectivo	21.000	1	21.001
Advérbio	1.390	0	1.390

Tabela 3: Distribuição dos items por categoria gramatical, no PAPEL 2.0

substantivos. Durante este processo, os casos das palavras flexionadas são também substituídos pelos seus lemas, quando essa informação é dada pelo Jspell.

3.2 Conteúdos

Após realizadas as quatro fases da sua construção, a versão actual do PAPEL, 2.0, contém perto de 100.000 items lexicais, cujas categorias gramaticais se distribuem de acordo com a tabela 3, e perto de 200.000 relações, distribuídas de acordo com a tabela 2. A sinonímia e a hiperonímia são as relações mais frequentes, e ainda podem ser aumentadas, como discutiremos abaixo, de uma forma semelhante ao feito no ReRelEM (Freitas et al., 2008; Freitas et al., 2009).

Como também podemos ver na tabela 3, a maior parte dos itens lexicais são expressões de uma única palavra. No entanto, o PAPEL também inclui expressões multipalavra, em casos como os seguintes:

- Substantivos seguidos das preposições *de/do/dos/da/das* e de uma outra palavra (e.g. *sistema de rodas*, *dispositivo de mira*);
- Verbos com o seu objecto directo (e.g. *abrir o apetite*, *produzir som*);

4 Avaliação do PAPEL

Aqui descrevemos uma avaliação inicial do PAPEL, feita de duas formas diferentes: as relações de sinonímia foram comparadas com as relações representadas num thesaurus para o português, enquanto que as restantes relações, apenas entre substantivos, foram validadas através das sua transformação em padrões textuais e procura em texto por esses padrões.

Esta avaliação foi incialmente feita sobre a primeira versão pública do PAPEL (1.0) e os seus resultados publicados em Gonçalo Oliveira, Santos e Gomes (2009b), junto com uma primeira motivação para o procedimento usado. No entanto na versão 1.0 do PAPEL existia apenas um tipo de relação de sinonímia, e não um tipo para cada categoria gramatical, o que não permitiu uma comparação com base neste ponto. Sendo assim, partes dessa avaliação foram repetidas para a versão 2.0 do PAPEL e completadas com uma avaliação da cobertura do PAPEL, com a qual iniciamos esta secção. No que diz respeito à avaliação das demais relações, tal como a parte não repetida da avaliação da sinonímia, calculamos que os valores não terão sofrido grandes alterações, por isso repetimos aqui os valores obtidos para a versão 1.0 do PAPEL.

Nas duas primeiras avaliações apresentadas a seguir, o TeP foi utilizado como recurso de referência, não só por ser possível levantá-lo na rede, mas também por se tratar de um recurso criado manualmente e que, tal como o PAPEL, pretende abranger toda a língua. Ainda assim, estamos conscientes das várias diferenças entre as variantes de português. O TeP 2.0 contém 19.888 nós, ou seja grupos de unidades lexicais com o mesmo sentido, correspondendo a 43.118 unidades lexicais (também designadas por termos neste artigo) ao todo.

4.1 Avaliação da cobertura

Esta avaliação teve como objectivo verificar quantos termos do TeP se encontravam também no PAPEL e vice-versa. Ao fim de comparar ambos os recursos, verificamos que existiam 28.971 termos comuns a ambos os recursos, o que corresponde a 30,0% dos termos do PAPEL e 68,2% dos termos do TeP.

Mais dados desta comparação podem ser consultados nas tabelas 4 e 5 onde, respectivamente, se encontram os resultados separados por termos simples e multipalavra, ou a proporção de termos comuns de acordo com a sua categoria gramatical. Tal como também é frisado por Santos et al. (2010), estes resultados revelam que, apesar de ambos os recursos terem o mesmo objectivo procurarem representar a mesma realidade, acabam por ser bastante complementares.

A título de curiosidade, indicamos ainda que os únicos três termos multipalavra comuns ao PAPEL e ao TeP são: *corrente de ar*, *pena de morte* e ainda *tremor de terra*.

Recurso	Simples		Multipalavra		Total
PAPEL	79.337	82,2%	17.201	17,8%	96.538
TeP	42.777	99,2%	341	0,8%	43.118
Ambos	28.971	100%	3	0%	28.974

Tabela 4: Termos no PAPEL 2.0 e TeP 2.0.

4.2 Avaliação da sinonímia

Para que a avaliação da sinonímia pudesse prosseguir sem enviesamento, começámos por retirar da comparação os termos do TeP que não estivessem

Grupo	Nome	Args.	Qnt.	Exemplos
Sinonímia	SINONIMO_N_DE	n,n	37.452	(*auxílio*, *contributo*)
	SINONIMO_V_DE	v,v	21.465	(*tributar*, *colectar*)
	SINONIMO_ADJ_DE	adj,adj	19.073	(*flexível*, *moldável*)
	SINONIMO_ADV_DE	adv,adv	1.171	(*após*, *seguidamente*)
Hiperonímia	HIPERONIMO_DE	n,n	62.591	(*planta*, *salva*)
Parte	PARTE_DE	n,n	2.805	(*cauda*, *cometa*)
	PARTE_DE_ALGO_COM_PROP	n,adj	3.721	(*tampa*, *coberto*)
Membro	MEMBRO_DE	n,n	5.929	(*ervilha*, *Leguminosas*)
	MEMBRO_DE_ALGO_COM_PROP	n,adj	34	(*pessoa*, *colectivo*)
	PROP_DE_ALGO_MEMBRO_DE	adj,n	883	(*celular*, *célula*)
Causa	CAUSADOR_DE	n,n	1.013	(*fricção*, *assadura*)
	CAUSADOR_DE_ALGO_COM_PROP	n,adj	17	(*paixão*, *passional*)
	PROP_DE_ALGO_QUE_CAUSA	adj,n	498	(*reactivo*, *reacção*)
	ACCAO_QUE_CAUSA	v,n	6.399	(*limpar*, *purgação*)
	CAUSADOR_DA_ACCAO	n,v	39	(*gases*, *fumigar*)
Produtor	PRODUTOR_DE	n,n	898	(*romãzeira*, *romã*)
	PRODUTOR_DE_ALGO_COM_PROP	n,adj	35	(*sublimação*, *sublimado*)
	PROP_DE_ALGO_PRODUTOR_DE	adj,n	359	(*fotógeno*, *luz*)
Finalidade	FINALIDADE_DE	n,n	2.886	(*defesa*, *armadura*)
	FINALIDADE_DE_ALGO_COM_PROP	n,adj	63	(*reprodução*, *reprodutor*)
	ACCAO_FINALIDADE_DE	v,n	5.192	(*fazer_rir*, *comédia*)
	ACC_FINALIDADE_DE_ALGO_COM_PROP	v,adj	260	(*corrigir*, *correccional*)
Localização	LOCAL_ORIGEM_DE	n,n	849	(*Japão*, *japonês*)
Maneira	MANEIRA_POR_MEIO_DE	adv,n	1.113	(*timidamente*, *timidez*)
	MANEIRA_SEM	adv,n	117	(*devagar*, *pressa*)
	MANEIRA_SEM_ACCAO	adv,v	11	(*assiduamente*, *faltar*)
Propriedade	PROP_DE_ALGO_REFERENTE_A	adj,n	6.518	(*dinâmico*, *movimento*)
	PROP_DO_QUE	adj,v	17.543	(*familiar*, *ser_conhecido*)

Tabela 2: As relações do PAPEL 2.0 e respectivas quantidades

Comparação	Substantivos	Verbos	Adjectivos	Advérbios	Total
PAPEL no TeP	19,8%	28,5%	33,9%	38,8%	30,0%
TeP no PAPEL	64,1%	62,9%	47,5%	47,5%	68,2%

Tabela 5: Termos comuns ao PAPEL 2.0 e TeP 2.0, por categoria gramatical.

presentes no PAPEL assim como todos os casos de relações do PAPEL que contivessem argumentos ausentes do TeP. Ficámos assim apenas com 68% dos triplos de sinonímia do PAPEL e com 62% das possíveis 202.980 relações do Tep[6]. A comparação de ambos os conjuntos de relações produziu os seguintes resultados: 50,1% das nossas relações estavam presentes no TeP, e 21,3% das relações do TeP estavam presentes no PAPEL[7]. A tabela 6 apresenta os resultados desta comparação distribuídos de acordo com a categoria gramatical dos argumentos de cada relação de sinonímia. Se no TeP cada grupo de sinónimos vem acompanhado da indicação da sua categoria gramatical, para a sinonímia, esta informação é disponibilizada apenas a partir do PAPEL 1.1, onde passou a existir uma relação de sinonímia para cada categoria gramatical, mais precisamente SINONIMO_N_DE (substantivos), SINONIMO_V_DE (verbos), SINONIMO_ADJ_DE (adjectivos) e SINONIMO_ADV_DE (advérbios).

Embora os valores apresentados possam ser surpreendentes, convém relembrar que as nossas relações tinham de ser encontradas directamente no dicionário, e não foram portanto ainda alvo de qualquer raciocínio. Em particular, a relação de transitividade parece ser óbvia: A SINONIMO_DE B $\wedge$ B SINONIMO_DE C $\rightarrow$ A SINONIMO_DE C. Esta regra foi aplicada uma vez só ao PAPEL 1.0 onde os 80.432 sinónimos iniciais deram origem a 689.073 sinónimos derivados.

Claro está que, como as definições (e as nossas regras) não separam entre sentidos distintos de uma mesma palavra, esta expansão poderá levar a muitas relações infelizes, tal como *queda* SINONIMO_DE *ruína* $\wedge$ *queda* SINONIMO_DE *habilidade* $\rightarrow$ *ruína* SINONIMO_DE *habilidade*. Após esta expansão, e como esperado, o número de casos atestado no TeP caiu para 14%,

[6]Para conversão do TeP todos os elementos de um grupo de sinónimos foram considerados como pertencendo a uma relação de sinonímia com todos os outros elementos do mesmo grupo.

[7]Outra das razões que nos levou a repetir esta parte da avaliação foi, na primeira avaliação descrita em Gonçalo Oliveira, Santos e Gomes (2009a) e Gonçalo Oliveira, Santos e Gomes (2009b), termos ignorados todos os termos do PAPEL que não constavam de pelo menos uma relação de sinonímia no PAPEL, o que levou à eliminação de mais termos do TeP e, consequentemente, a um valor superior para a percentagem de relações do TeP presentes no PAPEL.

contudo, 90% das relações no TeP puderam ser encontradas no PAPEL 1.0. Fica assim demonstrado que a combinação dos dois recursos permite não só melhorar ambos como separar o trigo do joio e mesmo alertar automaticamente para palavras com vários sentidos.

4.3 Avaliação das demais relações

Em relação às outras relações, e na impossibilidade de comparar automaticamente com outros recursos para o português, tivemos de desenvolver uma metodologia diferente, inspirada nos vários trabalhos de extracção automática de relações semânticas em texto, ou de validação das mesmas em texto.

Para os nossos testes usámos o CETEMPúblico (Rocha e Santos, 2000), através da interface do projecto AC/DC (Santos e Bick, 2000; Santos e Sarmento, 2003; Costa, Santos e Rocha, 2009; Santos, 2009). Este serviço permitiu-nos além disso ter acesso às frequências dos lemas respectivos. O trabalho realizado tem de ser considerado preliminar, já que, devido a limitações de ocorrência de muitas das unidades lexicais nos corpos que usámos, não tivemos possibilidade de as testar. Com efeito, não só muitas das palavras no PAPEL eram demasiado raras ou especializadas, como cedo nos demos conta que em texto jornalístico seria quase impossível encontrar num mesmo contexto (numa mesma frase) pares ou relações como *liquidar* ACCAO_QUE_CAUSA *liquidação*, *fósforo* PARTE_DE_ALGO_COM_PROPRIEDADE *fosforoso*, visto que são característicos de texto dicionarístico ou enciclopédico.

Restringimos assim o processo de validação, em primeiro lugar, apenas a relações entre substantivos, e, além disso, retirámos do teste as relações que envolvessem palavras cujos lemas estivessem ausentes do CETEMPúblico. Mesmo assim, e por questões de sobrecarga do serviço, para as duas relações mais populosas do PAPEL, hiperonímia e meronímia, ainda escolhemos uma amostra aleatória de relações a testar, correspondente respectivamente a 8% e 63% dos casos. Os resultados encontram-se na tabela 8.

Cerca de 20% destas relações parecem ser validadas ou confirmadas pelo corpo, enquanto que a percentagem é menor para as outras relações. Estes resultados parecem-nos satisfatórios, tendo em conta que: o corpo é bastante pequeno; os padrões usados foram muito simples (em texto real há uma miríade de outras possibilidades de indicar uma relação); e os nossos valores não se encontram demasiado longe daqueles apresentados na literatura de confirmação.

De qualquer maneira, e para mostrarmos que esta confirmação está longe de ser definitiva ou mesmo conclusiva, na tabela 7 apresentamos alguns exemplos, quer de confirmação certa quer de espúria (ou seja, parecem confirmar mas não o fazem).

Casos que não foram confirmados embora existam ambas as palavras no CETEMPúblico são, por exemplo, *fruto* HIPERONIMO_DE *alperce*, *algoritmia* PARTE_DE *matemática*, *ausência* CAUSADOR_DE *saudade*, *tamareira* PRODUTOR_DE *tâmara*, e *aquecimento* FINALIDADE_DE *salamandra*.

5 Ferramentas

Esta secção apresenta duas ferramentas associadas ao PAPEL, para a sua exploração e validação.

5.1 Folheador

O Folheador é uma interface na rede desenvolvida para navegar num conjunto de relações, como as do PAPEL, depois de carregadas numa base de dados. Este sistema encontra-se actualmente instalado no URL `http://sancho.dei.uc.pt/folheador/` e permite fazer procuras no PAPEL.

O seu funcionamento é muito simples: basta procurar por uma palavra e o sistema responde com uma lista de todas as relações onde essa palavra entra. Se a palavra tiver mais de uma categoria gramatical possível, as relações são separadas de acordo com a categoria gramatical. Além disso, é possível filtrar o resultado por tipo de relação e ainda, ao clicar nas palavras em argumentos das relações apresentadas, verificar todas as relações onde estas últimas estejam envolvidas. Na figura 3 é apresentada uma imagem do Folheador, depois de procurar pela palavra *vencedor*.

5.2 VARRA

Para permitir uma validação mais pormenorizada das relações presentes no PAPEL – e possivelmente noutros recursos – desenvolvemos o VARRA (Validação, Avaliação e Revisão de Relações no AC/DC), em conjunto com o projecto AC/DC, de forma a obter julgamentos mais completos em relação à seguinte questão: dado um triplo e uma possível frase que o ilustra e consequenetemente valida, obtida automaticamente dos corpos do AC/DC, pedimos às pessoas que escolham uma das seis possíveis alternativas:

1. Relação claramente incorrecta. Passe à frente
2. Relação possivelmente correcta. O texto ilustra a relação entre as duas palavras?

Comparação	Substantivos	Verbos	Adjectivos	Advérbios	Total
PAPEL no TeP	47,6%	52,0%	53,4%	54,3%	50,1%
TeP no PAPEL	28,4%	15,2%	24,4%	36,6%	21,3%

Tabela 6: Triplos comuns ao PAPEL 2.0 e TeP 2.0, por categoria gramatical.

Relação	Certa?	Justificação
língua HIPERONIMO_DE *italiano*	Sim	*As **línguas latinas, como o italiano** ou o português, tornam-se mais fáceis por causa das vogais.*
arbusto PARTE_DE *floresta*	Sim	*A **floresta é um conjunto de árvores, arbustos** e ervas de várias qualidades e tamanhos.*
cólera CAUSADOR_DE *diarreia*	Sim	*A **cólera provoca fortes diarreias** e vómitos e pode levar à desidratação e, consequentemente, à morte em poucas horas.*
oliveira PRODUTOR_DE *azeitona*	Sim	*Também a quantidade e tamanho das **azeitonas produzidas por uma oliveira** biológica é inferior, já que não são utilizados compostos de azoto que ajudam a planta a crescer.*
recrutamento FINALIDADE_DE *inspecção*	Sim	*Menos de metade dos jovens entre os 20 e os 22 anos apresentaram-se às **inspecções para recrutamento**, revelou o ministro da Defesa.*
músico PARTE_DE *música*	Não	*... um espectáculo baseado na obra "Cantos de Maldoror", de Lautréamont, com **música composta pelo músico** inglês Steven Severin...*
fim FINALIDADE_DE *sempre*	Não	*Sicilia aponta **sempre para o fim** do dia, para o fim da luz.*

Tabela 7: Exemplos de validação automática do PAPEL 1.0 através do CETEMPúblico (republicados de Gonçalo Oliveira, Santos e Gomes (2009a)).

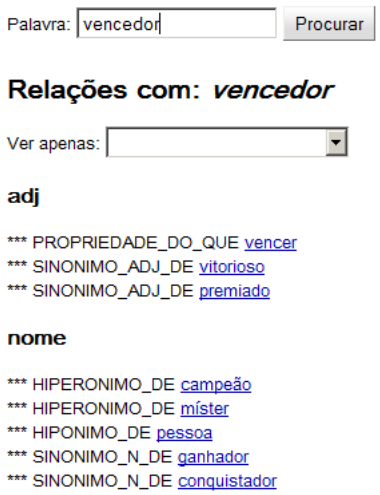


Figura 3: Resultados para a procura pela palavra *vencedor*, no Folheador.

(a) Sim

(b) Não... É compatível mas não exactamente.

(c) Não... O texto é completamente não relacionado.

(d) Não... Pelo contrário, invalida-a.

(e) Não sei

Esse serviço, acessível a partir de `http://www.linguateca.pt/ACDC/`, é ilustrado na figura 4 e encontra-se presentemente em fase de teste. Futuramente pretendemos alargá-lo de forma a que sirva também para avaliar outro tipo de recursos e de padrões de procura, além de poder ser usado pedagogicamente na formação de alunos na área de linguística com corpos.

6 *Considerações finais*

Apresentámos neste artigo um novo recurso lexical para o português, o PAPEL, que pode ser levantado integralmente no endereço acima citado, junto com ampla documentação sobre o mesmo. Também apresentámos algumas ferramentas relacionadas com este recurso. Esta primeira abordagem à avaliação do PAPEL, apesar de bastante preliminar, pode ser interessante como exemplo de avaliação, também para outros recursos.

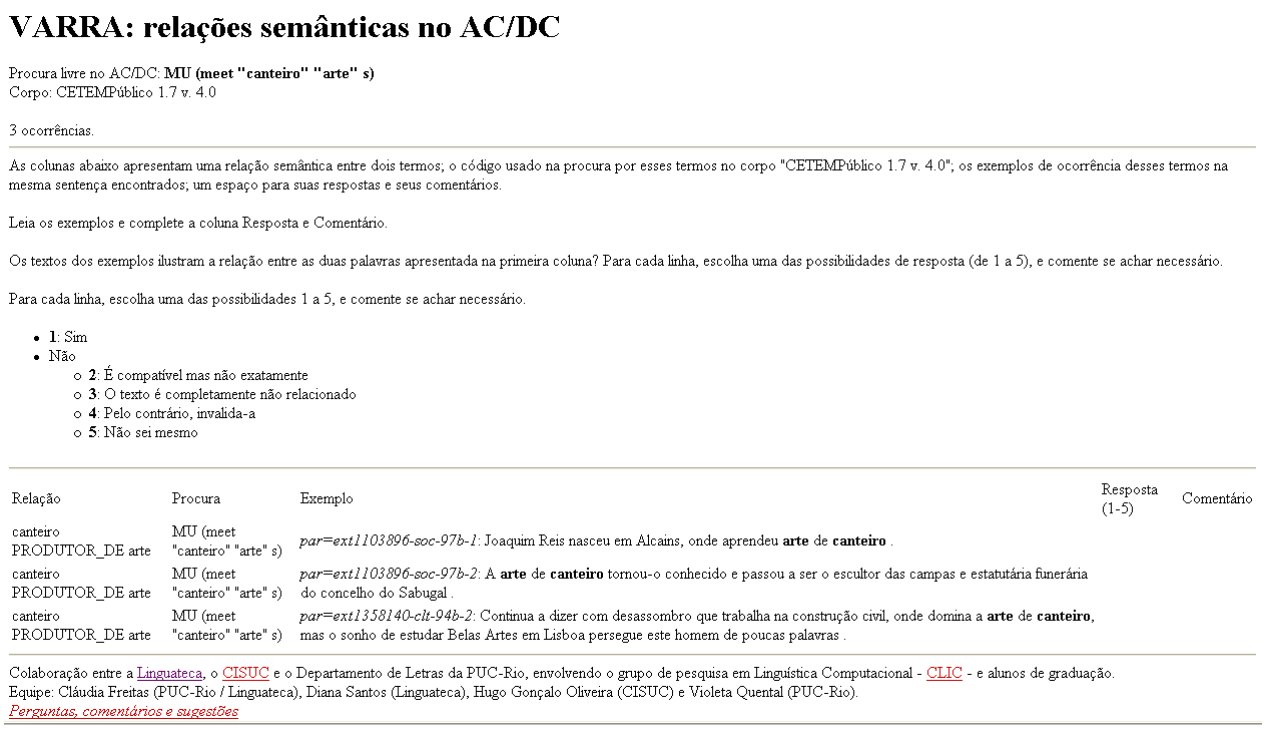
Esperamos que em breve possamos também referir trabalho de outros investigadores a usar e a melhorar este recurso, que é para ser propriedade comum de todos os investigadores e desenvolvedores na área do processamento da língua portuguesa. Para nossa satisfação podemos relatar que já foi levantado por vários grupos alguns dos quais nos deram retorno.

Salientamos novamente que o PAPEL não pretende ser um recurso final, mas sim um ponto de partida para futuros projectos, que o poderão enriquecer recorrendo a outras fontes de informação.

Por exemplo, e visto que o TeP foi criado à mão, um processo relativamente fácil de melhorar o PAPEL seria apenas juntar-lhe (ao PAPEL) as relações de sinonímia obtidas por transitividade (as 12%) que eram validadas no TeP, ou pelo menos a informação adicional de "concordância" com outros recursos.

Relação	Relações c/ args no CETEMPúblico	%	Amostra	%	Encontradas	%
Hiperonímia	40.079	63%	3.145	8%	560	18%
Meronímia	3.746	35%	2.343	63%	521	22%
Causa	557	50%	557	100%	20	4%
Produtor	414	44%	414	100%	12	3%
Finalidade	1.718	59%	1.718	100%	173	10%

Tabela 8: Resultados da validação das relações excepto sinonímia (republicados de Gonçalo Oliveira, Santos e Gomes (2009a)).



VARRA: relações semânticas no AC/DC

Procura livre no AC/DC: **MU (meet "canteiro" "arte" s)**
Corpo: CETEMPúblico 1.7 v. 4.0

3 ocorrências.

As colunas abaixo apresentam uma relação semântica entre dois termos; o código usado na procura por esses termos no corpo "CETEMPúblico 1.7 v. 4.0"; os exemplos de ocorrência desses termos na mesma sentença encontrados; um espaço para suas respostas e seus comentários.

Leia os exemplos e complete a coluna Resposta e Comentário.

Os textos dos exemplos ilustram a relação entre as duas palavras apresentada na primeira coluna? Para cada linha, escolha uma das possibilidades de resposta (de 1 a 5), e comente se achar necessário.

Para cada linha, escolha uma das possibilidades 1 a 5, e comente se achar necessário.

- **1**: Sim
- Não
  - **2**: É compatível mas não exatamente
  - **3**: O texto é completamente não relacionado
  - **4**: Pelo contrário, invalida-a
  - **5**: Não sei mesmo

Relação	Procura	Exemplo	Resposta (1-5)	Comentário
canteiro PRODUTOR_DE arte	MU (meet "canteiro" "arte" s)	*par=ext1103896-soc-97b-1*: Joaquim Reis nasceu em Alcains, onde aprendeu **arte** de **canteiro** .		
canteiro PRODUTOR_DE arte	MU (meet "canteiro" "arte" s)	*par=ext1103896-soc-97b-2*: A **arte** de **canteiro** tornou-o conhecido e passou a ser o escultor das campas e estatutária funerária do concelho do Sabugal .		
canteiro PRODUTOR_DE arte	MU (meet "canteiro" "arte" s)	*par=ext1358140-clt-94b-2*: Continua a dizer com desassombro que trabalha na construção civil, onde domina a **arte** de **canteiro**, mas o sonho de estudar Belas Artes em Lisboa persegue este homem de poucas palavras .		

Colaboração entre a Linguateca, o CISUC e o Departamento de Letras da PUC-Rio, envolvendo o grupo de pesquisa em Linguística Computacional - CLIC - e alunos de graduação.
Equipe: Cláudia Freitas (PUC-Rio / Linguateca), Diana Santos (Linguateca), Hugo Gonçalo Oliveira (CISUC) e Violeta Quental (PUC-Rio).
Perguntas, comentários e sugestões

Figura 4: Exemplo de resultados da invocação do VARRA

Pretendemos no futuro continuar a melhorar o conteúdo do PAPEL através da obtenção de novos dados na rede assim como aperfeiçoar a validação dos já presentes.

Uma melhoria óbvia que em breve implementaremos é a associação de um grau de certeza, assim como um conjunto de validação, a cada triplo, veja-se Wandmacher et al. (2007).

Outro caminho a explorar será adaptar ao português a metodologia proposta e testada em Rigau, Rodríguez e Agirre (1998) para o castelhano a partir da WordNet para o inglês e de um dicionário bilingue, conforme sugerido por Lluis Padró.

Finalmente, gostaríamos também de poder associar a novas versões do PAPEL um conjunto de ferramentas simples para o processar, conforme sugestão do Alberto Simões.

Agradecimentos

Agradecemos à Cláudia Freitas a colaboração preciosa no desenho do sistema VARRA, e a todos os co-autores do artigo de comparação de ontologias: Anabela Barreiro, Cláudia Freitas, José Carlos Medeiros, Luís Costa e Rosário Silva, as discussões férteis e o trabalho realizado.

Agradecemos também ao grupo de R&D da Porto Editora a colaboração na criação do PAPEL, ao CLIC e à Violeta Quental a colaboração com a PUC-Rio, assim como ao Nuno Seco a sua anterior participação no projecto.

O projecto PAPEL foi desenvolvido no âmbito da Linguateca, co-financiada pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN.

Hugo Gonçalo Oliveira é actualmente financiado pela FCT, bolsa SFRH/BD/44955/2008.

Agradecemos aos três parceristas da Linguamática, Lluis Padró, Gerardo Sierra e Alberto Simões, as suas recensões e comentários, que, se não foram todos levados em conta na presente versão por falta de tempo, muito contribuirão para uma melhoria significativa do projecto do PAPEL no futuro.

Referncias

Afonso, Susana. 2009. Uma FrameNet para o português, 29 de Junho - 3 de Julho, 2009. Apresentação na Escola de Verão Belinda Maia (Edv 2009), Porto, Portugal, `http://www.linguateca.pt/Repositorio/AfonsoFrameNetEdV2009.pdf`.

Alshawi, Hiyan. 1989. Analysing the dictionary definitions. Em Bran Boguraev e Ted Briscoe, editores, *Computational lexicography for natural language processing*, pp. 153–169, Nova Iorque, EUA. Longman Publishing Group.

Amsler, Robert A. 1981. A taxonomy for english nouns and verbs. Em *Proceedings of the 19th annual meeting on Association for Computational Linguistics*, pp. 133–138, Morristown, NJ, EUA. Association for Computational Linguistics.

Baker, Collin F., Charles J. Fillmore, e John B. Lowe. 1998. The Berkeley FrameNet Project. Em *Proceedings of the 17th International Conference on Computational linguistics*, pp. 86–90, Morristown, NJ, EUA. Association for Computational Linguistics.

Barreiro, Anabela. No prelo. Port4NooJ: an open source, ontology-driven Portuguese linguistic system with applications in machine translation. Em Max Silberztein e Tamas Varadi, editores, *Proceedings of the 2008 International NooJ Conference (NooJ'08)*, Cambridge, Reino Unido. Cambridge Scholars Publishing.

Barreiro, Anabela, Luzia Helena Wittmann, e Maria de Jesus Pereira. 1996. Lexical differences between European and Brazilian Portuguese. *INESC Journal of Research and Development*, 5(2):75–101.

Berland, Matthew e Eugene Charniak. 1999. Finding parts in very large corpora. Em *Proceedings of the 37th Annual Meeting of the ACL on Computational Linguistics*, pp. 57–64, Morristown, NJ, EUA. Association for Computational Linguistics.

Brank, Janez, Marko Grobelnik, e Dunja Mladenić. 2005. A survey of ontology evaluation techniques. Em *Proceedings of the 8th International Conference on Data Mining and Data Warehouses (SiKDD)*, pp. 166–169.

Brewster, Christopher, Harith Alani, Srinandan Dasmahapatra, e Yorick Wilks. 2004. Data-driven ontology evaluation. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, pp. 164–168, Lisboa, Portugal, 26-28 de Maio, 2004. European Language Resources Association.

Briscoe, Ted. 1991. Lexical issues in natural language processing. Em Ewan Klein e Frank Veltman, editores, *Natural Language and Speech: Symposium Proceedings*. Springer, Berlim e Heidelberg, Alemanha, pp. 39–68.

Calzolari, Nicoletta, Laura Pecchia, e Antonio Zampolli. 1973. Working on the Italian machine dictionary: a semantic approach. Em *Proceedings of the 5th conference on Computational linguistics*, pp. 49–52, Morristown, NJ, EUA. Association for Computational Linguistics.

Caraballo, Sharon A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. Em *Proceedings of the 37th annual meeting of the ACL on Computational Linguistics*, pp. 120–126, Morristown, NJ, EUA. Association for Computational Linguistics.

Chaves, Marcirio Silveira. 2009. *Uma Metodologia para Construção de Geo-Ontologias*. Tese de doutoramento, Faculdade de Ciências, Universidade de Lisboa, Setembro, 2009.

Chodorow, Martin S., Roy J. Byrd, e George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. Em *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pp. 299–304, Morristown, NJ, EUA. Association for Computational Linguistics.

Costa, Luís, Diana Santos, e Paulo Alexandre Rocha. 2009. Estudando o português tal como é usado: o serviço AC/DC. Em *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*, 8-11 de Setembro, 2009.

Costa, Rui P. e Nuno Seco. 2008. Hyponymy Extraction and Web Search Behavior Analysis Based on Query Reformulation. Em *Proceedings of the 11th Ibero-American Conference on Artificial Intelligence (IBERAMIA)*, LNAI, pp. 332–341. Springer.

Cuadros, Montse e German Rigau. 2006. Quality assessment of large scale knowledge resources. Em *Proceedings of the 2006 Conference on Empirical Methods in Natural Language*

Processing, pp. 534–541, Sydney, Austrália, Julho, 2006. Association for Computational Linguistics.

Dahlgren, Kathleen. 1995. A linguistic ontology. *International Journal Human-Computer Studies*, 43(5-6):809–818.

Demetriou, George e Eric Steven Atwell. 2001. A Domain-Independent Semantic Tagger for the Study of Meaning Associations in English Text. Em *Proceedings of the 4th International Workshop on Computational Semantics (IWCS-4)*, pp. 67–80, 10-12 de Janeiro, 2001.

Dias da Silva, Bento C., Mirna Oliveira, e Helio Moraes. 2002. Groundwork for the Development of the Brazilian Portuguese Wordnet. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing: Third International Conference, PorTAL 2002, Faro, Portugal, Junho 2002, Proceedings*, volume 2389 of *LNAI*, pp. 189–196. Springer.

Dias-Da-Silva, Bento Carlos e Helio Roberto de Moraes. 2003. A construção de um thesaurus eletrônico para o português do Brasil. *ALFA*, 47(2):101–115.

2005. *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto.

Dolan, William B. 1994. Word sense ambiguation: clustering related senses. Em *Proceedings of the 15th conference on Computational linguistics*, pp. 712–716, Morristown, NJ, EUA. Association for Computational Linguistics.

Dorow, Beate. 2006. *A Graph Model for Words and their Meanings*. Tese de doutoramento, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart.

Edmonds, Philip e Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Etzioni, Oren, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, e Alexander Yates. 2005. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Maio, 1998.

Fillmore, Charles J. 1982. Frame semantics. Em Linguistic Society of Korea, editor, *Linguistics in the morning calm*. Hanshin Publishing Co., Seoul, Coreia do Sul, pp. 111–137.

Freitas, Cláudia e Violeta Quental. 2007. Subsídios para a elaboração automática de taxonomias. Em *Actas do XXVII Congresso da SBC - V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 1585–1594.

Freitas, Cláudia, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira, e Paula Carvalho. 2009. Detection of relations between named entities: report of a shared task. Em *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, SEW-2009*, pp. 129–137, Boulder, Colorado, EUA, 4 de Junho, 2009.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota. 2008. Relações semânticas do ReRelEM: além das entidades no Segundo HAREM. Em Cristina Mota, Diana Santos, Cristina Mota, e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 77–96, 31 de Dezembro, 2008.

Girju, Roxana e Dan Moldovan. 2002. Text mining for causal relations. Em Susan M. Haller e Gene Simmons, editores, *Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 360–364.

Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes, e Nuno Seco. 2008. PAPEL: a dictionary-based lexical ontology for Portuguese. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume 5190 of *LNAI*, pp. 31–40. Springer.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2008a. Apresentação das relações extraídas do Dicionário da Porto Editora. Relatório técnico, CISUC, Dezembro, 2008. Relatório do PAPEL num. 4, `http://linguateca.dei.uc.pt/papel/GoncaloOliveiraetal2008relPAPEL4.pdf`.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2008b. Utilização do (analisador sintáctico) PEN para extracção de informação das definições de um dicionário. Relatório técnico, CISUC, Novembro, 2008. Relatório do PAPEL num.

3, `http://linguateca.dei.uc.pt/papel/GoncaloOliveiraetal2008relPAPEL3.pdf`.

Gonçalo Oliveira, Hugo, Diana Santos, e Paulo Gomes. 2009a. Avaliação da extracção de relações semânticas entre palavras portuguesas a partir de um dicionário. Em *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*, 8-11 de Setembro, 2009. Versão inicial do presente artigo.

Gonçalo Oliveira, Hugo, Diana Santos, e Paulo Gomes. 2009b. Relations extracted from a Portuguese dictionary: results and first evaluation. Em Luís Seabra Lopes, Nuno Lau, Pedro Mariano, e Luís M. Rocha, editores, *New Trends in Artificial Intelligence, Local Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA 2009)*, pp. 541–552, Aveiro, Portugal, 12-15 de Outubro, 2009.

Gruber, Thomas R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. Em *Proceedings of 14th conference on Computational linguistics*, pp. 539–545, Morristown, NJ, EUA. Association for Computational Linguistics.

Herbelot, Aurelie e Ann Copestake. 2006. Acquiring Ontological Relationships from Wikipedia Using RMRS. Em *Proceedings of the ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*, Athens, GA, EUA, 6 de Novembro, 2006.

Hirst, Graeme. 2004. Ontology and the lexicon. Em Steffen Staab e Rudi Studer, editores, *Handbook on Ontologies*. Springer, pp. 209–230.

Ide, Nancy e Jean Veronis. 1995. Knowledge extraction from machine-readable dictionaries: An evaluation. Em Petra Steffens, editor, *Proceedings of Machine Translation and the Lexicon, Third International EAMT Workshop, Heidelberg, Germany, 26-28 April, 1993*, pp. 19–34. Springer.

Kilgarriff, Adam. 1996. Word senses are not bona fide objects: implications for cognitive science, formal semantics, NLP. Em *Proceedings of the 5th International Conference on the Cognitive Science of Natural Language Processing*, pp. 193–200, Dublin, Irlanda.

Kilgarriff, Adam. 1997. "I don't believe in word senses". *Computing and the Humanities*, 31(2):91–113.

Lenat, Douglas B. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Liu, H. e P. Singh. 2004. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4):211–226.

Marcellino, Erasmo Roberto e Bento Dias da Silva. 2009. Sistematização linguístico-computacional do léxico do domínio conceitual Indústria do Bordado de Ibitinga. Em *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*, 8-11 de Setembro, 2009.

Marrafa, Palmira. 2002. Portuguese WordNet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo, e Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 390–392.

Medelyan, Olena, David Milne, Catherine Legg, e Ian H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, Setembro, 2009.

Montemagni, Simonetta e Lucy Vanderwende. 1992. Structural patterns vs. string patterns for extracting semantic information from dictionaries. Em *Proceedings of the 14th conference on Computational linguistics*, pp. 546–552, Morristown, NJ, EUA. Association for Computational Linguistics.

Navarro, Emmanuel, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Tzu Y. Kuo, Pierre Magistry, e Chu R. Huang. 2009. Wiktionary and NLP: Improving synonymy networks. Em Iryna Gurevych e Torsten Zesch, editores, *Proceedings of the Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 19–27, Suntec, Singapura. Association for Computational Linguistics.

Navigli, Roberto, Paola Velardi, Alessandro Cucchiarelli, e Francesca Neri. 2004. Quantitative and qualitative evaluation of the ontolearn ontology learning system. Em *Proceedings of*

the 20th International conference on Computational Linguistics, Morristown, NJ, EUA. Association for Computational Linguistics.

Nichols, Eric, Francis Bond, e Dan Flickinger. 2005. Robust ontology acquisition from machine-readable dictionaries. Em Leslie Pack Kaelbling e Alessandro Saffiotti, editores, *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1111–1116. Professional Book Center.

O'Hara, Thomas Paul. 2005. *Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions*. Tese de doutoramento, NMSU CS, Agosto, 2005.

Pianta, Emanuele, Lusia Bentivogli, e Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. Em *Proceedings of the 1st International WordNet Conference*, pp. 293–302, Mysore, Índia, 21-25 de Janeiro, 2002.

Raman, J. e Pushpak Bhattacharyya. 2008. Towards Automatic Evaluation of Wordnet Synsets. Em Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, e Piek Vossen, editores, *Proceedings of the 4th Global WordNet Conference (GWC 2008)*, Szeged, Hungria, 22-25 de Janeiro, 2008.

Richardson, Stephen D., William B. Dolan, e Lucy Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. Em *Proceedings of the 17th International Conference on Computational linguistics*, pp. 1098–1102, Morristown, NJ, EUA, 10-14 de Agosto, 1998. Association for Computational Linguistics.

Richardson, Stephen D., Lucy Vanderwende, e William Dolan. 1993. Combining dictionary-based and example-based methods for natural language analysis. Em *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 69–79, Kyoto, Japão.

Rigau, German, Horacio Rodríguez, e Eneko Agirre. 1998. Building Accurate Semantic Taxonomies from Monolingual MRDs. Em *Proceedings of COLING-ACL'98*, pp. 1103–1109.

Riloff, Ellen e Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. Em *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pp. 117–124.

Rocha, Paulo Alexandre e Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR)*, pp. 131–140, São Paulo. ICMC/USP.

Salomão, Maria M. M. 2009. Framenet Brasil: Um trabalho em progresso. *Calidoscópio*, 7(2).

Salton, G. e M. J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, Nova Iorque, EUA.

Sampson, Geoffrey. 2000. Review of (Fellbaum, 1998). *International Journal of Lexicography*, 13(1):54–59.

Santos, Diana. 2006. What is natural language? Differences compared to artificial languages, and consequences for natural language processing. Palestra convidada no SBLP2006 e no PROPOR'2006, Itatiaia, RJ, Brasil, 15 de Maio de 2006, `http://www.linguateca.pt/Diana/download/SantosPalestraSBLPPropor2006.pdf`.

Santos, Diana. 2007. Evaluation in natural language processing. Curso na European Summer School on Language, Logic and Information ESSLLI, Dublin, Irlanda, 6-17 de Agosto, `http://www.linguateca.pt/Diana/download/EvaluationESSLLI07.pdf`.

Santos, Diana. 2009. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. Apresentaçao no Workshop on research infrastructure for linguistic variation, Oslo, Noruega, 17-18 de Setembro, 2009, `http://www.hf.uio.no/tekstlab/rilivs/slides/SantosRILiVS2009workshop.pdf`.

Santos, Diana, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo Gonçalo Oliveira, José Carlos Medeiros, e Rosário Silva. 2009. O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o PAPEL. Apresentação no XXV Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, Portugal, 22-24 de Outubro, 2009, `http://www.linguateca.pt/Diana/download/aprSantosetalAPL2009.pdf`.

Santos, Diana, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes, e Rosário

Silva. 2010. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. Em *Textos seleccionados apresentados ao XXV Encontro Nacional da Associação Portuguesa de Linguística*. Enviado para apreciação.

Santos, Diana e Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis, e Gregory Stainhauer, editores, *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC)*, pp. 205–210, Atenas, Grécia, 31 de Maio - 2 de Junho, 2000.

Santos, Diana e Luís Sarmento. 2003. O projecto AC/DC: acesso a corpora/disponibilização de corpora. Em Amália Mendes e Tiago Freitas, editores, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)*, pp. 705–717, Lisboa. APL.

Saussure, Ferdinand de. 1916. *Cours de Linguistique Générale*. Payot, Paris, França. Edição empregue: 1972.

Scott, Bernard. 2003. The Logos Model: An Historical Perspective. *Machine Translation*, 18(1):1–72.

Silberztein, Max e Tamas Varadi, editores. No prelo. *Proceedings of 2008 International NooJ Conference (NooJ'08)*, Cambridge, Reino Unido. Cambridge Scholars Publishing.

Simões, Alberto M. e José João Almeida. 2002. Jspell.pm – um módulo de análise morfológica para uso em processamento de linguagem natural. Em *Actas do XVII Encontro da Associação Portuguesa de Linguística*, pp. 485–495, Lisboa, Portugal. APL.

Turney, Peter D. 2001. Mining the web for synonyms: PMI–IR versus LSA on TOEFL. Em Luc De Raedt e Peter Flach, editores, *Proceedings of the 12th European Conference on Machine Learning (ECML-2001)*, volume 2167, pp. 491–502. Springer.

Vanderwende, Lucy, Gary Kacmarcik, Hisami Suzuki, e Arul Menezes. 2005. MindNet: An Automatically-Created Lexical Resource. Em *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pp. 8–9. The Association for Computational Linguistics, 7 de Outubro, 2005.

Veale, Tony. 2007. Enriched lexical ontologies: Adding new knowledge and new scope to old linguistic resources. Curso na European Summer School on Language, Logic and Information ESSLLI, Dublin, Irlanda, 6-17 de Agosto, 2007, `http://afflatus.ucd.ie/papers/Essilli_EnrichedLexiOnto.pdf`.

Vossen, Piek. 1997. Eurowordnet: a multilingual database for information retrieval. Em *Proceedings of the DELOS workshop on Cross-Language Information Retrieval*, Zurique, Suíça, 5-7 de Março, 1997.

Vossen, Piek, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Wandmacher, Tonio, Ekaterina Ovchinnikova, Ulf Krumnack, e Henrik Dittmann. 2007. Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. Em Thomas Meyer e Abhaya C. Nayak, editores, *3rd Australasian Ontology Workshop (AOW 2007)*, volume 85 of *CRPIT*, pp. 61–69, Gold Coast, Austrália. ACS.

Zesch, Torsten, Christof Müller, e Iryna Gurevych. 2008. Using Wiktionary for computing semantic relatedness. Em A. Cohn, editor, *Proceedings of the 23rd national conference on Artificial intelligence, AAAI'08*, pp. 861–866, Chicago, Illinois, EUA, 13-17 de Julho, 2008. AAAI Press.

Estratégias de Seleção de Conteúdo com Base na CST (*Cross-document Structure Theory*) para Sumarização Automática Multidocumento

Maria Lucia del Rosario Castro Jorge, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400 - Centro
Caixa Postal: 668 - CEP: 13560-970 - São Carlos/SP, Brasil

{mluciacj,taspardo}@icmc.usp.br

Resumo

O presente trabalho apresenta a definição, formalização e avaliação de estratégias de seleção de conteúdo para sumarização automática multidocumento com base na teoria discursiva CST (*Cross-document Structure Theory*). A tarefa de seleção de conteúdo foi modelada por meio de operadores que representam possíveis preferências do usuário para a sumarização. Estes operadores são especificados em templates contendo regras e funções que relacionam essas preferências às relações CST. Em particular, definimos operadores para extrair a informação principal, apresentar informação de contexto, identificar autoria, tratar redundâncias e identificar informação contraditória. Nossos experimentos foram feitos usando um córpus jornalístico de textos escritos em português brasileiro e mostram que o uso da CST melhora a qualidade do conteúdo selecionado para os sumários, já que se exploram as relações entre os conteúdos dos diferentes textos.

1. Introdução

O uso e a disponibilidade cada vez maior de tecnologias de comunicação têm provocado um aumento considerável no volume de informação, principalmente on-line. Há muita informação redundante, complementar e contraditória, proveniente de diversas fontes. Conseqüentemente, o processamento dessa informação tem se tornado uma tarefa de difícil execução, tanto por humanos quanto por máquinas. Neste contexto, a sumarização multidocumento pode ser uma tarefa útil.

A sumarização automática multidocumento (SAM) consiste na produção automática de um único sumário (também chamado resumo) a partir de um grupo de textos sobre um mesmo tópico ou sobre tópicos relacionados (Mani, 2001). Imagine, por exemplo, que uma pessoa deseje se interar dos principais acontecimentos da recente crise econômica mundial. Em vez de ter que ler uma infinidade de textos sobre o assunto, o que seria inviável, um sistema de SAM poderia lhe fornecer um único sumário sintetizando os fatos relevantes. A Figura 1 mostra um exemplo de sumário multidocumento produzido manualmente a partir de três textos jornalísticos que reportavam diversos ataques criminosos organizados a várias regiões do estado de São Paulo, no Brasil.

> Uma nova série de ataques criminosos foi registrada na madrugada desta segunda-feira, dia 7, em São Paulo e municípios do interior paulista. Os bandidos atacaram agências bancárias, bases policiais e prédios públicos com bombas e tiros. As ações são atribuídas à facção criminosa Primeiro Comando da Capital (PCC), que já comandou outros ataques em duas ocasiões. Eles tinham prometido retomar os ataques no Estado de São Paulo no Dia dos Pais, no próximo domingo. A promessa aparentemente começou a ser cumprida na madrugada de hoje. Cidades do interior, como Jundiaí, foram alvo de ataques. Na região do ABC Paulista, pelo menos dez ônibus foram incendiados - sete em Mauá e três em Santo André. Na capital, houve ataques a outros quatro ônibus. Uma bomba caseira foi jogada contra o prédio do Ministério Público, na capital do estado. A Secretaria da Fazenda também foi atingida por uma bomba. Duas bases da Guarda Civil Metropolitana (GCM), sendo uma no Capão Redondo, Zona Sul de São Paulo, foram alvo dos criminosos.

Figura 1. Exemplo de sumário multidocumento

É interessante notar que um sumário multidocumento pode ser construído tendo-se

diferentes objetivos. Se o leitor deseja apenas uma visão geral do acontecimento, o sumário do exemplo é suficiente. Por outro lado, muitas vezes se quer informação contextual (no caso de um leitor que não sabe nada do assunto) ou se deseja visualizar a evolução de alguns fatos ocorridos em um determinado período de tempo (nesses casos, o histórico da facção PCC e como ela tem agido no estado de São Paulo são elementos importantes para o sumário). Ocasionalmente, pode-se querer confrontar diferentes versões de notícias para se detectar contradições entre elas (por exemplo, quais dos ataques são atribuídos pelas três fontes ao PCC). Portanto, sistemas de SAM devem ser capazes de produzir sumários que satisfaçam as preferências de sumarização do leitor/usuário.

A sumarização multidocumento, como grande parte dos sistemas de processamento multidocumento, tem que lidar, também, com diversos desafios provenientes da multiplicidade de informação. Por exemplo, dentre os fenômenos multidocumento, é necessário que se reconheça informação redundante, complementar e, como já mencionado, contraditória, que as correferências sejam resolvidas, que estilos variados de diferentes autores e fontes sejam uniformizados, e que a informação produzida para o usuário seja organizada/ordenada adequadamente, visando-se sempre a coerência e a coesão do texto produzido.

Mani e Maybury (1999), objetivando modelar a tarefa de sumarização e organizar seus diversos processos, sugerem que a sumarização envolva idealmente três tarefas: a análise dos textos-fonte, produzindo-se uma representação completa de seu conteúdo; a transformação desse conteúdo completo em um conteúdo condensado; e, finalmente, a síntese desse conteúdo condensado na forma de sumário, expresso em uma língua natural. Uma etapa completa de análise requer, por exemplo, o uso de léxicos, gramáticas e interpretadores de língua natural de níveis lingüísticos variados; a etapa de transformação deve realizar a seleção do conteúdo relevante, a agregação/fusão, a generalização e a substituição de informação, dentre outras operações; a etapa de síntese, por fim, deve ter capacidades de geração textual, envolvendo a escolha de expressões de referência, ordenação e organização da informação a ser apresentada, etc. Sistemas de SAM que adotam tal abordagem, privilegiando a manipulação e o uso de conhecimento lingüístico sofisticado, são ditos pertencerem à abordagem profunda, ou fundamental. Sistemas que fazem uso de pouco conhecimento lingüístico são ditos pertencerem à abordagem superficial. Apesar de serem mais custosos e exigirem mais recursos, sistemas da abordagem profunda são capazes de produzir sumários melhores.

Neste trabalho, foca-se na abordagem profunda, mais especificamente, em um dos processos mais importantes da etapa de transformação da SAM: a seleção de conteúdo. Assume-se que a etapa de análise é realizada previamente e corresponde unicamente à representação dos textos-fonte segundo a teoria/modelo lingüístico-computacional CST (*Cross-document Structure Theory*) (Radev, 2000), de natureza semântico-discursiva. Com base na CST, são exploradas estratégias de seleção de conteúdo que selecionam conteúdo relevante em função de preferências de sumarização do usuário. Por fim, a etapa de síntese realiza simplesmente a justaposição do conteúdo selecionado, produzindo o sumário final. A CST é, portanto, a base de desenvolvimento deste trabalho. Ela modela o relacionamento entre o conteúdo multidocumento, ou seja, estabelece relações semântico-dicursivas entre as partes dos textos sendo processados (por exemplo, relações de seqüência temporal, contradição, elaboração, etc.). A hipótese deste trabalho é que esse tipo de conhecimento é importante e, se manipulado adequadamente, pode produzir sumários multidocumento satisfatórios.

As estratégias de seleção de conteúdo propostas neste trabalho visam mapear as preferências de sumarização do usuário às relações previstas na CST, de forma que seja possível identificar nos textos-fonte o conteúdo relevante. Em particular, nossas estratégias são formalizadas e codificadas na forma de operadores de seleção de conteúdo, representados como templates contendo regras especificadas em termos de condições, restrições e operações primitivas de manipulação de informação. Neste trabalho, definimos operadores para extrair a informação principal dos textos-fonte, apresentar

informação de contexto, identificar autoria, tratar redundâncias e exibir informação contraditória dos textos-fonte. Nossos experimentos foram feitos usando um córpus jornalístico de textos escritos em português brasileiro e mostram que o uso da CST melhora a qualidade do conteúdo selecionado para os sumários, comprovando, desta forma, nossa hipótese.

Este trabalho dá continuidade a alguns trabalhos prévios na área para a língua portuguesa (Aleixo e Pardo, 2008a; Jorge e Pardo, 2009, 2010). Por se basear em um modelo semântico-discursivo, este trabalho alinha-se, portanto, à abordagem profunda da sumarização.

A seguir, na Seção 2, introduz-se a CST e apresentam-se os trabalhos relacionados. Na Seção 3, definimos e formalizamos nossos operadores de seleção de conteúdo. A avaliação e discussão dos resultados obtidos são apresentadas na Seção 4. Por fim, na Seção 5, fazem-se algumas considerações finais.

2. *Trabalhos Relacionados*

2.1 Cross-document Structure Theory

Inspirada na *Rhetorical Structure Theory* (RST) (Mann e Thompson, 1987) e nos trabalhos de Trigg (1983) e Trigg e Weiser (1987), a CST (Radev, 2000) é proposta como uma teoria para relacionar múltiplos documentos que versam sobre um mesmo assunto ou tópicos relacionados.

A CST foi originalmente proposta com um conjunto de 24 relações que representam os fenômenos multidocumento. As 24 relações são listadas na Tabela 1. Como exemplo de aplicação destas relações a um grupo de textos, a Figura 2 mostra alguns trechos de textos (de fontes diferentes) relacionados. Na figura, o primeiro par de sentenças está relacionado por meio da relação *Subsumption*, pois a segunda sentença contém toda a informação da primeira e outras informações adicionais. No segundo par de sentenças, as duas sentenças são iguais, portanto há uma relação *Identity*. Finalmente, o terceiro par de sentenças mostra uma contradição entre a distância ao aeroporto, o que caracteriza uma relação *Contradiction*.

Tabela 1. Conjunto original de relações propostas por Radev (2000)

Identity	*Judgment*
Equivalence (paraphrasing)	*Fulfilment*
Translation	*Description*
Subsumption	*Reader profile*
Contradiction	*Contrast*
Historical background	*Parallel*
Modality	*Cross-reference*
Attribution	*Citation*
Summary	*Refinement*
Follow-up	*Agreement*
Elaboration	*Generalization*
Indirect speech	*Change of perspective*

Relação: *Subsumption*

Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.

Relação: *Identity*

As vítimas do acidente foram 14 passageiros e três membros da tripulação.

As vítimas do acidente foram 14 passageiros e três membros da tripulação.

Relação: *Contradiction*

A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 10 quilômetros de distância da pista do aeroporto.

Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.

Figura 2. Exemplos de relações CST

A CST propõe um modelo geral em que as relações entre diferentes unidades de texto são representadas. Na Figura 3, ilustra-se este modelo, que assume a forma de um grafo. A figura foi reproduzida exatamente como aparece no trabalho de Radev (2000, p. 78) (em inglês, como no original). É importante notar que, em princípio, podem-se considerar diversas unidades textuais para análise, por exemplo, palavras, sintagmas, sentenças,

parágrafos ou, inclusive, todo o documento. As relações CST são estabelecidas em qualquer nível de análise. Nem todas as unidades textuais têm relações CST entre si, pois, em geral, existem partes dos textos que não estão diretamente relacionadas a um mesmo tópico. As relações estabelecidas também podem ter direcionalidade. Por exemplo, na Figura 2, as relações *Identity e Contradiction* não têm direcionalidade; por outro lado, a relação *Subsumption* tem direcionalidade, já que uma unidade textual está englobando outra.

Assim como sua antecessora RST, a CST está sujeita a ambigüidades na análise (Afantenos et al., 2004; Zhang et al., 2002), já que, como em toda análise subjetiva, pode haver mais de uma relação possível entre segmentos textuais. Com o objetivo de reduzir esta ambigüidade, Zhang et al. (2002) propuseram um refinamento das relações originais, em que são consideradas menos relações: 18. Para a língua portuguesa, o conjunto de relações foi ainda mais refinado (Aleixo e Pardo, 2008b), resultando em 14 relações. Esse refinamento foi feito pela eliminação de relações nunca verificadas experimentalmente e pela junção de relações com definições relacionadas. A Tabela 2 mostra as relações resultantes.

Tabela 2. Relações de Aleixo e Pardo (2008b)

Identity	*Attribution*
Equivalence	*Summary*
Translation	*Follow-up*
Subsumption	*Elaboration*
Contradiction	*Indirect speech*
Historical background	*Contradiction*
Modality	*Citation*

2.2 *Sumarização Multidocumento e CST*

Algumas pesquisas têm utilizado CST para fins de SAM, incluindo o trabalho do próprio Radev (2000), que, além de propor o modelo, também propôs uma metodologia de sumarização com CST de 4 etapas, as quais são ilustradas na Figura 4.

Na primeira etapa, os documentos são agrupados de acordo com a similaridade do conteúdo entre eles; na segunda etapa, os documentos são estruturados internamente, possivelmente envolvendo estruturas lexicais, sintáticas e semânticas; na terceira etapa, as relações CST são estabelecidas entre as partes dos textos e as unidades textuais relacionadas são organizadas em um grafo (que, deste ponto em diante, será referenciado por grafo CST) em que cada nó representa uma unidade informativa textual e as arestas representam as relações entre eles; finalmente, na quarta etapa, o conteúdo é selecionado de acordo com a informação dada pelas relações, para compor o sumário final. Para esta última etapa, Radev propõe a criação de operadores de preferência que representem possíveis preferências de sumarização para a seleção de conteúdo. Estas preferências estão associadas a certas relações estabelecidas pela CST. Por exemplo, um operador de contradição deveria selecionar informação relevante, além de apresentar principalmente as informações contraditórias entre os textos que estão sendo processados. Neste caso, sentenças relacionadas por meio da relação *Contradiction* terão uma preferência maior ao se selecionar o conteúdo para o sumário final. A proposta de Radev está baseada no trabalho prévio de Radev e McKeown (1998).

Outro trabalho importante baseado na CST foi o de Zhang et al. (2002). Considerando que após a seleção de conteúdo há um ranque de sentenças para compor o sumário (em função da relevância destas de acordo com uma métrica de importância qualquer), os autores propõem a alteração do ranque por meio do uso das relações CST. Sentenças que apresentam relações CST são preferidas em relação às sentenças que não apresentam tais relações e, portanto, obtêm melhores posições no ranque.

Otterbacher et al. (2002) investigam como o uso de relações CST ajuda a melhorar a coesão em sumários multidocumento. Eles propõem a seleção de sentenças de acordo com o conteúdo relevante e assumem que as sentenças relacionadas por meio de relações CST deveriam aparecer próximas no sumário final, podendo ser reorganizadas em função das restrições temporais impostas pelas próprias relações CST.

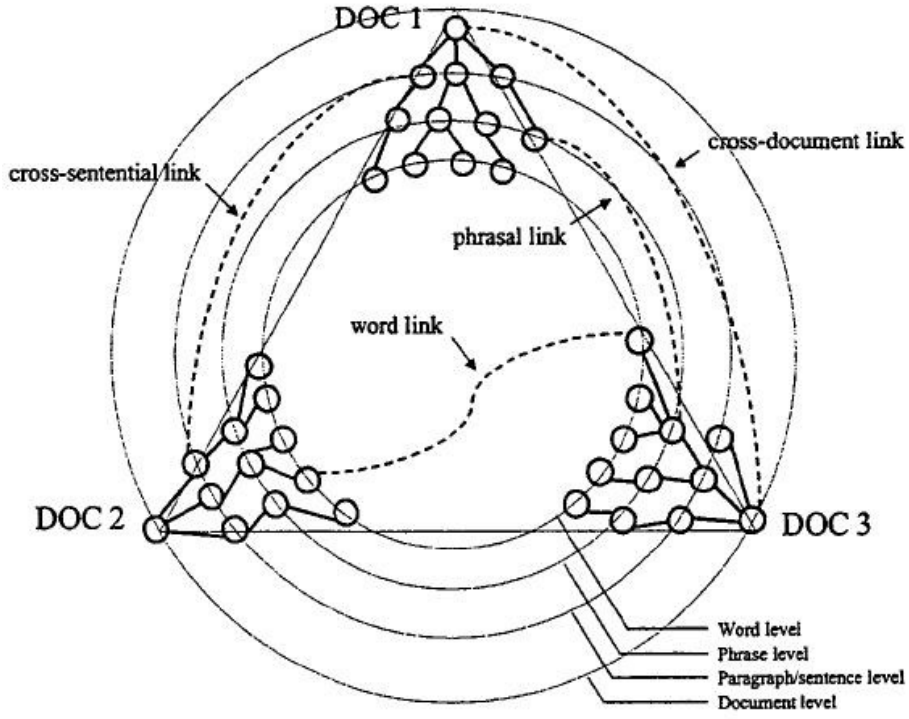


Figura 3. Modelo geral de representação via CST (Radev, 2000, p. 78)

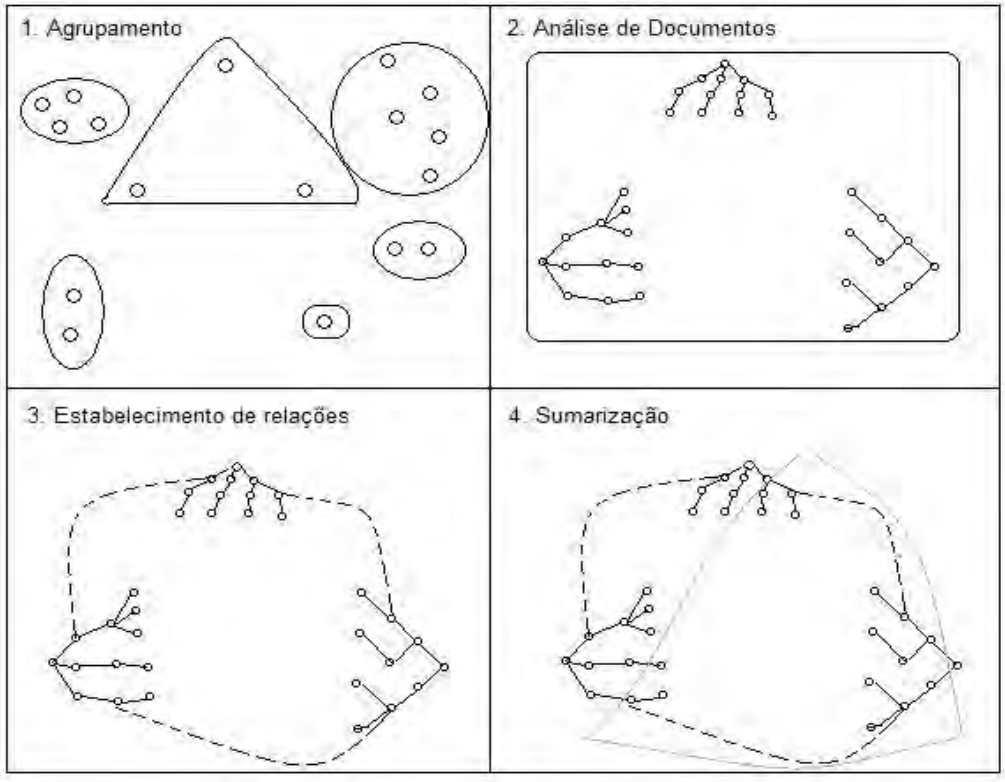


Figura 4. Etapas do processo de sumarização CST (Radev, 2000, p. 81)

Em uma linha um pouco diferente, Afantenos et al. (2004), com base na CST, propuseram uma nova classificação de relações entre textos. Os autores dividem as relações em duas categorias: sincrônicas e diacrônicas. As relações sincrônicas exploram o desenvolvimento de um evento descrito em varias fontes de informação, enquanto as relações diacrônicas exploram o

evento ao longo do tempo em uma mesma fonte de informação. De acordo com esta nova classificação, os autores propõem uma metodologia de sumarização que extrai mensagens dos textos (utilizando ferramentas de extração de informação) e as coloca em formato de *templates*, sendo que as mensagens são relacionadas pelas relações propostas. Com base nesses *templates* relacionados, os autores afirmam que é possível se produzir bons sumários. Os autores apenas apresentam essas idéias iniciais e mostram alguns exemplos para textos do domínio do esporte, mas não formalizam ou avaliam sua proposta.

A seguir, delineamos nossa proposta de seleção de conteúdo com base na CST.

3. Definição e Formalização de Operadores de Seleção de Conteúdo

Como discutido anteriormente, o objetivo deste trabalho é explorar estratégias de seleção de conteúdo para SAM, relacionando possíveis preferências de sumarização do usuário às relações da CST, modelo utilizado para representar os textos-fonte. Seguindo a proposta de Radev (2000), após definir cada estratégia de seleção de conteúdo, elas são representadas na forma de operadores.

Formalmente, definimos um operador de seleção de conteúdo como um artefato computacional que processa uma representação de conteúdo previamente fornecida e produz uma versão mais condensada contendo as informações mais relevantes segundo os critérios especificados. Em particular, neste trabalho, a representação de conteúdo consiste no conjunto de textos representados segundo a teoria CST. Portanto, os operadores são aplicados após os textos-fonte terem sido analisados segundo essa teoria (na etapa de análise). Atualmente, tal análise deve ser feita manualmente para a língua portuguesa, já que o primeiro analisador automático ainda está em desenvolvimento. Para a língua inglesa, já há um analisador disponível (Zhang et al., 2003), o qual poderia automatizar o processo para essa língua, apesar de ainda não ter grande precisão.

De fato, o dado de entrada para nossos operadores não é o grafo CST produzido na etapa de análise, mas um ranque inicial das unidades informativas contidas nele. Esse ranque inicial deve conter as unidades informativas do texto na ordem de preferência em que devem ser inseridas no sumário final. Quanto mais relevante for a unidade informativa, mais acima no ranque ela deve estar. A função de um operador é, a partir do ranque inicial, produzir um ranque refinado, de tal forma que as unidades informativas mais relevantes segundo o critério especificado pelo usuário melhorem de posição no ranque e, portanto, ganhem preferência para estar no sumário. Por fim, dada uma taxa de compressão (ou seja, o tamanho do sumário desejado em relação ao tamanho dos textos-fonte, em número de palavras), são selecionadas tantas sentenças do ranque quanto possível (a partir das sentenças mais bem posicionadas) para que a taxa seja respeitada.

O ranque inicial é construído considerando todas as unidades informativas contidas no grafo CST. A relevância das unidades informativas depende do número de relações CST que elas apresentam, pois se assume que as informações mais importantes são aquelas que se repetem e são elaboradas ao longo dos textos, apresentando, portanto, mais relações. Tal suposição é padrão na área de SAM (Mani, 2001) e, de fato, pode ser facilmente verificada.

Na Figura 5, mostra-se um exemplo hipotético de um grafo CST e o ranque inicial formado a partir deste. As relações CST extraídas do grafo também são incluídas no ranque, não sendo necessário que se consulte o grafo constantemente, portanto. Como se pode notar, a unidade informativa mais importante é a 4, pois apresenta 3 relações CST, seguida pelas unidades 2 e 1 (que apresentam a mesma quantidade de relações), que, por sua vez, são seguidas pela unidade 5 (com apenas 1 relação), terminando-se na unidade 3 (sem relação alguma). Note que a direcionalidade das relações (indicada pela direção das setas) não tem influência alguma no processo de construção do ranque inicial.

No momento, quando algumas unidades apresentam o mesmo numero de relações, elas são ranqueadas na ordem em que são lidas do grafo.

Neste trabalho, consideramos as sentenças como unidades informativas, pois em geral são bem formadas e autocontidas.

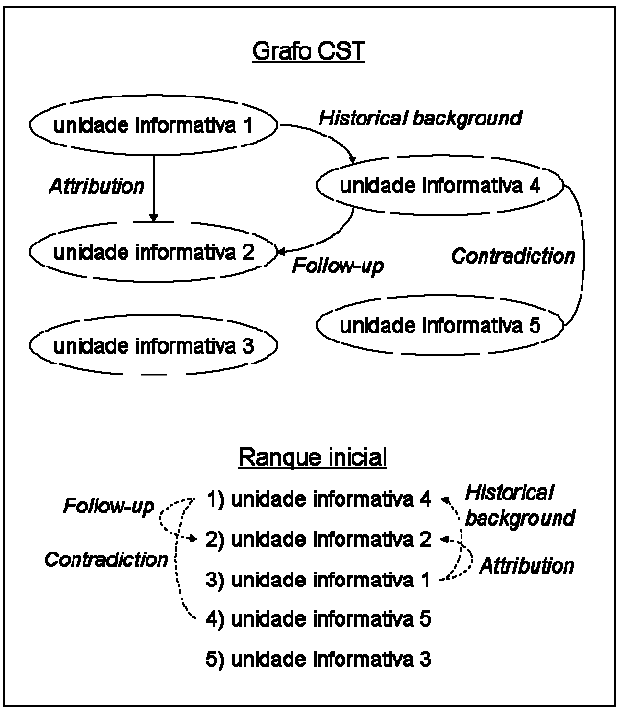


Figura 5. Exemplo de ranque inicial a partir de um grafo CST

Os operadores de seleção de conteúdo com base na CST estão definidos em formato de *templates*, contendo um conjunto de regras. As regras são especificadas por meio de condições e restrições, as quais, caso sejam satisfeitas, dispararão funções primitivas de manipulação da informação no ranque. Cada regra é definida da seguinte forma:

CONDIÇÕES, RESTRIÇÕES ⇒ AÇÕES

Cada condição tem o formato seguinte:

CONDIÇÃO(S_i, S_j, Direcionalidade, Relação)

Uma dada condição é satisfeita se existem a relação e a direcionalidade (de S_i até S_j: →; o caso oposto: ←; ou nenhuma direcionalidade: —) especificadas entre duas sentenças S_i e S_j, sendo que S_i aparece antes de S_j no texto. As restrições são opcionais, pois representam possíveis requisitos extras para que o operador seja aplicado. Atualmente, só usamos a restrição sobre o tamanho das sentenças, como será mostrado mais adiante.

Se todas as condições e restrições forem satisfeitas, então as ações serão aplicadas ao ranque inicial, produzindo assim uma versão refinada do ranque. As ações são definidas em termos de pelo menos uma das três funções primitivas definidas a seguir:

- SOBE(S_i,S_j): a sentença j é colocada em uma posição imediatamente após a sentença i no ranque; é importante notar que a sentença i sempre estará em uma posição superior a sentença j no ranque;
- TROCA(S_i,S_j): trocam-se as posições das sentenças i e j no ranque;
- ELIMINA(S_j): elimina-se a sentença j do ranque.

Para o presente trabalho, definimos e formalizamos 5 operadores que representam possíveis estratégias de seleção de conteúdo. São elas: apresentação de informação de contexto, exibição de informação contraditória, identificação de autoria, tratamento de redundância, e apresentação de eventos que evoluem com o tempo. O processo de construir o ranque inicial também pode ser representado como um operador, no qual a preferência é pela informação principal. Chamamos este último operador de "operador genérico" ou "operador de informação principal".

Cada operador é definido por três campos: um nome de referência, uma breve descrição e um conjunto de regras. Na Figura 6, mostra-se o operador para apresentação de informação contextual.

Nome
Apresentação de informação contextual

Descrição
Preferência por informações históricas e complementares

Regras
CONDIÇÃO(S_i, S_j, ←, *Elaboration*)
⇒ SOBE(S_i, S_j)
CONDIÇÃO(S_i, S_j, ←, *Historical background*)
⇒ SOBE(S_i, S_j)

Figura 6. Operador de apresentação de informação de contexto

Nesse operador, procuram-se por pares de sentenças (ao longo do ranque) que apresentem relações CST do tipo *Historical background* e *Elaboration*, já que essas relações são as que fornecem informação contextual. Caso essas informações sejam encontradas, elas sobem no ranque, obtendo, assim, maior preferência para estarem no sumário.

A aplicação deste operador ao ranque inicial da Figura 5 irá produzir o ranque refinado da Figura 7, na qual também se exibe o ranque inicial (para facilitar a comparação). É possível notar que a informação histórica da unidade informativa 1 sobe de posição no ranque, sendo posicionada imediatamente depois da sentença a qual se refere.

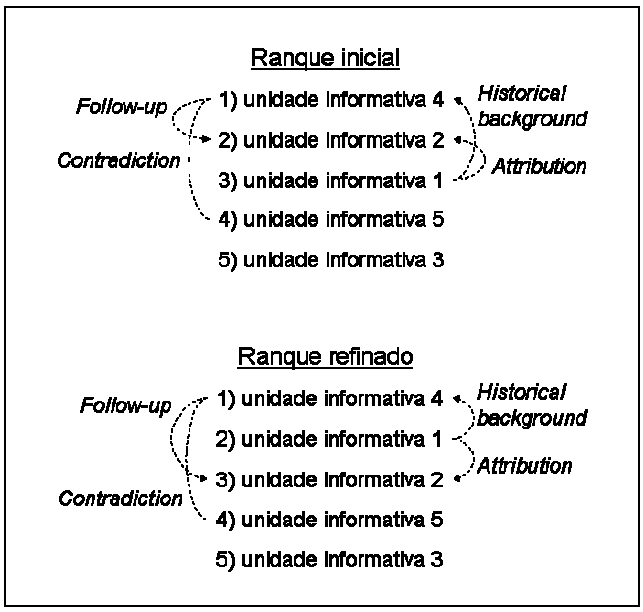


Figura 7. Ranque refinado

Na Figura 8 a seguir, é ilustrado um exemplo de sumário multidocumento usando o operador de apresentação de informação contextual. Como podemos ver na figura, a segunda e a terceira sentença (grifadas) contêm informação contextual e histórica, respectivamente, em relação a primeira sentença.

Pelo menos 80 pessoas morreram e mais de 165 ficaram feridas nesta segunda-feira após a colisão de dois trens de passageiros no delta do Nilo, ao norte do Cairo, informaram fontes policiais e médicas. O acidente ocorreu no delta do Nilo, ao norte de Cairo, no Egito. A maior tragédia ferroviária da história do Egito ocorreu em fevereiro de 2002, após o incêndio de um trem que cobria o trajeto entre Cairo e Luxor (sul), lotado de passageiros, e que deixou 376 mortos, segundo números oficiais.

Figura 8. Exemplo de sumário produzido pelo operador de apresentação de informação contextual

De fato, pode-se notar que a segunda sentença é redundante em relação a primeira, já que nenhum tratamento de redundância está sendo feito. Para resolver esse problema, faz-se necessário aplicar o operador de tratamento de redundância, detalhado posteriormente neste artigo.

O próximo operador prioriza a evolução de um evento no tempo. Esta evolução é modelada na CST por meio das relações *Historical background* e *Follow-up*. A Figura 9 mostra o operador correspondente. A forma de interpretação deste operador é a mesma do operador anterior. É interessante notar que, como a direcionalidade não importa neste caso, repetem-se regras para todas as possíveis direcionalidades.

Nome
Apresentação de eventos que evoluem no tempo

Descrição
Preferência por informações sobre eventos que evoluem no tempo

Regras
CONDIÇÃO(S_i, S_j, ←, *Historical background*)
$\Rightarrow$ SOBE(S_i, S_j)
CONDIÇÃO(S_i, S_j, →, *Historical background*)
$\Rightarrow$ SOBE(S_i, S_j)
CONDIÇÃO(S_i, S_j, ←, *Follow-up*)
$\Rightarrow$ SOBE(S_i, S_j)
CONDIÇÃO(S_i, S_j, →, *Follow-up*)
$\Rightarrow$ SOBE(S_i, S_j)

Figura 9. Operador de apresentação de eventos que evoluem no tempo

A Figura 10 mostra um sumário produzido pelo uso desse operador. Pode-se notar que a segunda sentença (grifada) contém informação sobre um fato anterior ao fato narrado na primeira sentença, foco dos textos-fonte.

A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de ouro da natação brasileira nos Jogos Pan-Americanos do Rio**.** Pouco antes Thiago Pereira já havia conquistado a segunda medalha de ouro brasileira no dia na final dos 400m medley, superando o norte-americano Robert Margalis e o canadense Keith Beavers.

Figura 10. Exemplo de sumário produzido pelo operador de apresentação de eventos que evoluem no tempo

A Figura 11 mostra o operador para exibir informações contraditórias, as quais são expressas por meio da relação *Contradiction*, enquanto a Figura 12 mostra o operador para

identificação de fonte/autoria, expressadas pelas relações *Attribution* e *Citation*. Pode-se perceber que as regras deste último operador contêm mais de uma condição, sendo que todas elas devem ser satisfeitas para que o operador seja aplicado. Este caso em particular se deve ao fato de que as relações *Attribution* e *Citation* sempre envolvem a presença de alguma outra relação, neste caso, a relação de conteúdo *Subsumption*.

Nome
Exibição de informações contraditórias

Descrição
Preferência por informações contraditórias

Regra
CONDIÇÃO(S_i, S_j, —, *Contradiction*)
$\Rightarrow$ SOBE(S_i, S_i)

Figura 11. Operador de exibição de informações contraditórias

Nome
Identificação de fonte/autoria

Descrição
Preferência por informações atribuídas a uma fonte

Regras
CONDIÇÃO(S_i, S_j, ←, *Attribution*),
CONDIÇÃO(S_i, S_j, ←, *Subsumption*)
$\Rightarrow$ TROCA(S_i,S_j), ELIMINA(S_i)
CONDIÇÃO(S_i, S_j, ←, *Citation*),
CONDIÇÃO(S_i, S_j, ←, *Subsumption*)
$\Rightarrow$ TROCA(S_i,S_j), ELIMINA(S_i)

Figura 12. Operador de identificação de fonte/autoria

As Figuras 13 e 14 mostram sumários produzidos por esses operadores, com a informação privilegiada grifada. Pode-se notar no sumário da Figura 13 que as duas últimas sentenças apresentam informações contraditórias entre si e também em relação a primeira sentença. A contradição, neste caso, tem origem da narração da notícia em momentos diferentes, quando números mais precisos vão surgindo conforme a passagem do tempo. No sumário da Figura 14, a segunda sentença apresenta o nome do diretor de uma organização, atribuindo a ele algumas informações ditas.

Finalmente, o operador de tratamento de redundância é mostrado na Figura 15. Em particular, neste operador, também são definidas algumas restrições em relação ao comprimento das unidades informativas (representado pelas barras verticais | |). Como a relação *Equivalence* indica que duas sentenças têm o mesmo conteúdo, elimina-se a sentença maior, mantendo-se a menor no sumário.

Cairo - O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo. No entanto, o ministro da Saúde, Hatem El-Gabaly, insistiu que até o momento foram recuperados apenas 36 cadáveres e que 133 feridos foram encaminhados a hospitais da região. Pelo menos 80 pessoas morreram e mais de 165 ficaram feridas nesta segunda-feira após a colisão de dois trens de passageiros no delta do Nilo, ao norte do Cairo, informaram fontes policiais e médicas.

Figura 13. Exemplo de sumário automático produzido pelo operador de apresentação de informações contraditórias

Quinze voluntários da ONG francesa Ação Contra a Fome (ACF) foram assassinados no nordeste do Sri Lanka, informou hoje um porta-voz da organização. O diretor da ACF no Sri Lanka, Benoit Miribel, confirmou a morte de seus funcionários e afirmou, comovido, que a ONG "não sofreu uma perda similar em seus mais de 25 anos de existência".

Figura 14. Exemplo de sumário automático produzido pelo operador de identificação de fonte/autoria

Nome
Tratamento de redundâncias

Descrição
Preferência por informações não redundantes

Regras
CONDIÇÃO(S_i, S_j, —, *Identity*)
$\Rightarrow$ ELIMINA(S_j)
CONDIÇÃO(S_i, S_j, —, *Equivalence*), $|S_i| \leq |S_j|$
$\Rightarrow$ ELIMINA(S_j)
CONDIÇÃO(S_i, S_j, —, *Equivalence*), $|S_i| > |S_j|$
$\Rightarrow$ TROCA(S_i,S_j), ELIMINA(S_i)
CONDIÇÃO(S_i, S_j, ←, *Subsumption*)
$\Rightarrow$ TROCA(S_i,S_j), ELIMINA(S_i)
CONDIÇÃO(S_i, S_j, →, *Subsumption*)
$\Rightarrow$ ELIMINA(S_j)

Figura 15. Operador de tratamento de redundâncias

É desejável que o operador de tratamento de redundâncias seja aplicado antes de qualquer outro operador (excetuando-se o operador genérico, logicamente, já que ele constrói o

ranque inicial), pois ele evita que conteúdo redundante seja incluído nos sumários. Ele evitaria, por exemplo, a sentença redundante (a segunda sentença) do sumário da Figura 8.

Na Figura 16, mostra-se o algoritmo geral para o procedimento de aplicação de operadores de seleção de conteúdo.

O procedimento tem como entrada o grafo CST e, como saída, o ranque refinado. Inicialmente, a partir do grafo CST, é construído o ranque inicial. Em seguida, lê-se a preferência de sumarização do usuário e, então, seleciona-se o operador correspondente, o qual é aplicado para todo par possível de sentenças no ranque, produzindo o ranque refinado.

Após esse processo, devem-se selecionar as sentenças mais bem ranqueadas que irão compor o sumário final, respeitando-se a taxa de compressão especificada pelo usuário. A etapa de síntese realiza a justaposição das sentenças selecionadas (não impondo nenhuma ordem em específico entre elas), exibindo o sumário final para o usuário.

De acordo com a forma que o método de seleção de conteúdo foi projetado, a partir do ranque inicial e da aplicação opcional do operador de tratamento de redundância, só se permite a aplicação de um dos demais operadores de seleção de conteúdo, a saber, de apresentação de informação de contexto, exibição de informação contraditória, identificação de autoria, e de apresentação de eventos que evoluem com o tempo. Ao permitir a aplicação de mais de um destes operadores, o ranqueamento feito pelo operador anterior pode ser alterado pelo novo operador. De fato, o último operador a ser aplicado vai fazer sua ordenação no ranque prevalecer.

Uma possibilidade para tornar possível ler mais de uma preferência de sumarização do usuário é ordenar as preferências em função de suas prioridades (que podem ser definidas pelo próprio usuário). Conseqüentemente, a aplicação dos operadores selecionados seria na ordem inversa, ou seja, deixando-se para o fim a aplicação dos operadores cujas preferências correspondentes têm maior prioridade, pois seriam essas que iriam prevalecer.

Outra possibilidade para lidar com várias preferências seria compor operadores mistos, considerando conjuntos maiores de relações em cada operador. Logicamente, ainda se teria que priorizar alguma informação, de forma que o operador possa produzir um ranque de informações que supra as expectativas do usuário. Para tal encaminhamento, acredita-se que estudos de caso com usuários sejam desejáveis, o que embasaria e tornaria possível o projeto de operadores mistos.

Nesse ponto, é interessante que se diga que a seleção de relações para a composição dos operadores atuais foi baseada nas bases teóricas da CST e na semântica de cada relação. Teoricamente, é possível compor novos operadores com relações diferentes, que poderiam, inclusive, incorporar outras preferências dos usuários que não são utilizadas nesse trabalho. Nesse trabalho, lidamos apenas com as preferências mais diretas e facilmente mapeadas para as relações da CST. Há relações não utilizadas nos operadores e que poderiam eventualmente produzir novos operadores ou serem incorporadas em alguns dos existentes.

Procedimento para a aplicação de operadores de seleção de conteúdo
Entrada: Grafo CST
Saída: Ranque refinado
Construir o ranque inicial a partir do grafo CST (usando o operador genérico/de informação principal)
Ler preferência de sumarização do usuário (se houver alguma)
Selecionar operador de seleção de conteúdo de acordo com a preferência de sumarização do usuário
Para cada regra do operador selecionado
  Para i=unidade informativa na primeira posição no ranque **até** a última posição do ranque
    Para j=unidade informativa na posição i+1 no ranque **até** a última posição no ranque
      Se as condições e restrições da regra são satisfeitas **então** aplicar as ações correspondentes nas sentenças i e j

Figura 16. Algoritmo de aplicação dos operadores de seleção de conteúdo

É interessante notar que nenhum dos operadores atuais lida com a relação *Overlap*. Esta relação indica que duas unidades informativas possuem informação em comum, além de informações particulares a cada uma. Veja, por exemplo, as duas sentenças abaixo de textos diferentes:

> Brasil e Finlândia se enfrentarão novamente neste sábado, às 12h30 (horário de Brasília), com transmissão ao vivo do canal de TV a cabo SporTV.
>
> Os dois times voltam a se enfrentar às 12h30 deste sábado, no mesmo ginásio, que normalmente é utilizado para competições de hóquei no gelo.

Há uma relação de *Overlap* entre elas, pois têm informação em comum (grifada), mas também têm informações extras: a primeira sentença informa por onde será feita a transmissão, enquanto a segunda dá mais detalhes do ginásio onde ocorrerá o evento. Há elementos redundantes que devem ser tratados no processo de seleção de conteúdo. Para tratar a redundância nesse caso, não se pode excluir uma das sentenças, como fizemos com as relações *Identity*, *Equivalence* e *Subsumption*, pois se estaria excluindo informações novas e que poderiam ser importantes. O que se precisa, de fato, é fundir as sentenças que apresentam relações *Overlap*, produzindo-se uma única sentença como a abaixo (dentre várias possibilidades):

> Com transmissão ao vivo do canal de TV a cabo SporTV, Brasil e Finlândia voltam a se enfrentar às 12h30 deste sábado (horário de Brasília), no mesmo ginásio, que normalmente é utilizado para competições de hóquei no gelo.

Para a língua portuguesa, poderia ser utilizado o sistema de fusão de Seno e Nunes (2009). Tal opção ainda não foi incorporada no estágio atual do método de seleção de conteúdo, pois implicaria em outros fatores a serem considerados, por exemplo, a gramaticalidade e o foco das sentenças fundidas, e a questão de se deixar de se produzir extratos (sumários formados pela justaposição de segmentos inalterados dos textos-fonte, os quais temos explorado aqui) para se produzir *abstracts* (em que há operações de reescrita textual).

A seguir apresentamos a avaliação das estratégias de seleção de conteúdo propostas.

4. Experimentos e Resultados

Para avaliar nossos operadores de seleção de conteúdo, construímos um protótipo de um sumarizador multidocumento, ao qual chamamos CSTSumm (*CST SUMMarizer*). Esse protótipo aplica o algoritmo da Figura 16 e realiza a síntese do sumário como explicado anteriormente. Os operadores propostos são armazenados de forma simples em um arquivo XML que pode ser facilmente manipulado, podendo-se adicionar, remover ou alterar operadores de maneira trivial. O conteúdo desse arquivo XML é carregado pelo protótipo no início de sua execução.

Para nossos experimentos, usamos um córpus composto de 50 coleções de textos jornalísticos escritos em Português Brasileiro (Aleixo e Pardo, 2008b), sendo que cada coleção tem 2 ou 3 textos sobre o mesmo tópico, e cada texto tem em média 20 sentenças. Esse córpus, chamado CSTNews, também contém a análise CST de cada coleção de textos e o sumário humano correspondente (genérico, com as informações mais importantes dos textos, sem preferências particulares), cujo tamanho corresponde a 30% do tamanho do maior texto da coleção (em número de palavras). Os textos do córpus foram coletados de vários jornais on-line brasileiros, como Folha de São Paulo, Estadão e Jornal do Brasil. O córpus foi analisado segundo a CST por 4 lingüistas computacionais previamente treinados nesse tipo de anotação, obtendo resultados de concordância satisfatórios.

A Figura 17 mostra a freqüência de ocorrência das relações CST no córpus. Pode-se notar que algumas relações ocorrem pouco (por exemplo, *Modality*, *Translation* e *Summary*), uma nunca ocorre (*Citation*) e outras ocorrem muito (por exemplo, *Elaboration* e *Overlap*).

Os sumários automáticos foram gerados para todos os operadores propostos neste trabalho, considerando a mesma taxa de compressão dos sumários humanos. Com exceção do operador genérico, o operador de tratamento de redundâncias foi aplicado antes dos demais operadores serem aplicados.

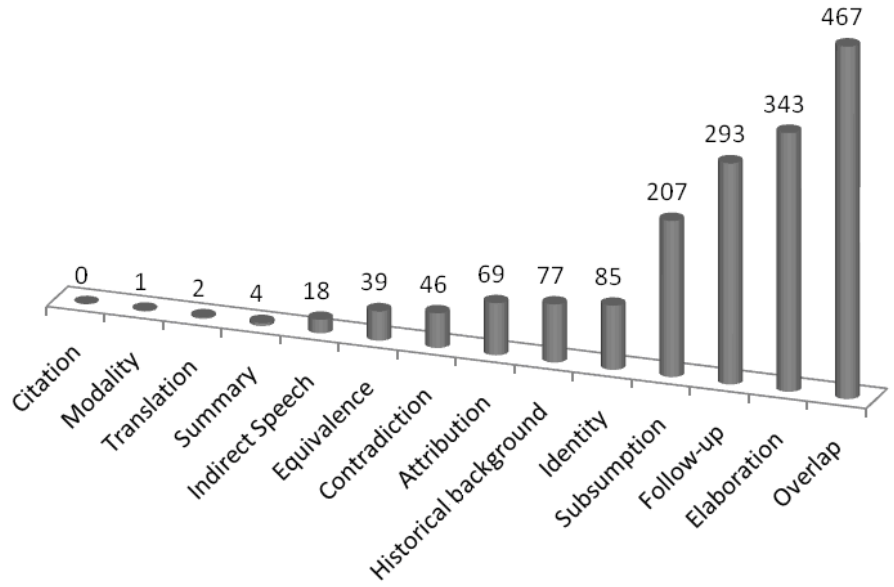


Figura 17. Relações CST no córpus

Neste trabalho consideramos dois métodos de avaliação: o automático, que é usado para medir a informatividade dos sumários, e o humano, que é usado para avaliar a coerência do sumário.

Para a avaliação automática, foi usada a medida ROUGE (Lin e Hovy, 2003), que é uma medida automática que computa o quão similar um sumário automático é em relação ao sumário humano correspondente. Basicamente, a similaridade é computada em função do número de n-gramas em comum entre os sumários, produzindo-se valores de precisão, cobertura e medida-f, tradicionais na área de pesquisa em questão. A precisão indica o quanto do sumário automático é, de fato, relevante; a cobertura indica o quanto do sumário humano é reproduzido no automático; a medida-f é uma medida única de desempenho, combinando precisão e cobertura. Apesar da comparação de n-gramas parecer simples demais para ser confiável, os autores da medida demonstraram que ela é tão boa quanto humanos em ranquear sumários em função de sua informatividade. De fato, tal medida foi amplamente aceita na comunidade de pesquisa e é usada até mesmo nas avaliações em larga escala organizadas anualmente (veja, por exemplo, as TACs – *Text Analysis Conferences* – principais competições mundiais na área de sumarização). Neste trabalho, utilizamos a ROUGE-1, ou seja, fazemos somente a comparação de unigramas, que, como os autores da medida mostraram, já basta para que se tenham resultados confiáveis.

Na avaliação humana, por enquanto, avaliamos somente o aspecto da redundância. O número de sentenças redundantes foi calculado para uma pequena amostra de sumários produzidos pelos diferentes operadores. O fato de se utilizar apenas uma amostra advém do custo e do tempo necessários para a avaliação humana.

Os resultados foram comparados com os resultados obtidos pelo único sumarizador multidocumento conhecido para a língua portuguesa, o GistSumm (Pardo et al., 2003, 2005). Este sumarizador concatena todos os textos de uma mesma coleção em um único arquivo e, posteriormente, sumariza-o utilizando um método de sumarização baseado nas palavras mais freqüentes. Esse método é muito simples, mas ainda assim robusto, correspondendo, portanto, a um ótimo *baseline*.

Na Tabela 3 são mostrados os resultados da avaliação para todos os operadores e para o GistSumm. Note que o operador de tratamento de redundância também foi avaliado de forma isolada, sem ser combinado com os demais.

Tabela 3. Resultados das avaliações

	Cobertura	Precisão	Medida-f	Sentenças redundantes
Informação Principal (operador genérico)	0.57218	0.52359	0.54384	3
Tratamento de Redundância	0.55137	0.54539	0.54299	0
Exibição de Informações Contraditórias	0.57108	0.51974	0.54114	1
Identificação de Autoria	0.56518	0.52368	0.53994	2
Apresentação de Eventos que Evoluem no Tempo	0.55136	0.49869	0.52110	3
Apresentação de Informação de Contexto	0.52079	0.48962	0.50171	4
GistSumm	0.66435	0.35997	0.45998	5

As primeiras 3 colunas da tabela reportam os resultados médios da ROUGE (que variam de 0 a 1 e, quanto maiores, melhores). Em geral podemos observar que todos os sumários produzidos pelos operadores têm melhores resultados do que o GistSumm em termos da medida-f. Logicamente, o operador genérico tem a maior medida-f dentre os operadores, já que os sumários de referência são genéricos também. Comparamos os sumários com preferências com os sumários genéricos para poder verificar seu nível de informatividade, independentemente do fato de terem priorizado outras informações. Em termos de precisão, o operador de tratamento de redundâncias é o melhor, pois elimina informações repetidas e pode, assim, incluir outras informações relevantes no sumário. É interessante notar também a alta cobertura do GistSumm e sua baixíssima precisão.

É importante notar que muitos operadores produziram resultados próximos do operador genérico e do de tratamento de redundância. Isso se deve ao fato de que alguns operadores têm poucas relações correspondentes disponíveis no ranque inicial, não alterando significativamente o sumário produzido. Por exemplo, há poucas relações *Contradiction* no córpus, de forma que há grandes chances de o sumário automático não ser muito alterado pelo operador de exibição de informações contraditórias.

O teste estatístico anova mostrou que os resultados da ROUGE obtidos são significantes com 95% de confiança.

A última coluna da tabela exibe o número de sentenças redundantes encontradas nos sumários. Pode-se observar que todos os operadores geraram sumários menos redundantes e, portanto, mais coerentes do que os sumários gerados pelo GistSumm. Como esperado, o operador de tratamento de redundâncias produziu sumários sem redundância alguma. Por outro lado, mesmo com a aplicação prévia do operador de tratamento de redundância, os operadores de preferência produziram redundâncias. Essas redundâncias são explicadas principalmente pela presença das relações *Contradiction* e *Overlap*: a primeira sempre traz alguma redundância consigo, enquanto a segunda não foi devidamente tratada neste trabalho (via fusão das sentenças envolvidas, por exemplo). Outra possibilidade é que existam sentenças nos textos que não tenham sido anotadas com relações CST, mas que de fato tenham relação entre si e contenham redundância.

A seguir, fazemos algumas considerações finais.

5. Considerações Finais

Neste trabalho, foram definidos, formalizados e avaliados um conjunto de operadores de seleção de conteúdo para SAM com base na CST. Mostramos que o uso da CST permite explorar o conhecimento entre vários textos que versam sobre um mesmo assunto, o que ajuda na seleção de conteúdo, melhorando a informatividade e coerência nos sumários finais.

Trabalhos futuros incluem a elaboração de novas estratégias de seleção de conteúdo com base na CST, incluindo possivelmente a criação de novos operadores de seleção de conteúdo. A avaliação do impacto da preferência do usuário, em particular, merece uma atenção maior. Neste artigo, tratou-se apenas da questão da informatividade, mas certamente alguma avaliação humana deverá ser conduzida, de tal forma que se possa mensurar a satisfação do usuário frente aos sumários gerados de acordo com suas preferências. Alternativamente,

sumários humanos com preferências específicas podem ser produzidos para serem considerados sumários de referência para avaliação automática dos sumários com preferências.

Acreditamos que a CST pode auxiliar em outros processos da sumarização, como ordenação das sentenças do sumário e resolução das correferências. A ordenação das sentenças, em especial, pode ter um grande efeito na coerência final do sumário, e deve ser foco de próximas pesquisas. Além disso, cremos também que a taxa de compressão utilizada interfere nos resultados obtidos, desde que, quanto maior a taxa, menos informação o sumário pode conter. Tal influência deve ser investigada em trabalhos futuros.

Por fim, é interessante notar que, em princípio, o trabalho apresentado é independente de língua e de gênero e domínio textual, já que a CST e, portanto, os operadores derivados dela são independentes de língua e genéricos o suficiente para serem aplicados a outros tipos de textos.

6. Agradecimentos

Os autores agradecem à FAPESP e ao CNPq pelo suporte a este trabalho.

7. Referências

Afantenos, S.D.; Doura, I.; Kapellou, E.; Karkaletsis, V. (2004). Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In the *Proceedings of SETN*, pp. 410-419.

Aleixo, P. and Pardo, T.A.S. (2008a). Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL, pp. 298-303. Vila Velha, Espírito Santo. October, 26-28.

Aleixo, P. and Pardo, T.A.S. (2008b). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 326.

Jorge, M.L.C and Pardo, T.A.S. (2009). Content Selection Operators for Multidocument Summarization based on Cross-document Structure Theory. In the *Brazilian Symposium in Information and Human Language Technology*. São Carlos, Brazil.

Jorge, M.L.C. and Pardo, T.A.S. (2010). Formalizing CST-based Content Selection Operations. In the *Proceedings of the International Conference on Computational Processing of Portuguese Language* - PROPOR. April 27-30, Porto Alegre/RS, Brazil.

Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In the *Proceedings of 2003 Language Technology Conference*. Edmonton, Canada.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co. Amsterdam.

Mani, I. and Maybury, M. T. (1999). *Advances in automatic text summarization*. MIT Press, Cambridge, MA.

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Otterbacher, J.C.; Radev, D.R.; Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. In the *Proceedings of the Workshop on Automatic Summarization*, pp 27-36.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In the *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal.

Pardo, T.A.S. (2005). *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP/Brasil.

Radev, D.R. and McKeown, K. (1998). Generating natural language summaries from multiple on-line sources.

Computational Linguistics, Vol. 24, N. 3, pp. 469-500.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

Seno, E.R.M. e Nunes, M.G.V. (2009). *Fusão Automática de Sentenças Similares em Português*. *Linguamática*, Vol. 1, pp. 71-87.

Trigg, R. (1983). *A Network-Based Approach to Text Handling for the Online Scientific Community*. Ph.D. Thesis. Department of Computer Science, University of Maryland.

Trigg, R. and Weiser, M. (1987). TEXTNET: A network-based approach to text handling. *ACM Transactions on Office Information Systems*, Vol. 4, N. 1, pp. 1-23.

Zhang, Z.; Goldenshon, S.B.; Radev, D.R. (2002). Towards CST-Enhanced Sumarization. In the *Proceedings of the 18th National Conference on Artificial Intelligence*.

Zhang, Z.; Otterbacher, J.C.; Radev, D.R. (2003). Learning Cross-document Structural Relationships Using Boosting. In the *Proceedings of Conference on Information and Knowledge Management*, pp. 124-130.

Um Analisador Semântico Inferencialista de Sentenças em Linguagem Natural

Vladia Pinheiro
Universidade Federal do Ceará
vladia@lia.ufc.br

Tarcisio Pequeno
Universidade Federal do Ceará
tarcisio@lia.ufc.br

Vasco Furtado
Universidade de Fortaleza e ETICE
vasco@unifor.br

Resumo

Este artigo descreve um raciocinador semântico para entendimento de linguagem natural que implementa um algoritmo que raciocina sobre o conteúdo inferencial de conceitos e padrões de sentenças – o Analisador Semântico Inferencialista (SIA). O SIA implementa um raciocínio material e holístico sobre a rede de potenciais inferências em que os conceitos de uma língua podem participar, considerando como os conceitos estão relacionados na sentença, de acordo com padrões de estruturas sintáticas. A medida de relacionamento inferencial e o processo de raciocínio do SIA são descritos. O SIA é usado como raciocinador semântico em um sistema de extração de informações sobre crimes – WikiCrimesIE. Os resultados obtidos e uma análise comparativa são apresentados e discutidos, servindo para a identificação de vantagens e oportunidades de melhoria para o SIA.

1. Introdução

Para o entendimento de linguagem natural por computadores, algumas questões de pesquisas são fundamentais e ainda estão em aberto: (i) *Qual o conhecimento semântico que deve ser expresso?* (ii) *Como se calcula ou infere o significado de uma expressão linguística?*.

Comumente, pesquisas e aplicações das áreas de Linguística Computacional (LC) e Processamento de Linguagem Natural (PLN) resolvem os problemas do nível semântico das linguagens naturais (responder perguntas sobre um texto, extrair informações, sumarizar textos, gerar textos etc) usando abordagens sintáticas. Dentre estas, podemos citar aquelas que consideram parâmetros morfossintáticos para identificar similaridade e relacionamento semânticos, por exemplo, a concordância de número para resolução de anáforas, freqüência de palavras em comum para fusão de textos, extração de informações a partir de padrões sintáticos de entidades nomeadas (endereços, cidades, empresas). Noutras abordagens, a intensão de um conceito (o "significado" de um conceito) é apreendida de sua extensão, expressa normalmente em um *corpus linguístico*. Em resumo, recorre-se a um processo de sintatização do nível semântico da linguagem que, claramente, é insuficiente para um completo entendimento de textos em linguagem natural.

Outros sistemas e aplicações usam conhecimento semântico onde a intensão dos conceitos é definida em bases de conhecimento (normalmente denominadas de ontologias) contendo classes, propriedades e atributos dos objetos referenciados pelos termos de uma língua natural. Outra característica destes sistemas é que eles normalmente adotam uma abordagem atomista para raciocínio semântico. Nesta abordagem, a interpretação semântica de um elemento é tratada de forma independente da atribuição semântica dos demais elementos de uma sentença. Estas características – a priorização de uma representação do mundo para definição do significado e um raciocínio semântico atomista - limitam a capacidade de entendimento de linguagem natural dos sistemas de PLN.

Frequentemente, as informações necessárias para o entendimento completo de textos por sistemas de PLN estão implícitas, e descobri-las requer a realização de inferências a partir do uso de conceitos em situações linguísticas. Por exemplo, quando lemos a notícia "*João assassinou sua esposa com dois tiros após uma discussão na Rua Solon Pinheiro.*", nós somos capazes de refutar uma afirmação que indicasse que o tipo de arma usada no crime foi "arma branca" (não foi usada arma de fogo), argumentar que o tipo de crime foi "homicídio" e que a causa do crime foi "crime passional". Estas conclusões são possíveis porque nós, usuários da língua natural, sabemos as condições nas quais os conceitos "tiro", "assassinar" e "esposa" podem ser usados e os compromissos que assumimos ao usá-los em uma sentença. Além disso, raciocinamos considerando o conteúdo individual dos conceitos de forma

conjunta com o conteúdo dos demais conceitos da sentença em que são usados.

O fato é que habilidades como argumentar sobre o texto, responder perguntas, extrair informacões explícitas e implícitas, refutar afirmações etc. são cada vez mais necessárias em tarefas de PLN que envolvem entendimento de linguagem natural.

Um caminho para melhoria da qualidade do processamento semântico de sistemas de PLN é buscar inspiração nas respostas que filósofos oferecem à questão *Em que consiste o significado de uma expressão linguística?*. Sellars (1980), Dummett (1973) e Brandom (1994)(2000) propuseram as teorias semânticas inferencialistas, que apresentam uma abordagem diferente para definir o conteúdo de conceitos e sentenças. Segundo estas teorias, a expressão do valor semântico de conceitos deve privilegiar o papel que estes desempenham em raciocínios, como premissas e conclusões, ao invés de seus referentes e suas características representacionais. Segundo Sellars (1980), compreender um conceito é ter o domínio prático sobre as inferências em que ele pode estar envolvido – saber o que segue da aplicabilidade do conceito e a partir de que situações ele pode ser aplicado.

Seguindo esta visão inferencialista, Pinheiro et al. (2008)(2009) propõem o *Semantic Inferentialism Model* (SIM) - um modelo computacional que define requisitos para expressão e raciocínio sobre conhecimento semântico linguístico. Suas bases de conhecimento semântico expressam conteúdo inferencial de conceitos e sentenças, ou seja, as condições e consequências de uso de conceitos e sentenças.

O componente principal do SIM é seu raciocinador semântico de textos em linguagem natural: Analisador Semântico Inferencialista – SIA. O SIA é responsável por gerar as premissas e conclusões das sentenças do texto de entrada. Estas premissas e conclusões habilitam os sistemas de PLN para dar razões sobre o texto, responder perguntas, extrair informacões explícitas e implícitas, refutar afirmações etc. As regras de inferência e a medida de relacionamento inferencial, implementadas pelo SIA, são responsáveis por um mecanismo de raciocínio semântico material e holístico. Raciocínio material no sentido de que as inferências são autorizadas e justificadas pelos conteúdos conceituais, e raciocínio holístico porque o SIA define a contribuição semântica dos conceitos considerando outros conceitos relacionados em uma sentença, de acordo com sua estrutura sintática.

Este artigo está estruturado da seguinte forma, A seção 2 discute os fundamentos teórico-filosóficos e apresenta a arquitetura e formalização do SIM. A seção 3 apresenta o algoritmo do SIA, seu processo de raciocínio, regras de inferência, e a medida de relacionamento inferencial. Na seção 4, tem-se a descrição de como o SIA é aplicado em um sistema para extração de informações sobre crimes – Extrator de Informações WikiCrimes (WikiCrimesIE), e a avaliação dos resultados obtidos. Na seção 5, os trabalhos relacionados e uma análise comparativa são discutidos e, finalmente, este artigo é concluído com a apresentação dos trabalhos em andamento e futuros.

2. Semantic Inferentialism Model (SIM)

2.1 Fundamentos do SIM

O *Semantic Inferentialism Model* (SIM) (Pinheiro et al, 2008) (Pinheiro et al, 2009) define os principais requisitos para expressar e manipular conhecimento semântico inferencialista de forma a capacitar os sistemas de linguagem natural para um entendimento mais completo de sentenças e textos.

SIM é fortemente inspirado nas teorias semânticas inferencialistas de Sellars (1980), Dummett (1973) e Brandom (1994)(2000). Para Dummett, saber o significado de uma sentença é saber a justificativa para o falante tê-la proferido: "Nós não explicamos o sentido de uma declaração estipulando seu valor-verdade em termos dos valores-verdade de seus constituintes, mas sim estipulando quando ela pode ser afirmada em termos das condições sobre as quais seus constituintes podem ser afirmados" (Dummett, 1978). Brandom (1994)(2000), por sua vez, sedimenta a visão inferencialista de Dummett e Sellars e reduz a visão pragmática da linguagem de Wittgenstein (1953) para um racionalismo pragmático, onde a tônica são os usos inferenciais de conceitos em jogos de pedir e dar razões (jogos racionais). Para Brandom, entendemos uma sentença quando sabemos defendê-la, argumentar a seu favor, dar explicações, e isto só é possível porque sabemos inferir as premissas que autorizaram seu proferimento e as conclusões de seu proferimento.

Seguindo esta visão inferencialista, SIM responde à questão (i) *Qual o conhecimento semântico que deve ser expresso?* definindo que expressar o conteúdo de um conceito requer expressar, tornando explícito, seus usos [do conceito] em inferências, como premissas ou conclusões de raciocínios. E, o que determina o uso de um conceito em inferências ou as potenciais inferências em que este conceito pode participar são:

- precondições ou premissas de uso do conceito – o que dá direito a alguém a usar o conceito e o que poderia excluir tal direito, servindo de premissas para proferimentos e raciocínios;
- pós-condições ou conclusões do uso do conceito – o que se segue ou as conseqüências do uso do conceito, as quais permitem saber com o que alguém se compromete ao usar um conceito, servindo de conclusões do proferimento em si e de premissas para futuros proferimentos e raciocínios.

Este conteúdo, denominado de *conteúdo inferencial*, define o importe ou competência inferencial de um conceito. Esta visão inferencialista de conteúdo conceitual se contrapõe à visão representacionalista, segundo a qual os sistemas de PLN deveriam expressar uma representação do mundo *a priori*. Eco (2001) assinala que qualquer classificação ou caracterização do mundo (qualquer representação do mundo) é conjectural e arbitrária, mesmo que consensual em uma comunidade ou área de conhecimento. Portanto, não se pode delimitar o poder de entendimento dos sistemas de PLN a este "muro ontológico" e a um método cartesiano de raciocínio semântico, no qual, a partir de hipóteses (uma representação do mundo), seguimos concluindo isso ou aquilo através de regras formais. Como conseqüência, a verdade de nossas conclusões herda as limitações da organização artificial do mundo, ou seja, tudo o que se pode entender de textos em linguagem natural já está condicionado *a priori* nas hipóteses assumidas.

Em contraposição, o que precisa ser expresso sobre um conceito deve ser expresso a partir de seus usos em práticas linguísticas. Isto é concernente com a idéia de que conceitos surgem dentro da prática linguística de uma comunidade, sociedade ou de uma área de conhecimento, e são apreendidos pelos usuários de uma língua a partir de seus usos e não porque existem *a priori* no mundo com tais e tais características. Para ilustrar a natureza do conteúdo inferencial de conceitos, imaginemos uma criança que, pela primeira vez, presencie o uso do conceito "egoísta" em uma discussão entre seus pais. Provavelmente, ela usará este conceito em uma situação de disputa com um colega de escola , ou seja, para ela, "alguém fazer algo que não gosto" é condição suficiente para que ela empregue o conceito. Na medida de seu amadurecimento na linguagem, perceberá que existem outras precondições e que, nem sempre, quando duas pessoas discutem, ela poderá usar o conceito "egoísta".

Em outro exemplo, tem-se o conceito "saidinha bancária" que se originou dentro da prática linguística de se descrever assaltos em que os clientes são abordados após realizarem saques em agências bancárias. Não se originou, prioritariamente, pelas representações deste tipo de crime, mas pelas circunstâncias e as conseqüências que ditaram seu uso. Os usuários deste conceito aprenderam em que situações usá-lo e o que se segue do seu uso. Embora existam os conceitos "saidinha" e "bancária", a nova expressão linguística "saidinha bancária" denota um conteúdo com valor semântico distinto que foi moldado a partir de seus usos em sentenças.

Brandom (2000) apresenta exemplos onde um papagaio pode falar "Esta bola é vermelha!" na presença de uma bola da cor vermelha, e um termostato pode ligar o compressor de um ar-condicionado quando a temperatura está acima de 20ºC. Brandom discute a natureza da distinção entre estes relatos e quando os mesmos relatos são feitos por humanos. A resposta dada, à luz das teorias inferencialistas, é que tanto o papagaio quanto o termostato não sabem defender, dar razões, explicar seus relatos em situações de raciocínio – e isto é porque não conhecem as circunstâncias e conseqüências do uso dos termos "vermelho" e "quente" em situações linguísticas, não conhecem as potenciais inferências em que estes conceitos podem participar. Da mesma forma, uma criança que escuta um termo específico de uma área de conhecimento, por exemplo "inteligência artificial", provavelmente não saberá quando usá-lo e, se usá-lo, que conclusões podem ser inferidas. Isto implica dizer que a criança não sabe participar de situações linguísticas e raciocinar com este termo – não saberá defendê-lo, explicá-lo etc - ou seja, não entende este termo. Mesmo ontologias simples, que definem taxonomias, ou base semânticas mais complexas, que expressam conhecimento causal, funcional ou relativo a eventos, devem ser consideradas sob o ponto de vista inferencial e pragmático. Ou seja, seus conteúdos devem ser manipulados e qualificados em termos de precondições ou pós-condições de uso dos conceitos, em situações linguísticas, capturando, desta forma, o viés pragmático da linguagem natural. O argumento do SIM é que um modelo semântico, inspirado nas teorias inferencialistas, possibilita melhor habilidade aos sistemas de PLN que se proponham a entender linguagem natural.

O SIM, baseado no paradigma semântico-inferencialista, também apresenta resposta à questão

(ii) *Como se calcula ou infere o significado de uma dada sentença?*.

Os raciocinadores dos sistemas lógicos, usuais em PLN (Lógica Descritiva, PROLOG e Lógicas Intensionais), se resumem a realizar inferências formais, as quais geram novos fatos considerando apenas a forma das expressões lógicas e um raciocínio logicamente autorizado.

Acontece que muitas conclusões e respostas que humanos dão ao ler um texto são justificadas pelo conteúdo dos conceitos relacionados. Por exemplo, considere a inferência de "João é irmão de Pedro" para "Pedro é irmão de João". O conteúdo do conceito "irmão" é que torna esta inferência correta. Se substituirmos na primeira sentença o conceito "irmão" pelo conceito "pai", a inferência não pode ser realizada. Da mesma forma, a inferência "um relâmpago é visto agora" para "um trovão será ouvido em breve" é autorizada pelo conteúdo dos conceitos "trovão" e "relâmpago". Em outro exemplo mais complexo, a inferência de "João assassinou sua esposa com dois tiros" para "o tipo de arma usada foi arma de fogo" é autorizada pelo conteúdo dos conceitos "assassinar" e "tiro", analisados conjuntamente.

Para se realizar inferências desta natureza não se deve ter unicamente um mecanismo de raciocínio sobre a forma das sentenças, mas principalmente deve-se ter domínio dos conteúdos dos conceitos articulados nas sentenças e de como estes [os conteúdos dos conceitos] contribuem para o significado das mesmas. Daí a importância da natureza do conteúdo dos conceitos, expresso em bases semânticas.

Outro fato observável é a tradição atomista na semântica formal. Na abordagem atomista a atribuição de uma interpretação semântica a um elemento é tratada de forma independente da atribuição semântica dos demais elementos de uma sentença. Ao contrário, a semântica inferencialista é essencialmente holista: não se pode definir o valor semântico de um elemento sem considerar os outros elementos relacionados em uma sentença e como todos estão estruturados. Define-se "essencialmente holista" porque esta característica é uma conseqüência direta e simples da concepção inferencial do conteúdo de conceitos - "ninguém pode ter qualquer conceito a menos que tenha muitos conceitos" (Brandom, 2000, p.15-16). Ao expressar as potenciais inferências em que um conceito pode estar envolvido nada mais fazemos que expressar as relações inferenciais deste com outros conceitos e, na medida em que conhecemos um conceito, conhecemos vários.

Em contraposição à predominância, nos sistemas de PLN, de inferências formais e de raciocínio atomista, o SIM propõe o Analisador Semântico Inferencialista (SIA). O SIA implementa um raciocínio material e holístico sobre a rede de [potenciais] inferências em que conceitos podem participar (conteúdo inferencial), considerando como os conceitos estão relacionados na sentença, de acordo com sua [da sentença] estrutura sintática.

O raciocínio material possibilita a realização de inferências autorizadas pelo conteúdo (p.ex. de "um relâmpago é visto agora" para "um trovão será ouvido em breve", autorizada pelo conteúdo dos conceitos "trovão" e "relâmpago"), e argumento para refutar e validar inferências (p.ex. "A água é vermelha" é refutada pela precondição de uso do conceito "água" que define que este conceito só pode ser usado em sentenças onde não são associados ao mesmo uma cor).

O raciocínio holístico, por sua vez, considera o todo (sentença) e como suas partes (elementos subsentenciais) estão estruturalmente relacionadas a fim de definir a contribuição semântica de cada parte para com o todo (sentença). Nesta abordagem holística, as estruturas de sentenças assumem um papel importante porque, para determinar o valor semântico de um elemento subsentencial, devem-se considerar os outros elementos relacionados e é imprescindível levar em conta a estrutura que os organiza na sentença e que define suas formas e funções sintáticas. Tem-se, portanto, uma abordagem de raciocínio semântico de cima para baixo (ou *top-down*). Por exemplo, seja a sentença "Geison Santos de Oliveira foi executado com vários tiros", a qual segue uma estrutura de sentença que relaciona os conceitos "executar" (assassinar) e "tiro". Pelo conteúdo inferencial dos conceitos "executar" (assassinar) e "tiro" e como eles são articulados na sentença, é possível identificar similaridade inferencial entre ambos e definir que o conceito usado na sentença foi "executar", que alude a assassinar, e não "executar" com acepção a realizar algo. É importante salientar que o raciocínio holístico mantém a característica de composicionalidade da linguagem, no sentido de que o significado da sentença é obtido com base no conteúdo semântico dos elementos subsentenciais. No entanto, ao considerar como a estrutura da sentença articula seus elementos subsentenciais a fim de definir a contribuição semântica desses para com a sentença, tem-se uma abordagem holista de raciocínio.

As duas qualidades de raciocínio semântico em linguagem natural do SIA – material e holístico – completam o diferencial deste analisador semântico de textos em linguagem natural.

2.2 Formalização do SIM

A Figura 1 apresenta a arquitetura do SIM. O SIM contém três bases para expressão de conhecimento semântico:

- Base Conceitual — que contém conceitos da língua natural e seus conteúdos inferenciais;
- Base de Sentenças-Padrão — que contém sentenças-padrão e seus conteúdos inferenciais; e
- Base de Regras para Raciocínio Prático — que contém a expressão de conhecimento prático oriundo da cultura de uma comunidade ou de uma área de conhecimento.

Além de bases de conhecimento, o SIM inclui um componente responsável pelo raciocínio semântico de sentenças e textos em linguagem natural:

- Analisador Semântico Inferencialista (SIA) — que recebe o texto de entrada em linguagem natural e a árvore de dependência sintática do texto, gerada por um analisador (ou *parser*) morfossintático, e, a partir do conteúdo expresso nas três bases semânticas, gera a rede inferencial das sentenças do texto.

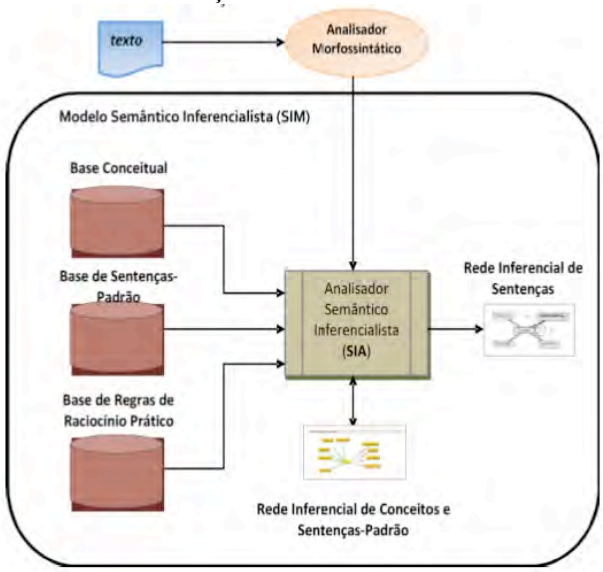


Figura 1: Arquitetura do *Semantic Inferentialism Model* - SIM.

A **Base Conceitual** contém o conteúdo inferencial de conceitos em língua natural, definidos e acordados em uma comunidade ou área de conhecimento. De acordo com a visão inferencialista, o conteúdo de um conceito *c* são as potenciais inferências em que *c* pode participar e o que determina esta participação são suas relações inferenciais com outros conceitos, na forma de suas precondições e pós-condições de uso. A base conceitual é um grafo direcionado $G_c(V,E)$, onde:

- V = conjunto não vazio de conceitos c_i (vértices do grafo). Um conceito em V pode ser representado na base conceitual por termos simples, que pertencem às classes abertas de palavras - nomes, verbos, adjetivos, advérbios (p.ex, 'crime', 'morte'); ou por expressões compostas de mais de um termo, ligados ou não por palavras das classes fechadas - preposições e conjunções (p.ex. 'prova de matemática', 'saidinha bancária').
- E = conjunto de arestas rotuladas por uma variável *tipo_rel* que expressa a relação binária entre conceitos de V. As relações *tipo_rel* são predefinidas como expressando as duas relações inferenciais: precondição ou pós-condição de uso de um conceito. Por exemplo, tem-se as relações "*CapazDe*" e "*EfeitoDesejávelDe*", onde a primeira define uma precondição de uso e a segunda uma pós-condição de uso de conceitos. Neste trabalho, usamos o formato $tipo\_rel(c_1,c_2)$ para expressar a relação inferencial *tipo_rel* entre c_1 e c_2, ambos conceitos em V, a qual pode ser interpretada como "c_1 possui *tipo_rel* em relação a c_2". Por exemplo, *CapazDe('crime','envolver violência')* é interpretada como "*crime é CapazDe envolver violência*".

Como se trata de um digrafo, $G_c(V,E)$ possui duas funções *s* e *t* onde:

- $s:E \rightarrow V$ é uma função que associa uma aresta de E ao seu conceito de origem em V;
- $t:E \rightarrow V$ é uma função que associa uma aresta de E ao seu conceito alvo em V.

A Base Conceitual permite expressar as relações inferenciais de um conceito com outros, obedecendo à visão holista de que conhecer um conceito é conhecer suas relações, na forma de premissas ou conclusões, com outros conceitos. A figura 2 apresenta o grafo inferencial G_{crime} do conceito '*crime*', o qual expressa as precondições e pós-condições de uso do conceito através de relações com outros conceitos.

A **Base de Sentenças-Padrão** contém sentenças genéricas que seguem uma dada estrutura sintática e que funcionam como *templates*, cujos *slots* podem ser preenchidos com termos de uma língua natural. Uma sentença-padrão segue certa estrutura de sentença, com algumas partes variáveis a serem

preenchidas (*slots*) por conceitos da base conceitual ou por outros elementos subsentenciais (nomes próprios, artigos, preposições, conjunções etc). Por exemplo, '*X ser assassinar por Y*' segue a estrutura <sentença> ::= <sn> <sv> <sp>[1]. O sintagma nominal (sn), representado por *X*, pode ser preenchido por 'uma mulher' e o sintagma complementar (sp), representado por 'por *Y*', pode ser complementado por 'seu amante'. Assim teremos a sentença '*uma mulher ser assassinar por seu amante*', gerada a partir do padrão '*X ser assassinar por Y*'.

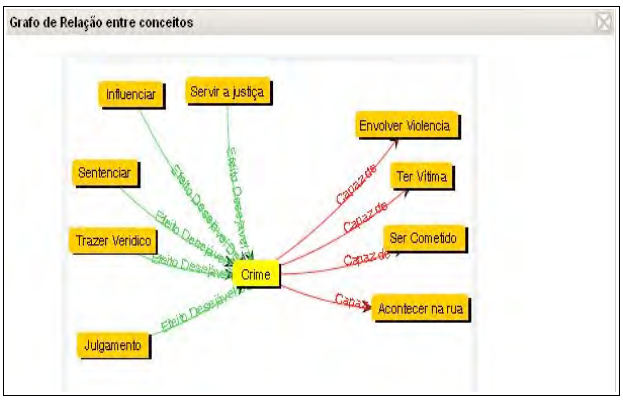


Figura 2: Grafo Inferencial do conceito 'crime'.

A importância de uma base de sentenças-padrão para análise semântica é a seguinte: algo do que se pode inferir ao se ler uma sentença não advém direta e unicamente, pelo menos de forma eficiente, do conteúdo dos conceitos da sentença, mas destes juntos e articulados sob determinada estrutura de sentença. Por exemplo, uma pessoa ao ler a sentença "*uma mulher foi assassinada por seu amante*" rapidamente é capaz de responder quem foi a vítima ('uma mulher') e quem foi o assassino ('seu amante'). Um mecanismo de raciocínio para gerar estas conclusões poderia até raciocinar sobre o conteúdo do conceito 'assassinar' - precondições de uso 'existir um assassino' e 'existir uma vítima' - porém, identificá-las de forma direta e eficiente na sentença exigiria conhecimento sobre como articular este conteúdo e sobre o elemento estruturador 'por'. Este conhecimento é justamente o que a base de sentenças-padrão provê: expressar conteúdo inferencial (premissas e conclusões) das sentenças-padrão. Este conteúdo inferencial consiste de conhecimento que não pode, pelo menos de forma eficiente, ser inferido do conteúdo dos conceitos. A base de sentenças-padrão consiste, portanto, de um grafo direcionado $G_s(V,E)$, onde:

- V = conjunto não vazio de sentenças-padrão s_j e conceitos c_i (vértices do grafo);
- E = conjunto de arestas rotuladas pela variável *tipo_rel*, que expressa uma relação binária entre uma parte (nominal, verbal ou complementar) da sentença-padrão p_j e um conceito c_i, ambos elementos de V, sempre na direção da sentença p_j para o conceito c_i. As relações *tipo_rel* são predefinidas como expressando as duas relações inferenciais: precondição ou pós-condição da sentença-padrão p_j. Neste trabalho, usamos o formato *tipo_rel(parte(p_j),c_i)* para expressar a relação inferencial *tipo_rel* entre uma parte de p_j e c_i. Esta relação é interpretada como "a parte de p_j possui *tipo_rel* em relação a c_i". Por exemplo, a precondição *ehUm(sn('X ser assassinar por Y'), 'pessoa')* é interpretada como "*X ehUm pessoa*", e a pós-condição *ehUm(sp('X ser assassinar por Y'), 'assassino')* é interpretada como "*Y ehUm assassino*".

Da mesma forma que no grafo da base conceitual (G_c), $G_s(V,E)$ possui duas funções *s* e *t* onde:

- $s:E \rightarrow V$ é uma função que associa uma aresta de E a uma sentença-padrão em V;
- $t:E \rightarrow V$ é uma função que associa uma aresta de E ao seu conceito alvo em V.

A **Base de Regras para Raciocínio Prático** possibilita a expressão de conhecimento prático oriundo da cultura de uma comunidade, através de regras. Cada regra ρ_i visa combinar as premissas e conclusões já geradas para uma sentença original s_i, gerando novas premissas e conclusões para s_i. As regras ρ_i são cláusulas Horn da forma ($A_1 \wedge A_2 \wedge \ldots \wedge A_n \rightarrow$ *tipo_rel*(s_i, s_j)). Por exemplo, há o consenso oriundo da cultura das grandes cidades, que o local provável do crime é o local onde o cadáver foi encontrado. Para expressar este conhecimento pode-se usar a seguinte regra:

$\forall x,y,z$ (((encontrar(x,y) $\wedge$ encontrarEm(y,z) $\wedge$ ehUm(y,'cadáver')) $\rightarrow$ ('Pos', s_i,ehUm(z,'local do crime')))

A construção destas bases de conhecimento semântico inferencialista é um desafio, principalmente pelo seu caráter inovador. Uma proposta de construção das bases está descrita em (Pinheiro et al., 2010) e deu origem ao primeiro recurso linguístico com conteúdo semântico inferencialista para língua portuguesa – InferenceNet.BR – contendo em torno de 190.000 conceitos, 700.000 relações inferenciais entre conceitos, 6000 sentenças-padrão e 1500 relações inferenciais de sentenças-padrão. O sítio www.inferencenet.org possui funcionalidades para consulta, evolução e disseminação deste recurso linguístico.

1 <SN>: sintagma nominal; <SV>: sintagma verbal; <SP>: sintagma complementar

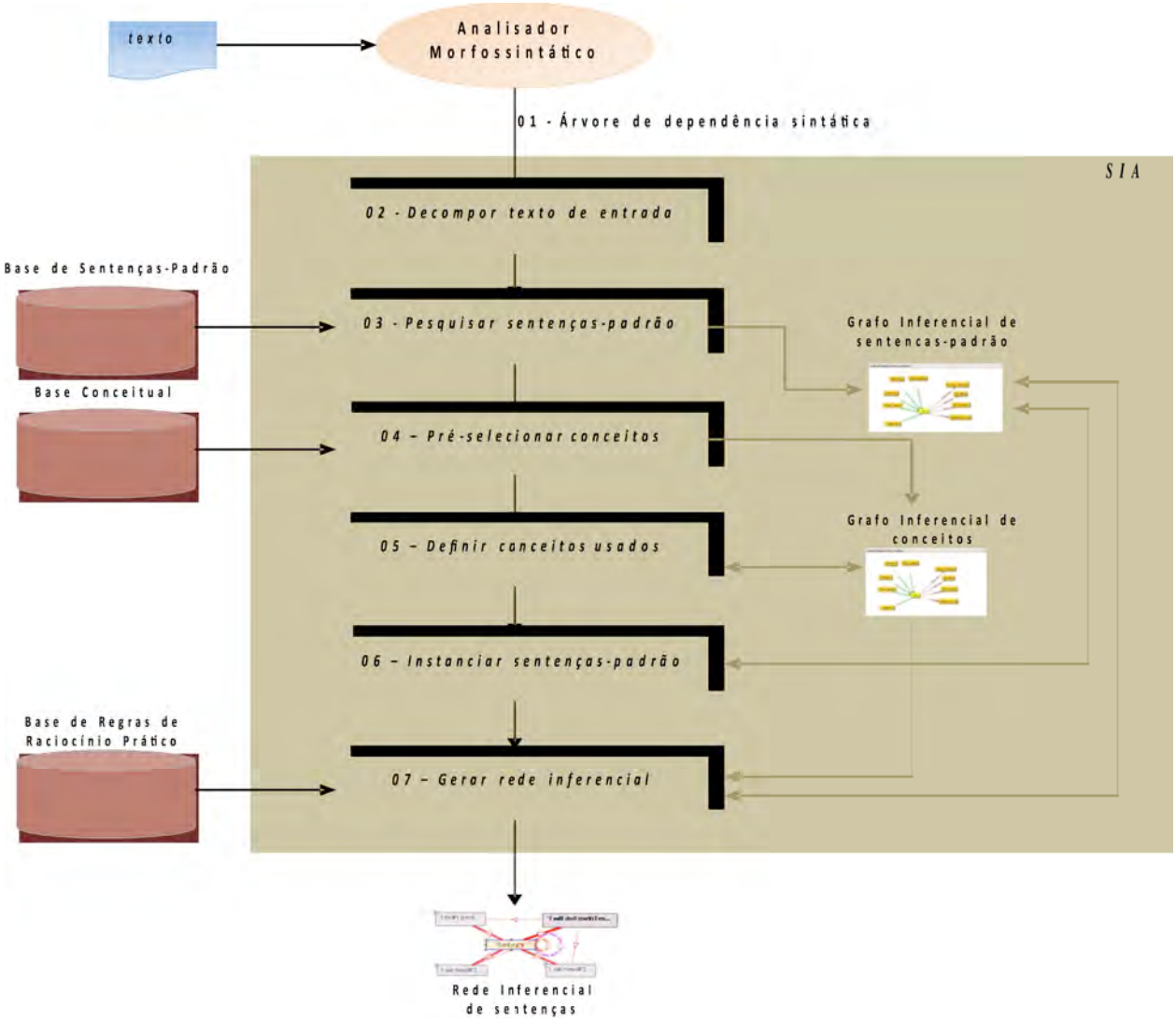


Figura 3: Visão gráfica do processo de raciocínio semântico do SIA.

3. Analisador Semântico Inferencialista (SIA)

O componente principal do SIM é seu raciocinador semântico de textos em linguagem natural - o Analisador Semântico Inferencialista – SIA. Em linhas gerais, um analisador semântico tem como objetivo descobrir o significado das expressões em linguagem natural e realizar o entendimento de sentenças em linguagem natural (Vieira e de Lima, 2001). De acordo com a teoria semântica do SIM (Pinheiro et al., 2008), o significado de uma sentença em linguagem natural é o conjunto de suas premissas (precondições) e conclusões (pós-condições), geradas a partir do conteúdo inferencial de seus conceitos articulados em uma dada estrutura de sentença (sentença-padrão).

SIA implementa um mecanismo de inferência sobre os grafos G_c e G_s (base conceitual e base de sentenças-padrão) e da base de regras ρ_i, com o objetivo de gerar a rede inferencial (premisssas e conclusões) das sentenças do texto, a qual consiste em um grafo direcionado $G_N(V,E)$, onde:

- V = conjunto de sentenças s_i (vértices do grafo);
- E = conjunto de arestas rotuladas por uma variável *tipo_rel* que indica o tipo de relação inferencial (precondição (*pre*) ou pós-condição (*post*)) entre uma sentença original do texto s_i e outra sentença s_j, que expressa uma premissa ou conclusão de s_i. Estas sentenças são inferidas a partir do processo de raciocínio do SIA (ver seção 3.2). Neste trabalho, usamos o formato *tipo_rel*(s_i, s_j) que é interpretado como "s_j é *tipo_rel* de s_i". Por exemplo, a pós-condição *post("João comeu","João ganhar energia")* é interpretada como " *'João*

ganhar energia' é pós-condição (ou conclusão) de s_j = 'João comeu' ".

Similarmente aos grafos G_c e G_s, o grafo $G_N(V,E)$ possui as funções *s* e *t* que associa, a uma aresta em *E*, seus elementos de origem e destino (sentenças) em *V*.

A figura 3 apresenta a visão gráfica do algoritmo implementado pelo SIA, que define os seguintes passos:

(1) Inicialmente, o algoritmo recebe a árvore de dependência sintática do texto de entrada, a qual foi gerada por um analisador morfossintático.

(2) É realizado um pré-processamento para decomposição dos períodos do texto em sentenças simples. Sentenças simples são sentenças que seguem a estrutura <sentença> ::= <SN> <SV> <SP>. Este pré-processamento é realizado pelo SIA com base na árvore de dependência sintática gerada pelo analisador morfossintático e gera uma sentença simples para cada ocorrência distinta de sujeito, verbo (ou locução verbal) e complemento verbal. Este passo é necessário porque os períodos compostos dificultam a combinação com sentenças-padrão. Por exemplo, o texto "*Na noite do último sábado, um jovem identificado como Geison Santos de Oliveira foi executado com vários tiros na Rua Titan, 33*" contém um período composto por três sentenças simples (s_1="*Um jovem identificado... foi executado na noite do último sábado*; s_2="*Um jovem identificado... foi executado com vários tiros*"; s_3="*Um jovem identificado... foi executado na Rua Titan, 33*").

(3) São pesquisadas e combinadas as estruturas das sentenças simples do texto com sentenças-padrão da Base de Sentenças-Padrão do SIM, gerando grafos G'_s (subgrafos de G_s) para cada sentença-padrão identificada.

(4) São pré-selecionados os conceitos da Base Conceitual do SIM que combinam literalmente com os termos usados nas sentenças do texto. Este passo é necessário porque existe um ou mais conceitos na base conceitual que são homônimos. Por exemplo, "executar" no sentido de realizar ou fazer, e "executar" com acepção a assassinar ou fuzilar.

(5) Neste passo, são definidos, dentre os conceitos pré-selecionados, quais conceitos foram usados, utilizando a ordem definida pela Medida de Relacionamento Inferencial, descrita na seção 3.1. Neste ponto, o algoritmo elimina os conceitos homônimos pré-selecionados que possuem menor proximidade inferencial com os demais conceitos da sentença s_i em que são usados. Para cada conceito *c* definido, gera grafos G'_c (subgrafo de G_c, a base conceitual do SIM, tal que $c \in V'$, conjunto de vértices de G'_c). Em seguida, é definida a contribuição semântica dos conceitos para a sentença s_i. A contribuição semântica de um conceito *c* usado em uma sentença s_i é o subgrafo de G'_c, gerado pela eliminação de G'_c das precondições e pós-condições que não influenciaram na proximidade inferencial de *c* com demais conceitos de s_i.

(6) Cada sentença-padrão em G'_s é instanciada com os elementos subsentenciais da sentença original (conceitos, preposições e outros elementos de ligação).

(7) Finalmente, neste passo é gerada a rede inferencial $G_N^{si}(V,E)$ de premissas e conclusões de cada sentença s_i do texto original. É o método principal do SIA, pois implementa formas de raciocínio que endossam inferências a partir de(a): (i) contribuição semântica dos conceitos usados em s_i, de acordo com a estrutura da sentença-padrão que os articula; (ii) contribuição semântica da sentença-padrão correspondente a s_i. Ambas as contribuições foram definidas a partir dos conteúdos inferenciais (pré e pós-condições) dos conceitos e sentenças-padrão e estão expressas nos subgrafos de G'_c e G'_s; (iii) regras pragmáticas da Base de Regras de Raciocínio Prático. As formas de raciocínio do SIA são detalhados na seção 3.2. Opcionalmente, objetivos da aplicação cliente são considerados para filtrar as premissas e conclusões geradas. A definição destes objetivos e como eles são usados pelo SIA são detalhadas na seção 3.2.

3.1 Medida de Relacionamento Inferencial

Cada vez mais, aplicações em Linguística Computacional requerem uma medida de relacionamento ou parentesco semântico entre dois conceitos e muitas abordagens têm sido sugeridas (Budanitsky e Hirst, 2001). A despeito de qualquer discussão filosófica e psicológica sobre a existência de uma medida numérica para a noção intuitiva de relacionamento semântico, a importância de uma

medida é que ela define uma relação de ordem (c_1 é mais similar a c_2 do que a c_3).

De acordo com a visão inferencialista e holística do SIM, o relacionamento entre conceitos não deve ser dissociado da sentença em que são usados e deve tomar como base o conteúdo inferencial compartilhado entre os conceitos articulados. Nesse sentido, dois conceitos usados em uma sentença estarão mais "inferencialmente relacionados" quanto mais o conjunto das precondições (ou das pós-condições) de um conceito é igual ao conjunto das precondições (ou das pós-condições) do outro conceito. A hipótese é que quanto mais as circunstâncias e consequências de uso de dois conceitos são semelhantes mais eles [os conceitos] podem ser usados em fluxos de raciocínio semelhantes.

São definidas, então, três formas de proximidade inferencial entre dois conceitos c_1 e c_2. Para cada uma das formas, tem-se um conjunto de relações inferenciais de c_1 e c_2 que satisfazem às condições de proximidade inferencial. As formas de proximidade inferencial são:

- **Proximidade por Relação Direta —** quando uma precondição (ou pós-condição) de c_1 expressa uma relação direta com c_2, ou vice-versa. Por exemplo, no caso dos conceitos c_1 = "crime" e c_2="roubo", e o conceito "roubo" possui a relação inferencial *éUm(roubo,crime)*.
- **Proximidade por Relação em Comum—** quando c_1 e c_2 expressam o mesmo tipo de relação semântica com um mesmo conceito. Por exemplo, no caso dos conceitos c_1 ="crime" e c_2="roubo" e ambos possuírem as relações inferenciais *capazDe(crime, ter vítima)* e *capazDe(roubo, ter vítima)*.
- **Proximidade por Relação de mesma Natureza —** quando c_1 e c_2 expressam relações inferenciais de mesma natureza (relações funcionais, causais, de eventos etc) com um mesmo conceito. Por exemplo, no caso dos conceitos c_1="tiro" e c_2="dedo" e ambos possuírem as relações inferenciais *usadoPara(tiro,ferir)* e *capazDeReceberAcao(dedo,ferir)*, onde as relações semânticas "usadoPara" e "capazDeReceberAcao" são de mesma natureza.

A medida de relacionamento inferencial $\theta_{c1,c2}$, entre dois conceitos c_1 e c_2 é calculada pela fórmula a seguir.

$$\theta_{(c_1,c_2)} = (F_1 w_1 + F_2 w_2 + F_3 w_3)\,\mu_{(c_1,c_2)}$$

Onde,

- F_1, F_2, F_3 são os somatórios das forças das relações inferenciais de c_1 e c_2 que satisfazem às três formas de proximidade inferencial, definidas acima;
- w_1, w_2, w_3 são os pesos, atribuídos por parâmetro, das três formas de proximidade inferencial, definidas acima; e
- $\mu_{(c1,c2)}$ é o fator de normalização entre os conceitos c_1 e c_2, calculado pela fórmula a seguir.

$$\mu_{(c_1,c_2)} = \frac{(n+m+p)}{|R_{(c_2)}|}$$

onde:

- (n+m+p) é o total de relações inferências de c_1 e c_2 que são semelhantes nas três formas de proximidade inferencial acima; e
- $|R_{c2}|$ é a cardinalidade do conjunto de relações inferenciais de c_2.

O fator de normalização serve para evitar que um conceito c_1 seja considerado mais inferencialmente relacionado a c_2 do que a c_3, somente porque c_2 possui maior número de relações inferenciais e, por isso, provavelmente maiores serão os valores de F_1, F_2 e F_3 , calculados entre c_1 e c_2.

A medida de relacionamento inferencial é utilizada no SIA para:

- desambiguação de termos homônimos;
- definição da contribuição semântica de um conceito *c* para a sentença *s*, pelo descarte de pré e pós condições de *c* que são irrelevantes para definição da proximidade inferencial de *c* com demais conceitos da sentença *s*;
- seleção de premissas e conclusões a serem geradas na rede inferencial da sentença *s*, a partir dos conceitos relacionados aos objetivos da aplicação cliente.

3.2 Raciocínio Inferencial do SIA

O SIA implementa três formas de raciocínio semântico para geração da rede inferencial $G_N^{si}(V,E)$, contendo premissas e conclusões das sentenças s_i do texto de entrada.

A primeira forma de raciocínio gera premissas e conclusões das sentenças do texto de entrada com base no conteúdo inferencial de conceitos usados nas sentenças. Definimos regras genéricas de introdução e eliminação de conceitos que podem ser instanciadas para cada conceito. A inspiração para estas regras vem do padrão de definição de conectivos lógicos de Gentzen (1935). Para Gentzen, um conectivo lógico é definido através de regras de introdução, que especificam sob quais circunstâncias o conectivo pode ser introduzido em um teorema; e através de regras de eliminação, que especificam sob quais condições o conectivo pode ser eliminado de um teorema. Dummett (1978) transpôs este modelo de definição para os conceitos de uma língua: um conceito é definido especificando-se regras de introdução do conceito (precondições de uso do conceito ou condições suficientes para uso do conceito) e regras de eliminação do conceito (pós-condições de uso do conceito ou conseqüências necessárias do uso do conceito).

A seguir, são apresentadas a interpretação e a sintaxe[2] das regras de introdução e de eliminação de conceitos em sentenças, e como estas regras são usadas pelo SIA para geração de premissas e conclusões das sentenças do texto de entrada.

(1) **A regra (I-c)** define que, se uma precondição de um conceito for satisfeita, a qual atende a uma precondição de uma sentença-padrão, então o conceito pode ser usado na parte da sentença que segue a estrutura da sentença-padrão (a parte é definida na precondição da sentença-padrão). Formalmente,

$$\frac{tipo\_rel(c_1, c_2), tipo\_rel(parte(p_1), c_2)}{s(parte(s)|c_2),\ p_1 \in P_s}(I-c)$$

Onde:

- p_1 é uma sentença-padrão;
- *parte(s)* é uma das partes nominal (sn), verbal (sv) ou complementar (sp) de *s*; e
- P_s é o conjunto das sentenças-padrão que determinam a estrutura sintática de *s*.

Exemplo 01:
Sejam
- c_1 = "jovem"
- precondição de c_1: éUm('jovem','pessoa')
- p_1 = "<X> <ser assassinar>"
- precondição de p1: éUm(sn(p_1), 'pessoa')
Logo, por (I-c), pode ser gerada sentença
s: <Um jovem> <ser assassinar>, a qual segue a estrutura sintática de p_1 e o conceito c_1 foi usado na parte nominal de s (sn(s)).

(2) **A regra (E_1-c)** define que, se um conceito é usado em uma sentença, então as precondições do conceito podem ser usadas para gerar precondições da sentença. A sentença $s(c_1|c_2)$ é a precondição na qual o conceito c_1 foi substituído por c_2. Formalmente,

$$\frac{tipo\_rel(c_1, c_2), s(c_1)}{('Pre', s(c_1), s(c_1|c_2))}(E_1-c)$$

A regra (E_1-c) autoriza o SIA a gerar sentenças s_j que expressam premissas das sentenças s_i do texto de entrada. A geração da premissa s_j ($s(c_1|c_2)$) depende da função sintática do conceito c_1 em s_i.

Se conceito c_1 = nucleo(sn(s_i)) (núcleo do sintagma nominal de s_i), então a premissa s_j é gerada da forma "*<reescrita(nome_relacao)> <c_2> <sv(s_i)> <sp(s_i)>*".

Exemplo 02:
Sejam
- s_1="O crime ocorreu na Rua Titan, 33"
- c_1="crime" = nucleo(sn(s_1))
- precondição de c_1: éUm('crime', 'violação da lei')
Logo, por (E_1-c), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2 = "<Um(a)> <violação da lei> <ocorreu> <na Rua Titan, 33>"

Se conceito c_1 = nucleo(sv(s_i)) (núcleo do sintagma verbal de s_i), então a premissa s_j é gerada da forma "*<sn(s_i)> <"realizou ação que"|"sofreu ação que"> < reescrita(nome_relacao)> <c_2>*".

Exemplo 03:
Sejam
- s_1=" Um jovem foi executado com vários tiros"
- c_1="executar" = nucleo(sv(s_1))
- precondição de c_1: usadoPara('executar','vingança')
Logo, por (E_1-c), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2: "<Um jovem> <sofreu ação que> <é usada para> <vingança>"

Se conceito c_1 = nucleo(sp(s_i)) (núcleo do sintagma complementar de s_i), então a premissa s_j é gerada da forma "*<sn(s_i)> <sv(s_i)> <preposicao> < reescrita(nome_relacao)> <c_2>*".

2 A formalização das regras de inferência do SIA segue o padrão de formalização das regras de inferência do sistema lógico de Dedução Natural de Prawitz (1965).

Exemplo 04:
Sejam
- s_1=" Um jovem foi executado com vários tiros"
- c_1="tiro" = nucleo(sp(s_1))
- precondição de c_1: usadoPara('tiro','ferir')
Logo, por (E_1-c), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2: "<Um jovem> <foi executado> <com> <algo usado para> <ferir>"

(3) **A regra (E_2-c)** define que, se um conceito é usado em uma sentença, as pós-condições do conceito podem ser usadas para gerar pós-condições da sentença. A sentença $s(c_1|c_2)$ é a pós-condição na qual o conceito c_1 foi substituído por c_2. Formalmente,

$$\frac{tipo\_rel(c_1, c_2), s(c_1)}{('Pos', s(c_1), s(c_1|c_2))}(E_2 - c)$$

A regra (E_2-c) autoriza o SIA a gerar sentenças s_j que expressam conclusões das sentenças s_i do texto de entrada. A geração da conclusão s_j ($s(c_1|c_2)$) depende da função sintática do conceito c_1 em s_i.

Se conceito c_1 = nucleo(sn(s_i)) (núcleo do sintagma nominal de s_i), então a conclusão s_j é gerada da forma "*<reescrita(nome_relacao)> <c_2> <sv(s_i)> <sp(s_i)>*".

Exemplo 05:
Sejam
- s_1="O crime ocorreu na Rua Titan, 33"
- c_1="crime" = nucleo(sn(s_1))
- pós-condição de c_1: efeitoDe('crime', 'sofrimento')
Logo, por (E_2-c), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2 = "<Algo que tem efeito de> <sofrimento> <ocorreu> <na Rua Titan, 33>".

Se conceito c_1 = nucleo(sv(s_i)) (núcleo do sintagma verbal de s_i), então a conclusão s_j é gerada da forma "*<sn(s_i)> <"realizou ação que"|"sofreu ação que"> < reescrita(nome_relacao)> <c_2>*".

Exemplo 06:
Sejam
- s_1=" Um jovem foi executado com vários tiros"
- c_1="executar" = nucleo(sv(s_1))
- pós-condição de c_1: efeitoDe('executar','morte')
Logo, por (E_2-c), pode ser gerada a relação ('Pos', s_1, s_2), onde s_2: "<Um jovem> <sofreu ação que> <tem efeito de> <morte>"

Se conceito c_1 = nucleo(sp(s_i)) (núcleo do sintagma complementar de s_i), então a conclusão s_j é gerada da forma "*<sn(s_i)> <sv(s_i)> <preposicao> < reescrita(nome_relacao)> <c_2>*".

Exemplo 07:
Sejam
- s_1=" Um jovem foi executado com vários tiros"
- c_1="tiro" = nucleo(sp(s_1))
- pós-condição de c_1: efeitoDe('tiro','ferir')
Logo, por (E_2-c), pode ser gerada a relação ('Pos', s_1, s_2), onde s_2: "<Um jovem> <foi executado> <com> <algo que tem efeito de> <ferir>"

A segunda forma de raciocínio gera premissas e conclusões das sentenças do texto de entrada com base no conteúdo inferencial das sentenças-padrão correspondentes. Definimos regras genéricas para premissas e conclusões de uma sentença-padrão, as quais podem ser instanciadas para cada sentença-padrão usada nas sentenças do texto de entrada. A seguir, são apresentadas a interpretação e a sintaxe das regras de premissa e de conclusão de sentenças-padrão, e como estas são usadas pelo SIA para geração de premissas e conclusões das sentenças do texto de entrada.

(4) **A regra (P-p)** define que, se uma sentença é usada conforme a estrutura de uma sentença-padrão, então as precondições da sentença-padrão podem ser usadas para gerar precondições da sentença. Formalmente,

$$\frac{tipo\_rel(parte(p_1), c_1), p_1 \in P_{si}}{('Pre', s_i, tipo\_rel(parte(s_i), c_1))}(P - p)$$

A regra (P-p) autoriza o SIA a gerar sentenças s_j que expressam premissas das sentenças s_i do texto de entrada. A geração da premissa s_j depende da parte de p_1 que é o domínio da precondição de p_1.

Se parte(p_1) = sn(p_1) (a parte nominal da sentença-padrão p_1 é o domínio da precondição), então a premissa s_j é gerada da forma "*<sn(s_i)> <reescrita(nome_relacao)> <c_1>*".

Exemplo 08:
Sejam
- s_1="Maria da Rocha foi assassinada por seu amante"
- p_1="<X> <ser assassinar> <por> <Y>"
- $p_1 \in P_{s1}$
- precondição de p_1: éUm (sn(p_1), 'pessoa')
Logo, por (P-p), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2: "<Maria da Rocha> <é um(a)> <pessoa>"

Se parte(p_1) = sp(p_1) (a parte complementar da sentença-padrão p_1 é o domínio da precondição), então a premissa s_j é gerada da forma "*<sp(s_i)> < reescrita(nome_relacao)> <c_1>*".

Exemplo 09:
Sejam
- s_1="Maria da Rocha foi assassinada por seu amante"
- p_1="<X> <ser assassinar> <por> <Y>"
- $p_1 \in P_{s1}$
- precondição de p_1: éUm(sp(p_1), 'pessoa')
Logo, por (P-p), pode ser gerada a relação ('Pre', s_1, s_2), onde s_2: "<Seu amante> <é um(a)> <pessoa>"

(5) **A regra (C-p)** define que, se uma sentença é usada conforme a estrutura de uma sentença-padrão, então as pós-condições da sentença-padrão podem ser usadas para gerar pós-condições da sentença. Formalmente,

$$\frac{tipo\_rel(parte(p_1), c_1),\ p_1 \in P_{si}}{('Pos', s_i, tipo\_rel(parte(s_i), c_1))}(C-p)$$

A regra (C-p) autoriza o SIA a gerar sentenças s_j que expressam conclusões das sentenças s_i do texto de entrada. A geração da conclusão s_j depende da parte de p_1 que é o domínio da pós-condição de p_1.

Se parte(p_1) = sn(p_1) (a parte nominal da sentença-padrão p_1 é o domínio da pós-condição), então a conclusão s_j é gerada da forma "*<sn(s_i)> < reescrita(nome_relacao)> <c_1>*".

Exemplo 10:
Sejam
- s_1="Maria da Rocha foi assassinada por seu amante"
- p_1="<X> <ser assassinar> <por> <Y>"
- $p_1 \in P_{s1}$
- pós-condição de p_1: éUm(sn(p_1), 'vítima')
Logo, por (C-p), pode ser gerada a relação ('Pos', s_1, s_2), onde s_2:
"<Maria da Rocha> <é um(a)> <vítima>"

Se parte(p_1) = sp(p_1) (a parte complementar da sentença-padrão p_1 é o domínio da pós-condição), então a conclusão s_j é gerada da forma "*<sp(s_i)> <reescrita(nome_relacao)> <c_1>*".

Exemplo 11:
Sejam
- s_1="Maria da Rocha foi assassinada por seu amante"
- p_1="<X> <ser assassinar> <por> <Y>"
- $p_1 \in P_{s1}$
- pós-condição de p_1: éUm(sp(p_1), 'assassino')
Logo, por (C-p), pode ser gerada a relação ('Pos', s_1, s_2), onde s_2: "<Seu amante> <é um(a)> <assassino>"

A terceira forma de raciocínio do SIA consiste na geração de premissas e conclusões das sentenças do texto de entrada com base na aplicação das regras expressas na Base de Regras de Raciocínio Prático do SIM. A seguir, são apresentadas a interpretação e a sintaxe de uma regra de raciocínio prático ρ_i, e como esta é usada pelo SIA para geração de premissas e conclusões das sentenças do texto.

(6) **A regra (I-r)** define que se os antecedentes de uma regra de raciocínio prático forem satisfeitos, então a conclusão da regra pode ser gerada para a sentenca do texto de entrada, que está sob análise. Os antecedentes da regra são comparados às premissas e conclusões da sentença do texto de entrada e às relações inferenciais de conceitos e sentenças-padrão, usados na sentença do texto de entrada. Formalmente, seja ρ_i uma cláusula de Horn da forma ($A_1 \wedge A_2 \wedge ... \wedge A_n \rightarrow tipo(s_i, s_j)$), então,

$$\frac{A_1, A_2, ..., A_n}{(tipo, s_i, s_j)}(I-r)$$

Exemplo 12:
Sejam
$\rho_1 = \forall x,y,z$ (((encontrar(x,y) $\wedge$ encontrarEm(y,z) $\wedge$ éUm(y,'cadáver'))$\rightarrow$ ("Pos",s,éUm(z,'local do crime'))).
s_1="<Os policiais Leandro e Vitor> <encontrar> <o corpo>"
s_2="<o corpo> <é um(a)> <cadáver>"
s_3="<o corpo> <foi encontrar> <Rua Titan,33>

Logo, por (I-r), os antecedentes de ρ_1 foram satisfeitos e pode ser gerada a seguinte conclusão de s_1: éUm('Rua Titan,33','local do crime').

Como visto, as três formas de raciocínio do SIA são responsáveis por gerar inferências endossadas pelo conteúdo que autoriza o uso dos conceitos e das conseqüências deste uso, bem como de premissas e conclusões de sentenças-padrão, as quais expressam conteúdo que não pode ser direta e eficientemente extraído dos conceitos, tomados individualmente. Todo este conteúdo prioritariamente inferencialista, tornado explícito, serve de base para responder perguntas, argumentar, refutar afirmações, extrair informações etc. Além

disso, como as bases semânticas do SIM são flexíveis para expressão de conhecimento de senso-comum e pragmático da língua natural, inferências mais interessantes sobre estes conteúdos são realizadas. Todas estas características completam o diferencial do uso do SIM em sistemas de PLN.

Outro componente particularmente importante no processo de raciocínio do SIA são os objetivos da aplicação cliente. Uma aplicação cliente é uma aplicação ou sistema de PLN que utiliza o SIA como raciocinador semântico de textos em linguagem natural. O uso de objetivos possibilita que o SIA direcione as premissas e conclusões geradas conforme necessidades de informações específicas, por exemplo, o local do crime. Portanto, somente as inferências relacionadas aos conceitos que expressam tais objetivos são potencialmente relevantes.

Os objetivos da aplicação cliente funcionam como *templates* que contém campos a serem preenchidos, com base nas inferências geradas pelo SIA. Cada *template* representa um objetivo. Por exemplo, o *template* "O local do crime é _____" representa o objetivo "Encontrar o local do crime". Cada objetivo é definido por: (i) um conceito que expressa o assunto do objetivo (por exemplo, o conceito *'crime');* (ii) uma lista de conceitos relacionados, que definem a informação requerida sobre o assunto do objetivo (por exemplo, o '*local*', a '*vítima*', o '*horário*', o '*tipo*' etc); (iii) um lista de conceitos que definem cada opção de resposta possível. Este último parâmetro de objetivos é opcional, pois algumas informações requeridas pela aplicação cliente possuem uma faixa de valores possíveis como resposta (por exemplo, o tipo de crime pode ser, alternativamente, um '*assassinato'*, um '*roubo'*, um '*furto'* etc).

Os conceitos envolvidos na definição de um objetivo são expressos na base conceitual do SIM. Todos eles são considerados para selecionar as premissas e conclusões, geradas pelo SIA, que são relevantes para responder ao objetivo. O critério de seleção é o melhor resultado da medida de relacionamento inferencial entre os conceitos das premissas/conclusões geradas e os conceitos relacionados ao objetivo.

4. Extrator de Informações para WikiCrimes

O SIA e as bases semânticas InferenceNet.Br (Pinheiro et al., 2010) são usadas em uma aplicação real: o Extrator de Informações para o sistema colaborativo WikiCrimes (Furtado et al., 2009). WikiCrimes[3] provê um ambiente colaborativo e interativo na Web para que as pessoas possam reportar e monitorar crimes ocorridos. Uma necessidade urgente do projeto WikiCrimes era fornecer a seus usuários uma ferramenta que os assistisse no registro de crimes a partir de notícias reportadas na Web e, desta forma, promovesse um estímulo à colaboração.

Esta necessidade existe para os sistemas colaborativos em geral. De um lado, tem-se a Web como uma fonte rica de informações sobre qualquer domínio, seu conteúdo é vasto e, em sua maioria, está na forma não estruturada e em linguagem natural. De outro lado, sistemas colaborativos dependem da iniciativa dos usuários para geração do conteúdo e de uma inteligência coletiva. No entanto, não é motivador deixar para os usuários a tarefa de ler os textos da Web, extrair as informações e ainda registrar manualmente no sistema. Portanto, existe a necessidade crescente de ferramentas que auxiliem a captura rápida, de forma simples, semi-automática e interativa de informações para registro em sistemas colaborativos e, além disso, que saibam manipular conteúdo em linguagem natural.

Para atender a esta necessidade, foi desenvolvido um sistema Extrator de Informações para o WikiCrimes - WikiCrimesIE – para extrair informações de crimes descritos em língua portuguesa, em jornais da Web, e gerar os registros do crime na base de dados de WikiCrimes.

O diferencial do uso do SIA como raciocinador semântico de WikiCrimesIE é sua melhor capacidade para entendimento completo de textos em linguagem natural. Algumas das informações requeridas pelo sistema WikiCrimes não estão comumente explícitas no texto, por exemplo, tipo e causa do crime, tipo de vítima e tipo de arma utilizada.

A figura 4 apresenta a interface de WikiCrimesIE, dividida em quadros, conforme segue:

A) Texto selecionado de um sítio da web, a partir do qual as informações sobre o crime relatado serão extraídas.
B) Mapa geoprocessado onde é localizado o endereço do crime.
C) Dados analíticos sobre o endereço do crime.
D) Dados do crime: data e horário da ocorrência, quantidade de criminosos e vítimas, relação do usuário com o crime e informação para polícia.
E) Dados especiais sobre o crime: tipo do crime, tipo de vítima, arma utilizada, motivos ou causas do crime.

3 www.wikicrimes.org, acessado em 18/12/2009

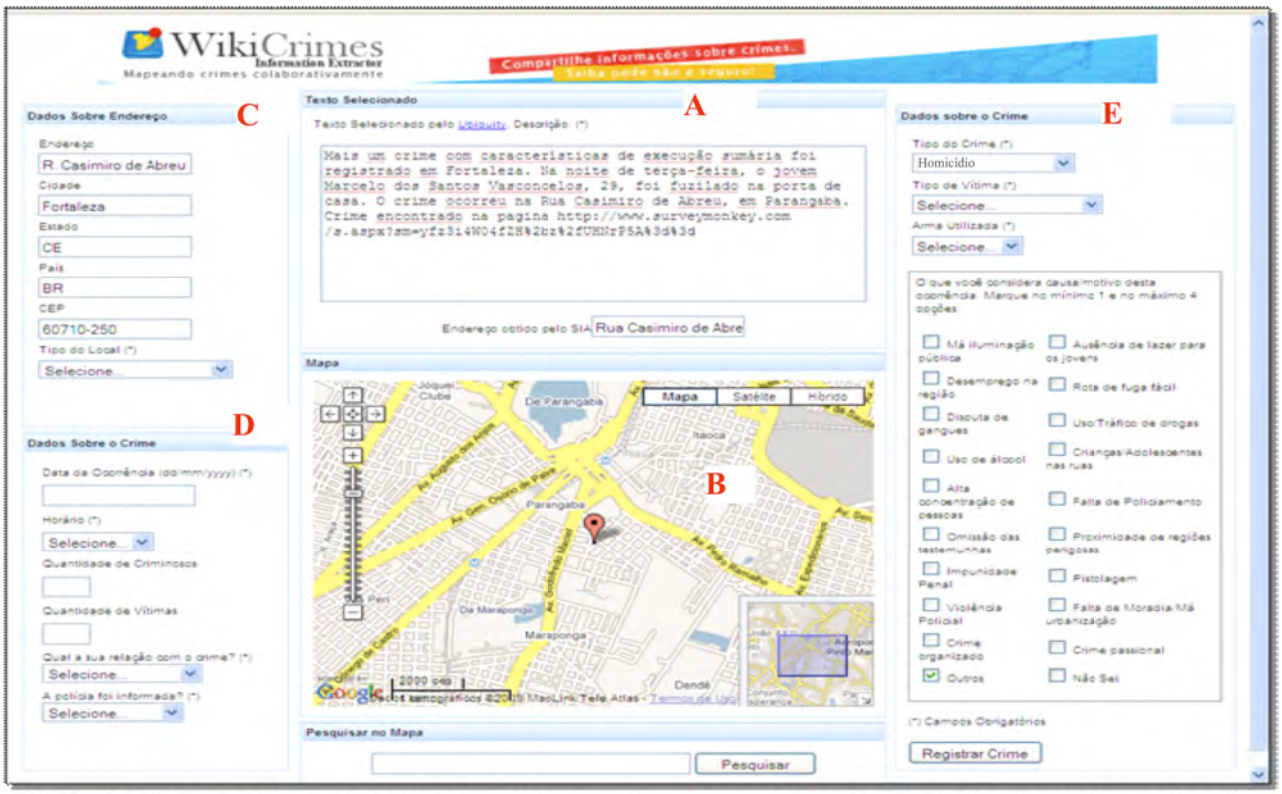


Figura 4: Interface do sistema WikiCrimesIE apresentando as informações extraídas do texto selecionado: local do crime ("Rua Casimiro de Abreu, Parangaba") e tipo do crime ("homicídio"). O endereço correspondente ao local do crime foi localizado no mapa geoprocessado.

4.1 Funcionamento de WikiCrimesIE

O processo de WikiCrimesIE para extração de informações sobre um dado crime, a partir de um texto descritivo em língua portuguesa, segue os seguintes passos:

(1) o usuário seleciona um texto de um sítio da web e executa o comando *mapcrimes*, desenvolvido na ferramenta *Ubiquity (*Nogueira et al., 2009). O Ubiquity é um plug-in do Mozilla Firefox e consiste em uma ferramenta de programação orientada ao usuário. Através de sua linguagem é possível implementar comandos que realizam a integração e *mashups* de aplicações Web. Nogueira et al. (2009) desenvolveram o comando *mapcrimes* que seleciona um texto de um sítio qualquer da Web e o envia ao sistema WikiCrimesIE;

(2) WikicrimesIE envia o texto selecionado ao analisador morfossintático PALAVRAS (Bick, 2000);

(3) WikicrimesIE instancia os objetivos da aplicação. Para cada objetivo devem ser especificados o conceito do assunto principal (por exemplo, 'crime'), conceitos relacionados à informação requerida (por exemplo, 'local', 'endereço' etc), e, no caso de existirem respostas alternativas, conceitos relacionados a cada opção de resposta (por exemplo, 'assassinato', 'roubo', 'furto');

(4) WikicrimesIE envia o texto analisado pelo parser PALAVRAS e os objetivos de extração de informação para o SIA;

(5) O SIA realiza a análise semântica das sentenças do texto e gera a rede de inferências para cada sentença, filtrando as premissas e conclusões pelos objetivos de extração. Para o caso de objetivos abertos, ou seja, objetivos sem opções de respostas predefinidas (por exemplo, local do crime), o SIA retorna ao WikicrimesIE a parte da sentença s_i (sintagma nominal, verbal ou complementar de s_i) que contém a resposta e a sentença s_j (premissa ou conclusão gerada pelo mecanismo de raciocínio do SIA) que justifica a resposta. Para o caso de objetivos com respostas predefinidas (por exemplo, tipo do crime), o SIA retorna a WikicrimesIE uma ou mais respostas selecionadas e as respectivas sentenças s_j (premissas ou conclusões geradas pelo

mecanismo de raciocínio do SIA), que justificam as respostas.

(6) WikicrimesIE interpreta as respostas dos objetivos, retornadas pelo SIA, e apresenta-as na interface (Figura 4);

(7) O usuário tem a opção de aceitar as respostas dadas pelo SIA ou alterá-las antes de registrar o crime na base de dados de Wikicrimes. Para fins de avaliação dos níveis de precisão e cobertura do SIA, é armazenado o log dos resultados retornados pelo SIA e as alterações realizadas pelos usuários.

4.1.1. Extração do Local e Tipo do Crime

A seguir será exemplificado o processo de raciocínio do SIA para extração do local do crime e do tipo de crime descrito no texto (ver quadro A da Figura 4): "*Mais um crime com características de execução sumária foi registrado em Fortaleza. Na noite de terça-feira, o jovem Marcelo dos Santos Vasconcelos, 29, foi fuzilado na porta de casa. O crime ocorreu na Rua Casimiro de Abreu, em Parangaba.*"

O sistema WikicrimesIE instancia dois objetivos.

OBJETIVO1. "*Encontrar o local do crime*":

i. conceito que expressa o assunto principal do objetivo: *'crime'*;
ii. lista de conceitos relacionados, que definem a informação requerida sobre o assunto principal: '*local*' , '*endereço*' , '*cidade*', '*bairro*';
iii. lista de respostas predefinidas: não se aplica para este objetivo.

OBJETIVO2. *"Encontrar o tipo do crime*":

i. conceito que expressa o assunto principal do objetivo: *'crime'*;
ii. lista de conceitos relacionados, que definem a informação requerida sobre o assunto principal: '*tipo*', '*espécie*';
iii. lista de respostas predefinidas[4]:
    1. **roubo** ('furto', 'violência')
    2. **tentativa de roubo** ('tentativa', 'furto', 'violência')
    3. **furto** ('furto')
    4. **tentativa de furto** ('tentativa, 'furto')
    5. **violência doméstica** ('violência', 'família', 'esposa', 'marido')
    6. **rixas ou brigas** ('luta')
    7. **homicídio** ('assassinato', 'morte')
    8. **tentativa de homicídio** ('tentativa', 'assassinato', 'morte')
    9. **latrocínio** ('roubo', 'morte', 'violência', 'furto')

O SIA executa os seguintes passos:

(1) Recebe a árvore sintática gerada pelo PALAVRAS.

(2) Decompõe as sentenças do texto em:
- s_1=*"Mais um **crime** com características de **execução sumária foi registrado** em Fortaleza."*
- s_2=*"O **jovem**... **foi fuzilado** na **noite** de **terça-feira**."*
- s_3=*"O **jovem**... **foi fuzilado** na **porta** de **casa**."*
- s_4=*"O **crime ocorreu** na Rua Casimiro de Abreu, Parangaba."*

(3) Combina sentenças-padrão com as sentenças decompostas. As respectivas sentenças-padrão são:
- p_1=*"X ser registrar em Y."*
- p_2=*"X ser fuzilar em Y."*
- p_2=*"X ser fuzilar em Y."*
- p_3=*"X ocorrer em Y."*

(4) Seleciona os conceitos possíveis da Base Conceitual que foram usados nas sentenças originais, correspondentes aos termos em negrito, destacados acima:
- *conceitos*(s_1)=*(crime, execução sumária, ser, registrar)*
- *conceitos*(s_2)=*(jovem, ser, fuzilar, noite, terça-feira)*
- *conceitos*(s_3)=*(jovem, ser, fuzilar, porta, casa)*
- *conceitos*(s_4)=*(crime, ocorrer)*

(5) Define todos os conceitos previamente selecionados, pois, neste exemplo, não há conceitos a desambiguar. Instancia, para cada sentença s_i, um grafo com os conteúdos inferenciais dos conceitos definidos (subgrafo G'_c , para cada conceito c) e outro grafo com os conteúdos inferenciais das sentenças-padrão definidas (subgrafo G'_s, para cada sentença-padrão p). No exemplo, não é possível eliminar pré e pós-condições porque não há conceitos a desambiguar. Por exemplo, para a sentença s_4, é gerado subgrafo $G'_{fuzilar}$ com aresta expressando a pós-condição *efeitoDe('fuzilar', 'morte')* e subgrafo G_{p3} com aresta expressando a precondição *ehUm(sp(p_3), 'local')*.

(6) Instancia as sentenças-padrão p_1 a p_3 com os elementos subsentenciais das respectivas

4 A lista de respostas predefinidas consiste de uma lista de opções da forma ***tipo_crime*** *('conceito$_1$', conceito$_2$, ..., conceito$_n$)*, onde *conceito$_i$* são os conceitos que definem o ***tipo_crime***.

sentenças originais s_1 *a* s_4. Por exemplo, a sentença-padrão p_3=*"X ocorrer em Y"* é instanciada com os elementos subsentenciais e respectivos conceitos da sentença s_4: *X="o crime"* e *Y="a Rua_Casimiro_de_Abreu"*. Com isso, tem-se a estrutura das sentenças originais s_1 a s_4 e seus respectivos elementos subsentenciais com conceitos associados.

(7) Gera a rede inferencial de G_N^{si} para cada sentença s_1 a s_4, aplicando as formas de raciocínio semântico sobre os subgrafos G'_c e G'_s. Para cada premissa/conclusão gerada é calculada a medida de relacionamento inferencial entre os conceitos do objetivo e o conceito relacionado na premissa e conclusão. Vejamos o detalhe das inferências geradas e o cálculo da medida de relacionamento inferencial em relação a cada objetivo:

OBJETIVO1.
A premissa de s_4 *ehUm('a Rua_Casimiro_de_Abreu', 'local')* foi gerada em G_N^{s4} pela regra de inferência (P-p) sobre o grafo G_{p3}. A medida ***θ('crime','crime')*** (c_1='*crime*', assunto principal do objetivo; e c_2='*crime*', núcleo do sintagma nominal de s'_4) e ***θ('local','local')*** (c_1='*local*', conceito que define a informação requerida sobre *crime*; e c_2='*local*', conceito relacionado na premissa de s'$_4$) apresentaram valor máximo, indicando que esta premissa responde melhor ao objetivo. Com isso, o SIA retorna a WikiCrimesIE a premissa "*a Rua_Casimiro_de_Abreu é um(a) local"* como resposta ao objetivo "*Encontrar local do crime*".

OBJETIVO2.
A conclusão de s_2 "*O jovem sofreu ação que tem efeito de morte"* foi gerada em G_N^{s2} pela regra de inferência (E_1-c) sobre o grafo $G_{fuzilar}$. A medida ***θ('crime','crime')*** (c_1='*crime*', assunto principal do objetivo; e c_2='*fuzilar*', núcleo do sintagma verbal de s'_2) e ***θ('morte','morte')*** (c_1='*morte*', conceito que define o tipo de crime = ***homicídio***('assassinato', 'morte')); e c_2='*morte*', conceito relacionado na premissa de s'$_4$) apresentaram valores maiores comparados às outras premissas/conclusões. Estes resultados indicam que esta conclusão responde melhor ao objetivo e o tipo do crime = homicídio. Com isso, o SIA retorna a WikicrimesIE a resposta selecionada *tipo_crime* = **homicidio** e a conclusão "*O jovem sofreu ação que tem efeito de morte"* como sentença inferida que justifica a resposta.

Nos quadros A e B da Figura 4, tem-se, respectivamente, a resposta do OBJETIVO1 identificada (**Rua Casimiro de Abreu**) e sua localização no mapa geoprocessado. No quadro E da Figura 4, tem-se a caixa de seleção do tipo do crime = **homicídio**, conforme OBJETIVO2.

4.2 Avaliação dos Resultados do SIM

Ao avaliarmos os resultados do sistema WikiCrimesIE na tarefa de extração de informações a partir de textos em linguagem natural, o que estamos avaliando, de fato, é o desempenho do SIM como modelo para expressão e raciocínio semântico em sistemas de entendimento de linguagem natural.

A metodologia de avaliação seguiu os passos delineados na sequência.

(1) Foi elaborada uma Coleção Dourada (CD) com 100 textos jornalísticos, publicados nas páginas policiais de jornais brasileiros, na Internet. Estes textos foram coletados por pessoas que não participavam do projeto e de forma aleatória.

(2) Os textos da CD foram lidos por duas pessoas adultas, proficientes em língua portuguesa, e foi solicitado a elas que respondessem às duas perguntas abaixo e registrassem a resposta, para cada texto. Antes, as pessoas receberam orientações sobre os tipos de crimes que deveriam ser considerados e acerca das respostas a serem dadas. Por exemplo, que a resposta para a pergunta sobre o local do crime deveria ser descritiva, contendo o maior número de informações sobre a localização exata do crime (endereço, bairro, ponto de referência, cidade, localidade etc).
   - Qual o local do crime?
   - Qual o tipo de crime?

(3) As respostas das duas pessoas participantes foram comparadas e, em caso de divergência, uma terceira pessoa foi consultada sobre qual das respostas era a correta. Ao final, apenas uma resposta de cada pergunta foi anotada para cada texto da CD.

(4) Os textos da CD foram submetidos ao sistema WikiCrimesIE e as informações extraídas pelo sistema sobre o local e tipo de crime foram registradas, para cada texto.

(5) As respostas dos especialistas humanos e do WikiCrimesIE foram comparadas e analisadas manualmente. Para uma avaliação quantitativa do SIM, foi atribuído, para cada informação extraída pelo WikiCrimesIE, de cada texto, um valor numérico que correspondia ao resultado da comparação, conforme Tabela 1.

Valor Atribuído	Resultado da comparação
1	***Informação CORRETA***
2	***Informação PARCIALMENTE CORRETA.*** ***Obs.: Este valor é atribuído quando o sistema não identificou o endereço completo do local do crime.***
3	***Informação INCORRETA*** ***Obs.: Este valor é atribuído quando o sistema identificou a sentença do texto que justificava a informação correta, porém não inferiu a informação correta.***
4	***Informação INCORRETA***
5	***Informação NÃO EXTRAÍDA***
6	***Informação NÃO EXTRAÍDA por erro de processamento***

Tabela 1: Valores atribuídos na comparação das informações extraídas pelo WikiCrimesIE em relação às respostas dadas pelos especialistas humanos.

As medidas usadas na avaliação do SIM, para cada informação extraída (local e tipo de crime), foram:

- **precisão**, que mede o quanto da informação extraída (casos em que A=1,2,3 ou 4) foi corretamente extraída (A=1 ou 2). Esta medida indica o quanto o sistema WikiCrimesIE é confiável em extrair a informação;
- **cobertura**, que mede o quanto da informação que deveria ter sido extraída (casos em que A=1,2,3,4 ou 5) foi corretamente extraída (A=1 ou 2). Esta medida indica o quanto o sistema WikiCrimesIE é abrangente em extrair a informação;
- **medida-F**, que é a média harmônica das medidas de precisão e cobertura;
- **percentual de erros de processamento**, que mede o percentual de textos não analisados por erro de processamento (A=6), ocasionado por problemas relacionados à estrutura sintática do texto: sentenças mal formadas (sem sujeito), períodos complexos etc; e
- **percentual de erros do analisador morfossintático**, que mede o percentual de erros de análise morfossintática do PALAVRAS. Esta medida indica o quanto a dependência da análise sintática prejudica os resultados do sistema.

A Tabela 2 apresenta os resultados das medidas de avaliação do WikiCrimesIE na tarefa de extração do local do crime, do tipo de crime e de ambos.

	Local do crime	Tipo do crime	Geral
Precisão	*87.00%*	*72.00%*	*79.00%*
Cobertura	*71.00%*	*68.00%*	*69.00%*
Medida F	*78.00%*	*70.00%*	*74.00%*
%Erros processamento	*3.00%*	*8.00%*	*8.00%*
%Erros Análise Morfossintática	*2.00%*	*7.00%*	*7.00%*

Tabela 2: Resultados do WikiCrimesIE na extração do "Local do Crime" e "Tipo do Crime".

5. Trabalhos Relacionados e Análise Comparativa

Nesta seção, serão citados alguns trabalhos relacionados a tarefas de PLN que envolvem entendimento de linguagem natural. Uma análise comparativa com a tarefa realizada pelo SIM no sistema WikiCrimesIE não é trivial, devido a diferença entre a natureza das informações requeridas por esse sistema e o foco dos sistemas atuais de Extração de Informação (EI). Segundo Grishman (2003), as pesquisas em EI evoluem em duas linhas: extração de nomes (*Named Entity Recognition* – NER) e extração de relações entre entidades participantes de eventos.

Na tarefa de NER, os sistemas da Priberam (Amaral et al., 2008) e REMBRANDT (Cardoso, 2008) apresentam um algoritmo para reconhecimento de entidades mencionadas (REM) para língua portuguesa. Tais sistemas foram os que apresentaram os melhores resultados em termos de cobertura e Medida-F para a tarefa de REM do Segundo HAREM (Mota e Santos, 2008). Respectivamente, tem-se os resultados de 51,46%

(cobertura) e 57,11% (Medida-F) do sistema da Priberam, e 50,36% (cobertura) e 56,74% (Medida-F) do sistema REMBRANDT. Para a língua inglesa, o melhor resultado foi registrado no evento MUC-7 (1997) com 87% de Medida-F. É importante salientar que a tarefa executada por estes sistemas restringe-se à identificação de entidades mencionadas nos textos e à classificação destas em categorias/tipos e subtipos semânticos predefinidos.

Na tarefa de extração de eventos, tem-se o melhor sistema avaliado na tarefa 4 do evento SemEval-2007 com 72,40% de Medida-F, para língua inglesa (Girju, 2007). Para língua portuguesa, tem-se o melhor sistema avaliado na tarefa de Reconhecimento de RELações entre Entidades Mencionadas (ReRelEM) do Segundo HAREM (Freitas et al, 2008) – o sistema REMBRANDT, com 45,02% de Medida-F.

Em uma análise quantitativa, WikicrimesIE, com medida-F = 74% (resultado geral da Tabela 2), apresentou melhor resultado dentre todos os sistemas mais bem avaliados para língua portuguesa. Para língua inglesa, ficou apenas abaixo do melhor sistema na tarefa de NER da MUC-7. É importante salientar que esta comparação é ainda injusta nos seguintes sentidos:

- nestes eventos de avaliação, requerem-se informações explícitas no texto. Argumentamos que os sistemas participantes destes eventos não foram avaliados na extração ou anotação de informações implícitas no texto. O tipo de crime, por exemplo, na maioria das vezes, não é mencionado. Na CD desta avaliação, havia 72% de textos nos quais o tipo de crime não era mencionado, requerendo o mínimo de raciocínio semântico para inferir o tipo de crime. WikiCrimesIE conseguiu extrair corretamente 69% dos tipos de crime, nestes casos;
- o raciocínio do SIA não se baseia em técnicas de aprendizagem de máquina ou regras gramaticais, como é o caso dos sistemas de EI aqui comparados.

Em uma análise qualitativa, foram estudados todos os casos de imprecisão do SIA na extração do local ou tipo de crime e identificados os principais problemas:

- 71% dos casos de imprecisão na extração do local do crime decorreram do mesmo estar de forma indireta em outras sentenças que relatam ações do criminoso ou da vítima;
- 12% dos casos de imprecisão na extração do local do crime decorreram de o SIM não realizar resolução de correferências;
- 50% dos casos de imprecisão na extração do tipo do crime tiveram origem na falta de conceitos na Base Conceitual do SIM; e
- 39% dos casos de imprecisão na extração do tipo do crime decorreram de problemas nas heurísticas de comparação de conceitos relacionados, implementadas pelo WikiCrimesIE

Esta análise evidenciou pontos de melhorias para o SIM: número de níveis da rede inferencial de conceitos a ser considerado pelo SIA; relações n-árias entre conceitos; relações com teor negativo; integração de uma solução de resolução de referência (anáfora pronominal, anáfora conceitual etc); pesquisa de expressões linguísticas com múltiplas palavras; definição de padrões gramaticais para conceitos da Base Conceitual.

Difícil encontrar sistemas que se propõem a realizar inferências de natureza complexa com base em textos. Textual Inference Logic (TIL) (de Paiva et al., 2007) (Bobrow et al., 2005) fornece mecanismos de raciocínio, baseados em um léxico unificado WordNet/VerbNet (Crouch e King, 2005) e em lógica descritiva. Por estes mecanismos, TIL consegue responder se uma sentença pode ser deduzida de outra ou é uma contradição de outra. Por exemplo, se a sentença "*A person arrived in the city*" é concluída de "*Ed arrived in the city*".

As inferências realizadas pelo SIA são materiais (autorizadas pelo conteúdo inferencial de conceitos e sentenças-padrão). Esta característica associada ao conteúdo semântico expresso nas bases do SIM possibilita realizar inferências para identificar/classificar o local específico do crime relatado no texto e não apenas identificar/classificar quaisquer endereços e locais mencionados no texto. Este é o caso dos sistemas de REM ou NER em geral, e de abordagens como a de Borges et al. (2007) para descoberta de localizações geográficas, a qual define seis padrões sintáticos de endereçamento. Esta abordagem apresentou uma precisão de 99,60% em reconhecer endereços explícitos no texto. O predomínio de abordagens sintáticas para a tarefa de EI evidencia a pressuposição de uma equivalência entre sintaxe e semântica das linguagens naturais.

Para a tarefa de extração do tipo de crime, nenhuma abordagem dos trabalhos relacionados se aplica a extração de informações desta natureza, principalmente por se tratar de informação comumente implícita em textos. O SIM ao explicitar o conteúdo inferencial dos conceitos, o

qual expressa as situações de uso dos mesmos, permite ao sistema uma base para explicações, argumentações, refutações, as quais permitem inferir conhecimento implícito.

6. Conclusão

Neste artigo foi descrito um analisador semântico de sentenças – o SIA, baseado nas teorias semânticas inferencialistas. O diferencial do SIA está em seu processo de raciocínio material e holístico sobre conceitos e sentenças e o uso de conhecimento inferencialista. O SIA implementa um processo sistemático para gerar as premissas e conclusões de sentenças, provendo a base para boas argumentações, respostas e explicações. A aplicação do SIA em um sistema de extração de informações sobre crimes – WikiCrimesIE - está sendo o seu cenário de avaliação. Os resultados obtidos até aqui foram motivadores, principalmente quando analisados os resultados da extração do tipo do crime. A extração de informações desta natureza exigem uma melhor capacidade para entendimento de textos em linguagem natural por parte de sistemas de PLN, principalmente, por não estarem, em sua maioria, explícitas no texto. Por exemplo, em 72% dos textos avaliados, as informações sobre o tipo do crime, causa do crime, tipo de vítima e tipo de arma utilizada não estavam explícitas, sendo necessário uma compreensão das notícias para que elas pudessem ser extraídas.

Como trabalhos em andamento, estamos otimizando a implementação do SIA, incluindo uma solução para resolução de anáforas e estendendo os objetivos de extração do WikiCrimesIE (causa do crime, tipo de vítima e tipo de arma utilizada). Um trabalho futuro será a divulgação do portal InferenceNet.org para que as bases semânticas do SIM possam ser usadas pela comunidade de PLN e, com isso, possibilitar a evolução do conhecimento inferencialista expresso. Outros trabalhos futuros incluem a investigação de: novas regras de inferência que combinem de forma diferente o conteúdo inferencial de conceitos e sentenças; técnicas de aprendizado automático e/ou semiautomático de conteúdo inferencialista; mecanismos de inferência com base no conteúdo inferencial de duas ou mais sentenças do texto, os quais gerem uma rede inferencial do texto; complexidade do algoritmo SIA.

Referências

Amaral, C. et al. 2008. Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM. Em Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca.

Bick, E. The Parsing System ”Palavras”. 2000. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press.

Bobrow, D.G. et al. 2005. A basic logic for textual inference. In: Proceedings of the AAAI Workshop on Inference for Textual Question Answering, Pittsburg, PA.

Borges, K. Laender, A., Medeiros, C. e Clodoveu, D.Jr. 2007. Discovering geographic locations in web pages using urban addresses. Proceedings of the 4th ACM workshop on Geographical Information Retrieval (GIR’07), p.31-36, Lisboa, Portugal.

Brandom, R.1994. Making it Explicit. Cambridge, MA, Harvard University Press.

Brandom, R.B. 2000. Articulating Reasons. In: An Introduction to Inferentialism. Harvard University Press, Cambridge.

Budanitsky, A e Hirst, G. 2001. Semantic distance in Wordnet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources, 2nd meeting of the NAACL, Pittsburgh, PA.

Cardoso, N. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Em Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca.

Crouch, R; King, T.H. 2005. Unifying lexical resources. In: Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarbruecken, Germany.

De Paiva, V. et al. 2007. Textual Inference Logic: Take Two. In: Proceedings of the Workshop on Contexts and Ontologies, Representation and Reasoning, CONTEXT 2007, 27-36.

Dummett, M. 1973. Frege’s Philosophy of Language. Harvard University Press.

Dummett, M. 1978. Truth and Other Enigmas. Duckworth, London.

Eco, U. 2001. A Busca da língua perfeita na cultura européia. EDUSC, São Paulo.

Freitas, C. et al. 2008. ReRelEM - Reconhecimento de Relações entre Entidades Mencionadas. Segundo HAREM: proposta de nova pista. In: Cristina Mota & Diana Santos (eds.). Desafios na avaliação conjunta do reconhecimento de

entidades mencionadas: O Segundo HAREM, 309-317, Linguateca.

Furtado, V et al. 2009. Collective intelligence in law enforcement – The WikiCrimes system. Information Sciences, In Press, Corrected Proof, Available online, August 2009. doi:10.1016/j.ins.2009.08.004.

Gentzen, G. 1935. Untersuchungen über das logische Schliessen. Mathematische Zeitschrift, 39, pp.176-210, pp. 405-431, 1935. Translated as 'Investigations into Logical Deduction',and printed in M. Szabo The Collected Papers of Gerhard Gentzen,Amsterdam: North-Holland,1969, 68–131.

Girju, R. 2007. SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).

Grishman, R. Information Extraction. 2003. In: Mitkov. R. (ed). Oxford Handbook of Computational Linguistics, Oxford University Press, 545-559.

Mota, C. e Santos, D. (eds.) 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca. (ISBN: 978-989-20-1656-6)

Nogueira, D., Pinheiro, V., Furtado, V., Pequeno, T. 2009. Desenvolvimento de Sistemas de Extração de Informações para Ambientes Colaborativos na Web. In proceedings of the II International Workshop on Web and Text Intelligence (WTI – 2009), co-located with STIL 2009.

Pinheiro,V., Pequeno, T., Furtado, V., Assunção, T. e Freitas, E. 2008. SIM: Um Modelo Semântico-Inferencialista para Sistemas de Linguagem Natural. VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2008), WebMedia, Brasil.

Pinheiro, V., Pequeno, T., Furtado, V., Nogueira, D. 2009. Information Extraction from Text Based on Semantic Inferentialism. T. Andreasen et al. (Eds.): FQAS 2009, Springer Berlin / Heidelberg, LNAI 5822, pp. 333–344.

Pinheiro, V., Pequeno, T., Furtado, V., Franco, W. 2010. InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language. International Conference on Computational Processing of Portuguese Language (PROPOR) (to appear).

Prawitz, D. 1965. Natural Deduction: A Proof Theoretical Study. Stockholm:Almqvist & Wiksell.

Sellars, W. 1980. Inference and meaning (1950). Reprinted in Pure Pragmatics and Possible Worlds. Ed. J.Sicha. Reseda, California. Ridgeview Publishing Co.

Vieira, R. e De Lima, V.L.S. 2001. Linguística Computacional: Princípios e Aplicações. Anais do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial. v.3. p. 47-86.

Wittgenstein, L. 1953. Philosophical Investigations. Tradução G.E.M.Anscombe, Oxford:Basil Blackwell.

Chamada de Artigos

A revista Linguamática pretende colmatar uma lacuna na comunidade de processamento de linguagem natural para as línguas ibéricas. Deste modo, serão publicados artigos que visem o processamento de alguma destas línguas.

A Linguamática é uma revista completamente aberta. Os artigos serão publicados de forma electrónica e disponibilizados abertamente para toda a comunidade científica sob licença *Creative Commons*.

Tópicos de interesse:

- Morfologia, sintaxe e semântica computacional
- Tradução automática e ferramentas de auxílio à tradução
- Terminologia e lexicografia computacional
- Síntese e reconhecimento de fala
- Recolha de informação
- Resposta automática a perguntas
- Linguística com corpora
- Bibliotecas digitais
- Avaliação de sistemas de processamento de linguagem natural
- Ferramentas e recursos públicos ou partilháveis
- Serviços linguísticos na rede
- Ontologias e representação do conhecimento
- Métodos estatísticos aplicados à língua
- Ferramentas de apoio ao ensino das línguas

Os artigos devem ser enviados em PDF através do sistema electrónico da revista. Embora o número de páginas dos artigos seja flexível sugere-se que não excedam 20 páginas. Os artigos devem ser devidamente identificados. Do mesmo modo, os comentários dos membros do comité científico serão devidamente assinados.

Em relação à língua usada para a escrita do artigo, sugere-se o uso de português, galego, castelhano ou catalão.

Os artigos devem seguir o formato gráfico da revista. Existem modelos LaTeX, Microsoft Word e OpenOffice.org na página da Linguamática.

Datas Importantes

- Envio de artigos até: 15 de Outubro de 2010
- Resultados da selecção até: 15 de Novembro de 2010
- Versão final até: 31 de Novembro de 2010
- Publicação da revista: Dezembro de 2010

Qualquer questão deve ser endereçada a: `editores@linguamatica.com`

Petición de Artigos

A revista Linguamática pretende cubrir unha lagoa na comunidade de procesamento de linguaxe natural para as linguas ibéricas. Deste xeito, han ser publicados artigos que traten o procesamento de calquera destas linguas.

Linguamática é unha revista completamente aberta. Os artigos publicaranse de forma electrónica e estarán ao libre dispor de toda a comunidade científica con licenza *Creative Commons*.

Temas de interese:

- Morfoloxía, sintaxe e semántica computacional
- Tradución automática e ferramentas de axuda á tradución
- Terminoloxía e lexicografía computacional
- Síntese e recoñecemento de fala
- Extracción de información
- Resposta automática a preguntas
- Lingüística de corpus
- Bibliotecas dixitais
- Avaliación de sistemas de procesamento de linguaxe natural
- Ferramentas e recursos públicos ou cooperativos
- Servizos lingüísticos na rede
- Ontoloxías e representación do coñecemento
- Métodos estatísticos aplicados á lingua
- Ferramentas de apoio ao ensino das linguas

Os artigos deben de enviarse en PDF mediante o sistema electrónico da revista. Aínda que o número de páxinas dos artigos sexa flexíbel suxírese que non excedan as 20 páxinas. Os artigos teñen que identificarse debidamente. Do mesmo modo, os comentarios dos membros do comité científico serán debidamente asinados.

En relación á lingua usada para a escrita do artigo, suxírese o uso de portugués, galego, castelán ou catalán.

Os artigos teñen que seguir o formato gráfico da revista. Existen modelos LaTeX, Microsoft Word e OpenOffice.org na páxina de Linguamática.

Datas Importantes

- Envío de artigos até: 15 de outubro de 2010
- Resultados da selección: 15 de novembro de 2010
- Versión final: 31 de novembro de 2010
- Publicación da revista: 15 de decembro de 2010

Para calquera cuestión, pode dirixirse a: `editores@linguamatica.com`

Petición de Artículos

La revista Linguamática pretende cubrir una laguna en la comunidad de procesamiento del lenguaje natural para las lenguas ibéricas. Con este fin, se publicarán artículos que traten el procesamiento de cualquiera de estas lenguas.

Linguamática es una revista completamente abierta. Los artículos se publicarán de forma electrónica y se pondrán a libre disposición de toda la comunidad científica con licencia *Creative Commons*.

Temas de interés:

- Morfología, sintaxis y semántica computacional
- Traducción automática y herramientas de ayuda a la traducción
- Terminología y lexicografía computacional
- Síntesis y reconocimiento del habla
- Extracción de información
- Respuesta automática a preguntas
- Lingüística de corpus
- Bibliotecas digitales
- Evaluación de sistemas de procesamiento del linguage natural
- Herramientas y recursos públicos o cooperativos
- Servicios lingüísticos en la red
- Ontologías y representación del conocimiento
- Métodos estadísticos aplicados a la lengua
- Herramientas de apoyo para la enseñanza de lenguas

Los artículos tienen que enviarse en PDF mediante el sistema electrónico de la revista. Aunque el número de páginas de los artículos sea flexible, se sugiere que no excedan las 20 páginas. Los artículos tienen que identificarse debidamente. Del mismo modo, los comentarios de los miembros del comité científico serán debidamente firmados.

En relación a la lengua usada para la escritura del artículo, se sugiere el uso del portugués, gallego, castellano o catalán.

Los artículos tienen que seguir el formato gráfico de la revista. Existen modelos LaTeX, Microsoft Word y OpenOffice.org en la página de Linguamática.

Fechas Importantes

- Envío de artículos hasta: 15 de octubre de 2010
- Resultados de la selección: 15 de noviembre de 2010
- Versión final: 31 de noviembre de 2010
- Publicación de la revista: diciembre de 2010

Para cualquier cuestión, puede dirigirse a: `editores@linguamatica.com`

Petició d'articles

La revista Linguamática pretén cobrir una llacuna en la comunitat del processament de llenguatge natural per a les llengües ibèriques. Així, es publicaran articles que tractin el processament de qualsevol d'aquestes llengües.

Linguamática és una revista completament oberta. Els articles es publicaran de forma electrònica i es distribuiran lliurement per a tota la comunitat científica amb llicència *Creative Commons*.

Temes d'interès:

- Morfologia, sintaxi i semàntica computacional
- Traducció automàtica i eines d'ajuda a la traducció
- Terminologia i lexicografia computacional
- Síntesi i reconeixement de parla
- Extracció d'informació
- Resposta automàtica a preguntes
- Lingüística de corpus
- Biblioteques digitals
- Evaluació de sistemes de processament del llenguatge natural
- Eines i recursos lingüístics públics o cooperatius
- Serveis lingüístics en xarxa
- Ontologies i representació del coneixement
- Mètodes estadístics aplicats a la llengua
- Eines d'ajut per a l'ensenyament de llengües

Els articles s'han d'enviar en PDF mitjançant el sistema electrònic de la revista. Tot i que el nombre de pàgines dels articles sigui flexible es suggereix que no ultrapassin les 20 pàgines. Els articles s'han d'identificar degudament. Igualement, els comentaris dels membres del comitè científic seràn degudament signats.

En relació a la llengua usada per l'escriptura de l'article, es suggereix l'ús del portuguès, gallec, castellà o català.

Els articles han de seguir el format gràfic de la revista. Es poden trobar models LaTeX, Microsoft Word i OpenOffice.org a la pàgina de Linguamática.

Dades Importants

- Enviament d'articles fins a: 15 d'octubre de 2010
- Resultats de la selecció: 15 de novembre de 2010
- Versió final: 31 de novembre de 2010
- Publicació de la revista: desembre de 2010

Per a qualsevol qüestió, pot adreçar-se a: `editores@linguamatica.com`

http://www.linguamatica.com/