Volume 6, Número 1- Julho 2014

# linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 6, Número 1 – Julho 2014

# LinguaMÁTICA

ISSN: 1647–0818

# Conteúdo

# Editorial

*Este ano de 2014 é iniciado com uma edição especial. Assim como em 2010, publicamos um conjunto de artigos alargados, seleccionados dos artigos aceites no nono Simpósio Brasileiro de Tecnologia da Informação e Linguagem Humana (STIL).*

*Portanto, esta edição abre com três artigos seleccionados da edição de 2013 do STIL que abordam diferentes aspectos da linguagem natural: em primeiro lugar a geração de texto de forma a descrever o ambiente virtual em que um utilizador se encontra; posteriormente será discutida a análise de coerência no uso de entidades em textos científicos; finalmente será apresentado um sistema para o reconhecimento de entidades mencionadas, ou nomeadas.*

*Para completar o volume, incluímos neste número especial dois artigos de investigação que não fazem parte do STIL: primeiro um trabalho relacionado com a previsão, usando diferentes tipos de regressão e dados extraídos de texto escrito em linguagem natural; e em seguida, um trabalho sobre a anotação e disponibilização de corpos paralelos criados a partir de trabalhos de tradução de alunos, de modo a serem úteis para, entre outras coisas, o próprio ensino de línguas.*

*Xavier Gómez Guinovart*
*José João Almeida*
*Alberto Simões*

# Prólogo

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

O Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL) é o principal evento nacional apoiado e organizado pela Comissão Especial de Processamento de Linguagem Natural (CE-PLN[1]) da Sociedade Brasileira de Computação (SBC[2]).

O evento foi concebido em 2003 com o nome TIL (Workshop de Tecnologia da Informação e da Linguagem Humana), tendo o propósito de estimular o desenvolvimento de uma área genuinamente multidisciplinar, procurando atrair pesquisadores, membros da comunidade acadêmica e da indústria que atuam nas áreas de Ciência da Computação, Linguística e Ciência da Informação, entre outras, pois o processamento computacional das línguas humanas requer a coordenação de esforços de diversas comunidades, que contribuem com conhecimentos específicos e metodologias de pesquisa próprias no desenvolvimento de técnicas e sistemas. O principal objetivo do STIL é fornecer o fórum adequado para a integração dessas comunidades.

Em 2003, foi realizado na USP-São Carlos/SP; em 2004 e 2005 foi hospedado pelo Congresso da SBC em Salvador/BA e São Leopoldo/RS, respectivamente; em 2006, o evento foi hospedado pela International Joint Conference IBERAMIA/SBIA/SBRN, em Ribeirão Preto/SP, que consistiu no maior evento de Inteligência Artificial já realizado no Brasil. A 5a. edição do evento foi hospedada novamente pelo XXVII Congresso da SBC no Rio de Janeiro/RJ, no Instituto Militar de Engenharia-IME. A 6a. edição do evento foi realizado em 2008 juntamente com o Webmedia, em Vila Velha-ES, e foi a última com o nome de TIL. A 7a. edição, já com o nome de STIL, foi realizada na USP-São Carlos/SP em 2009. A 8a edição ocorreu em Cuiabá/MT em 2011, na UFMT. A 9a. edição foi realizada em Fortaleza/CE em 2013, juntamente com o 2o. Brazilian Conference on Intelligent Systems (BRACIS-13) e o X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). Atualmente, o STIL é o maior evento da comunidade no Brasil, podendo ser considerado o único no país totalmente dedicado ao tema.

A edição de 2013 recebeu 65 submissões do Brasil, Grã-Bretanha, Peru, Alemanha, Estados Unidos e Portugal. Cada artigo foi revisado por, pelo menos, 3 membros do Comitê de Programa que contava com 64 membros de 7 países e 34 instituições de ensino superior.

[1] http://www.nilc.icmc.usp.br/cepln/
[2] http://www.sbc.org.br/

15 artigos foram selecionados para apresentação oral (taxa de aceitação de 23%) e 17 para apresentação na forma de pôster. Os melhores trabalhos foram convidados para apresentar versões estendidas a serem publicadas em duas revistas importantes da área da Computação e do Processamento Computacional das Línguas Humanas, respectivamente, o Journal of the Brazilian Computer Society (JBCS) e a Linguamática – Revista para o Processamento Automático das Línguas Ibéricas.

Nesta edição especial da Linguamática em homenagem ao STIL 2013, trazemos três versões estendidas dos seguintes artigos publicados na edição de 2013 do STIL:

- Geração de expressões de referência em ambientes virtuais interativos, por *Diego Silva e Ivandré Paraboni*;
- Usando grades de entidades na análise automática de coerência local em textos científicos, por *Alison Polpeta Freitas e Valéria Feltrim*;
- NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields, por *Daniela do Amaral e Renata Vieira*;

Aproveitamos a oportunidade para agradecer aos autores, membros do Comitê de Programa, Palestrantes Convidados, SBC e os comitês locais e gerais do STIL 2013.

Desejamos a todos uma leitura proveitosa destes trabalhos!

*Sandra Maria Aluísio*
*Valéria Delisandra Feltrim*

# 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**,
University of Surrey

**Anselmo Peñas**,
Univers. Nac. de Educación a Distancia

**Antón Santamarina**,
Universidade de Santiago de Compostela

**Antoni Oliver González**,
Universitat Oberta de Catalunya,

**Antonio Moreno Sandoval**,
Universidad Autónoma de Madrid

**António Teixeira**,
Universidade de Aveiro

**Arantza Díaz de Ilarraza**,
Euskal Herriko Unibertsitatea

**Arkaitz Zubiaga**,
Dublin Institute of Technology

**Belinda Maia**,
Universidade do Porto

**Carmen García Mateo**,
Universidade de Vigo

**Diana Santos**,
Linguateca/Universidade de Oslo

**Ferran Pla**,
Universitat Politècnica de València

**Gael Harry Dias**,
Universidade Beira Interior

**Gerardo Sierra**,
Univers. Nacional Autónoma de México

**German Rigau**,
Euskal Herriko Unibertsitatea

**Helena de Medeiros Caseli**,
Universidade Federal de São Carlos

**Horacio Saggion**,
University of Sheffield

**Hugo Gonçalo Oliveira**,
Universidade de Coimbra

**Iñaki Alegria**,
Euskal Herriko Unibertsitatea

**Irene Castellón Masalles**,
Universitat de Barcelona

**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

**Juan-Manuel Torres-Moreno**,
Lab. Informatique d'Avignon - UAPV

**Kepa Sarasola**,
Euskal Herriko Unibertsitatea

**Lluís Padró**,
Universitat Politècnica de Catalunya

**María Inés Torres**,
Euskal Herriko Unibertsitatea

**Maria das Graças Volpe Nunes**,
Universidade de São Paulo

**Mercè Lorente Casafont**,
Universitat Pompeu Fabra

**Mikel Forcada**,
Universitat d'Alacant

**Pablo Gamallo Otero**,
Universidade de Santiago de Compostela

**Patrícia Cunha França**,
Universidade do Minho

**Rui Pedro Marques**,
Universidade de Lisboa

**Salvador Climent Roca**,
Universitat Oberta de Catalunya

**Susana Afonso Cavadas**,
University of Sheffield

**Tony Berber Sardinha**,
Pontifícia Univ. Católica de São Paulo

**Xavier Gómez Guinovart**,
Universidade de Vigo

# Simpósio de Tecnologia da Informação e Linguagem Humana

# Geração de Expressões de Referência em Ambientes Virtuais Interativos*

**Referring Expression Generation in Interactive Virtual Worlds**

Diego dos Santos Silva
School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)
diego.silva@usp.br

Ivandré Paraboni
School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)
ivandre@usp.br

**Resumo**

Sistemas de geração de instruções em mundos virtuais interativos 3D possuem uma ampla gama de aplicações em áreas como educação, desenvolvimento de jogos etc. Neste artigo discutimos o problema computacional da geração de expressões de referência em ambientes deste tipo, enfocando a questão do uso de relações espaciais para descrever objetos do domínio.

**Palavras chave**

Geração de Língua Natural, Expressões de Referência

**Abstract**

Instruction-giving systems for virtual interactive 3D worlds have a wide range of applications in education, games and others. This paper discusses the computational task of referring expression generation for systems of this kind, focusing on the use of spatial relations to describe domain objects.

**Keywords**

Natural Language Generation, Referring Expressions

## 1 Introdução

A geração de expressões de referência (GER) é um dos componentes fundamentais de aplicações de geração de língua natural (GLN) a partir de dados de entrada não linguísticos. Algoritmos de GER tratam da tarefa de seleção do conteúdo semântico a ser realizado, por exemplo, na forma de descrições definidas[1] como a seguir:

a *O velho*

b *O homem de óculos, à esquerda*

c *O segundo homem, de casaco preto e ao lado do rapaz que está fumando*

A escolha de um determinado conjunto de propriedades semânticas para compor uma expressão de referência como nos exemplos a-c acima traz uma série de consequências para o leitor (ou ouvinte), tanto no que diz respeito à sua interpretação linguística como à sua resolução (aqui entendida como sendo a identificação do objeto-alvo da referência). Por exemplo, uma descrição excessivamente breve como (a) pode, em determinado contexto (e.g., uma multidão) dificultar a identificação do referente. Por outro lado, uma descrição muito extensa como (c) pode apresentar maior dificuldade de interpretação.

De forma mais ampla, pode-se dizer que a tarefa computacional de GER consiste em produzir descrições que sejam psicologicamente plausíveis, ou seja, o mais próximo possível das descrições que seriam produzidos por sujeitos humanos em condições semelhantes. Isso inclui, por exemplo, a necessidade de evitar a produção de descrições ambíguas, contendo falsas implicações lógicas, excessivamente breves ou extensas, ou que façam uso de propriedades incomuns para aquele tipo de contexto, dentre vários outros objetivos geralmente conflitantes.

GER é uma ativa linha de pesquisa em GLN, tendo sido inclusive objeto de uma série de competições (ou *shared tasks*) recentes (Belz e Gatt, 2007; Gatt, Belz e Kow, 2008; Gatt, Belz e Kow, 2009). Abordagens existentes, entretanto, tendem a considerar principalmente domínios simplificados e/ou bidimensionais. O problema de referência em domínios físicos mais realistas (e.g., com grande complexidade estrutural, tridimensionalidade etc.) permanece pouco explorado na pesquisa da área, possivelmente em virtude da própria dificuldade em criar bons modelos computacionais deste tipo.

*Este trabalho conta com apoio FAPESP.

[1]Em contraste à questão da *interpretação* (e.g., anafórica) de expressões existentes, (Paraboni, 1997; Cuevas e Paraboni, 2008).

Mais recentemente, entretanto, este cenário começou a mudar com iniciativas como o projeto GIVE (*Generating Instructions in Virtual Environments*) (Koller et al., 2009). GIVE é uma plataforma para desenvolvimento e teste de sistemas de GLN em mundos virtuais interativos, na qual o sistema encarrega-se de todo gerenciamento do ambiente gráfico e de interatividade, permitindo ao desenvolvedor concentrar-se apenas na tarefa de GLN e produzir rapidamente uma aplicação de teste. A Figura 1 ilustra este sistema.

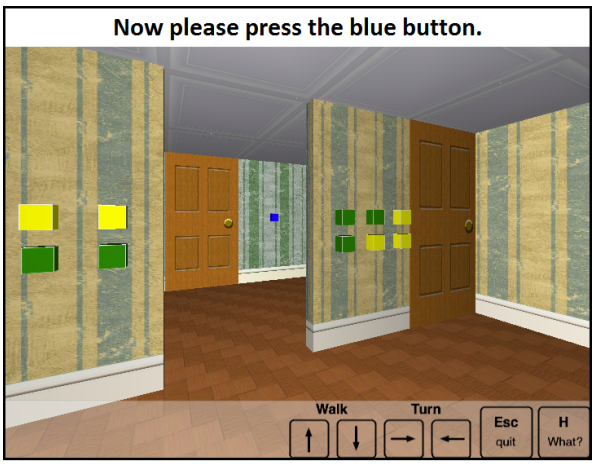


Figura 1: Exemplo de contexto GIVE.

O ambiente GIVE é composto de objetos manipuláveis do tipo botão, usados para abrir e fechar portas e outras funções de interação com o mundo virtual, e objetos maiores como peças de mobiliário e afins. Estes objetos podem, a critério do desenvolvedor de GLN, ser empregados como ponto de referência nas instruções fornecidas ao usuário, como em 'Aperte o botão *ao lado da porta*'. O uso de pontos de referência não é, entretanto, um recurso nativo do sistema, e sua implementação requer a extração e manipulação de relações espaciais (e.g., 'ao lado de') por parte do algoritmo de GER.

Este artigo discute o desenvolvimento de um algoritmo de GER que faz uso de relações espaciais em ambientes do tipo GIVE. Assim como em (Tenbrink, 2005; Baltaretu, Krahmer e Maes, 2013), assume-se que o propósito único da informação espacial seja a desambiguação do referente para fins de identificação, e não a localização de um objeto previamente identificado (Barclay e Galton, 2008). Será tratada especificamente a questão da seleção de conteúdo semântico, deixando de lado o problema da realização textual destas expressões (Pereira e Paraboni, 2007; Pereira e Paraboni, 2008; de Novais e Paraboni, 2012).

## 2 Trabalhos Relacionados

### 2.1 O problema computacional de GER

Um dos algoritmos mais conhecidos na área, e que ajudou a definir o próprio problema computacional de GER, é o algoritmo Incremental apresentado em (Dale e Reiter, 1995). Este algoritmo recebe como entrada um contexto $C$ formado por um grupo de objetos, o objeto-alvo ou referente $r$ que se deseja descrever, e suas propriedades semânticas na forma de pares (*atributo-valor*), como em (*cor-azul*).

O objetivo do algoritmo é produzir um conjunto $L$ de propriedades de $r$ tal que $L$ seja capaz de distinguir $r$ de todos os outros objetos em $C$. As propriedades são incluídas em $L$ incrementalmente (de onde provem o nome do algoritmo) segundo uma ordem $P$ predefinida, desde que contribuam para a desambiguação do referente (i.e., excluindo pelo menos um objeto do contexto $C$). O algoritmo termina quando um conjunto único (i.e., livre de ambiguidade) de propriedades é obtido (caso em que $L$ poderia ser realizada, por exemplo, como uma descrição definida), ou até que todas propriedades possíveis em $P$ tenham sido consideradas (caso em que $L$ permaneceria ambígua e poderia ser realizada, por exemplo, como uma descrição indefinida).

Considere o exemplo a seguir, ilustrando um contexto composto por três objetos: dois cachorros (um preto e um branco, sendo o preto de tamanho pequeno e o branco de tamanho grande), e um gato preto e pequeno.

- Obj1:
  (*tipo-cachorro*), (*cor-preto*), (*tamanho-pequeno*)
- Obj2:
  (*tipo-cachorro*), (*cor-branco*), (*tamanho-grande*)
- Obj3:
  (*tipo-gato*), (*cor-preto*), (*tamanho-pequeno*)

Seja:

- o objeto-alvo $r = Obj1$;
- o contexto $C = \{Obj2,\ Obj3\}$;

e a ordem preferencial

$$P =< tipo,\ cor,\ tamanho >$$

O algoritmo inicia com uma lista $L$ vazia e percorre $P$ na ordem estabelecida, inserindo em $L$ cada propriedade que exclua pelo menos um objeto em $C$. Neste exemplo, o algoritmo escolhe inicialmente a propriedade (*tipo-cachorro*) por excluir *Obj3*, que é *tipo* gato. A seguir (na ordem em $P$), a propriedade (*cor-preto*) exclui

*Obj2*, que é de *cor* branca. Como o contexto $C$ não possui mais elementos, o algoritmo retorna a expressão $L$, que poderia ser realizada, por exemplo, como 'o cachorro preto'.

A ordem preferencial de seleção de atributos da lista $P$ tem grande impacto sobre o tipo de expressão produzida pelo algoritmo. Por exemplo, se fosse considerada a ordem

$$P =< tipo, tamanho, cor >$$

para o domínio acima, a mesma referência a *Obj1* seria 'o cachorro pequeno'. Em obediência à máxima de brevidade de Grice (Grice, 1975), algoritmos de GER tendem a favorecer a seleção de atributos discriminatórios. Em anos recentes, no entanto, passaram a ser consideradas também questões como a naturalidade da expressão (ou *humanlikeness* em (Belz e Gatt, 2007)), dentre muitos outros objetivos desejáveis. Uma visão geral da área de GER e seus principais desafios é apresentada em (Krahmer e van Deemter, 2012).

### 2.2 Geração de descrições relacionais

Exemplos de algoritmos de GER que produzem descrições relacionais incluem (Dale e Haddock, 1991; Paraboni e van Deemter, 1999; Krahmer e Theune, 2002). De modo geral, entretanto, estes estudos não levam em conta as peculiaridades do fenômeno de referência em mundos virtuais.

Em sua proposta original, o algoritmo Incremental manipula apenas propriedades atômicas, mas com adaptações a serem discutidas na secção 4.2 pode também ser aplicado ao caso de propriedades relacionais. Esta modificação é de especial importância para o uso de relações espaciais de que trata este trabalho, como em (*acima*,*o*), no qual *o* é um objeto usado como ponto de referência para a descrição do objeto-alvo.

Alguns sistemas participantes da série de competições *GIVE Challenge* (Byron et al., 2009; Koller et al., 2010; Striegnitz et al., 2011) implementam certos recursos de manipulação de relações espaciais, ainda que de forma pouco documentada (Braunias et al., 2010; Schutte e Dethlefs, 2010; Garoufi e Koller, 2011; Akkersdijk et al., 2011). Entretanto, como estes sistemas foram avaliados apenas de forma extrínseca (i.e., medindo-se o desempenho global de usuários GIVE na tarefa de navegação) não é possível distinguir o eventual impacto do uso de relações espaciais das outras funcionalidades de cada sistema, as quais incluem, por exemplo, um grande número de melhorias não relacionadas à tarefa de GER.

### 2.3 Corpora de GER

Uma questão recorrente na pesquisa em GER é como obter evidência empírica sobre o fenômeno de referência. Uma opção natural, e amplamente empregada na grande área de Processamento de Língua Natural (PLN), é a avaliação baseada em corpus. Por exemplo, em aplicações de tradução automática (Aziz, Pardo e Paraboni, 2008), grandes coleções de textos paralelos (i.e., na língua-alvo e na língua-destino) são empregadas para treinamento de modelos estatísticos deste tipo.

O uso de corpora textuais para pesquisa em GER é entretanto limitado pelo fato de que o conhecimento de entrada que produziu o texto presente no corpus normalmente não está disponível. Ou seja, o conteúdo de um corpus textual típico é meramente o produto final de um processo de geração de língua natural realizado por humanos, e que não contém informações explícitas sobre o processo em si. Para diversos tipos de pesquisa em GER e áreas afins, faz-se necessário assim examinar não apenas o texto resultante do processo, mas também modelar as condições contextuais nas quais o texto foi produzida.

A forma usual de reproduzir estas condições na pesquisa em GER é a realização de experimentos controlados com uso de participantes humanos para coleta de corpus de GER, aqui entendido como sendo um conjunto de expressões de referência e os respectivos contextos (e.g., cenas) nos quais cada expressão foi produzida. Alguns exemplos de recursos deste tipo incluem TUNA (Gatt, van der Sluis e van Deemter, 2007), GRE3D3 (Dale e Viethen, 2009), GRE3D7 (Viethen e Dale, 2011) e Stars (Teixeira et al., 2014). De modo geral, entretanto, estes corpora tendem a representar apenas situações estáticas de referência (i.e., sem interatividade) e/ou envolvendo contextos visuais bidimensionais.

Uma exceção de especial interesse para esta pesquisa é o corpus GIVE-2 (Gargett et al., 2010) de instruções em mundos virtuais como 'dobre à esquerda', 'pressione o segundo botão, ao lado da porta' etc. GIVE-2 foi construído por meio de experimentos envolvendo 36 pares de participantes de língua inglesa e alemã alternando-se nas tarefas de instrutor e jogador. O corpus contém todas instruções fornecidas pelo instrutor e as respectivas decisões tomadas pelo jogador (e.g., movimentos, ações de pressionar botões etc.) em três mundos de exemplo. Este conjunto de dados multimodal pode ser visualizado na forma de animação com uso da ferramenta *Replay* apresentada em (Gargett et al., 2010).

## 3 Extração e Preparação de Dados

O presente trabalho faz uso de informações contextuais extraídas dos três mundos virtuais que compõe o corpus GIVE-2 (Gargett et al., 2010), e das expressões de referência produzidas nestes contextos. A extração e preparação destes dados é discutida individualmente nas secções a seguir.

### 3.1 Extração de relações espaciais

Objetos em um ambiente GIVE (Koller et al., 2009) possuem apenas uma propriedade atômica básica representando seu *tipo* (botões, portas, cadeiras etc.) e, no caso dos botões, uma propriedade *cor*. O primeiro passo deste trabalho foi assim a implementação de um conjunto de métodos básicos para computar relações espaciais de diversos tipos a partir de um mundo virtual.

As relações espaciais computadas para um dado objeto-alvo $r$ e ponto de referência $o$ são: *acima, abaixo, esquerda, direita, frente e atrás*. Para extração destas relações, foram utilizadas as funções propostas em (Kelleher e Costello, 2009), baseadas na posição angular de um objeto em relação ao outro no plano cartesiano.

O algoritmo de extração de relações espaciais utiliza uma constante de distância máxima $k$ única para cada tipo de relação. A posição física das entidades no ambiente GIVE é definida pela coordenada de seu ponto central, e assim entidades maiores como sofás, portas etc. possuem um ponto central mais distante das bordas. Uma entidade está próxima de outra entidade se a diferença entre os valores para os eixos $x$, $y$ e $z$ é no máximo $k$. Se esta condição for verdadeira, considera-se que há uma relação espacial válida do ponto de vista semântico.

Tendo em vista o propósito de gerar expressões de referência livres de ambiguidade, a questão da transitividade destas relações foi aqui desconsiderada. Assim, relações como *esquerda-o* devem ser entendidas como 'imediatamente à esquerda' do objeto $o$, e não contemplando objetos mais distantes que também possam estar à esquerda de $o$.

### 3.2 Preparação dos dados do corpus

As expressões de referência aqui consideradas são as que descrevem objetos do tipo botão no mundo GIVE. Botões são frequentemente referenciados nas instruções de navegação por serem os únicos elementos manipuláveis neste ambiente.

Foram extraídas do corpus todas instruções contendo a palavra 'button' e formas equivalentes como 'this' e 'box' inferidas pelo uso de verbos como 'press' ou 'click'. Como no entanto este procedimento não foi exaustivo, é possível que uma pequena parcela de descrições menos comuns tenha sido excluída da presente análise.

| Relação | Ocorrências | % |
|---|---|---|
| próximo | 217 | 21,87% |
| esquerda | 121 | 12,19% |
| direita | 92 | 9,27% |
| acima | 17 | 1,71% |
| canto | 14 | 1,41% |
| frente | 6 | 0,60% |

Tabela 1: Relações espaciais envolvendo objetos do *tipo* 'botão' em GIVE-2 (Gargett et al., 2010).

No total, foram identificadas 992 descrições de interesse. No caso da porção em alemão do corpus, as instruções foram previamente traduzidas para o inglês com uso da ferramenta *Google Translate*[2] de modo a facilitar sua interpretação.

Apenas dois tipos de propriedades atômicas foram observadas nas expressões de referência coletadas: *tipo* e *cor*. Das expressões coletadas, 467 (47,07%) utilizam algum tipo de relação espacial, sendo 248 (25%) do tipo topológica (e.g., 'O botão *perto* da planta'), e 219 (22,07%) do tipo projetiva (e.g., 'O botão *à esquerda* da planta') (Kelleher e Costello, 2009). A Tabela 1 sumariza os tipos de propriedades relacionais identificados.

No corpus GIVE-2 há três tipos de expressões que não foram consideradas neste trabalho: as que incluem relações com o jogador (e.g., 'o botão *à sua frente*'), propriedades comparativas (e.g., 'o botão *mais distante* da lâmpada') e envolvendo negações (e.g., 'o botão *que não está perto* da lâmpada'). Foram identificadas 68 descrições destes tipos, correspondendo a (6,8%) das expressões consideradas. Nestes casos é assumido um ônus para a solução proposta, que nem sempre será capaz de produzir descrições idênticas às observadas no corpus.

Um breve exame das expressões coletadas é suficiente para constatar que, em uma mesma situação de referência, pessoas diferentes podem usar ou não uma relação espacial, variação esta que pode ser tomada por um indicador da complexidade da tarefa de geração considerada. A Tabela 2 ilustra esta variação, representada pelo número de casos em que uma relação espacial foi ou não utilizada para cada referência nos três mundos (ou domínios) do corpus GIVE-2.

Pelos dados da Tabela 2 é possível observar,

[2] http://translate.google.com.br/

| **mundo 1** | | | **mundo 2** | | | **mundo 3** | | |
|---|---|---|---|---|---|---|---|---|
| r | sim | não | r | sim | não | r | sim | não |
| b19 | 5 | 32 | b18 | 32 | 6 | b1 | 29 | 4 |
| b18 | 26 | 0 | b11 | 1 | 6 | b12 | 1 | 38 |
| b3 | 2 | 36 | b12 | 26 | 10 | b5 | 74 | 2 |
| b20 | 30 | 0 | b2 | 0 | 47 | b4 | 39 | 2 |
| b11 | 5 | 30 | b5 | 19 | 17 | b10 | 32 | 0 |
| b5 | 0 | 72 | b10 | 5 | 75 | b15 | 39 | 33 |
| b4 | 0 | 1 | b6 | 6 | 39 | b36 | 3 | 33 |
| b9 | 9 | 29 | b9 | 26 | 0 | | | |
| b6 | 30 | 2 | b14 | 31 | 6 | | | |

Tabela 2: Uso de relações espaciais no corpus GIVE-2 (Gargett et al., 2010).

por exemplo, casos em que todos participantes usaram a mesma estratégia de referência, como na descrição do alvo $b20$ no mundo 1. Por outro lado, há também casos em que a divisão é de quase (50%), como no caso do alvo $b15$ no mundo 3. Casos deste último tipo ilustram algumas das dificuldades de modelagem computacional do problema de geração de relações espaciais.

A tarefa de preparação de dados produziu um conjunto de descrições anotadas com informações sobre seus atributos atômicos e relacionais, bem como suas informações contextuais (i.e., o objeto-alvo e demais objetos do contexto, seus atributos atômicos e relacionais). Este conjunto foi então dividido em um conjunto de treinamento (794 instâncias) e teste (198 instâncias) selecionadas aleatoriamente. O uso destes conjuntos é descrito nas secções a seguir.

## 4 A Abordagem Proposta

Nesta secção apresentamos uma proposta de solução para o problema computacional de GER que faz uso de relações espaciais para descrever objetos do domínio GIVE (Koller et al., 2009). A proposta é dividida em duas etapas: a tarefa de seleção de relações espaciais adequadas a partir do contexto de entrada, e o algoritmo de GER propriamente dito.

### 4.1 Seleção de pontos de referência

A extração de relações espaciais a partir do contexto visual (cf. secção 3.1) produz uma ampla gama de possibilidades de referência. Entretanto, o fato de que dois objetos mantêm uma relação espacial entre si não necessariamente significa que esta relação seja uma forma típica ou aceitável de referência a estes objetos. Por exemplo, uma expressão como "a caixa que contém um relógio" pode não ser apropriada em um contexto com várias caixas, e no qual não seja possível reconhecer a relação de forma imediata (e.g., porque um objeto oculta o outro). Em outras palavras, uma relação espacial pode ser perfeitamente válida do ponto de vista semântico, mas de uso limitado para fins de referência por questões pragmáticas variadas.

Uma forma de decidir o que constitui ou não uma relação espacial válida para fins de referência, e assim filtrar casos que não deveriam ser considerados na produção destas expressões, é pela observação de exemplos de uso real da língua. A primeira etapa da solução proposta trata assim da tarefa de determinar quais objetos mantêm relações válidas - para fins de referência - com outros objetos do mesmo contexto.

Como forma de manter um certo grau de independência do domínio, a implementação deste módulo faz uso de uma abordagem de aprendizado de máquina semelhante à adotada em (Viethen, 2010), porém descartando-se características que não se aplicam ao domínio GIVE[3], e acrescentando-se outras que capturam aspectos específicos da situação a ser tratada.

Dado um objeto-alvo $r$ e um candidato a ponto de referência $o$, utilizamos um classificador binário *use_relation* para determinar se $r$ pode ser referenciado via $o$ através de uma relação espacial. As características de aprendizagem consideradas foram extraídas da porção de treinamento do corpus, e são sumarizadas na Tabela 3.

As características *distractors*, *landmarks*, *ambiguity* e *distance* são auto explicativos. As demais, que são baseadas em observações feitas em (Viethen, 2010) para os corpora GRE3D3 e GRE3D7, são detalhadas a seguir.

A definição de *unique_relation* é motivada pela observação feita em (Viethen, 2010) de que a relação espacial entre o alvo e o ponto de re-

[3]Por exemplo, no ambiente GIVE todos objetos de um mesmo tipo possuem o mesmo tamanho, o que torna pouco útil a definição de características de aprendizagem baseadas neste tipo de informação.

| **Característica** | **Descrição** |
|---|---|
| *distractors* | quantidade de objetos do mesmo tipo que o alvo |
| *landmarks* | quantidade de objetos que mantêm relações espaciais com o alvo |
| *ambiguity* | quantidade de objetos iguais ao alvo e na mesma sala |
| *distance* | distância entre o alvo e o ponto de referência |
| *unique_relation* | indica se a relação entre o alvo e o ponto de referencia é única no contexto |
| *most_salient_landamark* | indica se o ponto de referência é mais saliente que o alvo |
| *equal_landmarks* | indica se o ponto de referência é o do mesmo tipo e cor que o alvo |

Tabela 3: Características de aprendizagem para a classe binária *use_relation*.

ferência tem mais chance de ser incluída em uma descrição quando for única, já que assim diminuise a complexidade de identificação do alvo. Um exemplo de aplicação deste princípio em um ambiente GIVE seria um contexto em que deseja-se descrever um botão, e este é o único objeto do *tipo* botão que se encontra à esquerda de uma cadeira. Neste caso, o uso da relação espacial seria altamente recomendado. O mesmo não ocorreria, entretanto, em um contexto em que houvesse outros botões à esquerda de alguma cadeira.

A definição de *most_salient_landamark* é baseada na observação feita em (Viethen, 2010) de que uma relação espacial entre o alvo e o ponto de referência é mais utilizada quando o ponto de referência é mais saliente que o alvo. Um exemplo de aplicação deste princípio em um ambiente GIVE seria um contexto em que deseja-se referenciar um botão (que é um objeto pequeno) próximo a uma porta (que é um objeto grande), resultando em uma expressão do tipo 'o botão ao lado da porta'. Para fins de implementação desta regra, a saliência das entidades é definida de acordo com seu tamanho físico. Outros fatores, como a distância entre o alvo e o ponto de referência, serão considerados na próxima secção, quando será discutido um ranking de relações espaciais. Deve-se ressaltar que as entidades consideradas próximas são as entidades que se encontram na mesma sala do respectivo mundo virtual.

Finalmente, a definição de *equal_landmarks* é também baseada em uma observação feita em (Viethen, 2010), segundo a qual uma relação espacial é preferível quando há similaridade visual entre o alvo e o ponto de referência. Um exemplo de aplicação deste princípio em um ambiente GIVE seria um contexto em que o alvo compartilha a mesma cor e mesmo tamanho que o ponto de referência, ou envolvendo objetos que não possuem variações de cor (e.g., sofás, cadeiras, portas, etc.), nos quais a similaridade visual se reduz à coincidência de tipos. Observa-se entretanto que no ambiente GIVE não existe variação de tamanho entre objetos do mesmo tipo, o que reduz a aplicabilidade da observação feita em (Viethen, 2010), a qual tinha um sentido original mais amplo já que naquele domínio havia variação de tamanho e cor.

As instâncias de aprendizagem para a classe *use_relation* foram geradas da seguinte forma. Para cada objeto-alvo $r$ do corpus, foi computada uma lista de $n$ objetos que seriam candidatos a pontos de referência em uma possível descrição de $r$ naquele contexto, utilizando-se o método de extração de relações espaciais do domínio descrito na secção 3.1.

A seguir, para cada par alvo-candidato ($r$,$o_{i=1..n}$), foi gerada uma instância de aprendizagem, totalizando assim $n$ instâncias para cada objeto-alvo. O conjunto de $n$ instâncias de cada objeto-alvo $r$ foi rotulado da seguinte forma: se a descrição de $r$ no corpus não usa uma relação espacial, então todas $n$ instâncias são rotuladas como negativas (*use_relation = não*), isso é, nenhum dos candidatos pode ser recomendado como ponto de referência para $r$. Por outro lado, se a descrição do objeto $r$ no corpus usou uma relação com um ponto de referência $o$, então a instância que representa o par ($r$,$o$) é rotulada como positiva (*use_relation = sim*), e todas outras $n$-1 instâncias como negativas.

Como resultado deste procedimento, foram geradas 3246 instâncias de aprendizagem compostas pelas características acima, sendo 335 positivas e 2911 negativas. A classificação propriamente dita foi realizada com o algoritmo de indução de árvores de decisão *J48* disponibilizado pelo pacote WEKA (Witten, Frank e Hall, 2011), utilizando-se *10-fold cross-validation* e demais parâmetros *default* do algoritmo (confiança $C$=0,25 e mínimo de $M$=2 instâncias por folha).

A Tabela 4 apresenta a matriz de confusão com base nos dados de treinamento. A Tabela 5 apresenta os resultados do treinamento de acordo com as medidas de precisão (P), cobertura (C) e medida $F_1$. Cabe reiterar que o resultado da aplicação deste classificador aos dados de teste será tratado na secção 5.

| | **sim** | **não** |
|---|---|---|
| usa relação | 213 | 122 |
| não usa relação | 97 | 2814 |

Tabela 4: Matriz de confusão para *use_relation* sobre dados de treinamento.

| **use_relation** | **P** | **C** | $F_1$ |
|---|---|---|---|
| sim | 0,68 | 0,63 | 0,66 |
| não | 0,95 | 0,96 | 0,96 |
| média | 0,93 | 0,93 | 0,93 |

Tabela 5: Resultados para *use_relation* sobre dados de treinamento.

Estes resultados podem ser considerados satisfatórios na medida em que o classificador objetiva modelar apenas o comportamento *médio* dos 72 participantes do experimento que deu origem ao corpus GIVE-2, os quais frequentemente adotam estratégias de referência conflitantes. Conforme discutido na secção 3.2, por exemplo, um mesmo botão pode ser descrito por participantes distintos como 'o botão azul' ou 'o botão ao lado da porta', dentre muitas outras possibilidades observadas no corpus.

O classificador foi incorporado ao ambiente GIVE e constitui o primeiro módulo a ser invocado na produção de uma descrição de um objeto-alvo $r$. Um exemplo completo é descrito a seguir, com base no contexto da Figura 2.

Neste exemplo, considera-se o objetivo de descrever o alvo $b2$ (um botão amarelo) em um contexto contendo também uma cadeira $c1$ e dois outros botões $b1$ (vermelho) e $b3$ (amarelo). Os rótulos associados a cada objeto, bem como a seta que aponta para o alvo $b2$ são meramente ilustrativos, e não fazem parte da imagem real.

Dado o objetivo de descrever $b2$, o módulo classificador recebe como entrada as propriedades relacionais extraídas conforme discutido na secção 3.1, conforme a Tabela 6.

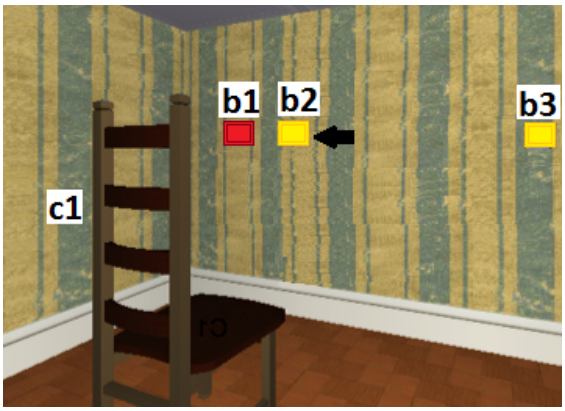


Figura 2: Exemplo de contexto visual.

| **Id** | **Propriedade** | **Distância** |
|---|---|---|
| $r1$ | *direita-b1* | 0,30 |
| $r2$ | *atrás-c1* | 1,25 |
| $r3$ | *esquerda-b3* | 1,80 |

Tabela 6: Propriedades relacionais do alvo $b2$.

A seguir, é construído um conjunto de instâncias de teste $i1..i3$ para verificar a possibilidade de uso de cada uma das relações $r1..r3$ computadas. As instâncias assim obtidas são relacionadas na Tabela 7.

As instâncias $i1..i3$ são submetidas à árvore de decisão, e o resultado da classificação determina se cada uma das relações correspondentes ($r1..r3$) deve ou não ser recomendada para uso pelo algoritmo de GER. Neste exemplo, apenas as relações $r1$ e $r2$ atendem aos critérios do classificador: a relação $r3$, a distância entre $b2$ e $b3$ foi considerada excessiva para fins de referência usando esta relação (ou, para ser mais exato, é superior à distância observada no corpus nos casos em que esta relação foi utilizada). As duas relações recomendadas serão fornecidas como entrada ao próximo módulo do sistema, a seguir.

### 4.2 Geração de expressões de referência usando relações espaciais

Utilizando-se o procedimento descrito na secção anterior é possível determinar, para uma determinada situação de referência a um objeto-alvo $r$, quais relações espaciais seriam adequadas para descrever $r$ naquele contexto. A etapa seguinte consiste então em gerar a descrição propriamente dita, a qual pode ou não incluir uma das relações espaciais sugeridas.

O algoritmo proposto para este fim é uma versão modificada do algoritmo Incremental (Dale e Reiter, 1995) para manipular propriedades relacionais, e também integrado ao sistema GIVE (Koller et al., 2009).

Uma propriedade é incluída na expressão resultante desde que elimine ao menos uma en-

| # | $i1$ | $i2$ | $i3$ |
|---|---|---|---|
| *distractors* | 3 | 3 | 3 |
| *landmarks* | 3 | 3 | 3 |
| *ambiguity* | 2 | 2 | 2 |
| *distance* | 0,30 | 1,25 | 1,80 |
| *unique_relation* | true | false | true |
| *most_salient_landmark* | false | true | false |
| *equal_landmarks* | true | false | true |

Tabela 7: Instâncias de teste para o exemplo da Figura 2.

| r | Propriedades selecionadas | Exemplo de realização possível |
|---|---|---|
| b1 | *Ref-b1*= [(*tipo-botão*), (*cor-vermelho*)] | o botão vermelho |
| b2 | *Ref-b2*= [(*tipo-botão*), (*cor-amarelo*), (*direita-b1*)]<br>*Ref-b1*= [(*tipo-botão*), (*cor-vermelho*)] | o botão amarelo, à direita do botão vermelho |
| b3 | *Ref-b3*= [(*tipo-botão*), (*cor-amarelo*)] | o botão amarelo (ambíguo) |
| c1 | *Ref-c1*= [(*tipo-cadeira*)] | a cadeira |

Tabela 8: Exemplos de descrições para os objetos da Figura 2.

tidade do contexto. Além disso, propriedades atômicas e relacionais são consideradas para inclusão em ordem de frequência, conforme observado na porção de treinamento do corpus GIVE-2, levando à definição da seguinte lista de preferência a ser usada como parâmetro $P$ do algoritmo. Esta lista é única para todos tipos de objetos, ou seja, tanto para o alvo (que é sempre um objeto do tipo botão) como para os diversos tipos de objetos usados como pontos de referência (outros botões, cadeiras, mesas etc.):

$$P =< tipo, cor, esq, dir, acima, frente, abaixo, atrás >$$

Observa-se também a ausência de relações espaciais de proximidade (e.g., *próximo*, *ao lado* etc.) na definição de $P$. No presente trabalho optamos por suprimir a geração de relações deste tipo em favor das formas mais específicas (e.g., *esquerda*, *direita*, *acima* etc.), as quais apresentam maior poder discriminatório e, consequentemente, podem levar à construção de expressões mais breves. Por exemplo, no caso da descrição de $b2$ no contexto da Figura 2, a expressão 'o objeto *à direita do* botão vermelho' seria preferível à forma ambígua 'o objeto *próximo ao* botão vermelho', que pode ser interpretada tanto como em referência a $b2$ como à cadeira $c1$.

Assim como em (Paraboni, 2000), no caso de expressões envolvendo um objeto e um ponto de referência, o algoritmo descreve cada objeto de forma independente. Isso pode, em alguns casos, acarretar superespecificação. Por exemplo, em um contexto com duas mesas e dois livros, em que apenas um dos livros está sobre uma das mesas, o presente algoritmo faria uso de propriedades adicionais, como em 'o livro vermelho, sobre a mesa da esquerda'[4]. Esta medida foi adotada para evitar possíveis problemas de identificação de objetos em domínios espaciais complexos, como os discutidos em (Paraboni e van Deemter, 2013).

Na Tabela 8 são apresentados exemplos de saídas do algoritmo para cada um dos botões do contexto da Figura 2, considerando-se as propriedades relacionais da Tabela 6 e a ordem de preferência $P$ acima, e observando-se que cada descrição é composta de uma série de listas de propriedades (cláusulas Ref-) que podem referenciar outros objetos.

A referência a $b3$ neste exemplo permanece ambígua, pois neste caso o contexto oferecido como entrada para o algoritmo não contém propriedades em número suficiente para que o objeto possa ser identificado de forma única. A modelagem de uma relação *extremidade-direita*, por exemplo, permitira a geração de descrições como 'o botão amarelo na extrema direita'.

## 5 Avaliação

### 5.1 Procedimento

Na avaliação do trabalho desenvolvido, considerou-se inicialmente a possibilidade de avaliação extrínseca, baseada no uso real do algoritmo de GER proposto em uma aplicação do tipo GIVE. O procedimento neste caso seria semelhante ao adotado em (Paraboni e van Deemter, 2013), medindo-se o desempenho de um usuário na tarefa de navegação ao reagir a uma série de instruções contendo expressões de referência produzidas pelo algoritmo proposto, ou por algum sistema de *baseline* usado como termo de comparação.

Uma análise dos diferentes tipos de expressões produzidas, e das limitações do ambiente GIVE, nos leva entretanto à conclusão de que este tipo de avaliação não seria factível. No ambiente GIVE, as únicas métricas de avaliação possíveis são aquelas baseadas na distância percorrida pelo usuário até aproximar-se do objeto-alvo, ou baseadas no tempo transcorrido entre a produção da descrição e a seleção do objeto referenciado. Estas métricas, entretanto, não nos permitiriam capturar de forma significativa a possível diferença entre alternativas como 'pressione o botão vermelho' e 'pressione o botão ao lado do botão amarelo', que são saídas típicas de um algoritmo

[4] Diferentemente de (Dale e Haddock, 1991), por exemplo, que neste caso permitiria mútua desambiguação como em 'o livro sobre a mesa'.

do tipo proposto. Em outras palavras, diferenças como esta exigiriam recursos de medição mais sofisticados, como técnicas de *eye-tracking* (Koller et al., 2012) ou uso de imagens cerebrais (Engelhardt, Demiral e Ferreira, 2011), as quais estão fora do escopo deste trabalho.

Além da inadequação para o tipo de avaliação aqui exigido, medidas de tempo e distância em ambientes GIVE são também altamente sujeitas a ruído. Por exemplo, o usuário pode ter dificuldades de manipulação da interface, ou pode interromper a navegação para considerar uma decisão com mais cautela, dentre muitas outras possibilidades que tornam as métricas de tempo e distância ainda menos confiáveis (embora naturalmente não inviabilize a avaliação de outros aspectos de um sistema deste tipo, como o caso da geração de instruções de navegação abordada na série de desafios GIVE (Byron et al., 2009; Koller et al., 2010; Striegnitz et al., 2011)).

Em razão destas dificuldades, diversos aspectos do trabalho proposto foram assim avaliados de forma intrínseca com base no conjunto de teste extraído do corpus GIVE-2 (Gargett et al., 2010) descrito na secção 3.2. Para este fim, três aspectos de interesse foram considerados: a política de seleção da relação espacial, a ordem de preferência para seleção de propriedades (o parâmetro $P$ do algoritmo de GER), e o tratamento de relações espaciais redundantes.

Com relação à política de seleção da relação espacial, consideramos duas alternativas: a proposta original - que seleciona a relação espacial *mais frequente* no corpus de treinamento - e uma alternativa na qual a propriedade espacial é selecionada de forma *aleatória*.

Com relação à ordem de preferência $P$ utilizada pelo algoritmo, consideramos também duas alternativas: a proposta original - que faz ordenação *por frequência* conforme observado no conjunto de treinamento - e uma estratégia *gulosa* na qual propriedades de maior poder discriminatório (ou seja, aquelas que diferenciam o objeto-alvo do maior número possível de objetos do contexto) têm preferência.

Finalmente, com relação ao tratamento de relações espaciais redundantes, consideramos duas alternativas: a proposta original - que *não* inclui propriedades redundantes (i.e., aquelas que não contribuem para a desambiguação do referente, como em (Dale e Reiter, 1995)) - e uma estratégia na qual a relação espacial é selecionada mesmo que seja *redundante*.

A avaliação destes três aspectos da solução leva ao enunciado de $(2 \times 2 \times 2)$ 8 algoritmos distintos, dos quais a proposta original corresponde à alternativa que seleciona a relação espacial mais frequente, ordena a lista de preferências $P$ também por frequência, e inclui uma propriedade relacional na expressão apenas se esta for discriminatória.

Por simplicidade, todas alternativas avaliadas consideram como contexto de referência o conjunto de objetos na mesma sala onde se encontra o objeto-alvo. Entretanto, cabe observar que, no caso do corpus GIVE-2, o contexto utilizado pelos participantes do experimento foi, em alguns casos, formado apenas pelos objetos visíveis naquele instante. Por exemplo, quando o jogador já estava muito próximo do objeto referenciado pelo instrutor, em alguns casos este optou por desconsiderar os objetos mais distantes e produzir uma descrição breve como 'o botão', mesmo havendo outros botões na mesma sala. Uma vez que estes casos não foram contemplados na presente avaliação, assume-se assim um ônus para todos os algoritmos avaliados.

Cada uma das 198 situações de referência constantes no corpus de teste - aqui denominado conjunto *Referência* - foi fornecida como entrada para cada um dos 8 algoritmos, resultando assim em 8 conjuntos de expressões denominados *Sistema 1..8*. A avaliação propriamente dita consistiu em comparar cada conjunto *Sistema 1..8* com o conjunto *Referência*.

Para a comparação de cada um dos 1584 pares $(8 \times 198)$ *Sistema-Referência*, utilizamos duas métricas amplamente utilizadas (e.g., (Belz e Gatt, 2007; de Lucena, Paraboni e Pereira, 2010)): o coeficiente *Dice* (Dice, 1945), que mede o grau de similaridade entre os dois conjuntos de propriedades, assumindo um valor entre 0 (totalmente distintos) e 1 (idênticos); e *MASI* (Passonneau, 2006), que possui correlação com *Dice*, porém atribuindo maior peso no caso de uma expressão ser subconjunto da outra.

Além de calcular coeficientes Dice e MASI, será considerada para fins ilustrativos a medida de exatidão, que é definida como o número de coincidência total entre cada par *Sistema-Referência* (número de vezes que o coeficiente Dice é igual a 1). Considerando-se entretanto que diversos aspectos da solução de GER estão fora do escopo deste trabalho (por exemplo, o tratamento de propriedades não espaciais que também fazem parte da expressão), a coincidência total em relação às descrições do corpus não seria um objetivo realista. Espera-se assim que os valores de exatidão sejam baixos para todos os algoritmos em questão, e que as métricas de menor granularidade Dice e MASI possam refletir de forma mais precisa o mérito de cada algoritmo.

### 5.2 Resultados

Antes de discutir a avaliação do sistema proposto e suas alternativas, será discutida a avaliação do classificador responsável pela seleção de possíveis pontos de referência (cf. secção 3.1) aplicado aos dados de teste. Este módulo é fixo para todas as versões do sistema consideradas nesta avaliação, e portanto seu desempenho não se reflete na análise a ser realizada.

A matriz de confusão obtida com base nos dados de teste é ilustrada na Tabela 9. O resultado obtido pelas etapas de extração e classificação de pontos de referência é sumarizado na Tabela 10.

Os 198 itens de teste apresentaram um total de 86 expressões contendo relações espaciais (43,43%). Destas, 36 itens (18,18%) faziam referência ao próprio ouvinte receptor da expressão, como em 'pressione o botão ao seu lado'. Conforme discutido na secção 3.2, estes casos não foram cobertos pela presente proposta e acarretam assim uma margem de erro a todas as implementações avaliadas. Os 8 algoritmos e seus resultados são sumarizados na Tabela 11. O algoritmo originalmente proposto é o primeiro (#1).

As seguintes comparações entre algoritmos foram realizadas utilizando-se o teste de *Wilcoxon* sobre coeficientes *Dice*. Primeiramente, observa-se que o algoritmo proposto apresenta os melhores resultados para todas as três métricas de avaliação. Entretanto, seus resultados são idênticos ao da alternativa #5, que difere apenas na estratégia de seleção. Em outras palavras, não houve diferença significativa entre a seleção da propriedade mais frequente e a simples seleção aleatória. Este resultado se explica pelo fato de que nos dados de teste a maioria dos objetos-alvo só possuir um ponto de referência possível.

Em segundo lugar, a ordenação por frequência é significativamente superior à ordenação gulosa para todos os pares de algoritmos avaliados (i.e., comparando-se #1 com #3, #2 com #4 etc.) ($W$=16110, $Z$=11,6, $p$ <0,001). Este resultado contrasta propostas de algoritmos para geração de descrições breves ou mínimas como em (Gardent, 2002). Por outro lado, a estratégia de não inserir redundância explícita (isso é, evitando a inclusão de atributos desnecessários para desambiguação) é significativamente superior ao seu uso para todos os pares de algoritmos (i.e., comparando-se #1 com #2, #3 com #4 etc.). A menor diferença observada, mas ainda altamente significativa, foi entre os algoritmos #7 e #8 ($W$=986, $Z$=5,75, $p$ <0,001).

Quanto à exatidão dos algoritmos propostos, observamos ainda que os algoritmos #1 e #5 relevaram-se superiores a todos os demais, sendo os únicos a gerar descrições completamente idênticas às do corpus de teste.

## 6 Discussão

Este artigo descreveu o desenvolvimento e avaliação de um algoritmo de GER que faz uso de relações espaciais em ambientes do tipo GIVE (Koller et al., 2009). Apesar da relativa simplicidade da proposta, este estudo contribui para o entendimento de três aspectos do problema: a política de seleção de propriedades espaciais, a ordenação das propriedades consideradas pelo algoritmo de GER, e o tratamento de propriedades redundantes.

Mesmo não tendo sido observada uma diferença significativa entre as alternativas de seleção de propriedade espacial, as estratégias de ordenação baseada em frequência e de inclusão de propriedades espaciais discriminatórias revelaram-se superiores às demais. Em outras palavras, não parece haver uma preferência geral por propriedades discriminatórias, mas as propriedades espaciais parecem ser empregadas predominantemente desta forma.

Como trabalho futuro planejamos expandir a presente análise para incluir aspectos adicionais da solução, tais como a prioridade da relação espacial considerada pelo algoritmo, e uma política de seleção de relações espaciais mais sofisticada. Além disso, esperamos testar a substituição do presente modelo de classificação baseado em árvores de decisão por outros, como máquinas de vetor de suporte recentemente aplicadas ao problema de GER em (Ferreira e Paraboni, 2014).

| | **sim** | **não** |
|---|---|---|
| usa relação | 13 | 4 |
| não usa relação | 9 | 172 |

Tabela 9: Matriz de confusão para *use_relation* sobre dados de teste.

| **use_relation** | **P** | **C** | $F_1$ |
|---|---|---|---|
| sim | 0,59 | 0,77 | 0,67 |
| não | 0,98 | 0,95 | 0,96 |
| média | 0,94 | 0,93 | 0,94 |

Tabela 10: Resultados para *use_relation* sobre dados de teste.

| # | Seleção | Ordenação | Redundância | Dice | MASI | Exatidão |
|---|---|---|---|---|---|---|
| 1 | mais frequente | por frequência | não | **0.73** | **0.49** | **0.34** |
| 2 | mais frequente | por frequência | sim | 0.58 | 0.22 | 0.00 |
| 3 | mais frequente | gulosa | não | 0.28 | 0.11 | 0.00 |
| 4 | mais frequente | gulosa | sim | 0.21 | 0.05 | 0.00 |
| 5 | aleatória | por frequência | não | **0.73** | **0.49** | **0.34** |
| 6 | aleatória | por frequência | sim | 0.58 | 0.22 | 0.00 |
| 7 | aleatória | gulosa | não | 0.25 | 0.10 | 0.00 |
| 8 | aleatória | gulosa | sim | 0.20 | 0.05 | 0.00 |

Tabela 11: Resultados.

## Referências


Akkersdijk, S., M. Langenbach, F. Loch, e M. Theune. 2011. The Thumbs Up! Twente system for GIVE 2.5. Em *Generation Challenges Session at ENLG-2011*, pp. 312–317.

Aziz, Wilker Ferreira, Thiago Alexandre Salgueiro Pardo, e Ivandré Paraboni. 2008. An experiment in Spanish-Portuguese statistical machine translation. *Advances in Artificial Intelligence-SBIA 2008*, LNAI 5249:248–257.

Baltaretu, Adriana Alexandra, Emiel Krahmer, e Alfons Maes. 2013. Factors influencing the choice of relatum in referring expressions generation: animacy vs. position. Em *CogSci workshop on the production of referring expressions: bridging the gap between cognitive and computational approaches to reference (PRE-CogSci-2013)*, pp. 1–6.

Barclay, Michael e Antony Galton. 2008. An influence model for reference object selection in spatially locative phrases. *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*, LNCS 5248:216–232.

Belz, A. e A. Gatt. 2007. The attribute selection for GRE challenge: Overview and evaluation results. Em *UCNLG+MT: Language Generation and Machine Translation*.

Braunias, J., U. Boltz, M. Drager, B. Fersing, e O. Nikitina. 2010. The GIVE-2 challenge: Saarland NLG system. Em *INLG-2010*.

Byron, Donna, Alexander Koller, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore, e Jon Oberlander. 2009. Report on the first NLG challenge on generating instructions in virtual environments (GIVE). Em *ENLG-2009*.

Cuevas, R. e Ivandré Paraboni. 2008. A machine learning approach to portuguese pronoun resolution. *Advances in Artificial Intelligence–IBERAMIA 2008*, LNAI 5290:262–271.

Dale, R. e N. J. Haddock. 1991. Content determination in the generation of referring expressions. *Computational Intelligence*, 7:252–265.

Dale, R. e E. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19.

Dale, Robert e Jette Viethen. 2009. Referring expression generation through attribute-based heuristics. Em *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pp. 58–65. Association for Computational Linguistics.

de Lucena, Diego Jesus, Ivandré Paraboni, e Daniel Bastos Pereira. 2010. From semantic properties to surface text: The generation of domain object descriptions. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 14(45):48–58.

de Novais, Eder Miranda e Ivandré Paraboni. 2012. Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2):135–146.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Engelhardt, Paul E., S. B. Demiral, e Fernanda Ferreira. 2011. Over-specified referring expressions impair comprehension: An erp study. *Brain and Cognition*.

Ferreira, Thiago Castro e Ivandré Paraboni. 2014. Classification-based referring expression generation. *Lecture Notes in Computer Science*, 8403:481–491.

Gardent, Claire. 2002. Generating minimal definite descriptions. Em *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 96–103, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gargett, A., K. Garoufi, Alexander Koller, e K. Striegnitz. 2010. The GIVE-2 corpus of giving instructions in virtual environments. Em *LREC-2010*.

Garoufi, K. e A. Koller. 2011. The Potsdam NLG systems at the GIVE-2.5 challenge. Em *Generation Challenges Session at ENLG-2011*, pp. 307–311.

Gatt, A., A. Belz, e E. Kow. 2008. The TUNA challenge 2008: Overview and evaluation results. Em *INLG-2008*, pp. 198–206.

Gatt, A., A. Belz, e E. Kow. 2009. The TUNA-REG challenge 2009: Overview and evaluation results. Em *ENLG-2009*, pp. 174–182.

Gatt, Albert, Ilka van der Sluis, e Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. Em *11th European Workshop on Natural Language Generation (ENLG-07)*.

Grice, H. P. 1975. Logic and conversation. Em Peter Cole e Jerry L. Morgan, editores, *Syntax and semantics*, volume 3. New York: Academic Press.

Kelleher, John D. e Fintan J. Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, June, 2009.

Koller, A., D. Byron, J. Cassell, R. Dale, K. Striegnitz, J. Moore, e J. Oberlander. 2009. The software architecture for the first challenge on generating instructions in virtual environments. Em *EACL-2009*.

Koller, Alexander, Konstantina Garoufi, Maria Staudte, , e Matthew Crocker. 2012. Enhancing referential success by tracking hearer gaze. Em *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 30–39.

Koller, Alexander, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, e Jon Oberlander. 2010. Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). Em *INLG-2010*.

Krahmer, E. e Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Krahmer, Emiel e Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. Em *Information sharing: Reference and presupposition in language generation and interpretation*, volume 143. CSLI Publications, California, pp. 223–263.

Paraboni, Ivandré. 1997. Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa. Tese de Mestrado, PUCRS.

Paraboni, Ivandré. 2000. An algorithm for generating document-deictic references. Em *Procs. of the INLG-2000 workshop Coherence in Generated Multimedia*, pp. 27–31, Mitzpe Ramon.

Paraboni, Ivandré e Kees van Deemter. 1999. Issues for the generation of document deixis. Em *Procs. of workshop on Deixis, Demonstration and Deictic Belief in Multimedia Contexts, in association with the 11th European Summers School in Logic, Language and Information (ESSLLI-99)*, pp. 44–48.

Paraboni, Ivandré e Kees van Deemter. 2013. Reference and the facilitation of search in spatial domains. *Language and Cognitive Processes*, online.

Passonneau, Rebecca. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. Em *LREC-2006*.

Pereira, D. B. e I. Paraboni. 2008. Statistical surface realisation of portuguese referring expressions. Em *Advances in Natural Language Processing*, volume LNAI 5221. Springer-Verlag, pp. 383–392.

Pereira, Daniel Bastos e Ivandré Paraboni. 2007. A language modelling tool for statistical NLP. Em *5th Workshop on Information and Human Language Technology (TIL-2007)*, pp. 1679–1688, Rio de Janeiro. Sociedade Brasileira de Computação.

Schutte, N. e N. Dethlefs. 2010. The Dublin-Bremen system for the GIVE-2 challenge. Em *INLG-2010*.

Striegnitz, K., A. Denis, A. Gargett, K. Garoufi, A. Koller, e M. Theune. 2011. Report on the second second challenge on generating instructions in virtual environments (GIVE-2.5). Em *Generation Challenges Session at ENLG-2011*, pp. 270–279.

Teixeira, Caio V. M., Ivandré Paraboni, Adriano S. R. da Silva, e Alan K. Yamasaki. 2014. Generating relational descriptions involving mutual disambiguation. *Lecture Notes in Computer Science*, 8403:492–502.

Tenbrink, T. 2005. Identifying objects on the basis of spatial contrast: An empirical study. Em C. Freksa, M. Knauff, B. Krieg-Bruckner, B Nebel, e T. Thomas Barkowsky, editores, *Spatial Cognition IV: Reasoning, Action, Interaction. International Conference Spatial Cognition 2004*. Springer, pp. 124–146.

Viethen, H. A. E. 2010. *The Generation of Natural Descriptions: Corpus-based Investigations of Referring Expressions in Visual Domains*. Tese de doutoramento, Macquarie University.

Viethen, J. e R. Dale. 2011. GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. Em *UCNLG+Eval: Language Generation and Evaluation Workshop*, pp. 12–22, Edinburgh, Scotland.

Witten, I. H., E. Frank, e M. A. Hall. 2011. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufman Publishers, Burlington, MA, 3rd edition.

# Usando Grades de Entidades na Análise Automática de Coerência Local em Textos Científicos

**Using Entity Grids to Automatically Evaluate Local Coherence in Scientific Texts**

Alison Rafael Polpeta Freitas
Universidade Estadual de Maringá
arpfreitas@gmail.com

Valéria Delisandra Feltrim
Universidade Estadual de Maringá
vfeltrim@din.uem.br

**Resumo**

Este artigo apresenta os resultados de uma investigação acerca da aplicabilidade do modelo grade de entidades proposto por Barzilay e Lapata (2008) na avaliação de coerência local em resumos científicos escritos em português. Mais especificamente, se buscou avaliar se tal modelo poderia ser empregado na implementação de um classificador capaz de detectar quebras de linearidade que afetam a coerência dos resumos. Os resultados experimentais se mostraram próximos aos do modelo original para a língua inglesa e semelhantes aos relatados por trabalhos relacionados para outras línguas. Nos experimentos com resumos científicos, os resultados foram próximos ao obtido por juízes humanos, mostrando que o modelo grade de entidades tem potencial para ser aplicado no contexto investigado.

**Palavras chave**

coerência local, modelo grade de entidades, texto científico

**Abstract**

In this paper we investigate the applicability of Barzilay and Lapata's (2008) entity-grid model in the evaluation of local coherence in scientific abstracts written in Portuguese. More specifically, we focused on assessing whether such model could be employed in the implementation of a classifier capable of detecting linearity breaks that affect text coherence. Our experimental results are close to those of the original entity-grid model for English and very similar to the results reported by related works for other languages. In experiments with scientific abstracts, results are close to those obtained by human judges, showing that the entity-grid model can be used in the investigated context.

**Keywords**

local coherence, entity-grid model, scientific writing

## 1 Introdução

Para uma grande variedade de aplicações na área de Processamento de Linguagem Natural, a avaliação da coerência textual tem sido uma parte importante do processo. De modo geral, qualquer aplicação que envolva geração automática de texto em algum nível de processamento pode se beneficiar de métodos que possibilitem avaliar a coerência do texto gerado. Um exemplo desse tipo de aplicação é a sumarização automática.

Outra categoria de aplicação que tem utilizado métodos de avaliação de coerência é a das ferramentas de auxílio à escrita, em especial aquelas com propósito educacional. Para a língua inglesa, são exemplos as ferramentas *Criterion* (Higgins et al., 2004; Burstein, Chodorow e Leacock, 2003), *Intelligent Essay Assessor* (Landauer, Laham e Foltz, 2003) e *Intellimetric* (Elliot, 2003). Essas ferramentas buscam avaliar a qualidade de redações (*essays*) escritas em inglês e para isso analisam um conjunto de aspectos relativos a qualidade do texto que inclui algum tipo de avaliação de coerência.

Para a língua portuguesa, um exemplo é o sistema SciPo (Feltrim et al., 2006), desenvolvido para ajudar escritores iniciantes na escrita científica, em especial na área da Ciência da Computação. Entre os vários recursos disponíveis, o SciPo possui um módulo de análise de coerência que detecta potenciais problemas de coerência semântica em resumos científicos (Souza e Feltrim, 2013). Atualmente, esse módulo é baseado na classificação de componentes retóricos e na similaridade semântica entre componentes medida por meio de *Latent Semantic Analysis* – LSA – (Landauer, Foltz e Laham, 1998). Especificamente, três tipos de relacionamentos semânticos ou dimensões são examinados: (1) Dimensão Título: verifica o relacionamento semântico entre o título do resumo e o componente Propósito; (2) Dimensão Propósito: verifica o relacionamento

semântico entre o componente Propósito e os componentes Metodologia, Resultado e Conclusão; e (3) Dimensão Lacuna-Contexto: verifica o relacionamento semântico entre o componente Lacuna e o componente Contexto.

Souza e Feltrim (2011) propõem ainda uma quarta dimensão, chamada Quebra de Linearidade, em que se verifica a existência de uma "quebra" entre sentenças adjacentes do resumo, que se caracteriza pela dificuldade em se estabelecer uma ligação clara da sentença atual com a sentença anterior, demandando maior esforço cognitivo para a interpretação do texto. No entanto, os resultados obtidos para essa dimensão foram pouco satisfatórios, mostrando que a LSA não foi capaz de capturar as quebras de linearidade por vezes sutis observadas no córpus de resumos científicos utilizado pelos autores. Conforme sugerido por Souza e Feltrim (2013), um modelo de coerência que fosse capaz de mapear o fluxo textual de forma mais refinada, como a grade de entidades proposta por Barzilay e Lapata (2008), poderia obter melhores resultados para essa dimensão.

Nesse contexto, este trabalho investigou a aplicabilidade do modelo grade de entidades (Barzilay e Lapata, 2008) na avaliação de coerência em resumos científicos escritos em português. Mais especificamente, se buscou avaliar se tal modelo poderia ser empregado na implementação de um classificador capaz de detectar quebras de linearidade que afetam a coerência dos resumos, de modo semelhante ao proposto por Souza e Feltrim (2013), visando a futura inclusão de tal classificador no módulo de análise de coerência do sistema SciPo.

Dois tipos de experimentos foram realizados. Primeiramente foram feitos experimentos com um córpus jornalístico em português, visando a comparação, ainda que indireta, com outros trabalhos que utilizam o modelo grade de entidades. Para avaliar o desempenho do modelo com textos científicos foram feitos experimentos com um córpus de resumos científicos escritos em português.

Os resultados experimentais com o córpus jornalístico se mostraram próximos aos do modelo original para a língua inglesa e semelhantes aos relatados por trabalhos relacionados para outras línguas. Nos experimentos com resumos científicos, os resultados obtidos com o modelo foram próximos ao obtido por dois juízes humanos, mostrando que o modelo tem potencial para ser aplicado no contexto do sistema SciPo.

O restante deste artigo está organizado da seguinte forma: o modelo grade de entidades é apresentado na Seção 2, assim como outros trabalhos relacionados. Na Seção 3 é descrita a implementação do modelo grade de entidades para a língua portuguesa e os resultados dos experimentos de avaliação são apresentados na Seção 4. Por fim, na Seção 5 são apresentadas as conclusões deste trabalho, bem como as sugestões de trabalhos futuros.

## 2 Modelo Grade de Entidades

Como explicitado pelo próprio nome, o modelo grade de entidades é baseado em uma grade (ou matriz) de entidades e busca aprender propriedades relativas à coerência local semelhantes às definidas pela Teoria de *Centering* (Grosz, Weinstein e Joshi, 1995). A teoria de *centering* preconiza que em um texto coerente o foco de atenção (uma entidade) tende a se manter em sentenças adjacentes e que certos tipos de transições entre focos de atenção são preferíveis a outros. O modelo grade de entidades generaliza essa teoria, modelando na grade todas as transições de todas as entidades de um texto e, posteriormente, calculando uma probabilidade para cada tipo de transição. Como na teoria de *centering*, o modelo grade de entidades assume que as entidades mais relevantes do discurso aparecerão em funções sintáticas importantes, como sujeito e objeto. Desse modo, o modelo seria capaz de apreender padrões de transições característicos de textos coerentes/incoerentes.

Cada texto é representado por uma grade em que as linhas correspondem às sentenças e as colunas às entidades. Por entidade se entende uma classe de sintagmas nominais correferentes. Para cada entidade, as células correspondentes da grade contêm informações sobre sua presença/ausência na sequência de sentenças, bem como informações sobre as suas funções sintáticas. Dessa forma, cada célula da grade é preenchida com uma letra representando se a entidade em questão aparece na função de sujeito ($S$), objeto ($O$) ou nenhuma das anteriores ($X$). A ausência de uma entidade na sentença é sinalizada pelo símbolo ($-$). A Figura 1(b) mostra a grade de entidades gerada para o texto de duas sentenças mostrado em (a).

Uma transição é uma sequência $\{S, O, X, -\}_n$ que representa as ocorrências de uma entidade e suas funções sintáticas em $n$ sentenças adjacentes. As transições podem ser obtidas a partir da grade de entidades como subsequências contínuas de cada coluna e possuem uma certa probabili-

dade de ocorrência na grade. Dessa forma, cada texto pode ser representado por um conjunto fixo de transições e suas probabilidades, usando a notação padrão de vetor de características. A Figura 2 exemplifica o vetor de características para dois documentos $d_1$ e $d_2$ considerando-se as transições de tamanho dois.

1. *[The Justice Department]S is conducting an [anti-trust trial]O against [Microsoft Corp.]X with [evidence]X that [the company]S is increasingle attempting to crush [competitors]O.*
2. *[Microsoft]O is accused of trying to forcefully buy into [markets]X where [its own products]S are not competitive enough to unseat [established brands]O.*

(a)

| | Department | Trial | Microsoft | Evidence | Competitors | Markets | Products | Brands |
|---|---|---|---|---|---|---|---|---|
| 1 | S | O | S | X | O | - | - | - |
| 2 | - | - | O | - | - | X | S | O |

(b)

Figura 1: (a) Exemplo de sentenças com anotações sintáticas e (b) grade de entidades correspondente – adaptado de Barzilay e Lapata (2008).

Outro aspecto incorporado ao modelo é a saliência. A saliência de uma entidade é definida com base na frequência de ocorrência da entidade no texto. Por exemplo, entidades mencionadas duas ou mais vezes são consideradas salientes. A partir dessa informação, uma grade apenas com entidades salientes pode ser construída e as probabilidades das transições podem ser calculadas separadamente para cada classe de saliência.

Vários trabalhos buscaram estender o modelo grade de entidades. Filippova e Strube (2007) modificaram o processo de seleção de entidades correferentes usando medidas de similaridade semântica em vez de resolução de correferência. Os experimentos foram realizados com textos jornalísticos escritos em alemão. Yokono e Okumura (2010) estenderam o modelo original visando sua aplicação para a língua japonesa por meio da adição de atributos baseados em mecanismos coesivos. A representação das entidades na grade por meio de funções sintáticas também foi refinada pela adição de marcadores de tópico específicos da língua japonesa. Os experimentos foram realizados com textos jornalísticos escritos em japonês. Burstein, Tetreault e Andreyev (2010) combinaram o modelo grade de entidades com atributos relacionados à qualidade de escrita, como erros gramaticais, uso de vocabulário e estilo, visando aplicar o modelo em redações (*essays*) escritas em inglês por estudantes de perfis variados. Também foram utilizados atributos do tipo *Type/Token* para medir a variedade léxica das entidades que ocorrem em cada função sintática. Elsner e Charniak (2011) estenderam o modelo por meio da adição de atributos entidade-específicos que buscam distinguir entre entidades importantes e entidades menos importantes. Também modificaram o processo de identificação de entidades, reconhecendo todo substantivo ou nome próprio como uma entidade em vez de usar apenas os núcleos dos sintagmas nominais. Os experimentos foram realizados com textos jornalísticos escritos em inglês. Lin, Ng e Kan (2011) combinaram o modelo grade de entidades com relações discursivas semelhantes as da RST (Mann e Thompson, 1988). Desse modo, em vez de serem representadas na grade por suas funções sintáticas, as entidades são representadas pela relação retórica em que aparecem. Os experimentos foram realizados com textos jornalísticos escritos em inglês.

## 3 Modelo Grade de Entidades para o Português

O modelo grade de entidades para o português foi implementado segundo a proposta original de Barzilay e Lapata (2008). Para extrair as entidades foi construído um sistema de pré-processamento que utiliza o *parser* PALAVRAS (Bick, 2002) como ferramenta principal para a identificação dos sintagmas nominais (SNs). Processamento adicional foi realizado para desmembrar os SNs complexos identificados pelo *parser* em SNs simples, a partir dos quais as entidades puderam ser extraídas para a construção da grade de entidades. Diferentemente do modelo original, não foi utilizado um resolvedor automático de correferência. Neste trabalho, a identificação de entidades seguiu uma abordagem similar a de Elsner e Charniak (2011), em que apenas sintagmas nominais que possuem o mesmo núcleo são considerados correferentes. Adicionalmente, para diminuir a duplicação de entidades, os SNs foram lematizados e agrupados por lemas antes de serem incluídos na grade. Uma visão geral das etapas de processamento para a construção da grade de entidades e extração do vetor de característica é mostrada na Figura 3.

| | SS | SO | SX | S- | OS | OO | OX | O- | XS | XO | XX | X- | -S | -O | -X | - - |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $d_1$ | .01 | .01 | 0 | .08 | .01 | 0 | 0 | .09 | 0 | 0 | 0 | .03 | .05 | .07 | .03 | .59 |
| $d_2$ | .02 | .01 | .01 | .02 | 0 | .07 | 0 | .02 | .14 | .14 | .06 | .04 | .03 | .07 | 0.1 | .36 |

Figura 2: Vetores de características para transições de tamanho dois dadas as categorias sintáticas {S, O, X, -} (Barzilay e Lapata, 2008).

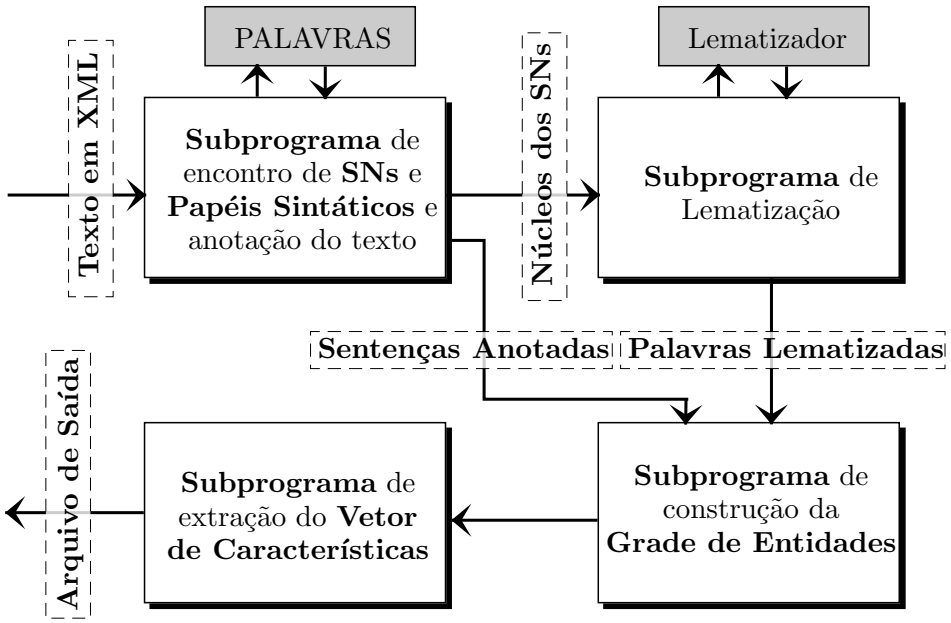


Figura 3: Etapas de processamento para a construção da grade de entidades e extração do vetor de características.

A partir da grade de entidades, o vetor de características é extraído de acordo com a configuração escolhida para o modelo. No modelo original, as configurações possíveis são definidas por `Correferência[+/-] Sintático[+/-] Saliência[+/-]`, representando a consideração (+) ou não (−) de tal conhecimento na construção do vetor. No caso do modelo implementado neste trabalho, como não foi empregada resolução de correferência, as configurações só variam nos aspectos sintático e saliência, sendo portanto representadas por `Sintático[+/-] Saliência[+/-]`.

Na configuração `Sintático+`, o vetor de características contém as probabilidades de todas as transições possíveis considerando-se as funções sintáticas $S$, $O$, $X$, $-$. No modelo original, o tamanho da transição é um parâmetro que pode ser ajustado conforme necessário. Neste trabalho foram consideradas apenas as transições de tamanho dois, uma vez que esse é o tamanho de transição comumente utilizado por outros trabalhos.

Barzilay e Lapata (2008) explicam que várias classes de saliência podem ser consideradas na configuração `Saliência+`. No entanto, como o tamanho do vetor de características aumenta conforme aumenta o número de classes de saliência, é comum usar apenas duas classes: entidades salientes e não salientes. Como no modelo original, neste trabalho foram consideradas salientes as entidades mencionadas duas ou mais vezes. Assim, na configuração `Saliência+`, a grade de entidades é dividida em duas – uma para entidades salientes e outra para entidades não salientes. As probabilidades das transições são computadas separadamente para cada grade e depois incluídas no vetor de características.

Além dos atributos previstos no modelo original, neste trabalho também foram extraídos atributos do tipo *Type/Token* (TT) semelhantes aos utilizados por Burstein, Tetreault e Andreyev (2010), que buscam medir a variedade léxica das entidades que ocorrem em cada função sintática. Quando a configuração é `Sintático+`, quatro atributos TT são calculados: um para cada função sintática (S_TT, O_TT, X_TT), mais um para a combinação das três funções (SOX_TT). O atributo S_TT representa a proporção de entidades que aparecem como sujeito (S) em relação ao número total de sujeitos observados na grade de entidades. O mesmo tipo de proporção

é calculada para as outras funções sintáticas e para a combinação de todas as funções. Quando a configuração é `Sintático-`, apenas um atributo TT é calculado, representando o número de entidades diferentes na grade dividido pelo número de ocorrências das entidades nas sentenças.

## 4 Experimentos e Resultados

Para avaliar o modelo grade de entidades para o português foram realizados dois tipos de experimentos: (1) um experimento de ordenação de sentenças usando um córpus jornalístico e (2) um experimento de classificação baseado no julgamento de juízes humanos usando um córpus de resumos científicos. O experimento (1) buscou replicar os experimentos realizados para outras línguas, mais especificamente para o inglês (Barzilay e Lapata, 2008; Elsner e Charniak, 2011), para o alemão (Filippova e Strube, 2007) e para o japonês (Yokono e Okumura, 2010). O objetivo, nesse caso, foi validar a implementação feita neste trabalho, bem como avaliar se o comportamento do modelo aplicado à língua portuguesa é semelhante ao observado para outras línguas. O experimento (2) buscou avaliar o desempenho do modelo no contexto de um classificador capaz de detectar problemas de coerência local em resumos científicos, uma vez que a motivação para este estudo está na melhoria do módulo de análise de coerência da ferramenta SciPo e na implementação da Dimensão Quebra de Linearidade. Os experimentos (1) e (2) e os respectivos resultados são apresentados a seguir, nas Seções 4.1 e 4.2, respectivamente.

### 4.1 Experimento 1: Ordenação de Sentenças

Nesse experimento foi utilizado um córpus de 286 textos jornalísticos extraídos dos córpora CSTNews (Cardoso et al., 2011) – 136 textos –, Summ-it (Collovini et al., 2007) – 50 textos – e Temário (Rino e Pardo, 2007) – 100 textos. A preparação do córpus seguiu o mesmo procedimento descrito em Barzilay e Lapata (2008). Para cada texto foram geradas aproximadamente 20 versões sintéticas em que a ordem original das sentenças foi permutada aleatoriamente e assumiu-se que o texto com as sentenças na ordem original deve ser mais coerente que a maioria dos textos com as sentenças permutadas. Desse modo, os textos originais foram marcados como "sem problemas" de coerência e as versões permutadas como "com problemas". Como resultado foi obtido um conjunto de 5.720 pares {*texto_original, versão_permutada*} (286 textos × 20 versões permutadas), que foi separado aleatoriamente em conjuntos de treinamento (2/3 dos pares) e teste (1/3 dos pares). A descrição completa desse córpus está disponível em Freitas (2013).

Como em Barzilay e Lapata (2008), a ordenação de sentenças foi tratada como um problema de ranqueamento, em que o modelo é usado para ranquear diferentes versões de um mesmo texto, esperando que as versões mais coerentes fiquem no topo do *ranking*. Desse modo, para treinar e testar o modelo foi utilizado o sistema SVM$^{rank}$ (Joachims, 2006), que implementa o algoritmo SVM (*Support Vector Machine*) para problemas de ranqueamento.

Como *baseline* foi utilizado um modelo baseado em LSA semelhante ao implementado por Barzilay e Lapata (2008). Esse modelo estima um valor de coerência V para um texto T por meio da média dos valores de similaridade para todos os pares de sentenças de T. Para o cálculo da LSA foi utilizada a implementação de Souza e Feltrim (2013) e os dois córpus compilados para este trabalho foram utilizados na criação do espaço semântico.

A métrica de avaliação seguiu a de Barzilay e Lapata (2008), em que dadas todas as comparações entre pares, a acurácia é medida como a quantidade de predições corretas feitas pelo modelo dividida pelo número de pares existentes no conjunto de teste. Na Tabela 1 é mostrado o percentual de acertos da *baseline* (LSA) e do modelo grade de entidades com os atributos *Type/Token*, representado na tabela por suas quatro configurações possíveis (`Sintático[+/-] Saliência[+/-]`). Como os textos jornalísticos são provenientes de córpus diferentes, os resultados são mostrados considerando-se o córpus de origem dos textos originais, além dos resultados calculados para o córpus jornalístico como um todo (coluna "Todos juntos"). Os melhores resultados obtidos por cada modelo estão destacados em negrito.

Conforme pode ser observado na Tabela 1, o modelo grade de entidades superou a *baseline* em todos os casos, com exceção do córpus Temário, em que a *baseline* foi superior em 4%. De fato, os textos do córpus Temário são maiores do que os textos dos outros dois córpus (média de sentenças por texto: CSTNews e Summit ≈ 16; Temário ≈ 29) e isso pode ter influenciado o resultado da *baseline*, que é baseada na média de similaridade entre pares de sentenças. Embora não seja possível uma comparação direta, no

| Modelo | Cstnews | Summit | Temário | Todos juntos |
|---|---|---|---|---|
| LSA | 61,429 % | 56,000 % | **79,000 %** | 67,000 % |
| Sintático+ Saliência− | 64,000 % | 48,235 % | 60,455 % | 62,105 % |
| Sintático+ Saliência+ | **74,444 %** | 50,294 % | 59,242 % | 58,105 % |
| Sintático− Saliência− | 69,444 % | 63,824 % | **74,848 %** | **68,579 %** |
| Sintático− Saliência+ | 70,889 % | **72,059 %** | 65,455 % | 67,368 % |

Tabela 1: Percentual de acertos da *baseline* e do modelo grade de entidades para o experimento 1.

| Trabalho | Língua | Córpus | | Melhor resultado |
|---|---|---|---|---|
| Barzilay e Lapata (2008) | Inglês | 100 textos originais | (T) | 83,0 % |
| | | 100 textos originais | (A) | 89,9 % |
| Elsner e Charniak (2011) | Inglês | 1004 textos originais | | 84,0 % |
| Filippova e Strube (2007) | Alemão | 100 textos originais | | 69,0 % |
| Yokono e Okumura (2010) | Japonês | 100 textos originais | | 59,4 % |
| | | 300 textos originais | | 77,3 % |

Tabela 2: Resumo dos resultados obtidos por trabalhos relacionados.

geral, os resultados obtidos pelo modelo grade de entidades para o português são semelhantes aos relatados pelos trabalhos desenvolvidos para outras línguas, ficando abaixo apenas dos resultados obtidos pelo modelo original. A Tabela 2 apresenta um resumo dos melhores resultados relatados pelos trabalhos relacionados considerando-se sempre a configuração `Correferência-`.

Ainda na Tabela 1, vale observar que a variação na configuração `Sintático[+/-] Saliência[+/-]` não resultou em um padrão de resultados que permitisse a julgar sobre a melhor configuração, variando conforme o córpus utilizado. Esse mesmo comportamento pode ser observado nos trabalhos relacionados e no modelo original.

### 4.2 Experimento 2: Classificação Baseada no Julgamento Humano

Nesse experimento foi utilizado um córpus de 139 resumos científicos: 99 resumos extraídos do córpus de resumos de Trabalhos de Conclusão de Curso em Ciência da Computação compilado por Souza e Feltrim (2013), e 40 resumos experimentais coletados diretamente com os autores – alunos formandos do curso de graduação em Ciência da Computação da Universidade Estadual de Maringá. O córpus foi preparado para experimentos utilizando o julgamento humano acerca do nível de coerência dos resumos nos mesmos moldes do trabalho de Burstein, Tetreault e Andreyev (2010). Para a anotação manual, os anotadores foram instruídos a marcar o resumo como "com problemas", caso fossem encontradas barreiras na leitura (por exemplo, dificuldade de se estabelecer uma ligação semântica entre sentenças), caracterizando a quebra de linearidade; caso contrário, os anotadores foram instruídos a marcar o resumo como "sem problemas".

A concordância entre anotadores foi avaliada por meio de um experimento em que dois anotadores treinados anotaram separadamente os 40 resumos experimentais. A concordância medida por meio da estatística *Kappa* foi de 0,70%, valor próximo ao obtido por Burstein, Tetreault e Andreyev (2010) ($K$ = 0,68%) em experimento semelhante. O restante do córpus foi anotado por apenas um dos anotadores treinados no experimento. No total, a anotação manual resultou em 117 (84%) resumos marcados como "sem problemas" e 22 (16%) como "com problemas". Vale ressaltar que esse desbalanceamento é característico de córpus manualmente anotados e que o mesmo nível de desbalanceamento foi observado nos córpus de redações (*essays*) utilizados por Burstein, Tetreault e Andreyev (2010).

Nesse experimento, a tarefa foi modelada como um problema de classificação binária. O treinamento e o teste do modelo foram realizados no ambiente WEKA (Witten e Frank, 2005) utilizando-se os seguintes algoritmos de aprendizado de máquina: SMO – *Sequential Minimal Optimization* – (Platt, 1998), uma implementação do algoritmo SVM para classificação; J48, uma implementação em Java e de código aberto do algoritmo C4.5 (Quinlan, 1993) que gera árvores de decisão; e *Naïve Bayes*, um algoritmo probabilístico baseado na regra da probabilidade condicional de *Bayes*. A escolha pelo algoritmo SMO se deu por ele ser uma implementação de SVM, que é o algoritmo utilizado no Experimento 1; o J48 foi escolhido por ser uma implementação do

C4.5, que é o algoritmo de aprendizado utilizado por (Burstein, Tetreault e Andreyev, 2010); e o *Naïve Bayes* foi escolhido por ser um algoritmo de aprendizado simples, rápido e de larga utilização em tarefas que envolvem classificação textual.

Os resultados foram calculados aplicando-se 10-*fold cross-validation* ao córpus de 139 resumos. Os resultados em termos das medidas *F-measure* ($F_1$) e *Kappa* são mostrados para cada algoritmo de aprendizado e configuração do modelo na Tabela 3. Os valores de *F-measure* representam a média das *F-measures* calculadas para as duas classes, ponderada pelo número de exemplos de cada classe. Os resultados listados como TT+ foram calculados adicionando-se ao modelo os atributos do tipo *Type/Token*. Os resultados listados como TT− foram calculados utilizando apenas o modelo grade de entidades.

Conforme pode ser observado na Tabela 3, os melhores resultados foram obtidos com o algoritmo J48, sendo que o melhor resultado ($K = 0,65$) se aproxima do valor obtido pelos juízes humanos ($K = 0,70$) e ultrapassa o melhor sistema de Burstein, Tetreault e Andreyev (2010) ($K = 0,61$) em experimento semelhante. O algoritmo SMO apresentou os piores resultados. Também é possível observar que enquanto os valores de *F-measure* são relativamente altos (acima de 0,8 para o algoritmo J48), os valores da medida *Kappa* são mais baixos e apresentam maior variação entre os diferentes modelos. Isso pode ser atribuído ao forte desbalanceamento do córpus (84%/16%), que eleva o desempenho dos classificadores induzidos para a classe majoritária ("sem problema"), elevando por consequência os valores da medida *F-measure*. A medida *Kappa*, por sua vez, prioriza os acertos para a classe minoritária, em que a probabilidade de acerto "ao acaso" é menor, fornecendo assim uma medida mais realista do desempenho do classificador nesse contexto de desbalanceamento. Os resultados completos, detalhados por algoritmo de aprendizado e por classe, expressos nas cinco medidas de avaliação utilizadas, estão disponíveis em Freitas (2013).

Analisando as diferentes configurações do modelo com base no algoritmo J48, fica evidente a contribuição do aspecto saliência. O melhor resultado ($K = 0,65$) foi obtido com a configuração `Sintático- Saliência+` e o segundo melhor ($K = 0,52$) com a configuração `Sintático+ Saliência+`. Curiosamente, neste caso, o modelo mais simples, que não considera a função sintática das entidades (`Sintático-`), se saiu melhor do que o modelo mais rico (`Sintático+`). De fato, esse comportamento também foi observado por (Filippova e Strube, 2007) e pode ser atribuído ao tamanho reduzido do córpus de treinamento. Uma vez que a configuração `Sintático+` gera um vetor de características quatro vezes maior que a configuração `Sintático-`, um número maior de exemplos de treinamento pode ser necessário para que o modelo possa se beneficiar das informações relativas ao aspecto sintático. Quanto aos atributos *Type/Token* (TT), observa-se que a sua inclusão no vetor de características teve pouca influência nos resultados, sendo que, em alguns casos, os valores com TT+ permaneceram iguais aos valores com TT−. Uma discreta contribuição dos atributos TT pode ser notada nos resultados calculados com o algoritmo *Naïve Bayes*.

Na Tabela 4, os resultados obtidos com o melhor modelo (`Sintático- Saliência+` treinado com J48) são detalhados por classe e comparados com os obtidos por uma *baseline* simples que classifica todos os textos como "sem problemas". Como pode ser observado, o modelo é superior a *baseline* para as duas classes.

Para avaliar o efeito do desbalanceamento do córpus nos resultados, os experimentos com os três algoritmos de aprendizado foram refeitos utilizando-se a técnica de balanceamento SMOTE (Chawla et al., 2002) – *Synthetic Minority Oversampling Technique* – (Chawla et al., 2002), também disponível no WEKA (Witten e Frank, 2005), a qual realiza *oversampling*, isto é, adiciona ao conjunto novos casos da classe minoritária gerados sinteticamente a partir de casos já existentes. Esses novos casos são gerados na vizinhança de cada caso da classe minoritária. Segundo Chawla et al. (2002), esse método produz resultados melhores do que a simples replicação de casos existentes, uma vez que essa prática pode levar a modelos muito específicos, prejudicando o poder de generalização do modelo (*overfitting*).

A Tabela 5 apresenta os resultados após a classe minoritária ter sido aumentada em 400%, valor que deixa o córpus com um balanceamento próximo ao perfeito. Assim como na Tabela 3, os resultados são mostrados em termos das medidas *F-measure* e *Kappa* para cada algoritmo de aprendizado e configuração do modelo grade de entidades.

Conforme pode ser observado na Tabela 5, os resultados usando *oversampling* foram melhores para os três algoritmos testados, sendo que o J48 continuou apresentando o melhor resultado, especialmente em termos da

| | Naïve Bayes | | SMO | | J48 | |
|---|---|---|---|---|---|---|
| **TT−** | $F_1$ | **Kappa** | $F_1$ | **Kappa** | $F_1$ | **Kappa** |
| Sintático+ Saliência− | 0,663 | 0,211 | 0,769 | 0,000 | 0,810 | 0,256 |
| Sintático+ Saliência+ | 0,741 | 0,053 | 0,802 | 0,144 | 0,882 | 0,515 |
| Sintático− Saliência− | 0,707 | 0,211 | 0,769 | 0,000 | 0,804 | 0,183 |
| Sintático− Saliência+ | 0,799 | 0,168 | 0,766 | -0,014 | **0,910** | **0,650** |
| **TT+** | | | | | | |
| Sintático+ Saliência− | 0,731 | 0,262 | 0,766 | -0,014 | 0,809 | 0,271 |
| Sintático+ Saliência+ | 0,770 | 0,114 | 0,802 | 0,144 | 0,876 | 0,494 |
| Sintático− Saliência− | 0,740 | 0,223 | 0,769 | 0,000 | 0,804 | 0,183 |
| Sintático− Saliência+ | 0,799 | 0,168 | 0,797 | 0,127 | **0,910** | **0,650** |

Tabela 3: Resultados do modelo grade de entidades para o experimento 2.

| | Sem Problema (117) | | | Com problema (22) | | | Média ponderada (139) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | $F_1$ | **Precision** | **Recall** | $F_1$ | **Precision** | **Recall** | $F_1$ |
| Melhor modelo | 0,934 | 0,966 | 0,950 | 0,778 | 0,636 | 0,700 | 0,909 | 0,914 | 0,910 |
| *Baseline* | 0,842 | 1,000 | 0,914 | 0,000 | 0,000 | 0,000 | 0,708 | 0,841 | 0,769 |

Tabela 4: Melhor modelo (`Sintático- Saliência+` treinado com J48) vs. *baseline*.

medida *Kappa*. A configuração `Sintático- Saliência+` continuou sendo a melhor e a contribuição dos atributos *_TT ficou mais evidente, elevando drasticamente o valor *Kappa* da configuração `Sintático- Saliência+` para os algoritmos *Naïve Bayes* e SMO. Vale destacar que enquanto o valor da *F-measure* ponderada teve pouca variação em relação aos valores sem *oversampling* (Tabela 3), o valor da medida *Kappa* melhorou significativamente, especialmente para os algoritmos *Naïve Bayes* e SMO. Isso se deve ao fato da medida *Kappa* refletir de forma mais apropriada o desempenho nas duas classes consideradas. A variação do valor da medida *Kappa* de acordo com o percentual de *oversampling* da classe minoritária é mostrada na Figura 4.

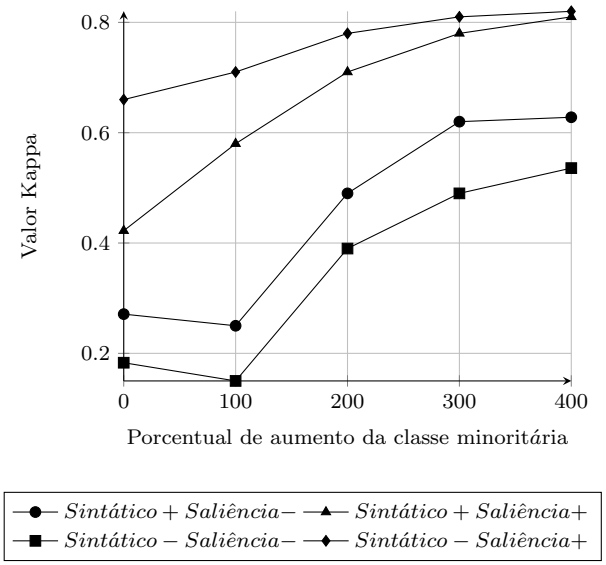


Figura 4: Variação dos valores da medida *Kappa* de acordo com o percentual de *oversampling* (SMOTE).

Quando se considera as medidas *Precision*, *Recall* e *F-measure* para cada classe, é possível notar que os resultados com *oversampling* ficaram mais uniformes, uma vez que os valores para a classe minoritária melhoraram, alcançando valores semelhantes aos obtidos para a classe majoritária. A Tabela 6 mostra os resultados em termos de *Precision*, *Recall* e *F-measure* para cada classe. Os valores sem *oversampling* são mostrados na primeira metade da tabela e os valores com *oversampling* são mostrados na segunda parte.

Com o objetivo de observar a influência do tamanho do córpus nos resultados foi realizado um experimento em que se aumentou artificialmente e gradativamente o tamanho do córpus. Para isso, foi utilizado novamente o algoritmo SMOTE sobre o córpus já balanceado pelo *oversampling*. Nesse caso, como o córpus já estava balanceado, a cada execução o SMOTE selecionava aleatoriamente uma das classes para a criação de novos exemplos. Dessa forma, para cada vez que o tamanho foi aumentado, o algoritmo foi aplicado duas vezes para que o balanceamento fosse mantido.

Os resultados desse experimento foram calculados para dois cenários usados nos experimentos anteriores: (1) o modelo na configuração `Sintático- Saliência+ *_TT+` treinado e testado com o J48, por ser o cenário que apresentou os melhores resultados, e (2) o modelo na configuração `Sintático- Saliência+ *_TT−` treinado e testado com o SMO, por ser o cenário que

| | **Naïve Bayes** | | **SMO** | | **J48** | |
|---|---|---|---|---|---|---|
| ***_TT−** | $F_1$ | **Kappa** | $F_1$ | **Kappa** | $F_1$ | **Kappa** |
| Sintático+ Saliência− | 0,718 | 0,460 | 0,725 | 0,478 | 0,806 | 0,612 |
| Sintático+ Saliência+ | 0,762 | 0,525 | 0,830 | 0,663 | 0,904 | 0,808 |
| Sintático− Saliência− | 0,631 | 0,294 | 0,670 | 0,383 | 0,769 | 0,544 |
| Sintático− Saliência+ | 0,615 | 0,290 | 0,587 | 0,253 | 0,912 | 0,824 |
| ***_TT+** | | | | | | |
| Sintático+ Saliência− | 0,772 | 0,554 | 0,832 | 0,667 | 0,806 | 0,612 |
| Sintático+ Saliência+ | 0,797 | 0,596 | 0,802 | 0,144 | 0,904 | 0,808 |
| Sintático− Saliência− | 0,706 | 0,433 | 0,729 | 0,466 | 0,764 | 0,536 |
| Sintático− Saliência+ | 0,890 | 0,780 | 0,868 | 0,735 | 0,916 | 0,833 |

Tabela 5: Resultados do modelo grade de entidades para o córpus de resumos científicos balanceado com SMOTE em termos de *F-measure* e *Kappa*.

**Sem *oversampling***

| | **Sem Problema (117)** | | | **Com problema (22)** | | |
|---|---|---|---|---|---|---|
| *_TT− | **Precision** | **Recall** | $F_1$ | **Precision** | **Recall** | $F_1$ |
| Sintático+ Saliência− | 0,877 | 0,915 | 0,895 | 0,412 | 0,318 | 0,359 |
| Sintático+ Saliência+ | 0,906 | 0,975 | 0,939 | 0,769 | 0,455 | 0,571 |
| Sintático− Saliência− | 0,862 | 0,957 | 0,907 | 0,444 | 0,182 | 0,258 |
| Sintático− Saliência+ | 0,934 | 0,966 | 0,950 | 0,778 | 0,636 | 0,700 |
| *_TT+ | | | | | | |
| Sintático+ Saliência− | 0,882 | 0,897 | 0,890 | 0,400 | 0,364 | 0,381 |
| Sintático+ Saliência+ | 0,906 | 0,966 | 0,935 | 0,714 | 0,455 | 0,556 |
| Sintático− Saliência− | 0,862 | 0,957 | 0,907 | 0,444 | 0,182 | 0,258 |
| Sintático− Saliência+ | 0,934 | 0,966 | 0,950 | 0,778 | 0,636 | 0,700 |

**Com *oversampling* (SMOTE)**

| | **Sem Problema (117)** | | | **Com problema (110)** | | |
|---|---|---|---|---|---|---|
| *_TT− | **Precision** | **Recall** | $F_1$ | **Precision** | **Recall** | $F_1$ |
| Sintático+ Saliência− | 0,812 | 0,812 | 0,812 | 0,800 | 0,800 | 0,800 |
| Sintático+ Saliência+ | 0,915 | 0,899 | 0,907 | 0,893 | 0,909 | 0,901 |
| Sintático− Saliência− | 0,849 | 0,675 | 0,752 | 0,716 | 0,873 | 0,787 |
| Sintático− Saliência+ | 0,929 | 0,897 | 0,913 | 0,895 | 0,927 | 0,911 |
| *_TT+ | | | | | | |
| Sintático+ Saliência− | 0,807 | 0,821 | 0,814 | 0,806 | 0,791 | 0,798 |
| Sintático+ Saliência+ | 0,915 | 0,899 | 0,907 | 0,893 | 0,909 | 0,901 |
| Sintático− Saliência− | 0,848 | 0,667 | 0,746 | 0,711 | 0,873 | 0,784 |
| Sintático− Saliência+ | 0,922 | 0,915 | 0,918 | 0,910 | 0,918 | 0,914 |

Tabela 6: Resultados usando o J48 sem e com *oversampling* (SMOTE).

apresentou os piores resultados. Como medidas de avaliação foram utilizadas a medida Kappa e a acurácia (percentagem de acerto), uma vez que o aumento gradativo do córpus buscou mantê-lo balanceado. O gráfico mostrando os resultados para os dois cenários é apresentado na Figura 5.

Como esperado, o aumento do tamanho do córpus influenciou positivamente os valores de acurácia e *Kappa* nos dois cenários avaliados, porém de maneiras diferentes. Conforme pode ser observado na Figura 5, os resultados calculados para o cenário (1) – melhor cenário – permaneceram os mesmos após o balanceamento (227), aumentaram significativamente no primeiro aumento de tamanho (454) e se estabilizaram a partir desse ponto. Já os resultados para o cenário (2) – pior cenário – aumentaram de forma acentuada e contínua desde o balanceamento e pelos consecutivos aumentos de tamanho, começando a estabilizar a partir do terceiro aumento no tamanho (908). Ainda sim, os resultados para o cenário (2) permaneceram abaixo dos resultados para o cenário (1) em todas as avaliações. Esse experimento confirma o que já havia sido observado por Barzilay e Lapata (2008), que a partir de um certo número de exemplos – e considerando-se o balanceamento entre as

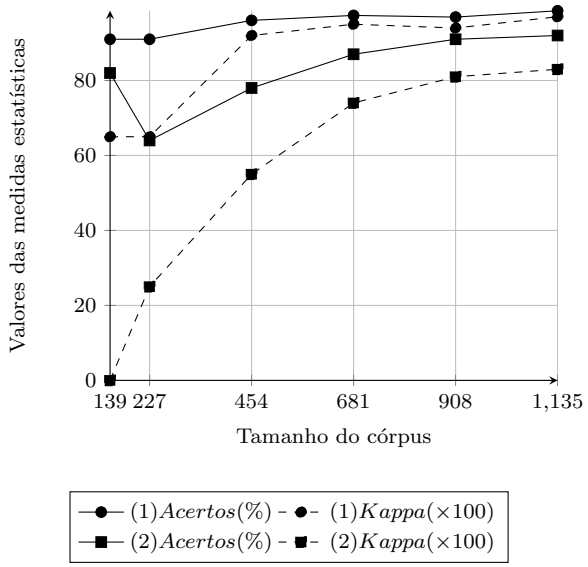


Figura 5: Variação dos valores de acurácia e medida *Kappa* de acordo com o aumento artificial do tamanho do córpus.

classes – o desempenho do modelo grade de entidades se estabiliza. Esses resultados também reforçam a configuração utilizada no cenário (1) como a melhor configuração para esse córpus, independentemente do seu tamanho.

## 5 Conclusões e Trabalhos Futuros

Este trabalho teve por objetivo avaliar o modelo grade de entidades proposto por Barzilay e Lapata (2008) na avaliação de coerência em textos científicos. A motivação está em encontrar um modelo de coerência capaz de mapear o fluxo textual de forma mais refinada do que o modelo baseado em LSA proposto por Souza e Feltrim (2013), visando melhorar os resultados obtidos no âmbito da detecção de quebras de linearidades entre sentenças adjacentes de um resumo.

O modelo grade de entidades para o português foi implementado segundo a proposta original, com exceção do tratamento automático de correferências, que não foi realizado neste trabalho. Além dos atributos previstos no modelo original, neste trabalho também foram avaliados atributos do tipo *Type/Token* de forma similar a realizada por Burstein, Tetreault e Andreyev (2010).

A avaliação do modelo foi feita de dois modos. Primeiramente buscou-se reproduzir o mesmo cenário de testes empregado pelos trabalhos encontrados na literatura. Isso permitiu a comparação, ainda que indireta, dos resultados deste trabalho com os obtidos para outras línguas, mostrando que os resultados são próximos aos relatados para a língua inglesa e superiores ao relatado para a língua japonesa e alemã. Em um segundo momento, buscou-se avaliar o modelo na tarefa de avaliação de coerência em resumos científicos. Os resultados mostraram que o uso do modelo grade de entidades é viável nesse contexto. O melhor resultado ($K = 0,65$), alcançado com o algoritmo J48 com o modelo na configuração `Sintático- Saliência+`, é próximo ao obtido por juízes humanos ($K = 0,70$) na classificação dos resumos usando duas classes: "sem problemas"/"com problemas".

Um desdobramento natural deste trabalho é a aplicação efetiva do modelo grade de entidades no módulo de análise de coerência do sistema SciPo, possibilitando a avaliação extrínseca do modelo no contexto de uma ferramenta de auxílio à escrita científica. Para isso é preciso encontrar formas de se mapear os resultados do modelo em um *feedback* que seja útil ao usuário do sistema. Também pretende-se avaliar o modelo no contexto de outras aplicações que possam se beneficiar de um modelo de coerência, como é o caso da sumarização automática. Outra linha de trabalhos futuros aborda a melhoria dos resultados obtidos com modelo grade de entidades por meio da combinação do modelo original com conhecimentos provenientes de outras fontes, como os índices calculados pela Coh-Metrix-Port (Scarton e Aluísio, 2010; Scarton, Almeida e Aluísio, 2009). A inclusão de um sistema de resolução automática de correferência no modelo atual também será explorada em trabalhos futuros, já que os melhores resultados da literatura foram obtidos utilizando-se esse tipo de conhecimento.

## Agradecimentos

Os autores agradecem a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro na realização deste trabalho.

## Referências

Barzilay, Regina e Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34:1–34, March, 2008.

Bick, Eckhard. 2002. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. Tese de doutoramento, Department of Linguistics – Aarhus: Aarhus University Press – DK.

Burstein, Jill, Martin Chodorow, e Claudia Leacock. 2003. Criterion online essay evaluation: An application for automated evaluation of student essays. Em *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, pp. 3–10.

Burstein, Jill, Joel Tetreault, e Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. Em *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 681–684.

Cardoso, Paula Christina Figueira, Erick Galani Maziero, Maria Lucia del Rosario Castro Jorge, Eloize Rossi Marques Seno, Ariani Di Felippo, Lúcia Helena Machado Rino, Maria das Graças Volpe Nunes, e Thiago Alexandre Salgueiro Pardo. 2011. Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. Em *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88–105, Cuiabá/MT, Brazil.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, e W. Philip Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Collovini, Sandra, Thiago Ianez Carbonel, Juliana Thiesen Fuchs, Jorge César Barbosa Coelho, Lúcia Helena Machado Rino, e Renata Vieira. 2007. Summ-it: um corpus anotado com informações discursivas visando sumarização automática. Em *Anais do XXVII Congresso da SBC: V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2007)*.

Elliot, Scott. 2003. Intellimetric: From here to validity. Em *Shermis, M.; Burstein, J., eds. Automatic Essay Scoring: A Cross-Disciplinary Perspective.*, pp. 71–86, Hillsdale, NJ. Lawrence Erlbaum Associates.

Elsner, Micha e Eugene Charniak. 2011. Extending the entity grid with entity-specific features. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pp. 125–129.

Feltrim, Valéria Delisandra, Simone Teufel, Maria das Graças Volpe Nunes, e Sandra Maria Aluísio. 2006. Argumentative zoning applied to criquing novices scientific abstracts. Em James G. Shanahan, Yan Qu, e Janyce Wiebe, editores, *Computing Attitude and Affect in Text: Theory and Applications*, pp. 233–246, Dordrecht, The Netherlands. Springer.

Filippova, Katja e Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. Em *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pp. 139–142.

Freitas, Alison Rafael Polpeta. 2013. *Análise Automática de Coerência Usando o Modelo Grade de Entidades para o Português*. Dissertação de Mestrado, Departamento de Informática – Universidade Estadual de Maringá, Maringá/PR - Brasil.

Grosz, Barbara J., Scott Weinstein, e Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225.

Higgins, Derrick, Jill Burstein, Daniel Marcu, e Cláudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. Em *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 185–192.

Joachims, Thorsten. 2006. Training linear svms in linear time. Em *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pp. 217–226, New York, NY, USA. ACM.

Landauer, Thomas K., Peter W. Foltz, e Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25:259–284.

Landauer, Thomas K., Darrell Laham, e Peter W. Foltz. 2003. *Automated essay scoring and annotation of essays with the intelligent essay assessor.* Lawrence Erlbaum Associates, Mahwah, NJ.

Lin, Ziheng, Hwee Tou Ng, e Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 997–1006.

Mann, William C. e Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Platt, John. 1998. Fast training of support vector machines using sequential minimal optimization. Em *B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning.*

Quinlan, Ross. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA.

Rino, Lúcia Helena Machado e Thiago Alexandre Salgueiro Pardo. 2007. *A coleção TeMário e a avaliação de sumarização automática*, volume 1. IST Press, Lisboa, Portugal.

Scarton, Carolina Evaristo, Daniel Machado de Almeida, e Sandra Maria Aluísio. 2009. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. Em *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*. 1 CD-ROM v1.

Scarton, Carolina Evaristo e Sandra Maria Aluísio. 2010. Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. Em *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings*, PROPOR '10. 1 CD-ROM v1.

Souza, Vinícius Mourão Alves e Valéria Delisandra Feltrim. 2011. An analysis of textual coherence in academic abstracts written in portuguese. Em *Proceedings of the Sixth Corpus Linguistics Conference (CL 2011)*, pp. 1–13, Birmingham, UK.

Souza, Vinícius Mourão Alves e Valéria Delisandra Feltrim. 2013. A coherence analysis module for scipo: providing suggestions for scientific abstracts written in portuguese. *Journal of the Brazilian Computer Society*, 19:59–73.

Witten, Ian H. e Eibe Frank. 2005. *Data Mining - Practical Machine Learning Tools and Techniques.* Morgan Kaufmann – Elsevier.

Yokono, Hikaru e Manabu Okumura. 2010. Incorporating cohesive devices into entity grid model in evaluating local coherence of japanese text. Em Alexander Gelbukh, editor, *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010)*, number 6008 in Lecture Notes in Computer Science. Springer Berlin/Heidelberg, pp. 303–314.

# NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de *Conditional Random Fields*

**A tool for the named entity recognition using conditional random fields**

Daniela Oliveira F. do Amaral
Pontifícia Universidade Católica
do Rio Grande do Sul
daniela.amaral@acad.pucrs.br

Renata Vieira
Pontifícia Universidade Católica
do Rio Grande do Sul
renata.vieira@pucrs.br

**Resumo**

*Conditional Random Fields* (CRF) é um método probabilístico de predição estruturada que tem sido amplamente aplicado em diversas áreas, tais como a de Processamento da Linguagem Natural (PLN), incluindo o Reconhecimento de Entidades Nomeadas (REN), visão computacional e bioinformática. Nesse sentido, propõe-se a realização da tarefa de REN aplicando o método CRF e, sequencialmente, é feita uma avaliação do seu desempenho com base no corpus do HAREM. Conclui-se que, nos testes realizados, o sistema NERP-CRF obteve os melhores resultados de Precisão quando comparado com os sistemas avaliados no mesmo corpus, com plenas condições de ser um sistema competitivo e eficaz.

**Palavras chave**

Reconhecimento de Entidades Nomeadas, Conditional Random Fields, Processamento da Linguagem Natural, Língua Portuguesa.

**Abstract**

Conditional Random Fields (CRF) is a probabilistic method for structured prediction which has been widely applied in various areas such as Natural Language Processing (NLP), including the Named Entity Recognition (NER), computer vision, and bioinformatics. Therefore, this paper proposes to perform the task of applying the method CRF NER and an evaluation of its performance based on the corpus of HAREM. In summary, the system NERP-CRF achieved the best Precision results when compared to the systems evaluated in the same corpus, proving to be a competitive and effective system.

**Keywords**

Named Entity Recognition, Conditional Random Fields, Natural Language Processing.

## 1 Introdução

A Extração da Informação (EI) é uma importante tarefa na mineração de texto e tem sido amplamente estudada em vários grupos de pesquisa, incluindo nos de processamento da linguagem natural, de recuperação de informação e de mineração na Web. O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa primordial na área de EI, juntamente com a extração de relação entre Entidades Nomeadas (EN) (Jing, 2012).

Dentro desse contexto, o REN em textos tem sido amplamente estudado por meio de métodos como aprendizagem supervisionada para classificar entidades do tipo Pessoa, Lugar e Organização em textos ou, ainda, doenças e genes nos resumos das áreas médicas e biológicas (Chinchor, 1994). Esses métodos dependem de recursos caros e extensos para a etiquetagem manual, a qual realiza a identificação das entidades. Os dados etiquetados e o conjunto de *features* extraídas automaticamente são então usados para treinar modelos tais como os Modelos de Markov de Máxima Entropia (MEMMs) (McCallum, 2000) ou *Conditional Random Fields* (Lafferty, 2001).

Os MEMMs são modelos de uma sequência probabilística condicional, (McCallum, 2000), em que cada estado inicial tem um modelo exponencial que captura as características de observação e a distribuição sobre os próximos estados possíveis. Esses modelos exponenciais são treinados por um método apropriado de dimensionamento iterativo no *framework* de máxima entropia.

O modelo denominado *Conditional Random Fields* (CRF) é um *framework* de modelagem de sequência de dados, que tem todas as vantagens do MEMM e, além disso, resolve o problema a partir do viés dos rótulos. A diferença crítica entre CRF e MMEM é que o MMEM utiliza modelos exponenciais por estados para as probabilidades condicionais dos próximos estados, dado o estado atual. Já o CRF tem um modelo exponencial único para uma probabilidade conjunta de uma sequência de entrada de rótulos, dada uma sequência de observação. Portanto, as influências das diferentes

características em estados distintos podem ser tratadas independentemente umas das outras (Lafferty, 2001). Os resultados da Conferência Internacional de Aprendizado de Máquina (*International Conference on Machine Learning* - ICML) no ano de 2001 (LAF01), seguido de outros trabalhos sobre *Conditional Random Fields* (Suakkaphong, 2011), (Lee, 2011), (Lishuang, 2011), indicam que o algoritmo de CRF apresenta um dos melhores desempenhos para o REN.

Sendo assim, o método escolhido foi o CRF e o corpus que receberá a classificação por ele é o do HAREM. O HAREM é um evento de avaliação conjunta da língua portuguesa, organizado pela Linguateca (Santos, 2007). Seu objetivo é o de realizar a avaliação de sistemas reconhecedores de ENs (Santos, 2009). Entre as edições do HAREM temos: o Primeiro HAREM, ocorrido no ano de 2004, e o Segundo HAREM, em 2008. A Coleção Dourada (CD) é um subconjunto da coleção do HAREM, sendo utilizada para tarefa de avaliação dos sistemas que tratam REN. O corpus do HAREM é considerado a principal referência na área de PLN, e caracteriza-se por ter um conjunto de textos anotados e validados por humanos (CD), o que facilita a avaliação do método em estudo.

Com isso, este trabalho teve como motivação o fato de: (i) o REN ter sido pouco explorado utilizando o método de aprendizagem supervisionada CRF para a língua portuguesa; (ii) não existirem propostas de REN aplicando o CRF para identificar as ENs e classificá-las de acordo com as dez categorias do HAREM; e (iii) o método de CRF poder ajudar a identificar um maior número de ENs, o que poderá ser verificado por meio da comparação com outros sistemas.

Portanto, o objetivo geral do presente artigo é utilizar o aprendizado de máquina, ou seja, aplicar CRF para a tarefa de REN em corpora da língua portuguesa e avaliar comparativamente o desempenho desse método com outros sistemas que realizam REN, tendo como base o corpus do HAREM.

Este artigo é estruturado como segue: a Seção 2 elucida o assunto REN e CRF. A Seção 3 expõe uma revisão dos trabalhos relacionados à pesquisa proposta. A Seção 4 descreve o desenvolvimento do sistema NERP-CRF, sua modelagem, implementação e o processo de avaliação. A Seção 5 apresenta os resultados obtidos. Sequencialmente, a análise de erros é efetuada na Seção 6. Por fim, a Seção 7 aponta as conclusões e os trabalhos futuros.

## 2 Reconhecimento de Entidades Nomeadas e Conditional Random Fields

O REN consiste na tarefa de identificar as ENs, na sua maioria nomes próprios, a partir de textos de forma livre e classificá-las dentro de um conjunto de tipos de categorias pré-definidas, tais como Pessoa, Organização e Local, as quais remetem a um referente específico (Mota, 2007). Adicionalmente, o REN em textos que abordam os mais variados domínios, além do emprego de extração de relações entre ENs, é uma das tarefas primordiais dentro do trabalho de EI.

Segundo Sureka (2009), o REN é uma técnica amplamente utilizada no PLN e consiste na identificação de nomes de entidades-chave presentes na forma livre de dados textuais. A entrada para o sistema de extração de entidade nomeada é um texto de forma livre, e a saída é um conjunto desses textos anotados, ou seja, uma representação estruturada a partir da entrada de um texto não estruturado.

As três principais abordagens para extração de ENs são: sistemas baseados em regras, sistemas baseados em aprendizado de máquina e abordagens híbridas. Sistemas baseados em regras ou sistemas baseados no conhecimento consistem em definir heurísticas na forma de expressões regulares ou de padrões linguísticos. Sistemas baseados em aprendizado de máquina utilizam algoritmos e técnicas que permitam ao computador aprender.

Já o CRF são modelos matemáticos probabilísticos, baseados numa abordagem condicional, utilizados com o objetivo de etiquetar e segmentar dados sequenciais (Lafferty, 2001). O CRF é uma forma de modelo grafo não direcionado que define uma única distribuição logaritmicamente linear sobre sequências de rótulos, dada uma sequência de observação particular. A vantagem primária dos modelos de CRF sobre outros formalismos, como os *Hidden Markov Model* (HMM) (Lafferty, 2001), é a sua natureza condicional, pois resulta no abrandamento de pressupostos sobre a independência dos estados, necessários para os modelos HMM, a fim de assegurar uma inferência tratável.

## 3 Trabalhos Relacionados

Os trabalhos de Sutton e McCallum (2005), Lafferty (2001) e Chatzis e Demiris (2012), apresentam um *framework* para a construção de modelos probabilísticos para segmentação e etiquetagem de dados sequenciais baseados em CRF. Nesse sentido, temos assistido, durante os últimos anos, a uma explosão de vantagens nos modelos de CRF (Chatzis, 2012), à medida que tais

modelos conseguem alcançar uma previsão de desempenho excelente em uma variedade de cenários. Sendo assim, o processamento de texto por meio da técnica de aprendizado de máquina CRF, é uma das abordagens de maior sucesso para o problema de predição de saída estruturada, com aplicações bem sucedidas em áreas como a bioinformática e o processamento da linguagem natural (PLN). Não obstante, o trabalho de Ratinov e Roth (2009) aponta que o REN pode ser obtido a partir de modelos de classes de palavras, os quais aprendem a partir de rótulos não estruturados. Os autores investigaram a aplicação de REN a partir da necessidade de usar o conhecimento prévio e decisões não locais para a identificação de tais ENs em um texto. Logo, esse modelo que detecta e classifica ENs pode ser uma alternativa para o paradigma de aprendizado supervisionado tradicional como o CRF.

Existem diversos trabalhos que também usam CRF e outras abordagens estatísticas para extração de informação textual em PLN e, especificamente, para a tarefa de REN (Finkel, 2005; McCallum e Li 2003).

Sendo assim, a importância de aplicar o CRF para o REN, especialmente, em textos da língua portuguesa deve-se ao fato de que essa técnica de aprendizado de máquina possibilita a extração automática de EN a partir de um grande conjunto de dados com uma capacidade de resposta mais rápida do que outras técnicas já utilizadas, como a implantação de heurísticas ou de sistemas baseados em regras. Além disso, o CRF tem sido muito pouco explorado em corpora do nosso idioma, uma vez que trabalhos que visam o processo de identificação e de classificação de EN para o português são raros na literatura. O sistema *Hendrix* (Batista, 2010), por exemplo, foi elaborado com o propósito de extrair entidades geográficas de documentos em português e produzir o seu resumo geográfico. O processo dividiu-se em três partes: (i) reconhecer Entidades Geográficas em um documento, ou seja, nomes de ruas, rios, serras, utilizando CRF; (ii) desambiguar significados geográficos a fim de eliminar nomes idênticos aos extraídos dos textos; (iii) gerar um resumo geográfico por meio da criação de uma lista de entidades geográficas descoberta em uma base de conhecimento externa, por exemplo, em uma ontologia.

Tanto quanto o sistema *Hendrix*, os sistemas Priberam ao HAREM, R3M, REMBRANDT, SEI-Geo e CaGE realizam REN para textos da língua portuguesa (Mota, 2008). Com exceção do Hendrix os demais sistemas participaram da trilha do Segundo HAREM e foram comparados com o sistema que desenvolvemos para este trabalho. O Priberam ao HAREM é baseado em um léxico com classificação morfossintática e semântica. Cada entrada do léxico corresponde a uma ligação com um ou mais níveis de uma ontologia multilíngue (Amaral, 2004), podendo corresponder a um ou mais sentidos, os quais possuem diferentes valores morfológicos e semânticos. Para a construção do sistema foram utilizadas regras contextuais. As regras para a tarefa de REN consideram as sequências de nomes próprios, separadas ou não por algumas preposições e o contexto em que as EN são encontradas. Por exemplo, uma EN "João Pedro", classificada como Pessoa, poderá ser classificada como Organização se esta for precedida por uma expressão como "instituto".

Já o R3M aplica aprendizagem supervisionada, utilizando um algoritmo de co-training para inferir regras de classificação (Collins, 1999) no REN. A escolha do algoritmo de *co-training* deve-se ao fato de que este tem grande probabilidade de obter bons resultados de classificação que se aproximam dos 80% de precisão, usando um número muito reduzido de exemplos previamente anotados. As ENs que o R3M classifica compreendem as categorias Pessoa, Organização e Local. A opção por essas três categorias deve-se ao fato de que essas, de uma forma geral, têm sido estudadas mais amplamente dentro da área de extração da informação. Além disso, os desenvolvedores do R3M não tiveram disponibilidade de dedicar mais tempo a esse sistema. Mesmo assim, o R3M foi projetado de modo que permita estender-se ao reconhecimento de outras categorias, assim como incluir o reconhecimento de relações de EN. Esse sistema é uma reimplementação do sistema criado por Mota (Mota, 2009), apresentando várias melhorias.

Além da tarefa de REN realizada pelos sistemas REMBRANDT e SEI-Geo, ambos detectam o Reconhecimento de Relações entre ENs (ReRelEN). O REMBRANDT - Reconhecimento de ENs Baseado em Relações e Análise Detalhada do Texto - por sua vez, utiliza a Wikipédia como base de conhecimento a fim de classificar as ENs, além de um conjunto de regras gramaticais para extrair o seu significado. O REMBRANDT surgiu da necessidade de se criar um sistema de marcação de textos que indique as ENs relacionadas a locais geográficos de forma semântica, como por exemplo, nomes de países, rios, universidades. Seu funcionamento divide-se em três fases primordiais: 1) o reconhecimento de expressões numéricas e geração de candidatas a EN; 2) a classificação de EN e 3) repescagem de ENs sem classificação. Já o SEI-Geo tem o objetivo de fazer REN classificando somente a categoria Local e suas relações. Dentre as características que compõem o SEI-Geo

destacam-se: (i) a incorporação na arquitetura global do sistema *GKBGeographic Knowlegde Base*, o qual estabelece o gerenciamento de conhecimento geográfico; e (ii) a utilização das Geo-ontologias, que exploram as relações entre locais identificados em textos a partir de relações presentes na ontologia. O domínio Organização ajudou significativamente o bom desenvolvimento do SEI-Geo no reconhecimento de relações entre ENs, pois, nos textos, Locais estão situados próximos a Organizações.

Por fim, o sistema CaGE trata do problema do reconhecimento e desambiguação de nomes de locais. Essa é uma tarefa muito importante na geo-codificação de documentos textuais (Martins, 2009). O objetivo principal do sistema CaGE é atribuir a área geográfica e o âmbito temporal aos documentos de modo geral, combinando a informação diferente extraída do texto. As categorias que tal sistema classifica são: Pessoa, Local, Organização e Tempo. O CaGe caracteriza-se por ser um método híbrido, o qual utiliza dicionários e regras de desambiguação. Quatro etapas resumem uma sequência de operações de processamento que compõem o algoritmo do sistema: 1) identificação inicial das ENs; 2) classificação das entidades mencionadas e tratamento da ambiguidade; 3) desambiguação completa de entidades geográficas e temporais; e 4) atribuição de âmbitos geográficos e temporais aos documentos.

Dessa forma, o nosso trabalho difere dos demais na aplicabilidade do modelo de CRF, o qual vem demonstrando bons resultados frente a outros métodos que utilizam aprendizado de máquina para a tarefa de REN. Além disso, a literatura apresenta muitos poucos trabalhos que identificam e classificam ENs, utilizando as dez categorias do HAREM, em corpus da língua portuguesa por meio de modelos probabilísticos.

## 4 NERP-CRF

Esta seção descreve o desenvolvimento do sistema NERP-CRF (Amaral, 2013) desde o pré-processamento dos textos, o modelo gerado pelo CRF para o REN até a avaliação empregada.

### 4.1 Modelagem do Sistema

A elaboração do modelo consiste em duas etapas: treino e teste. Dessa forma, adotamos um corpus que é dividido em um conjunto de textos para treino e um conjunto de textos para teste. A CD do HAREM foi o corpus utilizado para tarefa de avaliação dos sistemas que tratam REN. As ENs foram identificadas e classificadas por todos os sistemas participantes do evento, sendo que a sua classificação foi dividida em categorias, tipos e subtipos. Destacam-se para essa pesquisa dez categorias: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro.

Os textos, utilizados como entrada para o NERP-CRF, estão no formato XML com a marcação das entidades e sofreram dois procedimentos, os quais pertencem ao pré-processamento do sistema: primeiro, a etiquetagem de cada palavra por meio do *Part-of-Speech* (POS) *tagging* (Schmid, 1994) (Bick, 2000) e segundo, a segmentação em sentenças a fim de que a complexidade seja menor ao aplicar o algoritmo de CRF nos textos de entrada.

Após a conclusão da etiquetagem POS e da segmentação das sentenças, determinou-se como as ENs seriam identificadas. Para tal, foi feito um estudo de duas notações citadas na literatura: BIO e BILOU (Ratinov, 2009). A primeira possui o seguinte significado: B (*Begin*) significa a primeira palavra da EN; I (*Inside*) uma ou mais palavras que se localizam entre as entidades; e O (*Outside*) a palavra não é uma EN. Já a segunda notação tem a mesma descrição do BIO, acrescentando-se as seguintes particularidades: L (*Last*) a última palavra reconhecida como EN e U (*Unit*) quando a EN for uma única palavra.

Para o presente trabalho, utilizou-se a notação BILOU por dois motivos: (i) testes aplicados sob a CD do Segundo HAREM, empregando ambas as notações, demonstraram que a notação BILOU se equivale à BIO, conforme os resultados apresentados. Isso porque o BILOU facilita o processo de classificação feito pelo sistema desenvolvido por possuir mais duas identificações: L(*Last*) e U(*Unit*); e (ii) os autores (Ratinov, 2009) também fizeram testes com as duas notações, concluindo também que, apesar do formalismo BIO ser amplamente adotado, o BILOU o supera significativamente.

Depois da identificação das EN por meio do BILOU, foi gerado o vetor de *features*. Tal vetor corresponde aos dados de entrada que serão aplicados ao sistema de aprendizado do CRF. As *features* têm o objetivo de caracterizar todas as palavras do corpus escolhido para esse processo, direcionando o CRF na identificação e na classificação das ENs. A Tabela 1 apresenta a lista de *features* criadas.

| Features | Descrição das features |
|---|---|
| 1) tag | Etiqueta POS de cada palavra de acordo com a sua classe gramatical. Ex.: artigo, adjetivo, verbo. |
| 2) word | A própria palavra do texto, ignorando letras maiúsculas e minúsculas; |
| 3) prevW | A palavra anterior a que está sendo analisada no texto, ignorando letras maiúsculas e minúsculas. |
| 4) prevT | Classe gramatical da palavra anterior. Ex.: artigo, adjetivo, verbo. |
| 5) prevCap | A palavra anterior totalmente formada por letras minúsculas, formada por letras minúsculas e maiúsculas ou por letras maiúsculas. Cada uma dessas palavras pode receber um dos atributos: 'min', 'maxmin' ou 'max'. |
| 6) prev2W | Igual a *feature* 3, porém considerando a palavra que está na posição p-2; |
| 7) prev2T | O mesmo que a *feature* 4, considerando a palavra que está na posição p-2; |
| 8) prev2Cap | Igual a *feature* 5, porém considerando a palavra que está na posição p-2; |
| 9) nextW | A palavra subsequente àquela que está sendo analisada, ignorando maiúsculas e minúsculas; |
| 10) nextT | A classe gramatical da palavra subsequente à que está sendo analisada; |
| 11) nextCap | O mesmo que a *feature* 5, levando em consideração a palavra subsequente àquela que está sendo analisada; |
| 12)next2W, next2T, next2Cap | Semelhante as *features* 3, 4 e 5, mas para a palavra na posição p + 2; |
| 13) cap | O mesmo que a *feature* 5, mas para palavra atual que está sendo analisada; |
| 14) ini | Se a palavra iniciar com letra maiúscula, minúscula ou símbolos. Essas palavras podem receber um dos atributos: 'max', 'min' ou 'sim'. |
| 15) simb | Caso a palavra seja composta por símbolos, dígitos ou letras. Tais palavras recebem o atributo 'alfa'. |

Tabela 1: Features implementadas no NERP-CRF.

Dois vetores são considerados como entrada para o CRF na etapa de treino: primeiro, o vetor contendo a etiquetagem POS, as categorias estabelecidas pela Conferência do HAREM e a notação BILOU, e segundo, o vetor de *features*.

Na etapa de teste um conjunto de textos é enviado ao NERP-CRF. O referido sistema (a) cria o vetor de POS e o vetor de *features*; (b) envia esses vetores para o modelo de CRF gerado que, por sua vez, (c) treina e (d) classifica as ENs do corpus trabalhado. Por fim, são apresentados aos usuários do sistema as ENs extraídas e as métricas precisão e abrangência. O sistema é concluído com o vetor de saída, o qual classifica o texto com a notação BILOU e com as dez categorias conforme o Segundo HAREM.

### 4.2 Descrição dos Testes Realizados

Dois testes foram realizados utilizando o sistema NERP-CRF, com as seguintes características:

'Teste 1': empregou a CD do Segundo HAREM para treinar e testar o modelo de CRF, o qual faz a classificação de dez categorias. A avaliação do desempenho do modelo treinado para o 'teste 1' utilizou a técnica de *Cross Validation* (Arlot, 2010), com cinco repetições (*5 – fold cross validation)*. Trabalhou-se com 5 *folds* porque foi empregada uma pequena quantidade de textos, 129, para os testes iniciais, incluindo 670.610 palavras. Esse procedimento resultou em 7.610 ENs identificadas pelo NERP-CRF num valor máximo de 17.767 ENs identificadas por humanos nessa mesma CD. Dado o conjunto de textos da CD do Segundo HAREM, utilizou-se, a cada *fold,* 80% do conjunto de textos para treino e 20% para teste, de modo que, a cada repetição do *Cross Validation,* não se empregasse o mesmo conjunto de teste das *folds* anteriores e, assim, não reduzisse significativamente o número de casos para teste. A finalidade de executar esse experimento foi para verificar o desempenho do NERP-CRF utilizando apenas o corpus citado.

'Teste 2': caracteriza-se por trabalhar com a CD do Primeiro HAREM para treino, a qual abrange 129 textos, e a CD do Segundo HAREM para teste, formada por mais 129 textos. Os dois conjuntos somam 258 textos e aproximadamente 804.179 palavras. O novo corpus recebe a classificação do CRF abordando as dez categorias do HAREM, citadas no "Teste 1". Essa estrutura foi arquitetada com o objetivo de verificar o desempenho do NERP-CRF em um maior número de textos e avaliá-lo perante os resultados obtidos por ele com os outros sistemas participantes do Segundo HAREM.

## 5 Resultados

A comparação dos resultados do NERP-CRF com os sistemas que participaram da Conferência do Segundo HAREM foram obtidos por meio do SAHARA (Mota, 2008), o qual determinou as métricas Precisão, Abrangência e Medida-F a cada um deles nas tarefas de REN.

O NERP-CRF, no 'Teste 1', apresentou os melhores resultados para as medidas de Precisão e de Medida-F em relação aos outros sistemas, respectivamente, 83,48% e 57,92% (Tabela 2) (Figura 1). Esse resultado é baseado em um único corpus para treino e teste, apesar de validá-lo com *Cross-validation*.

| Sistemas | Precisão | Abrangência | Medida-F |
|---|---|---|---|
| NERP-CRF | 83,48% | 44,35% | 57,92% |
| Priberam | 64,17% | 51,46% | 57,11% |
| Rembrandt | 64,97% | 50,36% | 56,74% |
| R3M | 76,44% | 25,20% | 37,90% |
| CaGE | 44,99% | 27,57% | 34,19% |
| SEI-Geo | 74,85% | 11,66% | 20,17% |

Tabela 2: Resultados do NERP-CRF comparado com os sistemas apresentados para o 'Teste 1'.

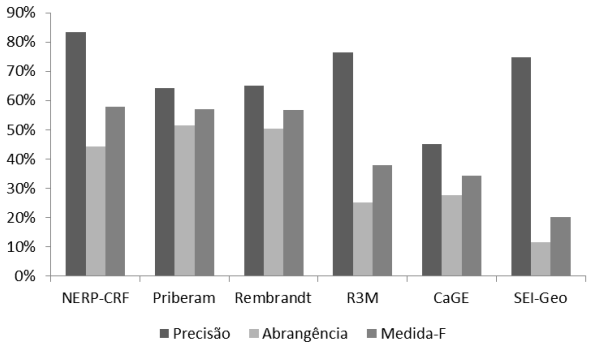


Figura 1: Desempenho do NERP-CRF comparado graficamente com os sistemas no 'Teste 1'.

Com a finalidade de ver o comportamento do aprendizado em outro corpus, realizamos o 'Teste 2', o qual apresentou 80,77% de Precisão como o melhor resultado do NERP-CRF (Tabela 3). A Medida-F ocupou a terceira posição em relação aos sistemas em comparação, 48,43%. Essa última métrica não alcançou a melhor posição como no 'Teste 1' devido a uma baixa Abrangência de classificação, 34,59% (Figura 2).

A desigualdade dos resultados entre os dois testes ocorreu, principalmente, por dois motivos: a mudança do corpus de treino e de validação além do número reduzido de exemplos para determinadas categorias, por exemplo, Coisa e Abstração. Isso fez com que o NERP-CRF treinasse menos com essas categorias e gerasse um modelo menos abrangente. Nesse cenário, consideram-se os nossos resultados muito positivos, principalmente no que tange ao valor de Precisão alcançado pelo NERP-CRF.

| Sistemas | Precisão | Abrangência | Medida-F |
|---|---|---|---|
| Priberam | 64,17% | 51,46% | 57,11% |
| Rembrandt | 64,97% | 50,36% | 56,74% |
| NERP-CRF | 80,77% | 34,59% | 48,43% |
| R3M | 76,44% | 25,20% | 37,90% |
| CaGE | 44,99% | 27,57% | 34,19% |
| SEI-Geo | 74,85% | 11,66% | 20,17% |

Tabela 3: Resultados do NERP-CRF comparado com os sistemas apresentados para o 'Teste 2'.

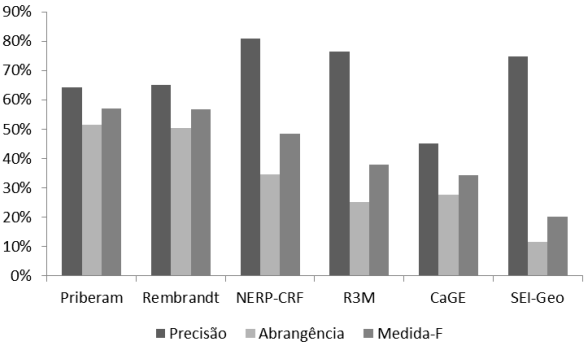


Figura 2: Desempenho do NERP-CRF comparado graficamente com os sistemas no 'Teste 2'.

A Tabela 4 reporta os resultados das métricas de Precisão, Abrangência e Medida-F de acordo com as dez categorias estabelecidas pelo evento do HAREM. Logo, pode-se observar de acordo com os valores da Medida-F, que o NERP-CRF classificou melhor as categorias Tempo, Pessoa, Valor e Local. Em contra partida, houve um maior número de erros com as categorias Abstração, Outra, Coisa e Acontecimento, devido a poucos exemplos no corpus de treino.

| CATEGORIAS | Abrangência | Precisão | Medida-F |
|---|---|---|---|
| TEMPO | 68,05% | 83,99% | 75,18% |
| PESSOA | 71,89% | 61,57% | 66,33% |
| VALOR | 54,42% | 78,23% | 64,19% |
| LOCAL | 57,22% | 52,06% | 54,51% |
| ORGANIZACAO | 43,44% | 44,75% | 44,08% |
| OBRA | 28,48% | 40,71% | 33,52% |
| ACONTECIMENTO | 22,76% | 50,83% | 31,44% |
| COISA | 7,36% | 26,80% | 11,55% |
| OUTRA | 4,74% | 43,49% | 8,55% |
| ABSTRAÇÃO | 3,65% | 16,45% | 5,97% |

Tabela 4: Resultados com as dez categorias.

## 6 Análise de Erros

Com base em uma análise dos textos utilizados como entrada para testar o NERP-CRF, constata-se que o sistema, tanto para o 'Teste 1' quanto para o 'Teste 2', não identificou determinadas ENs ou não as classificou corretamente. A Tabela 4 apresenta

alguns erros encontrados após a execução do NERP-CRF. A notação apresentada (Tabela 5) pela saída desse sistema refere-se ao POS tagger de cada EN, seguido da notação BILOU e da classificação dessas entidades. Por exemplo, substantivo, etiquetado como <n>; preposição <prp>; nome próprio <prop>; verbo finito <v-fin>; numeral <num>; artigo <art>. Quanto à identificação e a classificação das ENs o NERP-CRF apresentou a notação conforme alguns exemplos: <I-Obra> EN identificada como Inside e classificada como Obra; <L-Obra> Last e Obra; < B-Pessoa> Begin e classificação Pessoa; <U - Org> Unit e Organização. As ENs que não foram identificadas e não receberam classificação, foram marcadas pelo sistema como: <O-OUT>.

Percebeu-se que a má formatação de alguns textos, como por exemplo, a falta de pontuação e a anotação incorreta pelo POS tagger afetaram os resultados. A delimitação errônea de ENs, como em "Diário de Notícias", marcado pelo NERP-CRF como O I L, mas identificado pelo corpus de referência como B I L, prejudicou também o resultado do sistema. Outro erro em destaque foi a não identificação da preposição 'de' e de suas combinações com artigos, como I (Inside), no caso de ENs compostas, como "Fenando de Bulhões" e "Igreja dos Mártires". Esses erros podem ser sanados com a aplicação de algoritmos de classificação como o de Viterbi (Finkel, 2005), abordagem utilizada em ferramentas com propósitos similares (FreeLing User Manual, 2013). Outra alternativa seria o AdaBoosting (Carreras, 2003).

| NERP-CRF | CD do HAREM |
|---|---|
| **Diário** <n, O-OUT><br>**de** <prp, I-Obra><br>**Notícias** <n, L-Obra> | **Diário**<n, B-Org><br>**de** <prp, I-Org><br>**Notícias** <n, L-Org> |
| **Fernando**<prop,B-Pessoa><br>**de** <prp, O-Out<br>**Bulhões** <prop, L-Local> | **Fernando**<prop,B-Pessoa><br>**de** <prp, I-Pessoa><br>**Bulhões** <prop, L-Pessoa> |
| **Igreja** <v-fin, O -OUT><br>**dos** <n, O - OUT><br>**Mártires**<prop,U-Pessoa> | **Igreja** <v-fin, B -Local><br>**dos** <n, I - Local><br>**Mártires** <prop, L -Local> |
| **RF** <prop, U- Org> | **RF** <prop, U-Coisa> |
| **IFF** <prop,U- Org> | **IFF** <prop, U-Coisa> |
| **Friendly**<br><prop, U-Local> | **Friendly**<br><prop,U- Abstração> |
| **em** <prp, O-OUT><br>**1973** <num, U-Tempo> | **em** <prp, B-Tempo<br>**1973** <num, L-Tempo> |
| **desde** <prp, O-OUT><br>**os** <art, I-Tempo><br>**anos** <n, I-Tempo><br>**1990** <num, L-Tempo> | **desde** <prp, B-Tempo><br>**os** <art, I-Tempo><br>**anos** <n, I-TempoO><br>**1990** <num, L-Tempo> |

Tabela 5: Alguns erros apresentados pelo NERP-CRF.

Outro ponto relevante foram os erros de classificação das ENs. Podemos citar as siglas "RF" e "IFF", consideradas como ENs, as quais deveriam ter sido classificadas como "Coisa", porém o sistema considerou-as como "Organização". As palavras estrangeiras sofreram o mesmo tipo de erro, como a EN "Friendly" que foi classificada como "Local", ao passo que deveria ter recebido "Abstração" como classificação correta. Percebeu-se também que houve pouco contexto para classificar corretamente certas ENs, como ocorreu com a categoria "Abstração", a qual tem pouca exemplificação no corpus de referência. Além disso, são ENs que não seguem padrão algum de escrita, ou seja, não há uma sintaxe própria para essa categoria que faça com que o sistema aprenda corretamente a identificá-la. Já a categoria "Tempo" apresenta-se num formato que a identifica com mais clareza, isto é, possui um padrão bem rígido de sintaxe como <um número> de <outro número>, indicando data, ou até mesmo outras palavras indicativas de tempo como "desde", "enquanto" e "quando". Mesmo assim, o sistema teve dificuldade de classificá-la, pois esse tipo de EN pode não iniciar com letra maiúscula, o que prejudicou o aprendizado feito pelo NERP-CRF. Por exemplo, na EN "em 1973", a preposição "em" não foi identificada como EN. O correto seria que o NERP-CRF a tivesse classificado como B-Tempo. Situação semelhante também ocorreu com outra EN de Tempo, "desde os anos 1990". O sistema não reconheceu a preposição "desde" como EN e, consequentemente, não a classificou.

## 7 Conclusões e Trabalhos Futuros

CRF oferece uma combinação única de propriedades: modelos treinados para etiquetar e segmentar sequências de dados, combinação de arbitrariedade, *features* de observação aglomeradas, decodificação e treinamento eficiente baseado em programação dinâmica e estimativa de parâmetro garantida para encontrar o ótimo global (Lafferty, 2001) (Ratinov, 2009).

O NERP-CRF foi o sistema desenvolvido para executar duas funções: a identificação de ENs e a classificação dessas com base nas dez categorias do HAREM: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro.

Dois testes foram realizados. Um deles utilizou a CD do Segundo HAREM para treino e teste, obtendo Medida-F de 57,92%. Outro teste empregou a CD do Primeiro HAREM para treinar o modelo de CRF e a CD do Segundo HAREM para testar o mesmo modelo gerado. Nesse caso, as métricas obtidas foram: 80,77% de Precisão, 34,59% de Abrangência e 48,43% de Medida-F. A

Precisão foi o melhor resultado quando comparado com os outros sistemas. Já a Medida-F apresentou o terceiro melhor resultado, ficando abaixo dos sistemas Priberam e Rembrandt, que apresentaram maior abrangência. O modelo proposto, baseado em CRF e no conjunto de *features* estabelecidas, gerou um sistema eficaz, competitivo, sendo ainda passível de fácil adaptação e modificação.

A análise de erros mostrou que o NERP-CRF precisa melhorar a identificação e a classificação das EN. Dentre os erros que ocorreram, aqueles mais frequentes foram: marcação pela notação BILOU, erros de classificação entre as categorias Local e Pessoa, classificação de sigla e de palavras estrangeiras, identificação e classificação de EN de Tempo.

Tomando por base a análise de erros, sugere-se um trabalho futuro com experimentos que utilizem algoritmos de meta aprendizagem, como combinação de classificadores, para aumentar a efetividade do NERP-CRF. Resultados interessantes baseados em anotações BIO foram obtidos com o uso de AdaBoosting (Carreras et al. 2003). A atual versão do NERP-CRF já utiliza anotações BILOU, logo, acredita-se que tanto a abrangência como a precisão do processo proposto possa ser melhorada com esse tipo de abordagem. Especificamente, busca-se melhorar a qualidade da anotação BILOU, induzir *features* e classificar ENs consideradas ambíguas. Adicionalmente, sugere-se também experiências com outros *parsers* e eventual comparações com o desempenho obtido para outras línguas.

Acredita-se que o teste com outros *parsers* possibilitará um melhor resultado de Abrangência pelo NERP-CRF. O FreeLing (Padró, 2010) e o PALAVRAS (Bick, 2000) são os *parsers* que serão utilizados para etiquetar o mesmo corpus empregado na fase de pré-processamento.

O CRF pode implementar, eficientemente, a seleção de *features* e de algoritmos de indução de *features*. Isso quer dizer que, em vez de especificar antecipadamente quais *features* serão utilizadas, pode-se iniciar a partir de regras que geram *features* e avaliam o benefício dessas geradas automaticamente sobre os dados (Lafferty, 2001).

Outra abordagem de pesquisa futura é a classificação correta de uma mesma EN apresentada de formas diferentes, por exemplo: a EN 'Pontifícia Universidade Católica do Rio Grande do Sul' pode receber a mesma classificação ou ser categorizada como Organização e Local, dependendo do contexto no qual essas entidades estão inseridas. Outra situação que pode ocorrer é que ENs que possuem como acrônimo, 'Pontifícia Universidade Católica do Rio Grande do Sul' e 'PUCRS' devem ser identificadas como a mesma entidade. Portanto, essas devem receber a mesma classificação. As soluções para a correta categorização de ENs, nesse caso, pode ser a aplicabilidade, como da Correferência (Black, 1998) (Lee, 2011) e de recursos externos, como o emprego de *Gazetters* (Ratinov, 2009).

## Referência


Amaral, C.; Figueira, H.; Mendes, A.; Mendes, P.; Pinto, C. 2004. A workbench for developing natural language processing tools. In: *1st Workshop on International Proofing Tools and Language Technologies*, Patras, Greece, July 1-2.

Amaral, D.O.F. 2012. O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa. M.Sc. dissertation, PUCRS.

Arlot, S.; Celisse, A. 2010. A survey of cross-validation procedures for model selection. Statistics Surveys, vol. 04, p. 4, 40.

Batista, S.; Silva, J.; Couto, F. e Behera, B. 2010 Geographic Signatures for Semantic Retrieval, In: *Proceedings of the 6th Workshop on Geographic Information Retrieval*, ACM, p.18-19.

Bick, E. 2000. The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. *Aarhus University Press.*

Black, W. J., Rinaldi, F. e Mowatt, D. 1998. Facile: Description of the NE system used for MUC-7, In: *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

Carreras, X.; Màrquez, L.; Padró, L. 2003. A simple named entity extractor using adaboost. In *Proceedings of CoNLL-2003 Shared Task Edmonton, Canada*.

Chinchor, N.; Hirschman, L. e Lewis, D. 1994. Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3), In: *Computational Linguistics*, p. 409-449.

Chatzis, Sotirio P. e Demiris, Yiannis. 2012. The echo state conditional random field model for sequential data modeling. In: *International Journal of Expert Systems with Applications.*

Collins, M.; Singer, Y. 1999. Unsupervised models for named entity classification. In: *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, p.100–110.

Finkel, Jenny R.; Grenager, T.; Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL*, p. 363–370.

FreeLing User Manual, October 2013. In *http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf*

Jing, J. (2012) "Information extraction from text", In *Mining Text Data*, p. 11-41.

Lafferty, J.; McCallum, A. e Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *Proceedings of the 18th International Conference on Machine Learning.*

Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M. e Jurafsky, D. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In: *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task,* p. 28-34.

Lishuang L.; Degen H.; Dan L. 2011. Recognizing Chinese Person Names based on Hybrid Models. *Advanced Intelligence*, vol. 3: 219-228.

Mansouri, A.; Affendey, Lilly S. e Mamat, A. 2008. Named Entity Recognition Approache**,** In *International Journal of Computer Science and Network Security*, vol. 8 Nº.2.

Martins, B. 2009. Geographically aware Web text mining. Tese de Doutorado, Faculdade de Ciências, Universidade de Lisboa, p. 155-157.

McCallum, A.; Freitag, D. e Pereira, F. 2000. Maximum entropy Markov models for information extraction and segmentation. In: *International Conference on Machine Learning.*

McCallum, A.; Li, W. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of 7th conference on natural language learning, CoNLL.*

Mota, C.; Santos, D. e Ranchhod, E. 2007. Avaliação de reconhecimento de entidades mencionadas: Princípio de Harem. In: *Diana Santos, editor, Avaliação Conjunta: Um novo paradigma no processamento computacional da língua portuguesa*, capítulo 14, IST Press, p. 161–176.

Mota, C. e Santos, D. 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. http://www.linguateca.pt/LivroSegundoHAREM/, Dezembro.

Mota, C. 2009. How to keep up with language dynamics: A case study on named entity recognition. Tese de doutoramento, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Padró, L.; Collado, M.; Reese, S.; Lloberes, M.; Castellon, I. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference, LREC, ELRA, La Valletta, Malta.*

Ratinov, L.; Roth, D. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In: *13th Conference on Computational Natural Language Learning*, CONLL, p. 147-155.

Santos, D.; Cardoso, N. 2007. Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, capítulo 1, p. 1–16.

Santos, D.; Cabral, L. M. 2009. GikiCLEF: Crosscultural issues in an international setting: asking non-english-centered questions to wikipedia. *Cross Language Evaluation Forum: Working notes for CLEF*.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In: *International Conference on New Methods in Language Processing*, 1994, p. 44-49.

Suakkaphong, N.; Zhang, Z.; Chen, H. 2011. Disease Named Entity Recognition Using Semisupervised Learning and Conditional Random Fields. *Journal of the American Society for Information Science and Technology*, vol. 62, p. 727-737.

Sureka, Ashish S.; Pranav, P. M.; Kishore, I. V. 2009. Polarity Classification of Subjective Words Using Common-Sense Knowledge-Base. In: *12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, p. 486-493.

# Artigos de Investigação

# Realização de Previsões com Conteúdos Textuais em Português

**Making Predictions with Textual Contents in Portuguese**

Indira Mascarenhas Brito Bruno Martins
Instituto Superior Técnico, INESC-ID
{indira.brito,bruno.g.martins}@tecnico.ulisboa.pt

**Resumo**

A previsão de quantidades do mundo real com base em informação textual atraiu recentemente um interesse significativo, embora os estudos anteriores se tenham concentrado em aplicações que envolvem apenas textos em inglês. Este artigo apresenta um estudo experimental sobre a realização de previsões com base em textos em português, envolvendo o uso de documentos associados a três domínios distintos. Relatamos experiências utilizando diferentes tipos de modelos de regressão, usando esquemas de ponderação para as caraterísticas descritivas do atual estado da arte, e usando caraterísticas descritivas derivadas de representações para as palavras baseadas no agrupamento automático das mesmas. Através de experiências, demonstramos que modelos de regressão usando a informação textual atingem melhores resultados, quando comparados com abordagens simples tais como realizar as previsões com base no valor médio dos dados de treino. Demonstramos ainda que as representações de documentos mais ricas (e.g., usando o algoritmo de Brown para o agrupamento automático de palavras, e o esquema de ponderação das caraterísticas denominado Delta-TF-IDF) resultam em ligeiras melhorias no desempenho.

**Palavras chave**

Previsões com Base em Textos, Modelos de Regressão, Agrupamento Automático de Palavras, Engenharia de Caraterísticas em Aplicações de PLN

**Abstract**

Forecasting real-world quantities, from information on textual descriptions, has recently attracted significant interest as a research problem, although previous studies have focused on applications involving only the English language. This paper presents an experimental study on the subject of making predictions with textual contents in Portuguese, using documents from three distinct domains. We specifically report on experiments using different types of regression models, using state-of-the-art feature weighting schemes, and using features derived from cluster-based word representation. Our experiments show that regression models using the textual information achieve better results than simple baselines such as the average value in the training data, and that richer document representations (i.e., using Brown clusters and the Delta-BM25 feature weighting scheme) results in slight performance improvements.

**Keywords**

Text-Driven Forecasting, Learning Regression Models, Word Clustering, Feature Engineering for NLP

## 1 Introdução

A realização de previsões com base em textos atraiu recentemente um interesse significativo nas comunidades internacionais de Extração de Informação, Recuperação de Informação, Aprendizagem Automática, e Processamento de Língua Natural (Smith, 2010; Radinsky, 2012). Exemplos bem conhecidos de estudos anteriores incluem o uso de conteúdos textuais para fazer previsões sobre o comportamento de mercados financeiros (Luo, Zhang e Duan, 2013; Lerman et al., 2008; Tirunillai e Tellis, 2012; Schumaker e Chen, 2009; Bollen, Mao e Zeng, 2011), os resultados de mercados de apostas desportivas (Hong e Skiena, 2010), padrões de vendas de produtos e serviços (Chahuneau et al., 2012; Joshi et al., 2010), eleições governamentais, atividades legislativas e inclinações políticas no geral (Yano, Smith e Wilkerson, 2012; Dahllöf, 2012), ou sondagens de opinião pública em diversos temas (Mitchell et al., 2013; O'Connory et al., 2010; Schwartz et al., 2013).

Este trabalho apresenta um estudo experimental no âmbito da realização de previsões com base em textos em português, usando documentos de três domínios distintos, nomeadamente (i) descrições de hotéis em Portugal recolhidos desde um portal Web[1], associadas aos preços médios dos quartos nas épocas altas e baixas para os turistas, (ii) descrições de restaurantes

[1] http://www.lifecooler.com

e os menus correspondentes, também recolhidos do mesmo portal Web, associados aos preços médios das refeições, e (iii) comentários sobre filmes recolhidos a partir de um site Web especializado[2], em conjunto com os respetivos resultados de bilheteira para a primeira semana de exibição, tal como disponibilizados pelo Instituto do Cinema e do Audiovisual[3]. O nosso estudo incidiu sobre o uso de métodos de aprendizagem automática do atual estado-da-arte (e.g., regressão com florestas aleatórias, ou regressão linear com regularização dada pela método da rede elástica), implementados numa biblioteca Python para aprendizagem automática, chamada *scikit-learn*[4]. Além da questão dos conteúdos em português, o nosso estudo também apresenta algumas novidades técnicas em relação à maior parte do trabalho anterior na área, nomeadamente através de (i) experiências com esquemas de ponderação de caraterísticas do atual estado-da-arte, como o Delta-TF-IDF ou o Delta-BM25, e (ii) experiências com caraterísticas derivadas de representações com base no algoritmo de Brown para o agrupamento automático de palavras.

O restante conteúdo do artigo está organizado da seguinte forma: a Seção 2 apresenta trabalhos anteriores relevantes. A Seção 3 detalha as principais contribuições do artigo, apresentando as técnicas de regressão que foram consideradas, bem como as abordagens para a representação dos conteúdos textuais como vetores de caraterísticas descritivas. A Seção 4 apresenta a avaliação dos métodos propostos, descrevendo os conjuntos de dados dos três domínios diferentes, e discutindo os resultados obtidos. Finalmente, a Seção 5 apresenta as principais conclusões e aponta possíveis direções para trabalho futuro.

## 2 Trabalhos Relacionados

Inspirando-se em trabalhos recentes sobre análise de sentimentos (Pang e Lee, 2008), em que técnicas de aprendizagem automática são usadas para interpretar textos com base na atitude subjetiva dos autores, Noah Smith e os seus colegas têm abordado várias tarefas de prospeção de texto relacionadas, onde os documentos textuais são interpretados para prever os valores de variáveis de interesse do mundo real. Este é um dos grupos de investigadores com mais atividade nesta área específica. Um artigo relativamente recente, que resume o trabalho que estes investigadores têm vindo a desenvolver, foi publicado on-line por Smith (2010). Exemplos específicos para as tarefas de previsão com base em textos, que estes autores abordaram, incluem:

1. A interpretação de relatórios financeiros anuais, publicados por empresas aos seus acionistas, a fim de tentar prever o risco incorrido ao investir numa empresa, no ano seguinte (Kogan et al., 2009);
2. A interpretação de comentários sobre filmes, feitos por críticos de cinema, com o objetivo de tentar prever o resultado de bilheteira dos filmes (Joshi et al., 2010);
3. Interpretar mensagens colocadas em blogues políticos, para tentar prever a resposta recolhida dos leitores (Yano e Smith, 2010);
4. A interpretação de mensagens diárias em *microblogs*, a fim de prever opinião pública e a confiança dos consumidores (Yano, Cohen e Smith, 2009; Yano e Smith, 2010);
5. Interpretar textos correspondentes a descrições de restaurantes, menus de restaurantes, e comentários de clientes, para tentar prever o preço médio de refeições e as avaliações dos clientes (Chahuneau et al., 2012).

Em todos os casos acima, um aspeto do significado do texto é observável a partir de dados objetivos do mundo real, embora talvez não imediatamente no momento em que o texto é publicado (ou seja, respetivamente, observa-se a volatilidade dos retornos, a receita bruta, os comentários dos utilizadores, resultados de questionários de opinião tradicionais, e os preços médios das refeições, nos cinco problemas que foram anteriormente enumerados). Smith (2010) propôs uma abordagem genérica para a previsão baseada em textos, suportada em modelos de regressão que utilizam caraterísticas descritivas derivadas do texto, as quais são geralmente ruidosas e esparsas. Este autor argumentou que a previsão com base em textos, como um problema de investigação, pode ser abordada por meio de metodologias baseadas em aprendizagem que são neutras a diferentes teorias linguísticas (Smith, 2010).

Por exemplo no que diz respeito às críticas e receitas de filmes, e sumarizando o trabalho anterior de Joshi et al. (2010), Smith mencionou que antes da estreia de um filme, os críticos assistem e publicam comentários sobre o mesmo. Os autores procuraram realizar previsões sobre os resultados de bilheteira com base nos comentários, à medida que estes são produzidos pelos críticos. Consideraram-se 1,351 filmes lançados entre Janeiro de 2005 e Junho de 2009. Para cada filme, foram obtidos dois tipos de dados:

[2] `http://www.portal-cinema.com`
[3] `http://www.ica-ip.pt`
[4] `http://scikit-learn.org`

1. Metadados descritivos recolhidas a partir do site *Metacritic*[5], incluindo o nome, a produtora, o(s) gênero(s), realizador(es), diretor(es), os atores principais, e o país de origem, entre outros dados. Metadados de um site chamado *The Numbers*[6] foram também recolhidos, contendo informação sobre o orçamento de produção, as receitas brutas do final de semana da estreia do filme, e o número de salas de cinema em que o filme foi exibido nesse final de semana.
2. Comentários extraídos dos seis sites de análise de filmes mais referenciados no site *Metacritic*, e apenas comentários publicados antes da data de estreia do filme.

Os autores descrevem a aplicação de um modelo de regressão linear com regularização dada pelo método da rede elástica (Zou e Hastie, 2005; Fridman, Hastie e Tibshirani, 2008). O modelo foi treinado em 988 exemplos lançados entre 2005 e 2007, e foi avaliado na previsão da receita de bilheteira para cada filme lançado entre Setembro de 2008 e Junho de 2009 (i.e., um total de 180 filmes). Foi calculado o erro absoluto médio (MAE) sobre o conjunto de teste, analisando a diferença entre a receita estimada para cada filme durante a sua semana de lançamento, e os ganhos brutos reais, por ecrã. Modelos que usam apenas o texto (MAE de 6,729$), ou o texto em adição a metadados (MAE de 6,725$), foram melhores do que os modelos que usam apenas os metadados (MAE de 7,313$). O texto reduz o erro por 8% em relação aos metadados, e por 5% quando comparado com uma forte base de previsão dos resultados de bilheteira, dada pelo valor médio dos filmes nos de dados de treino.

Mais recentemente, Chahuneau et al. (2012) exploraram as interações no uso da linguagem que ocorrem entre os preços dos menus de restaurantes, descrições de itens do menu, e sentimentos expressos em comentários de utilizadores, a partir de dados extraídos do site *Allmenus*[7]. Deste site, os autores recolheram menus de restaurantes em sete cidades norte-americanas, nomeadamente *Boston*, *Chicago*, *São Francisco*, *Los Angels*, *Nova Iorque*, *Washington DC* e *Filadélfia*. Cada menu contém uma lista de nomes de itens, com descrições textuais opcionais e preços. Metadados adicionais (e.g., gama de preço, localização, e ambiente) e comentários dos clientes (i.e., descrições textuais associadas a uma classificação numa escala de 0 a 5 estrelas), para a maioria dos restaurantes, foram recolhidos a partir de um conhecido site Web chamado *Yelp*[8].

Os autores consideraram diversas tarefas de previsão, tais como a previsão de preços individuais de itens, a previsão da gama de preços para cada restaurante, e a previsão em conjunto do preço médio e da opinião dos clientes. Para as duas primeiras tarefas, os autores utilizaram modelos de regressão linear, e para a terceira tarefa, usaram modelos de regressão logística, em todos os casos com regularização $l_1$. Para a avaliação, os autores usaram métricas como o erro absoluto médio (MAE) ou o erro relativo médio (MRE).

Para prever o preço individual de cada item num menu, Chahuneau et al. (2012) utilizaram o logaritmo do preço como o valor a modelar e a prever, pois a distribuição dos preços é mais simétrica numa escala logarítmica. Os autores avaliaram várias estratégias simples que fazem previsões independentes para cada nome diferente dos itens nos menus. Duas destas estratégias usam a média e a mediana do preço, no conjunto de treino e dado o nome do item. Uma terceira estratégia utiliza um modelo de regressão linear com regularização $l_1$, que foi treinado com múltiplas caraterísticas binárias. Os autores realizaram uma simples normalização dos nomes de itens em todas estas estratégias, devido à sua grande variação no conjunto de dados (i.e., mais de 400 mil nomes distintos). A normalização consiste na remoção das palavras mais frequentes nos nomes dos itens, ordenando de seguida as palavras em cada nome lexicograficamente. Esta normalização reduziu o número de nomes por 40%.

Os autores exploraram diferentes modelos ricos em caraterísticas descritivas, com base em regressão regularizada, considerando (i) características binárias para cada propriedade de metadados do restaurante, (ii) $n$-gramas em nomes de itens do menu, em que $n$-gramas correspondem a sequências de $n$ palavras (i.e., com $n \in \{1, 2, 3\}$) extraídas de um determinado nome, (iii) $n$-gramas nas descrições de itens do menu, e (iv) $n$-gramas de menções a itens do menu nos comentários correspondentes. Ao usar o conjunto completo de caraterísticas descritivas, os autores relatam uma redução final de 0,5 na métrica MAE, e de cerca de 10% no MRE. Os autores reportam assim bons resultados para esta técnica, quando comparados com os resultados das estratégias elementares.

Na tarefa de prever a gama de preços, os valores alvo eram números inteiros de 1 a 4 que indicam o custo de uma refeição típica do restaurante. Na avaliação desta tarefa, os autores

[5] `http://www.metacritic.com`
[6] `http://www.the-numbers.com`
[7] `http://www.allmenus.com`
[8] `http://www.yelp.com`

arredondaram os valores previstos para inteiros, e usaram o erro absoluto médio (MAE) e a exatidão, como medidas da qualidade dos resultados. Os autores notaram uma pequena melhoria ao comparar o modelo de regressão linear com um modelo de regressão ordinal (i.e., um modelo que atribui, para cada caso, um valor de classificação entre 1 e 4, e que leva em consideração a ordenação dos valores alvo (McCullagh, 1980)), medindo 77,32% de exatidão na regressão ordinal, contra 77,15%, para modelos com os metadados. Os autores também usaram caraterísticas descritivas do texto completo dos comentários, além das caraterísticas utilizadas para a tarefa de previsão de preços individuais de itens do menu. Ao combinar os metadados e as caraterísticas descritivas dos comentários, a exatidão medida excedeu o valor de 80%.

Para a tarefa de analisar as opiniões expressas nos comentários, os autores treinaram um modelo de regressão logística, prevendo a polaridade da opinião expressa em cada comentário. A polaridade de um comentário foi determinada pela correspondente pontuação de classificação em estrelas, ou seja, se está acima ou abaixo da média. A exatidão obtida foi de 87%.

Finalmente, Chahuneau et al. (2012) consideraram a tarefa de prever, em conjunto, o preço médio e a opinião agregada para um restaurante. Para fazer isso, os autores tentaram modelar, ao mesmo tempo, a polaridade dos comentários $\overline{r}$ e o preço dos itens $\overline{p}$. Os autores calcularam, para cada restaurante no conjunto de dados, a média do preço dos itens e a média da pontuação em estrelas. Um plano ($\overline{r}$, $\overline{p}$) foi dividido em quatro seções, com os pontos médios dos dois atributos no conjunto de dados como as coordenadas de origem, ou seja, 8,69$ para $\overline{p}$ e 3,55 estrelas para $\overline{r}$. Esta divisão permitiu aos autores treinar um modelo de regressão logística de 4 classes, usando as caraterísticas descritivas extraídas das avaliações para cada restaurante. A exatidão obtida foi, neste caso, de 65%.

## 3 Realização de Previsões com Base em Conteúdos Textuais

Neste estudo, à semelhança do trabalho anterior de Noah Smith e seus colegas, abordámos o problema de fazer previsões, com conteúdos textuais, como uma tarefa de regressão. Reportamos resultados para experiências em três domínios distintos, com textos escritos em português, e considerando algumas inovações, tais como o uso de agrupamentos de palavras ou de diferentes esquemas de ponderação das caraterísticas.

Cada documento é modelado como um vetor de caraterísticas descritivas num determinado espaço vetorial, em que a dimensionalidade corresponde ao número de caraterísticas descritivas diferentes. Esta representação está associada a um modelo bem conhecido para o processamento e representação de documentos na área da recuperação de informação, normalmente referido como o modelo do espaço vetorial. Formalmente, tem-se que cada documento é representado como um vetor de caraterísticas descritivas $\overrightarrow{d_j} =< w_{1,j}, w_{2,j}, ..., w_{k,j} >$, em que $k$ é o número de caraterísticas, e em que cada $w_{i,j}$ corresponde a um peso que reflete a importância da caraterística $i$ para a descrição dos conteúdos do documento $j$. As caraterísticas descritivas são, essencialmente, as palavras que ocorrem na coleção de documentos. No entanto, em algumas das nossas experiências, também usámos outras caraterísticas, tais como propriedades de metadados referentes à localização geográfica (i.e., os distritos administrativos) associados aos exemplos, os tipos de restaurantes, ou agrupamentos de palavras associados aos termos que ocorrem no documento correspondente.

### 3.1 Agrupamento Automático de Palavras

O agrupamento automático de palavras semelhantes permite abordar o problema da esparsidade dos dados, ao proporcionar uma representação de dimensionalidade inferior para as palavras de uma coleção de documentos. Neste trabalho, foi utilizado o algoritmo de agrupamento de palavras proposto por Brown et al. (1992), que induz representações generalizadas de palavras individuais. Este algoritmo é essencialmente um processo de agrupamento hierárquico que forma grupos de palavras com caraterísticas comuns, a fim de maximizar a informação mútua de bi-gramas. A entrada para o algoritmo é um corpus de texto, que pode ser visto como uma sequência de $N$ palavras $w_1, \cdots, w_N$. O resultado é uma árvore binária, na qual as folhas são as palavras. O processo de agrupamento está relacionado com um modelo de linguagem baseado em bi-gramas e classes:

$$\mathrm{P}(w_1^N|C) = \prod_{i=1}^{N} \mathrm{P}\left(\mathrm{C}(w_i)|\mathrm{C}(w_{i-1})\right) \times \mathrm{P}\left(w_i|\mathrm{C}(w_i)\right)$$

Na fórmula, $\mathrm{P}(c|c')$ corresponde a probabilidade de transição da classe $c$ dada a sua classe antecessora $c'$, e $\mathrm{P}(w|c)$ é a probabilidade de emissão da palavra $w$ numa dada classe $c$. As probabilidades do modelo podem ser calculadas por

contagem de frequências relativas de unigramas e bi-gramas. Para determinar as classes ótimas $C$ para um número de classes $M$, podemos adotar uma abordagem de máxima verosimilhança do tipo $\mathrm{C} = \arg\max_{\mathrm{C}} \mathrm{P}(W_1^N|\mathrm{C})$. Brown et al. (1992) demonstraram que o melhor agrupamento é o que resulta da maximização da informação mútua entre as classes adjacentes, dada por:

$$\sum_{c,c'} \mathrm{P}(c,c') \times \log\left(\frac{\mathrm{P}(c,c')}{\mathrm{P}(c) \times \mathrm{P}(c')}\right)$$

A estimativa do modelo de linguagem é, portanto, baseada num procedimento de agrupamento automático aglomerativo, que é usado para construir uma estrutura hierárquica sobre as distribuições de contextos de cada palavra. O algoritmo começa com um conjunto de nós folha, um para cada uma das classes de palavras (i.e., inicialmente, uma classe para cada uma das palavras). Em seguida, de forma iterativa, são selecionados pares de nós a fundir, otimizando de forma gananciosa um critério de qualidade baseado na informação mútua entre agrupamentos adjacentes (Brown et al., 1992). Cada palavra é, portanto, inicialmente atribuída ao seu próprio grupo/classe, e o algoritmo funde pares de classes, de modo a induzir a redução mínima na informação mútua, parando quando o número de classes é reduzido para o número predefinido $|C|$.

Neste trabalho, para induzir representações para as palavras, utilizámos uma implementação *open source*[9] do algoritmo de Brown, que seque a descrição dada por Turian, Ratinov e Bengio (2010). Este *software* foi usado em conjunto com uma grande coleção de textos escritos em português, à qual tínhamos acesso e que representa diversos tipos de temas e géneros linguísticos. Este textos correspondem a um conjunto de frases que combina o corpus CINTIL do português moderno (Barreto et al., 2006), com artigos noticiosos publicados no jornal *Público*[10], durante um período de 10 anos. Induzimos um total de mil grupos de palavras, onde cada grupo tem um identificador único a usar nas representações.

### 3.2 Ponderação das Caraterísticas

Neste trabalho, experimentámos diferentes formas de calcular o peso das caraterísticas descritivas a serem usadas nos nossos modelos, para os termos textuais e os agrupamentos de palavras semelhantes. Estas formas incluem o uso de valores binários, a frequência dos termos, TF-IDF, e também esquemas de ponderação mais sofisticados, tais como os esquemas Delta-TF-IDF e Delta-BM25 discutidos por Martineau e Finin (2009) e por Paltoglou e Thelwall (2010).

No caso de pesos binários, cada parâmetro $w_{i,j}$ toma o valor zero ou um, dependendo se o elemento $i$ está presente ou não no documento $j$.

Outra abordagem comum é usar a frequência de ocorrência de cada elemento $i$ no documento $j$, tendo-se que é comum considerar uma penalização logarítmica para estes valores.

TF-IDF é, talvez, o esquema de ponderação de caraterísticas mais popular, combinando a frequência individual de cada elemento $i$ num documento $j$ (i.e., a componente *Term Frequency*, ou TF), com a frequência inversa do elemento $i$ na coleção de documentos (i.e., a componente *Inverse Document Frequency*, ou IDF). Quando $n_i > 0$, o peso TF–IDF de um elemento $i$ para um documento $j$ é dado pela seguinte fórmula:

$$\mathrm{TF\text{–}IDF}_{i,j} = \log_2(1 + \mathrm{TF}_{i,j}) \times \log_2\left(\frac{N}{n_i}\right)$$

Na fórmula, $N$ é o número total de documentos na colecção, e $n_i$ é o número dos documentos contendo o elemento $i$. TF–IDF é zero se $n_i \leq 0$.

Os esquemas de ponderação Delta–TF–IDF e Delta–BM25 medem a importância relativa de um termo em duas classes distintas. No contexto dos nossos problemas de regressão, não temos classificações binárias associadas a cada uma das instâncias, mas sim valores reais. No entanto, considerámos duas classes a fim de determinar os pesos das caraterísticas descritivas de acordo com estes esquemas, ao dividir as instâncias entre aquelas que tem um valor superior ou igual à mediana dos valores nos dados de treino, e aquelas que são menores ou iguais à mediana.

O esquema Delta–TF–IDF estende a fórmula do TF–IDF, localizando a estimação das pontuações de IDF para os documentos associados a cada uma das duas classes, subtraindo depois os dois valores. O peso de um elemento $i$ num documento $j$ pode ser obtido como se mostra na seguinte equação, quando $\mathrm{TF}_{i,j} > 0$:

$$\Delta\mathrm{TF\text{–}IDF}_{i,j} = \log_2(1 + \mathrm{TF}_{i,j}) \times \log_2\left(1 + \frac{N_{pos} \times n_{i,neg}}{n_{i,pos} \times N_{neg}}\right)$$

Cada parâmetro $N_c$ corresponde ao número de documentos de treino na colecção $c$, e $n_{i,c}$ é o número de documentos da coleção $c$ no qual o termo $i$ ocorre. No contexto deste trabalho, $c$

[9] `http://github.com/percyliang/brown-cluster`
[10] `http://www.publico.pt`

pode ser *positivo* para os exemplos com o valor superior à mediana, e *negativo* para os casos inferiores à mediana. De acordo com um grande conjunto de experiências relacionadas com a classificação binária de opiniões em textos, a abordagem Delta–TF–IDF é significativamente melhor do que esquemas baseados em TF ou numa ponderação binária (Martineau e Finin, 2009).

Paltoglou e Thelwall (2010) concluíram que ao introduzir, adicionalmente, variantes localizadas e suavizadas das funções IDF, em conjunto com esquemas de ponderação TF escalados, a exatidão pode ser aumentada ainda mais. No esquema de ponderação Delta–BM25, o peso de um elemento $i$ para um documento $j$ é dado pela seguinte fórmula, onde $s$ é uma constante de suavização que é normalmente definida como 0,5:

$$\begin{aligned}\Delta\text{BM25}_{i,j} &= \log_2(1+\text{TF}_{i,j}) \\ &\quad\times \log_2\left(1+\frac{(N_{pos}-n_{i,pos}+s)\times n_{i,neg}+s}{(N_{neg}-n_{i,neg}+s)\times n_{i,pos}+s}\right)\end{aligned}$$

### 3.3 Modelos de Regressão

Poderiam ter sido usados vários tipos de modelos de regressão, a fim de abordar as tarefas de previsão com base em textos que foram por nós consideradas. Neste trabalho, comparámos modelos de regressão linear, usando diferentes tipos de regularização, com modelos baseados na combinação de várias árvores de regressão.

#### *3.3.1 Métodos de Regressão Linear*

Considerando um conjunto de dados $\{y_i, x_{i1}, ..., x_{ik}\}_{i=1}^{n}$ com $n$ instâncias, e assumindo que a relação entre a variável dependente $y_i$ e um vetor de $k$ caraterísticas descritivas $x_i$ é linear, tem-se que uma regressão linear assume a seguinte forma:

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix} \times \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \cdots \\ b_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix}$$

Na fórmula, $x_{i,j}$ corresponde à *i–ésima* caraterística descritiva do *j–ésimo* exemplo, os parâmetros $b_i$ correspondem aos coeficientes de regressão, e $e_j$ é um erro que capta a diferença entre a forma efetiva das respostas observadas $y_i$, e os resultados da previsão do modelo de regressão. A fórmula pode ser re-escrita usando uma notação matricial, ficando $y = Xb + e$.

Vários procedimentos têm sido desenvolvidos para a estimativa de parâmetros em modelos de regressão linear. O método dos mínimos quadrados (i.e., *linear least squares regression*, ou LSR) é a forma mais simples e mais utilizada para estimar os parâmetros desconhecidos num modelo de regressão linear. O método LSR minimiza a soma dos quadrados dos resíduos $\text{S}(b)$ entre os dados e o modelo, ou seja, minimiza a soma $\sum_{i=1}^{n} e_i^2$. O quadrado dos resíduos pode ser re-escrito em notação matricial como $e'e$, onde o apóstrofo significa que a matriz foi transposta. Substituindo $e$ por $y - Xb$, temos que:

$$\begin{aligned}\text{S}(b) &= \sum_{i=1}^{n} e_i^2 \\ &= (y-Xb)'(y-Xb) \\ &= y'y - y'Xb - b'X'y + b'X'Xb\end{aligned}$$

A condição para a $\text{S}(b)$ estar no mínimo é ter as derivadas $\frac{\partial \text{S}(b)}{\partial b} = 0$. O primeiro termo da equação acima não depende de $b$, enquanto que o segundo e o terceiro termos são iguais nas suas derivadas, e o último termo é uma forma quadrática dos elementos $b$. Assim, temos que:

$$\frac{\partial \text{S}(b)}{\partial b} = -2X'y + 2X'Xb$$

Igualando a equação diferencial a zero, temos:

$$\begin{aligned}-2X'y + 2X'Xb &= 0 \\ X'Xb &= X'y \\ b &= (X'X)^{-1}X'y \\ b &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) \\ b &= \arg\min_b ||y-Xb||^2\end{aligned}$$

Na regressão com o método dos mínimos quadrados, é bem conhecido que quando o número de instâncias de treino $n$ for menor do que o número de caraterísticas $k$, ou se existirem muitas caraterísticas correlacionadas, os coeficientes de regressão podem apresentar uma alta variação, e tendem a ser instáveis. Neste caso, são necessários métodos de regularização para melhor ajustar o modelo aos dados, e para manter a variação dos coeficientes de regressão sob controlo.

A abordagem da regressão em crista (i.e., *ridge regression*) penaliza o tamanho dos coeficientes, por adição de uma penalização $l_2$ correspondente a um termo $||b||_2^2$ no modelo. Assim, os

coeficientes de regressão são estimados através do seguinte problema de otimização:

$$b = \arg\min_b ||y - Xb||^2 + \lambda||b||_2^2$$

Na fórmula, $\lambda \geq 0$ é um parâmetro de ajuste, o qual controla a força do termo de regularização. Quando $\lambda = 0$ obtemos a estimativa de regressão linear regular, e quando $\lambda = \infty$ obtemos $b = 0$.

Tibshirani (1996) propôs um outro método de regularização, chamado *Least Absolute Shrinkage and Selection Operator* (Lasso), que encolhe alguns coeficientes e fixa outros a zero. O método Lasso utiliza uma penalização $l_1$ dada por $||b||_1$ como forma de regularização, e tenta estimar os parâmetros $b$ através do seguinte problema:

$$b = \arg\min_b ||y - Xb||^2 + \lambda||b||_1$$

Uma das principais diferenças entre o Lasso e a regressão em crista é que na regressão em crista, quando a penalização é aumentada, todos os parâmetros são reduzidos mas permanecem ainda diferentes de zero, enquanto que com a regularização Lasso, aumentar a penalização conduz alguns dos parâmetros a zero. Assim, os modelos Lasso tendem a ser esparsos, na medida em que utilizam menos caraterísticas descritivas. No entanto, uma limitação do Lasso é que se $k > n$, então este método seleciona no máximo $n$ variáveis, ou seja, o número de variáveis selecionadas é limitado pelo número de exemplos de treino. Outra limitação do método Lasso ocorre quando existe um grupo de variáveis altamente correlacionadas. Neste caso, o Lasso tende a selecionar uma variável apenas a partir deste grupo, ignorando as outras. Para resolver estes problemas, Zou e Hastie (2005) propuseram a abordagem de regularização conhecida como o método da rede elástica (i.e., *elastic net*), a qual combina as regularizações $l_1$ e $l_2$ com pesos $\lambda_1$ e $\lambda_2$, respetivamente. As estimativas deste método são definidas pela seguinte fórmula:

$$b = \arg\min_b ||y - Xb||^2 + \lambda_1||b||_1 + \lambda_2||b||_2^2$$

O método da rede elástica tende a produzir melhores estimativas quando as variáveis são consideravelmente correlacionadas. O método elimina a limitação do número de variáveis selecionadas, incentiva efeitos de agrupamento, e estabiliza a abordagem de regularização $l_1$.

Vários métodos têm sido propostos para estimar modelos de regressão linear com regularização *ridge*, Lasso, e *elastic net*, incluindo a pesquisa cíclica por coordenadas no sentido descendente do gradiente (i.e., *cyclical coordinate descent*) (Kim e Kim, 2006), ou outros métodos de otimização convexa com base em cálculos iterativos (Boyd e Vandenberghe, 2004), tais como o SpaRSA (Wright, Nowak e Figueiredo, 2009). Neste estudo, foi utilizada a implementação disponível a partir do pacote Python de aprendizagem automática denominado *scikit-learn*.

### *3.3.2 Métodos de Aprendizagem por Combinação*

Os métodos de aprendizagem por combinação (i.e., *ensemble learning*) tentam combinar simultaneamente vários modelos, que são geralmente modelos simples baseados em árvores de decisão, para obter um melhor desempenho em problemas de previsão (Sewell, 2011). Neste trabalho, foram utilizados dois tipos de métodos de aprendizagem por combinação, nomeadamente o método da floresta aleatória (i.e., *random forest*) de árvores de regressão, e um modelo de regressão baseado em múltiplas árvores de regressão obtidas por reforço em relação ao gradiente de uma dada função (i.e., *gradient boosted regression trees*), mais uma vez tal como implementados no pacote Python denominado *scikit-learn*.

A floresta aleatória de árvores de regressão combina a ideia de *bagging*, desenvolvida por Breiman (1996), com uma seleção aleatória de caraterísticas descritivas (Breiman, 2001). Em suma, temos que este método constrói uma coleção de árvores de-correlacionadas, e produz como resultado o seu valor médio. O principal objetivo deste método é reduzir a variância no modelo final de regressão, através da redução da correlação entre as variáveis. Isto é conseguido pela seleção aleatória de variáveis. Seja $N$ o número de casos de treino, e seja $M$ o número de instâncias utilizadas para treino do modelo. O algoritmo procede da seguinte forma:

1. Escolher um conjunto de treino para cada árvore, amostrando $n$ vezes com substituição de todos os $N$ casos de treino disponíveis. Usar as instâncias restantes para estimar o erro da árvore, ao fazer as previsões.

2. Para cada nó da árvore, selecionar $m$ variáveis aleatoriamente para apoiar a decisão naquele nó, e calcular a melhor divisão para a árvore com base nessas $m$ variáveis e no conjunto de treino da etapa anterior.

3. Cada árvore cresce até atingir a maior extensão possível, e nenhuma poda é realizada. O algoritmo CART tal como proposto

por Breiman et al. (1984) é utilizado para a geração das várias árvores.

Os passos acima são iterados para gerar um conjunto de árvores. Ao fazer uma previsão, a média das previsões de todas as árvores é relatada. Cada árvore vota com um peso correspondente ao seu desempenho no subconjunto de dados que foi deixado de parte durante o treino.

Quanto ao método *gradient boosted regression trees* (GBRT), este é por sua vez baseado na ideia de *boosting*, suportando funções de otimização específicas para problemas de regressão, como a soma dos erros quadráticos (Friedman, 2001). Um modelo GBRT consiste de um conjunto de modelos fracos, normalmente árvores de decisão (i.e., árvores CART, semelhantes às que são utilizadas no caso de regressão com florestas aleatórias). O modelo toma a seguinte forma, onde $\mathrm{h}_m(X)$ são as funções de base do modelo:

$$y = \mathrm{F}(X) = \sum_{m=1}^{M} \mathrm{h}_m(X)$$

De acordo com o princípio da minimização do risco empírico, o método tenta minimizar o valor médio de uma função de perda em relação ao conjunto de treino. Isto faz-se começando com um modelo, que consiste de uma função constante $F_0$, e incrementalmente expandindo-o de forma gananciosa, de acordo com a seguinte equação:

$$\mathrm{F}_m = \mathrm{F}_{m-1}(X) - \gamma_m \mathrm{h}_m(X)$$

Em cada estágio, uma árvore de decisão $\mathrm{h}_m(X)$ é escolhida para minimizar a função de perda considerada (i.e., a soma dos quadrados dos erros), dado o modelo atual $\mathrm{F}_{m-1}$ e as suas previsões $\mathrm{F}_{m-1}(X)$ (i.e., as árvores $\mathrm{h}_m$ aprendidas em cada etapa são geradas usando pseudo-resíduos como o conjunto de treino). A inicialização para o modelo $F_0$ é escolhida frequentemente com base na média dos valores alvo, e o multiplicador $\gamma_m$ é encontrado em cada fase, resolvendo o seguinte problema de otimização, onde L é a função de perda considerada:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} \mathrm{L}(y_i, \mathrm{F}_{m-1}(x_i) + \gamma \mathrm{h}_m(x_i))$$

O problema de otimização é normalmente abordado através de um procedimento de descida ao longo do gradiente (i.e., *gradient descent*) da função (Boyd e Vandenberghe, 2004).

## 4 Validação Experimental

Este trabalho explorou a realização de previsões com base em textos de três domínios distintos com caraterísticas diferentes, ou seja, considerando informação sobre hotéis, restaurantes e os seus preços em Portugal, ou sobre comentários de filmes, em conjunto com os resultados de bilheteira correspondentes. Para os hotéis e os restaurantes, recolhemos descrições textuais do site Lifecooler[11]. Para cada restaurante, a informação disponibilizada por este site inclui o nome, a descrição textual, o menu, as especialidades, o tipo de restaurante, o preço médio de uma refeição, e a localização, que inclui o nome da cidade e o respetivo distrito. Para os hotéis, a informação disponível inclui o nome, a descrição textual, a localização, e o preço dos quartos nas épocas altas e baixas. Para o caso dos filmes, usamos comentários do site Portal do Cinema[12], e metadados provenientes do Instituto do Cinema e do Audiovisual[13], para os filmes que foram lançados entre 2008 e 2013 em Portugal. A informação disponível inclui o nome do filme, o distribuidor, o produtor, o número de salas em que o filme foi exibido durante a semana de estreia, a receita bruta da primeira semana, o número de espectadores, e uma classificação por estrelas numa escala de 0 a 5. Considerámos apenas os filmes encontrados nos dois sites, e assim acabámos por utilizar um conjunto de 502 filmes. As principais caraterísticas descritivas, dos conjuntos de dados associados aos três domínios, estão apresentadas na Tabela 1. Para os hotéis e os restaurantes, os valores alvo são mostrados em Euros, enquanto que para o caso dos filmes são apresentados valores em milhares de Euros.

Os conjuntos de dados diferem em vários aspetos, tais como no número de documentos disponíveis e na distribuição dos valores a serem previstos. A distribuição dos valores alvo para os três domínios é apresentada na Figura 1.

Todo o texto disponível foi utilizado nas nossas experiências (e.g., para os hotéis, utilizámos o nome do hotel e a descrição textual, enquanto que para os restaurantes utilizámos o nome do restaurante, a descrição textual, as especialidades, e os menus. Para os filmes, usámos o nome do filme e o comentário). Para além das características textuais, em algumas experiências utilizámos a localização, para hotéis e restaurantes, os tipos dos restaurantes, ou o número de salas em que o filme foi exibido. A localização (i.e., os

[11] `http://www.lifecooler.com`
[12] `http://www.portal-cinema.com`
[13] `http://www.ica-ip.pt`

| | Hotéis Alta | Hotéis Baixa | Restaurantes | Filmes |
|---|---|---|---|---|
| Número de descrições textuais | 2656 | 2656 | 4677 | 502 |
| Tamanho do vocabulário | 9932 | 9932 | 19421 | 28720 |
| Número médio de termos por texto | 35 | 35 | 47 | 346 |
| Valor alvo mínimo | 10,00 | 10,00 | 4,50 | 0,93 |
| Valor alvo máximo | 3000,00 | 1200,00 | 100,00 | 1437,71 |
| Média dos valores alvo | 95,92 | 71,48 | 18,10 | 162,75 |
| Mediana dos valores alvo | 75,00 | 60,00 | 15,00 | 73,56 |
| Desvio padrão nos valores alvo | 93,42 | 51,67 | 8,41 | 229,21 |

Tabela 1: Caraterização estatística dos três conjuntos de dados que foram considerados.

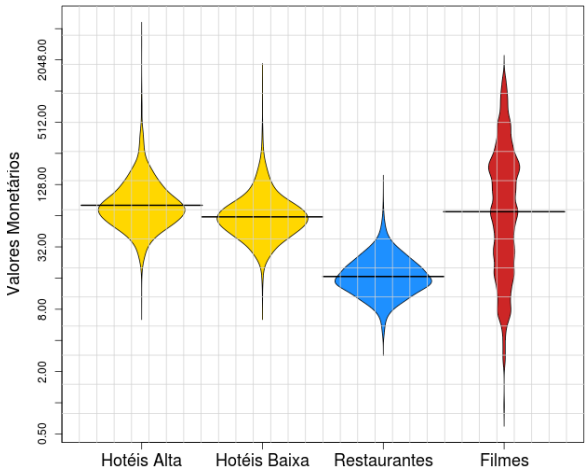


Figura 1: Distribuição dos valores a serem previstos, nos vários conjuntos de dados.

distritos administrativos) pode naturalmente influenciar quão caro é um hotel ou um restaurante. Por exemplo, como podemos ver na Figura 2, os distritos com os hotéis e os restaurantes mais caros são Lisboa e Faro, e estes mesmos distritos são os que apresentam a maior variação de preços.

### 4.1 Metodologia Experimental

Todas as experiências foram realizadas com uma metodologia de validação cruzada usando 10 desdobramentos (i.e., *folds*). A qualidade dos resultados foi aferida usando métricas como o erro absoluto médio (i.e., o *mean absolute error*) e o desvio padrão empírico generalizado (i.e., a métrica *root mean squared error*).

O erro absoluto médio (MAE) é uma medida que compara as previsões com os valores reais, e que corresponde a uma média dos valores de erro absoluto, como mostra a Equação 1. O desvio padrão empírico generalizado é outra medida da exatidão de modelos de previsão, calculada com base na raiz quadrada da média dos quadrados dos erros, como mostrado na Equação 2.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

Considerando um conjunto de dados $\{y_i, x_{i1}, ..., x_{ik}\}_{i=1}^{n}$, onde os valores $x_{ik}$ correspondem às caraterísticas, onde $y_i$ corresponde aos verdadeiros valores alvo, e tendo $\hat{y}_i$ como os resultados previstos, pode-se facilmente ver que as métricas anteriores estimam erros nas mesmas unidades de medida que os valores alvo, ou seja, em Euros ou em milhares de Euros, no caso das experiências relatadas neste trabalho.

Além das métricas MAE e RMSE, reportamos também alguns resultados em termos de uma variante normalizada do erro médio, correspondendo ao erro relativo médio (i.e., *mean relative error*) tal como apresentado na equação abaixo:

$$\text{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

O erro relativo médio (MRE) permite estabelecer comparações entre tarefas de previsão diferentes, dado que esta medida é independente das unidades e da escala em que as variáveis a ser previstas se encontram expressas.

### 4.2 Resultados Obtidos

Num primeiro teste, tentámos prever os preços dos quartos de hotéis, o preço médio dos pratos em restaurantes, ou as receitas de bilheteira de filmes, usando apenas conteúdos textuais. Nesta tarefa, comparámos modelos de regressão com abordagens simples, tais como realizar as previsões com base no valor médio e na mediana nos dados de treino. Foram consideradas representações baseadas no esquema de ponderação de termos mais popular, ou seja, TF-IDF. Como podemos ver na Tabela 2, os modelos de regressão com caraterísticas derivadas dos textos atingiram

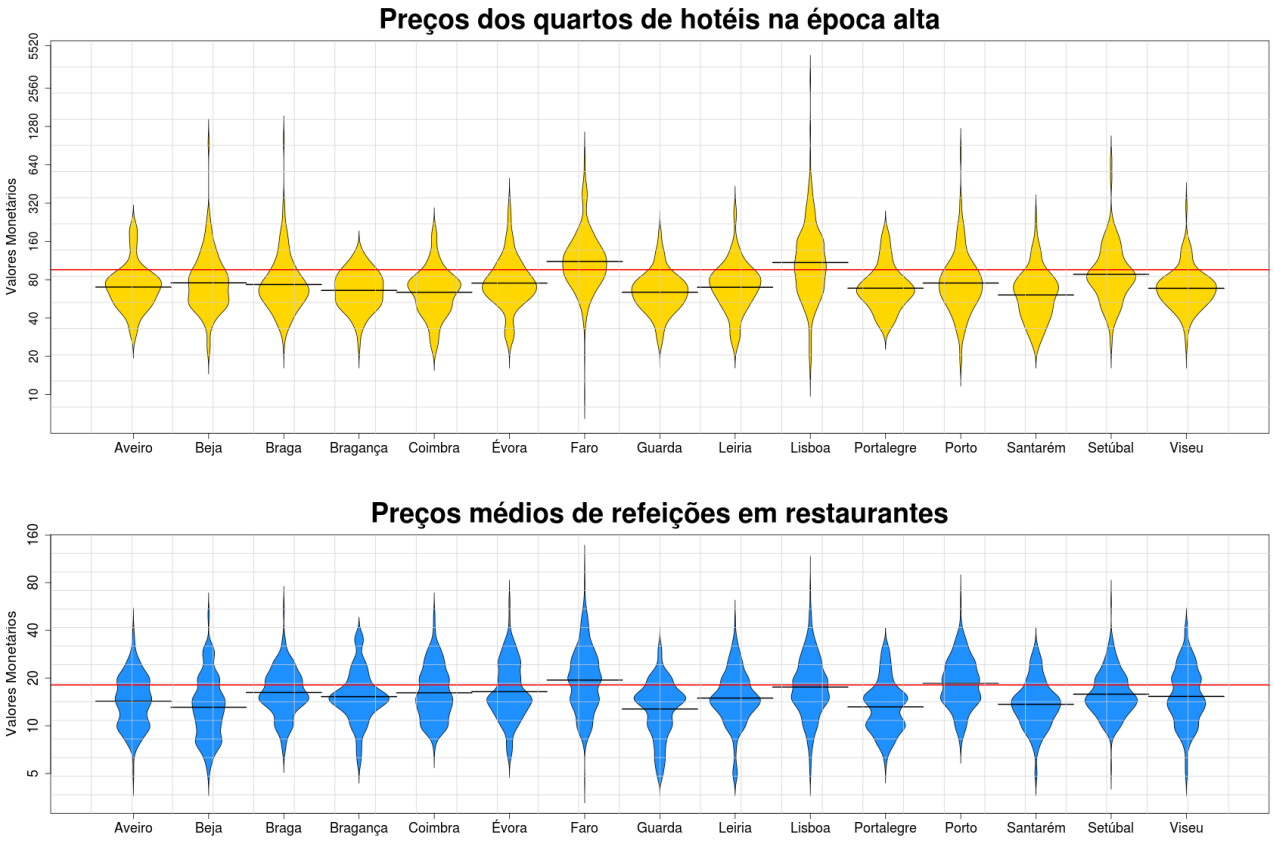


Figura 2: Distribuição para os valores a serem previstos, por distrito em Portugal Continental.

melhores resultados que as abordagens mais simples. Verificamos também que, de todos os modelos considerados, os melhores resultados foram obtidos com o método de regularização da rede elástica. O melhor modelo baseado em aprendizagem por combinação corresponde à abordagem da floresta aleatória de árvores de regressão.

Num conjunto separado de experiências, procurámos analisar a importância das diferentes caraterísticas descritivas correspondentes aos termos, vendo as suas diferenças relativas em termos da contribuição para prever os valores alvo. Isto foi feito para o caso de modelos com base na técnica da floresta aleatória de árvores de regressão, ou com base em regressão linear com regularização dada pelo método da rede elástica, usando pesos para as caraterísticas descritivas calculados com a abordagem TF-IDF.

No caso de modelos de regressão linear com regularização dada pelo método da rede elástica, inspecionamos os pesos das caraterísticas descritivas (i.e., os coeficientes de regressão) dos modelos aprendidos, e calculámos a média dos pesos de cada caraterística sobre as múltiplas *folds* dos nossos testes de validação cruzada. No caso de modelos baseados na técnica da floresta aleatória de árvores de regressão, a posição relativa (i.e., a profundidade) de uma caraterística que seja usada como nó de decisão numa árvore pode servir para avaliar a importância relativa dessa caraterística, no que diz respeito à previsão. As caraterísticas que são utilizadas na parte superior das árvores contribuem para a decisão final de uma maior fração dos exemplos e, deste modo, a fração esperada dos exemplos, para as quais a caraterística contribui, pode ser utilizada como uma estimativa da importância relativa das caraterísticas. Pela média dessas taxas de atividade esperada, ao longo das várias árvores, e pela média também sobre as múltiplas *folds* das experiências de validação cruzada, pode-se estimar uma importância para cada caraterística.

A Figura 3 ilustra as 20 caraterísticas descritivas mais importantes em termos dos coeficientes de regressão linear (i.e., as 10 caraterísticas descritivas com os maiores valores positivos ou negativos), ou em termos da posição relativa dos nós de decisão, para o caso de modelos de previsão de preços de quartos de hotel nas épocas altas e baixas. Como esperado, termos como *sheraton, luxo* ou *resort* parecem indicar preços mais altos, enquanto que termos como *pensão* ou *hostel* estão associados a preços mais baixos.

| | Hotéis Alta | | Hotéis Baixa | | Restaurantes | | Filmes | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| **Média** | 44,06 | 93,65 | 28,58 | 51,86 | 6,09 | 8,41 | 161,15 | 247,87 |
| **Mediana** | 39,96 | 95,69 | 26,59 | 52,92 | 5,80 | 8,97 | 140,48 | 264,29 |
| **Regressão em Crista** | 45,87 | 72,94 | 30,38 | 42,41 | 6,40 | 6,83 | 127,70 | 201,52 |
| **Lasso** | 35,78 | 72,96 | 24,27 | 43,60 | 4,59 | 6,57 | 183,01 | 268,33 |
| **Rede Elástica** | **34,63** | **70,86** | **23,25** | **41,97** | **4,27** | **6,20** | **127,55** | **192,70** |
| **Floresta Aleatória** | **34,25** | **74,13** | **23,17** | **44,25** | **4,40** | **6,56** | **135,89** | **211,77** |
| **Gradient Boosting** | 37,91 | 79,94 | 25,18 | 47,09 | 4,65 | 7,02 | 166,74 | 269,95 |

Tabela 2: Resultados da primeira experiência, com uma representação baseada em TF-IDF.

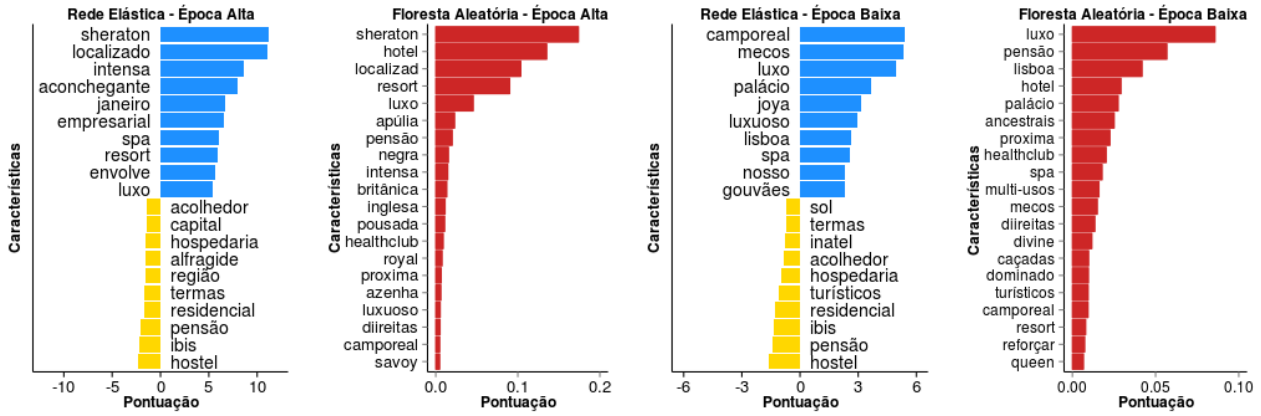


Figura 3: As 20 caraterísticas mais importantes para a previsão dos preços de quartos de hotéis.

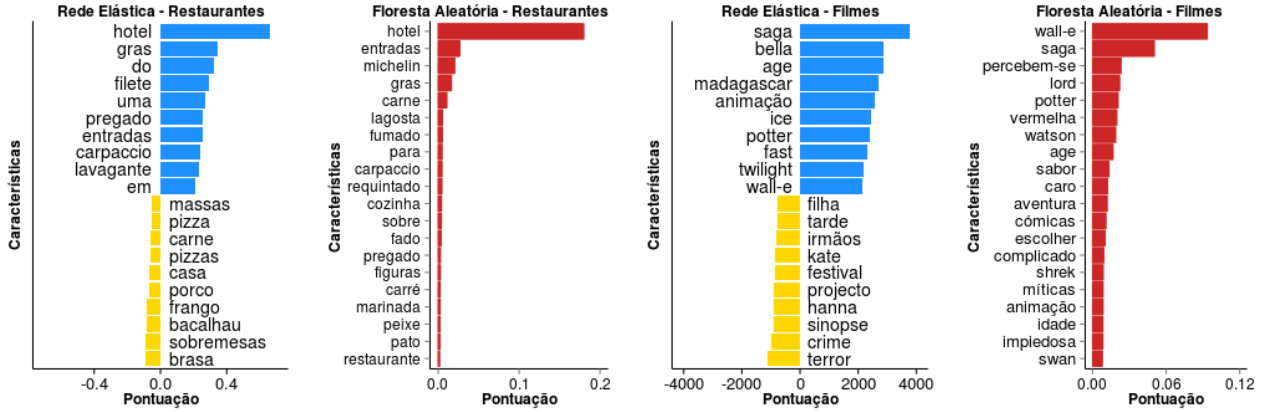


Figura 4: As 20 caraterísticas mais importantes para a previsão dos preços médios dos pratos em restaurantes, ou para a previsão dos resultados de bilheteira de filmes.

A Figura 4 mostra as 20 caraterísticas descritivas mais importantes, mas neste caso para os modelos de previsão de preços de refeições em restaurantes (i.e., os gráficos do lado esquerdo), e para os modelos de previsão de resultados de bilheteira de filmes. No caso de restaurantes, termos como *hotel* ou *michelin* parecem ser muito discriminativos, enquanto que no caso dos filmes, termos como *saga* ou *terror* parecem fornecer as melhores pistas. Na Figura 4 é possível notar que algumas palavras correspondentes a *stop-words* (e.g., as palavras *do* ou *uma*), ou correspondentes a nomes de filmes específicos (e.g., *wall-e*) estão associadas a importâncias altas nos modelos de regressão. Nos nossos testes, não usámos nenhuma estratégia de remoção das *stop-words*, e temos que estes valores podem indicar efeitos de sobre-ajustamento (i.e., *over-fitting*) dos modelos aos dados de treino. Importa realçar que os testes foram realizados apenas com conjuntos de dados relativamente pequenos.

As Tabelas 3 e 4 mostram uma comparação dos resultados obtidos com os melhores modelos, ou seja, a regressão linear com regularização com o método da rede elástica, e regressão com florestas aleatórias, respetivamente, utilizando cada

| | **Hoteis Alta** | | **Hotéis Baixa** | | **Restaurantes** | | **Filmes** | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| **Binário** | 40,91 | 77,49 | 27,14 | 46,64 | 5,94 | 8,22 | 133,01 | 199,34 |
| **Frequência do Termo** | 51,18 | 86,10 | 30,34 | 48,64 | 5,42 | 6,31 | 209,50 | 279,51 |
| **TF-IDF** | 34,63 | 70,86 | 23,25 | 41,97 | 4,27 | 6,20 | 127,55 | 192,70 |
| **Delta-TF-IDF** | **34,55** | **70,63** | 24,33 | 41,77 | 4,36 | 6,62 | 131,59 | 194,37 |
| **Delta-BM25** | 34,70 | 72,82 | **23,21** | **40,24** | **4,22** | **6,14** | **127,41** | **191,08** |

Tabela 3: Resultados com modelos de regressão com regularização dada pelo método da rede elástica, usando diferentes esquemas de ponderação das caraterísticas.

| | **Hotéis Alta** | | **Hotéis Baixa** | | **Restaurantes** | | **Filmes** | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| **Binário** | 36,56 | 79,06 | 24,28 | 46,19 | 4,83 | 7,08 | 135,68 | 214,60 |
| **Frequência do Termo** | 36,62 | 77,09 | 25,53 | 46,69 | 5,37 | 6,57 | 137,92 | 207,82 |
| **TF-IDF** | 34,25 | 74,13 | **23,17** | 44,25 | 4,40 | 6,56 | 135,89 | 211,77 |
| **Delta-TF-IDF** | **34,12** | **73,43** | 24,91 | 44,04 | **4,32** | 6,85 | **130,59** | **209,49** |
| **Delta-BM25** | 34,47 | 73,55 | 23,19 | **43,45** | 4,63 | **6,52** | 134,84 | 210,24 |

Tabela 4: Resultados com modelos baseados em florestas aleatórias de árvores de regressão, usando diferentes esquemas de ponderação das caraterísticas.

uma das representações para os conteúdos textuais. A representação mais rica é, talvez, dada pelo esquema Delta-BM25, embora o esquema Delta-TF-IDF tenha obtido resultados muito semelhantes ou até melhores, no caso dos modelos aprendidos com base na abordagem da floresta aleatória de árvores de regressão.

Além dos conteúdos textuais, considerámos caraterísticas derivadas de outros elementos de metadados, tais como a localização geográfica de hotéis e restaurantes, o tipo de restaurante, ou o número de salas em que o filme foi exibido no fim de semana de estreia. Também experimentámos medir os resultados após a adição de caraterísticas derivadas de agrupamentos automáticos de palavras semelhantes (i.e., *word clusters* gerados com o algoritmo de Brown), para a representação dos conteúdos textuais.

As propriedades de metadados, tais como a localização ou o tipo de restaurante, são representadas como valores binários (por exemplo, uma caraterística para cada distrito administrativo), assumindo o valor de um ou zero, dependendo da localização ou do tipo correspondente à descrição textual. Para o caso dos filmes, adicionámos ainda o número de salas como uma das dimensões do vetor de caraterísticas.

A Tabela 5 lista os resultados correspondentes à previsão de preços de quartos de hotéis com diferentes conjuntos de caraterísticas descritivas. A combinação de conteúdos textuais, metadados e agrupamentos de palavras obteve um melhor desempenho, no caso do uso da abordagem de regularização da rede elástica. Ao utilizar a técnica da floresta aleatória de árvores de regressão, a combinação de texto e localização geográfica atingiu melhores resultados.

A Tabela 6 apresenta os resultados para a previsão do preço médio das refeições em restaurantes, usando diferentes conjuntos de caraterísticas. Com o método da rede elástica, a experiência que considera apenas as caraterísticas descritivas baseadas nas palavras produziu melhores resultados. Com a abordagem da floresta aleatória de árvores de regressão, a combinação de caraterísticas derivadas do texto e localização atingiu melhores resultados.

Finalmente, a Tabela 7 mostra os resultados para a previsão de receitas de bilheteira de filmes. O número de salas tem uma forte influência nos resultados. Com ambos os tipos de modelos de regressão, a execução que obteve melhores resultados é claramente aquela que envolve a combinação de caraterísticas derivadas do texto com o número de salas. O número de salas e as receitas de bilheteira são, de facto, altamente correlacionadas, como mostra a Figura 5.

Para restaurantes e hotéis, também comparámos os resultados obtidos através dos nossos modelos de regressão contra abordagens mais simples, como prever com base no valor médio dos dados de treino, ou com base no valor médio e mediano por localização. No caso dos filmes, tentámos também usar apenas o número de salas. Estes resultados estão apresentados na Tabela 8. Conseguimos melhores resultados ao fazer as previsões com base no valor médio por localização, comparando com o uso da média ao longo de todo o conjunto de dados de treino. Na Tabela 8, apresentamos também os resultados em termos da

| | | Texto | | +WClusters | | +Localização | | Todas | |
|---|---|---|---|---|---|---|---|---|---|
| | | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| **Época Baixa** | **Rede Elástica** | 23,21 | 40,24 | 25,52 | 45,06 | 23,57 | 43,05 | **23,17** | **40,19** |
| | **Floresta Aleatória** | 23,19 | 43,45 | 25,33 | 46,06 | **23,18** | **43,06** | 23,76 | 44,23 |
| **Época Alta** | **Rede Elástica** | 34,70 | 72,82 | 39,39 | 75,47 | 34,75 | 70,46 | **34,00** | **70,13** |
| | **Floresta Aleatória** | 34,47 | 73,55 | 38,87 | 77,99 | 34,38 | 76,46 | **34,02** | **73,30** |

Tabela 5: Resultados da previsão de preços de quartos em hotéis, com diferentes conjuntos de características descritivas.

| | Texto | | +WClusters | | +Tipo | | +Localização | | Todas | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| **Rede Elástica** | **4,22** | **6,14** | 5,52 | 7,82 | 4,88 | 7,03 | 4,87 | 7,02 | 4,71 | 6,79 |
| **Floresta Aleatória** | 4,63 | 6,52 | 5,11 | 7,55 | 4,36 | 6,59 | **4,33** | **6,52** | 4,34 | 6,44 |

Tabela 6: Resultados da previsão de preços de refeições em restaurantes, com diferentes conjuntos de caraterísticas descritivas.

| | Texto | | +WClusters | | +Salas | | Todas | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| **Rede Elástica** | 127,91 | 191,08 | 126,48 | 191,30 | 79,06 | 125,45 | **72,95** | **122,29** |
| **Floresta Aleatória** | 127,41 | 191,08 | 138,46 | 220,45 | **66,02** | **137,78** | 72,49 | 135,14 |

Tabela 7: Resultados da previsão para os resultados de bilheteira associados a filmes, com diferentes conjuntos de caraterísticas descritivas.

métrica MRE, por forma a suportar comparação entre diferentes tarefas. A previsão de receitas de bilheteira de filmes apresenta-se como um problema mais difícil, com resultados ligeiramente piores em termos da medida MRE. Este problema em concreto é também aquele onde temos um menor volume de dados disponíveis.

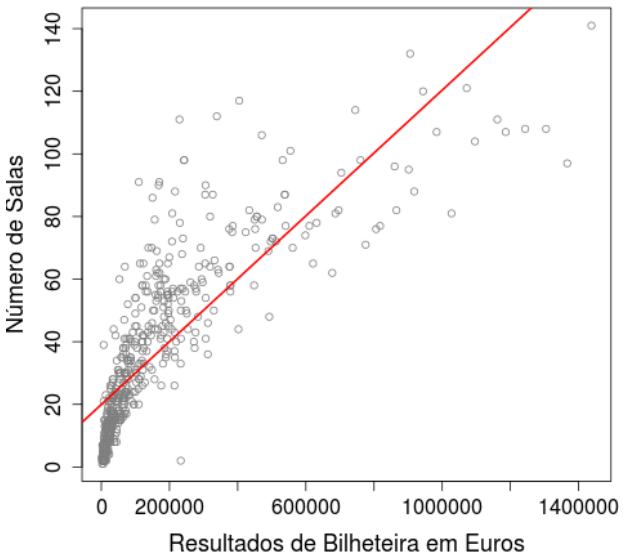


Figura 5: Resultados de bilheteira versus o número de salas no qual o filme foi apresentado.

Em suma, e como relatado na Tabela 8, podemos concluir que os modelos de regressão que usam caraterísticas derivadas do conteúdo textual de fato resultam em ganhos de exatidão para as tarefas de previsão consideradas. A adição de caraterísticas descritivas dos metadados, tais como a localização, resulta apenas em ligeiras melhorias sob modelos de regressão que usam caraterísticas baseadas no texto.

## 5 Conclusões e Trabalho Futuro

Este artigo apresentou um estudo experimental sobre a realização de previsões com base em conteúdos textuais escritos em português, e usando documentos associados a três domínios diferentes. As tarefas específicas abordadas no nosso trabalho envolveram (i) a previsão de preços de quartos para hotéis em Portugal, nas épocas altas e baixas para os turistas, usando descrições textuais recolhidas a partir de um portal Web conhecido, (ii) a previsão de preços médios de refeições em restaurantes localizados em Portugal, com descrições textuais para os restaurantes e os seus menus, recolhidos também a partir do mesmo portal Web, e (iii) a previsão de resultados de bilheteira de filmes, na primeira semana de exibição, conforme relatado pelo Instituto do Cinema e do Audiovisual, para filmes

| | Hotéis Alta | | | Hotéis Baixa | | | Restaurantes | | | Filmes | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MRE | MAE | RMSE | MRE | MAE | RMSE | MRE | MAE | RMSE | MRE |
| **Média** | 44,06 | 93,65 | 0,54 | 28,58 | 51,86 | 0,46 | 6,09 | 8,41 | 0,38 | 161,15 | 247,87 | 5,38 |
| **Média por Localização** | 40,57 | 90,80 | 0,48 | 28,11 | 50,91 | 0,46 | 5,81 | 8,11 | 0,36 | – | – | – |
| **Mediana por Localização** | 37,45 | 92,11 | 0,38 | 26,48 | 51,80 | 0,37 | 5,78 | 8,45 | 0,33 | – | – | – |
| **Número de Salas** | – | – | – | – | – | – | – | – | – | 83,45 | 131,89 | 2.01 |
| **Melhor Modelo** | **34,00** | **70,13** | **0,37** | **23,17** | **40,19** | **0,35** | **4,22** | **6,14** | **0,31** | **66,02** | **122,29** | **0,63** |

Tabela 8: Resultados gerais para as diferentes tarefas de previsão.

exibidos em Portugal e usando comentários textuais de outro portal Web bem conhecido. Relatámos especificamente experiências utilizando diferentes tipos de modelos de regressão, usando esquemas de ponderação para as caraterísticas do actual estado da arte, e usando caraterísticas derivadas de representações para as palavras baseadas no agrupamento automático das mesmas. Através das nossas experiências, conseguimos demonstrar claramente que os modelos de previsão usando a informação textual alcançam melhores resultados do que abordagens mais simples, tais como realizar previsões com base no valor médio dos dados de treino. Demonstrámos ainda que o uso de representações de documentos mais ricas (e.g., usando o algoritmo de Brown para o agrupamento automático de palavras, e o esquema de ponderação das caraterísticas Delta-TF-IDF) resultara em ligeiras melhorias no desempenho.

Apesar dos resultados interessantes, há muitas ideias para trabalho futuro. Dado que só temos acesso a conjuntos de dados de treino relativamente pequenos, acreditamos que um caminho interessante se relaciona com a avaliação de técnicas de aprendizagem semi-supervisionada, capazes de aproveitar grandes quantidades de dados não anotados. Também nos parece razoável supor que as pistas para estimar corretamente um determinado valor alvo, com base num documento textual, podem estar contidas num subconjunto pequeno das frases do documento. Yogatama e Smith (2014) introduziram um algoritmo de aprendizagem que explora esta intuição, através de uma abordagem de regularização cuidadosamente projetada, mostrando que o método resultante pode superar significativamente outras abordagens (e.g., regularizadores padrão como os métodos *ridge*, Lasso, e a rede elástica) em diversos problemas de categorização de texto. Finalmente, dado o sucesso parcial das representações de documentos feitas com base no algoritmo de agrupamento automático de palavras desenvolvido por Brown, gostaríamos de experimentar com outros tipos de representações baseadas em similaridade distribucional, tais como as representações de palavras enquanto vetores densos de baixa dimensionalidade, propostas no estudo de Mikolov et al. (2013).

## Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e a Tecnologia (FCT), através do projeto com referência UTA-Est/MAI/0006/2009 (REACTION), bem como através do financiamento plurianual do laboratório associado INESC-ID, com a referência PEst-OE/EEI/LA0021/2013.

Agradecemos particularmente aos nossos colegas do projeto REACTION acima mencionado, pela sua ajuda e pelas observações pertinentes.

## Referências

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fern, A Bacelar Do Nascimento, Filipe Nunes, e João Ricardo Silva. 2006. Open resources and tools for the shallow processing of Portuguese: The TagShare project. Em *Proceedings of the International Conference on Language Resources and Evaluation*.

Bollen, Johan, Huina Mao, e Xiao-Jun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 1(2).

Boyd, Stephen e Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Breiman, Leo. 1996. Bagging Predictors. *Machine Learning*, 24(2).

Breiman, Leo. 2001. Random Forests. *Machine Learning*, 45(1).

Breiman, Leo, J. H. Friedman, R. A. Olshen, e C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth and Brooks.

Brown, Peter F., Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra, e Jenifer C. Lai. 1992. Class-based $n$-gram models of natural language. *Computational Linguistics*, 18(4).

Chahuneau, Victor, Kevin Gimpel, Bryan R. Routledge, Lily Scherlis, e Noah A. Smith. 2012. Word salad: Relating food prices and

descriptions. Em *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Dahllöf, Mats. 2012. Automatic prediction of gender, political affiliation, and age in swedish politicians from the wording of their speeches—a comparative study of classifiability. *Literary and Linguistic Computing*, 27(2).

Fridman, Jerome, Trevor Hastie, e Rob Tibshirani. 2008. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5).

Hong, Yancheng e Steven Skiena. 2010. The wisdom of bookies? sentiment analysis vs. the NFL point spread. Em *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, e Noah A. Smith. 2010. Movie reviews and revenues: An experiment in text regression. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Kim, Yongdai e Jinseog Kim. 2006. Gradient Lasso for feature selection. Em *Proceedings of the International Conference on Machine Learning*.

Kogan, Shimon, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, e Noah A. Smith. 2009. Predicting risk from financial reports with regression. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Lerman, Kevin, Ari Gilder, Mark Dredze, e Fernando Pereira. 2008. Reading the markets: forecasting public opinion of political candidates by news analysis. Em *Proceedings of the International Conference on Computational Linguistics*.

Luo, Xueming, Jie Zhang, e Wenjing Duan. 2013. Social media and firm equity value. *Information Systems Research*, 24(1).

Martineau, Justin e Tim Finin. 2009. Delta TF-IDF: An improved feature space for sentiment analysis. Em *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of Royal Statistical society*, 42(2).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, e Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Proceedings of the Conference on Neural Information Processing Systems*.

Mitchell, Lewis, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, e Christopher M. Danforth. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, 8(5).

O'Connory, Brendan, Ramnath Balasubramanyany, Bryan R. Routledge, e Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. Em *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

Paltoglou, Georgios e Mike Thelwall. 2010. A study of information retrieval weighting schemes for sentiment analysis. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Pang, Bo e Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2).

Radinsky, Kira. 2012. Learning to predict the future using Web knowledge and dynamics. *ACM SIGIR Forum*, 46(2).

Schumaker, Robert P. e Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems*, 27(12).

Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Megha Agrawal Gregory J. Park, Shrinidhi K. Lakshmikanth, Sneha Jha, Martin E. P. Seligman, Lyle Ungar, e Richard E. Lucas. 2013. Characterizing geographic variation in well-being using tweets. Em *Proceeding of the AAAI International Conference on Weblogs and Social Media*.

Sewell, Martin. 2011. Ensemble methods. Relatório Técnico RN/11/02, University College London Departament of Computer Science.

Smith, Noah A. 2010. Text-Driven Forecasting.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58(1).

Tirunillai, Seshadri e Gerard J. Tellis. 2012. Does chatter really matter? Dynamics of

User-Generated Content and Stock Performance. *Information Systems Research*, 31(2).

Turian, Joseph, Lev Ratinov, e Yoshua Bengio. 2010. Word representation: a simple and general method for semi-supervised learning. Em *Proceedings of the Annual Meeting of the Association for Computional Linguistics*.

Wright, Stephen J., Robert D. Nowak, e Mário A. T. Figueiredo. 2009. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7).

Yano, Tae, William W. Cohen, e Noah A. Smith. 2009. Predicting response to political blog posts with topic models. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Yano, Tae e Noah A. Smith. 2010. What's worthy of comment? Content and Comment Volume in Political Blogs. Em *Proceedings of the AAAI International Conference on Weblogs and Social Media*.

Yano, Tae, Noah A. Smith, e John D. Wilkerson. 2012. Textual predictors of bill survival in congressional committees. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Yogatama, Dani e Noah A. Smith. 2014. Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. Em *Proceedings of the International Conference on Machine Learning*.

Zou, Hui e Trevor Hastie. 2005. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B*, 67(5).

# PoNTE: apontando para corpos de aprendizes de tradução avançados

**PoNTE: a bridge for advanced translation learners**

Diana Santos
Linguateca, Universidade de Oslo
d.s.m.santos@ilos.uio.no

**Resumo**

É possível ensinar usando os materiais criados pelos próprios alunos, e ao mesmo tempo anotá-los para obter mais material que fique público para mais professores e estudantes? É possível desenvolver o DISPARA, inicialmente concebido para disponibilizar corpos de tradução "tradicionais", de forma a conter mais um nível de anotação de "crítica de tradução"? O projeto PoNTE pretende ser uma abordagem inicial a estas duas questões.

Neste artigo, descrevo o tipo de comentários e anotação crítica que seria desejável ter codificado num corpo deste género, a sua primeira implementação e estudos realizados, e os problemas técnicos que ainda se põem na gestão de um corpo sempre crescente.

**Palavras chave**

Ensino de português como língua estrangeira, corpos paralelos, ferramentas de apoio ao ensino, ferramentas de corpos, tradução, linguística contrastiva

**Abstract**

Is it possible to teach using materials created by the students themselves, while at the same time annotate them to raise more material available to other teachers and students? Is it possible to further develop DISPARA, initially deployed for "traditional" translation corpora, so that it contains an extra annotation level with "translation critique"? The PoNTE project ("*ponte*" means bridge in Portuguese, and PoNTE stands for "Portuguese-Norwegian Translation Examples") offers an initial approach to these two questions.

In this paper, I describe the type of comments and critical annotations aimed for, together with a first implementation, a short reference to some studies performed, and discuss the technical and philosophical problems involved in the management of a dynamic and always growing corpus.

**Keywords**

Portuguese for foreigners, parallel corpora, computer aided language learning, teaching tools, corpora tools, translation, contrastive linguistics

## 1 Apresentação

A esmagadora maioria dos corpos paralelos envolvendo o português tem o inglês como A outra língua, ou como a língua pivô, o que se explica naturalmente pelo peso que o inglês tem a nível internacional e o facto de corresponder à língua mais traduzida para português e vice-versa (a língua na qual se encontram mais textos traduzidos de um original em português, cf. Rosa (2006)).

Contudo, e sobretudo numa situação de ensino da língua, é conhecida a importância da língua materna assim como a importância do treino na tradução e na retroversão associada à língua materna. (Estou plenamente consciente de que existem várias opiniões divergentes sobre o assunto, mas eu alinho com a fação, possivelmente minoritária presentemente, que afirma que esse é um assunto que apaixona os alunos como poucos, em completo acordo com o que escreve Kåre Nilsson (1997), o meu predecessor na Universidade de Oslo.)

Na Universidade de Oslo, a esmagadora maioria dos alunos tem o norueguês como língua materna, e os poucos que têm o português ou outra têm excelentes conhecimentos do norueguês, o que tornou óbvio um ensino da gramática (e cultura) relacionado – também – com a tradução. Assim surgiu o PoNTE (Portuguese-Norwegian Translation Examples), um corpo de múltiplas traduções entre as línguas portuguesa e norueguesa, que usa as traduções dos alunos como textos de chegada.

Ao contrário do que é geralmente assumido pelos pessoas de fora do ambiente do ensino das línguas, os professores de línguas na sua prática pedagógica não têm geralmente tempo para criar ou recorrer a um corpo criado a partir dos materiais dos seus próprios alunos. Quando criam esses corpos, é com a ideia de que virão um dia a ser úteis na sua prática diária, ou para os tempos vindouros.

Um bom indicador do irrealismo desta esperança foi o recente inquérito de Sylviane Granger na lista corpora[1], a 22 de outubro de 2013:

> I'm looking for as much information as possible on concrete uses of learner corpus data for: (...) Many publications describe the potential use of learner corpus data, but I would like to collect information on actual use.[2]

Sylviane Granger é talvez a mais famosa especialista de corpos de aprendentes, como se pode ver por Granger, Gilquin e Meunier (2013). O seu pedido na lista teve resposta de nove investigadores diferentes (sumarizado a 31 de dezembro de 2013), e apenas um semelhante ao PoNTE (Kutuzov et al., 2012), que discutiremos mais adiante.

Por essa razão, pode ser interessante realçar que praticamente desde o primeiro instante eu usei o PoNTE no meu ensino, ou melhor, os resultados da invocação da sua primeira versão, como auxiliar pedagógico na sala de aula.

De qualquer forma, não me distingo dos outros investigadores quando reconheço que a criação última deste corpo pretende vir a permitir – mais tarde, quando tiver chegado a um maior tamanho, ou no caminho enquanto é incrementado – identificar, e permitir estudar:

- o que é difícil de compreender em português para alunos noruegueses,
- o que é difícil de traduzir para norueguês, e
- o que é difícil de exprimir em português (vindo do norueguês).

É importante realçar, neste contexto, que a própria construção do corpo e a simples comparação das traduções por si só não é necessariamente útil para os "tradutores" (alunos de bacharelato), que podem até ser inundados de piores traduções do que as suas próprias. Como sempre insisti ao apresentar o material, e eles também pedem, o mais importante para a sua aprendizagem é o comentário por parte do professor.

Assim, em 2012 passei a disponibilizar, aos tradutores de cada texto, o conjunto das traduções, marcando a negrito os casos em que tinha havido clara diferença entre as traduções, e nesse caso escolhendo o trecho (que pode ir de uma palavra à unidade de alinhamento total) que me parecia melhor. Nos casos de nenhuma tradução ser satisfatória, marquei a negrito o (trecho do) original.

Este material serve assim como lembrança da aula em que estiveram e ouviram os meus comentários, mas naturalmente não permite a subsequente recuperação, por parte de outro utilizador do corpo, dos problemas e soluções anotados, até porque apenas sublinha partes do texto.

Por isso a nossa intenção de anotar o PoNTE com informação mais explícita e estruturada para subsequente procura, informação que chamo, a partir daqui, "informação crítica".

Embora o PoNTE seja um trabalho em progresso, porque a anotação crítica ainda não foi incorporada nem fixado o sistema subjacente para essa mesma anotação, a mera existência de um corpo paralelo nas duas línguas, com abundância de traduções de um mesmo texto, já permitiu e permite alguma investigação interessante, como tentaremos demonstrar no presente artigo.

Assim, após descrever brevemente o conteúdo presente na secção 2 e a implementação atual na secção 3, faço uma breve menção a alguns estudos já efetuados sobre (versões anteriores de) o PoNTE na secção 4 e discuto finalmente a questão da anotação crítica na secção 5, documentando brevemente também duas anotações piloto.

A publicação deste texto tem, assim, dois objetivos: anunciar e documentar um novo recurso, e explicar as opções tomadas e as que ainda estão por tomar.

## 2 Breve descrição do conteúdo do corpo

O PoNTE é um corpo em permanente evolução visto que deriva de uma atividade didatica continuada.

Os originais são textos curtos, selecionados para serem traduzidos, por vezes correspondendo a excertos de textos. Dado que as aulas têm lugar todos os anos, prevê-se que o número de textos diferentes, assim como o número das suas traduções, vá aumentando a um ritmo constante.

Em janeiro de 2014, os textos presentes no corpo, assim como o número de traduções distintas, estão indicados na Tabela 1 para originais em português e na Tabela 2 para originais em norueguês.

[1] Lista de discussão `corpora@uib.no`

[2] A minha tradução: Estou à procura de exemplos concretos de corpos de aprendentes para : (...) Muitas publicações descrevem o uso potencial de corpos de aprendizes, mas eu gostava de recolher informação sobre o uso real.

A lista das fontes encontra-se acessível do sítio do projeto[3].

Os textos são de variantes diferentes, de géneros diferentes, e publicados em suportes diferentes, só tendo em comum possuirem um tamanho razoavelmente pequeno (ver Figura 1 para o número de palavras).

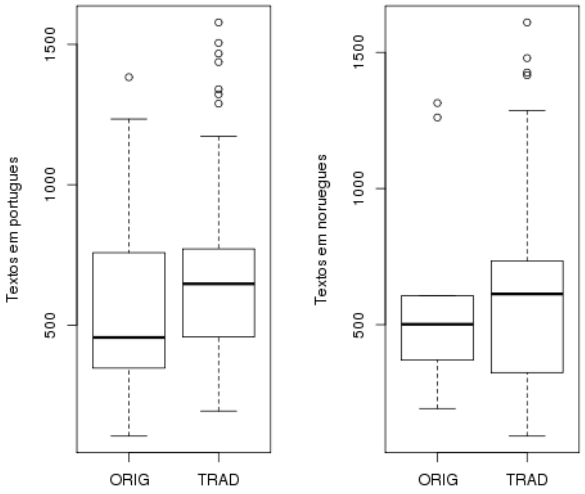


Figura 1: Tamanho dos textos do PoNTE em número de palavras. Repare-se que TRAD se refere às traduções dos ORIG da outra língua.

Assim, e sem entrar em pormenores, temos textos brasileiros, moçambicanos, portugueses e angolanos, e contos, reportagens, blogues, humor, burocracia, promessas eleitorais e texto técnico no que toca aos originais em português, e textos jornalísticos, literários e cartas no que se refere ao norueguês, que também é representado em três variantes (bokmaal, nynorsk e riksmaal[4]).

De uma forma mais condensada, podemos dizer que existem 266 pares de tradução, 195 traduções para norueguês e 71 para português. A razão desse desiquilíbrio é clara: há muito mais alunos nas cadeiras iniciais (que traduzem do português para o norueguês) do que nas mais avançadas, em que fazem retroversão para português.

Mais importante do que a constituição presente, contudo, é a análise das potencialidades deste material para o ensino da tradução e para a linguística contrastiva, que serão tornadas possíveis pela sua anotação crítica. Depois de descrever o estado atual do projeto, e a sua implementação, debruçar-nos-emos sobre esta.

[3] `http://www.linguateca.pt/PoNTE/`

[4] Para quem não conheça a situação linguística da Noruega, menciono simplesmente que existem duas grafias modernas, e uma antiga, próxima do dinamarquês atual.

| Texto | Trads | Formas | Tipos |
|---|---|---|---|
| AMAZ | 10 | 161 | 113 |
| BP | 5 | 276 | 162 |
| BRI | 16 | 234 | 142 |
| CAMP | 9 | 456 | 233 |
| CAR | 18 | 646 | 299 |
| CIE | 16 | 1234 | 454 |
| DDS | 3 | 687 | 272 |
| DIL | 11 | 958 | 436 |
| DSC | 17 | 795 | 400 |
| EDS | 6 | 305 | 156 |
| ELEI | 7 | 734 | 322 |
| EPA | 4 | 1383 | 525 |
| EXA | 9 | 105 | 70 |
| JP | 4 | 346 | 196 |
| LOG | 2 | 417 | 202 |
| MEC | 1 | 857 | 306 |
| MIA | 5 | 834 | 403 |
| MRC | 14 | 758 | 326 |
| MUL | 3 | 448 | 264 |
| OC | 18 | 746 | 346 |
| PIB | 2 | 379 | 190 |
| SEM | 6 | 371 | 212 |
| SPG | 1 | 557 | 270 |
| TED | 6 | 368 | 209 |
| VAN | 2 | 347 | 156 |

Tabela 1: Composição atual do PoNTE, em termos dos originais em português.

| Texto | Trads | Formas | Tipos |
|---|---|---|---|
| BEB | 2 | 1314 | 501 |
| BRS | 9 | 523 | 262 |
| CLI | 11 | 192 | 130 |
| DN | 3 | 362 | 222 |
| KB | 2 | 370 | 222 |
| MOB | 13 | 480 | 258 |
| MUS | 3 | 389 | 224 |
| QUEI | 8 | 1261 | 450 |
| SAU | 14 | 606 | 324 |
| VES | 6 | 570 | 282 |

Tabela 2: Composição atual do PoNTE, em termos dos originais em norueguês.

## 3 Implementação

Como seria natural, a implementação do PoNTE é mais uma adaptação do DISPARA (Santos, 2002) originalmente criada para o COMPARA (Frankenberg-Garcia e Santos, 2002), mas que tem sido expandida e modificada para outros corpos paralelos como o Squirrel (Borin, Carlson e Santos, 2001), o CorTrad (Tagnin, Teixeira e

Santos, 2009; Teixeira, Santos e Tagnin, 2011), e agora o PoNTE e o PANTERA[5].

### 3.1 Interface

De facto, e do ponto de vista do utilizador, o PoNTE tem duas interfaces diferentes: uma que, a cada texto original, alinha todas as traduções existentes (e que é constituída por casos vazios quando o número de traduções é inferior ao máximo existente), e outra que alinha tantas vezes quantas um texto original foi traduzido, contendo assim o número total de pares original-tradução.

A razão da existência destas duas interfaces é que um utilizador pode estar interessado na comparação de traduções de um mesmo texto (caso em que usará a primeira interface), ou na contabilização de fenómenos de tradução efetuados por tradutores independentes (e, para isso, a segunda é mais natural).

Noto que a ordem das traduções não é relevante no PoNTE. Por exemplo, a segunda tradução corresponde a tradutores diferentes em textos diferentes (a numeração foi feita pela ordem de chegada das traduções, e nem todos os alunos as fizeram, nem as enviaram pela mesma ordem). Por razões de anonimização não é possível saber quem efetuou qual tradução, embora eu tenha alguma meta-informação sobre os tradutores, por exemplo a sua língua materna e a variante de português que privilegiam. A cada tradutor/aluno é atribuído um identificador, o que pode ser importante para a compreensão de alguns problemas, e esse identificador (que permitirá recuperar a língua materna, etc.) poderá vir a ser adicionado numa futura versão para refinar as procuras.

Nas figuras 2 e 3 apresenta-se a forma das duas interfaces, enquanto que as figuras 4 e 5 mostram os respetivos tipos de resultado.

### 3.2 Conceção teórica

De uma forma mais teórica, podemos reapresentar a implementação assim: Embora se conceba o PoNTE, abstratamente, como um único corpo, na prática ele é codificado através de uma série de diferentes corpos CWB, como aliás é ou foi o caso na implementação do COMPARA e do CorTrad, cujas interfaces invocam de facto vários corpos diferentes – conforme a direção da procura, no COMPARA, e também conforme o género, no CorTrad.

[5]Veja-se `http://www.linguateca.pt/PANTERA/` para este último, em fase de arranque.

No entanto, a conceção do PoNTE, além de também distinguir a direção, foi mais radical porque, além de conter um corpo "simples" por cada texto original, engloba também um corpo com as múltiplas traduções de cada texto, o chamado "PoNTE condensado", e um corpo com todos os pares de traduções concebidos como um novo caso, o chamado "PoNTE distribuído". (Este caso é semelhante aos dois únicos exemplos de traduções múltiplas no COMPARA).

Nesse aspeto, assemelha-se ao Águia (Santos, 2003), sistema de procura na Floresta Sintática, que, numa mesma página de interface na rede, junta dois corpos organizados de forma completamente diferente: um por texto, e outro por classificação gramatical (tipo de sintagma).

Ao adicionar um novo par de tradução ao PoNTE, diversos programas em Perl são ativados e (re)criam vários corpos.

### 3.3 Integração com o resto dos corpos da Linguateca

Outra questão extremamente importante, em que nunca é demais insistir, é o facto de que os textos em português recebem o mesmo tratamento, em termos de anotação e de revisão, que o já extenso material corpóreo da Linguateca contém: assim, tanto a anotação pelo PALAVRAS (Bick, 2000), como a anotação com a cor e outros campos semânticos, assim como a sua revisão – para um exemplo desta, cf. Maia e Santos (2012).

A possibilidade de comparar com, evocando, muito maiores quantidades de texto em português é assim algo que é oferecido automaticamente pela integração do DISPARA com o AC/DC (Costa, Santos e Rocha, 2009).

## 4 Alguns estudos usando o PoNTE

Aqui apresentamos brevemente dois trabalhos que já fizeram uso do PoNTE – usando diferentes versões, correspondendo a diferentes estágios do seu desenvolvimento, para motivar o seu uso e para mostrar que é possível já tirar algum proveito da sua existência, mesmo em fases preliminares de desenvolvimento.

### 4.1 Dativos possessivos

Um fenómeno relativamente frequente em português, e que tem merecido pouco interesse a nível monolingue, é a questão do dativo possessivo, que se torna contudo imediatamente relevante se considerarmos o seu ensino a falantes de línguas germânicas (como o norueguês ou

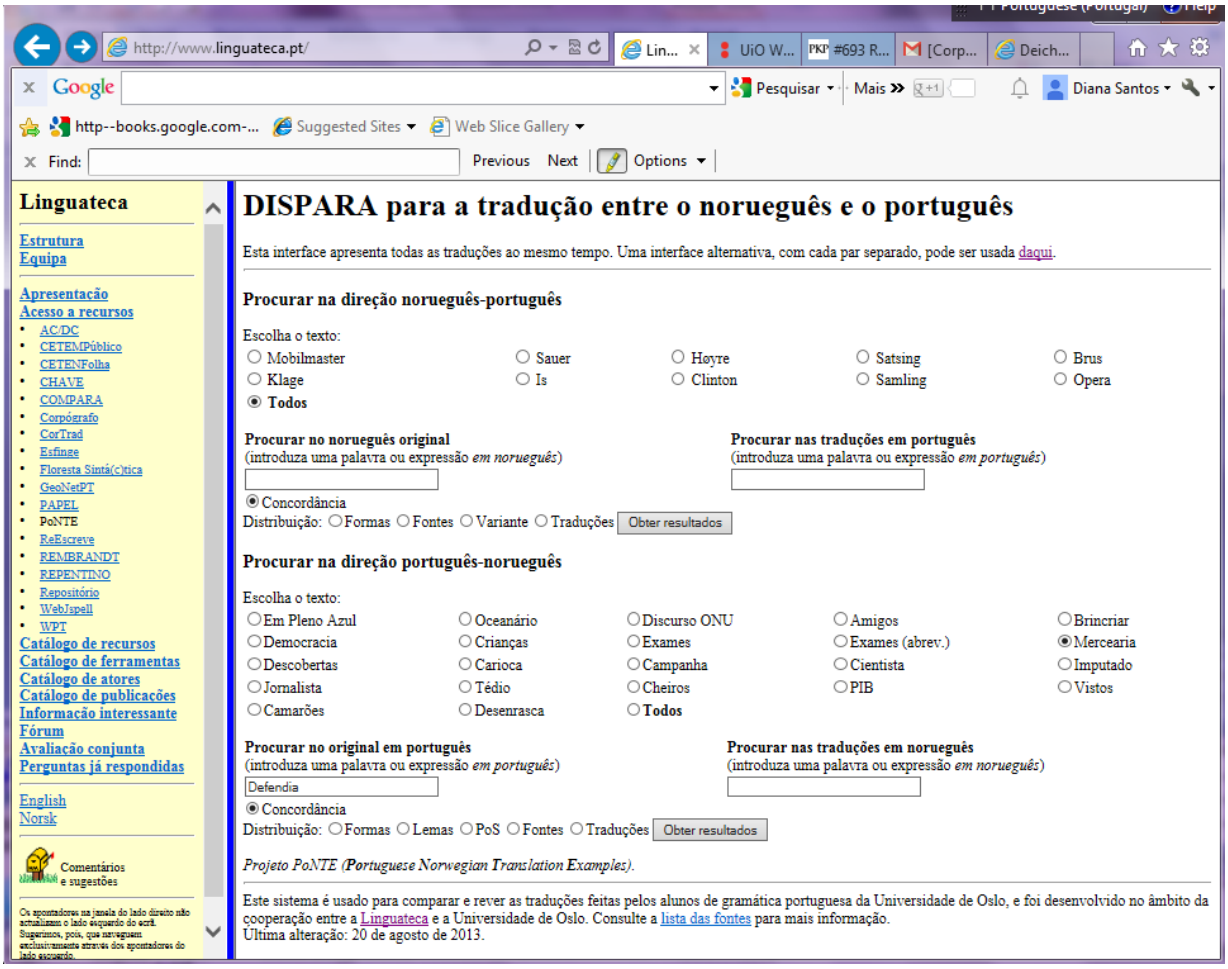


Figura 2: Exemplo de invocação do PoNTE compacto.

o inglês), devido à quase sistemática diferença entre as línguas em relação à descrição dessas situações.

Veja-se Santos (2012; Santos (2014d) para a descrição do fenómeno e da sua extensão contrastiva, usando corpos paralelos com o inglês.

Dado que o PoNTE é muito mais pequeno, que o fenómeno pressupõe um comando nativo ou quase nativo da língua, e que os aprendizes de tradutores tendem a manter a estrutura do original e não procurar traduções mais idiomáticas, a previsão que fazemos é a de que os tradutores aprendizes noruegueses muito raramente usarão dativos possessivos em português, e que, se aparecerem no original, dado ainda não terem sido alertados para esta diferença, talvez os traduzam literalmente (ou seja, como um pronome pessoal dativo).

Contudo, conseguimos identificar um caso destes no PoNTE (na tradução de português para norueguês) em que o aluno usou um pronome possessivo na tradução.

(1) Exceto que você não estará «estudando»; você estará trabalhando, gerando conhecimento, e contribuindo para as universidades publicarem os artigos científicos que **lhes** servem como base de avaliação no cenário mundial.
Bortsett fra at du ikke skal «studere»; du skal arbeide, skaffe deg kunnskap og bidra til universitetenes publikasjoner av vitenskapelige artikler som legger grunnlaget for **deres** verdsettelse i verdenssamfunnet.
. . . a avaliação delas no mundo.

Este é, além disso, um caso interessante precisamente porque foi identificado em português do Brasil, onde tal fenómeno é mais raro, devido entre outras coisas à tendência para eliminação dos clíticos, bem patente nesta variante do português (Bakkejord, 2008).

O importante a reter deste caso, contudo, é que a existência de dados contrastivos para este par de línguas permite reforçar a convicção de que este assunto deve ser ensinado a alunos de português como língua estrangeira.

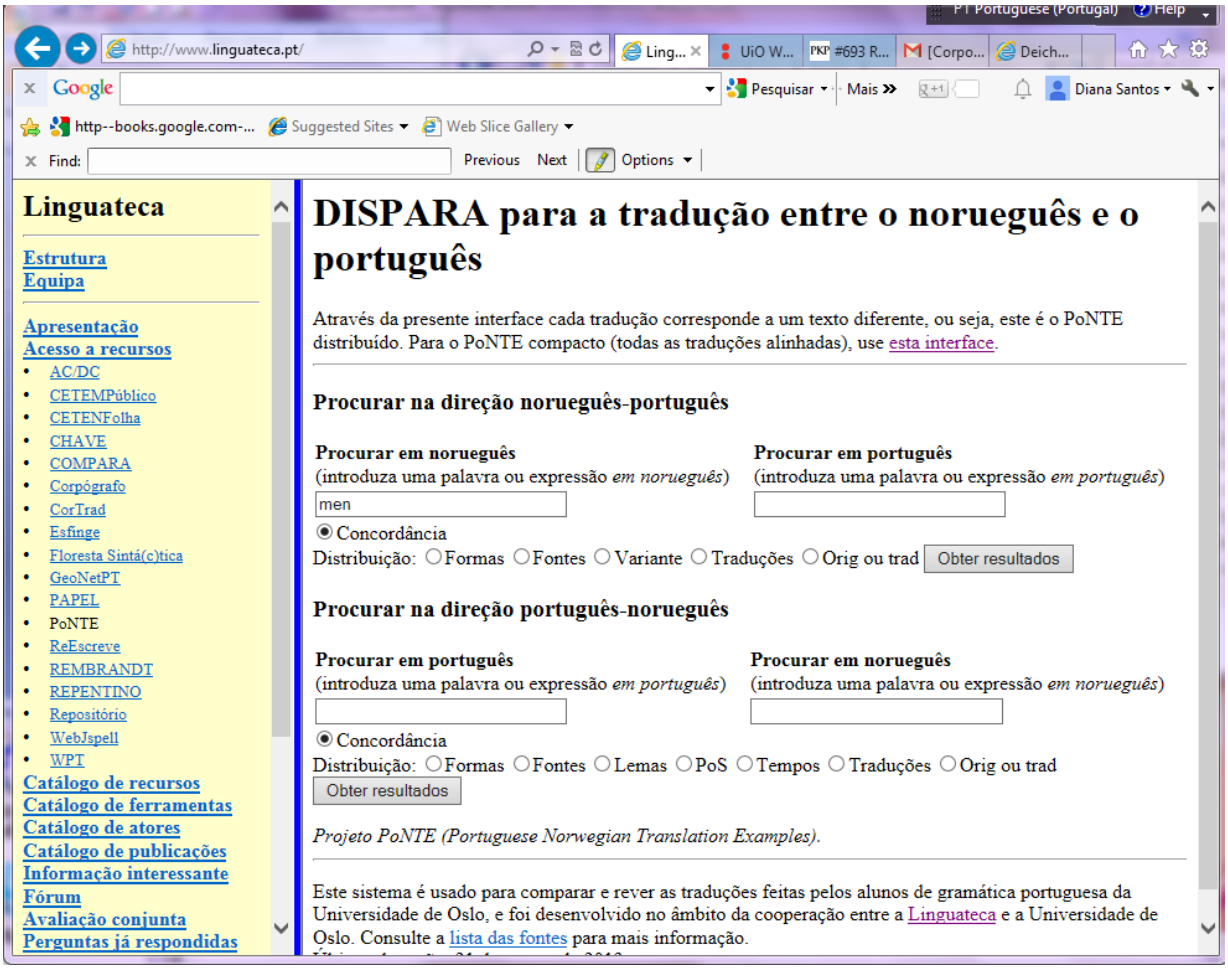


Figura 3: Exemplo de invocação do PoNTE distribuído.

### 4.2 Comparação de verbos estativos

Outra área em que o PoNTE já foi utilizado, agora como a principal fonte de dados, é o contraste dos verbos leves e frequentes como *ser*, *estar*, *ficar*, *ter* e *haver* (Santos, 2013; Santos, 2014e) com os seus correspondentes em norueguês.

Visto que estes verbos são dos mais frequentes nas duas línguas, além da identificação de exemplos interessantes também foi possível fazer uso da comparação de diversas traduções, e distinguir diferentes perfis entre texto traduzido e texto original, como a Figura 6 de (Santos, 2014e) ilustra.

Podemos assim afirmar, com base no PoNTE, que os textos em português têm mais menções que são traduzidas por *bli* e *være* do que os originais em norueguês.

## 5 Anotação crítica

Seja como for, considero que o mais importante do PoNTE ainda não foi implementado, e que é preciso alguma reflexão em torno das funcionalidades desejáveis. Este artigo é, pois, uma forma de partilhar essas dúvidas e questões com a comunidade científica, e talvez com a comunidade de futuros utilizadores do recurso ou de recursos semelhantes.

Em primeiro lugar, **o que** pretendemos anotar?

Há dois tipos de anotação que se referem a níveis distintos da crítica de tradução, correspondentes às bem conhecidas dimensões de preservação do sentido e de fluência do texto final:

- marcar aquilo que não foi compreendido pelos tradutores (e portanto erro de conteúdo);
- apontar aquilo que foi mal expresso para o público alvo (e portanto erros de adequação ou de formulação).

A pergunta seguinte é **onde** anotar essas questões. Do lado da língua fonte, ou do lado da língua alvo?

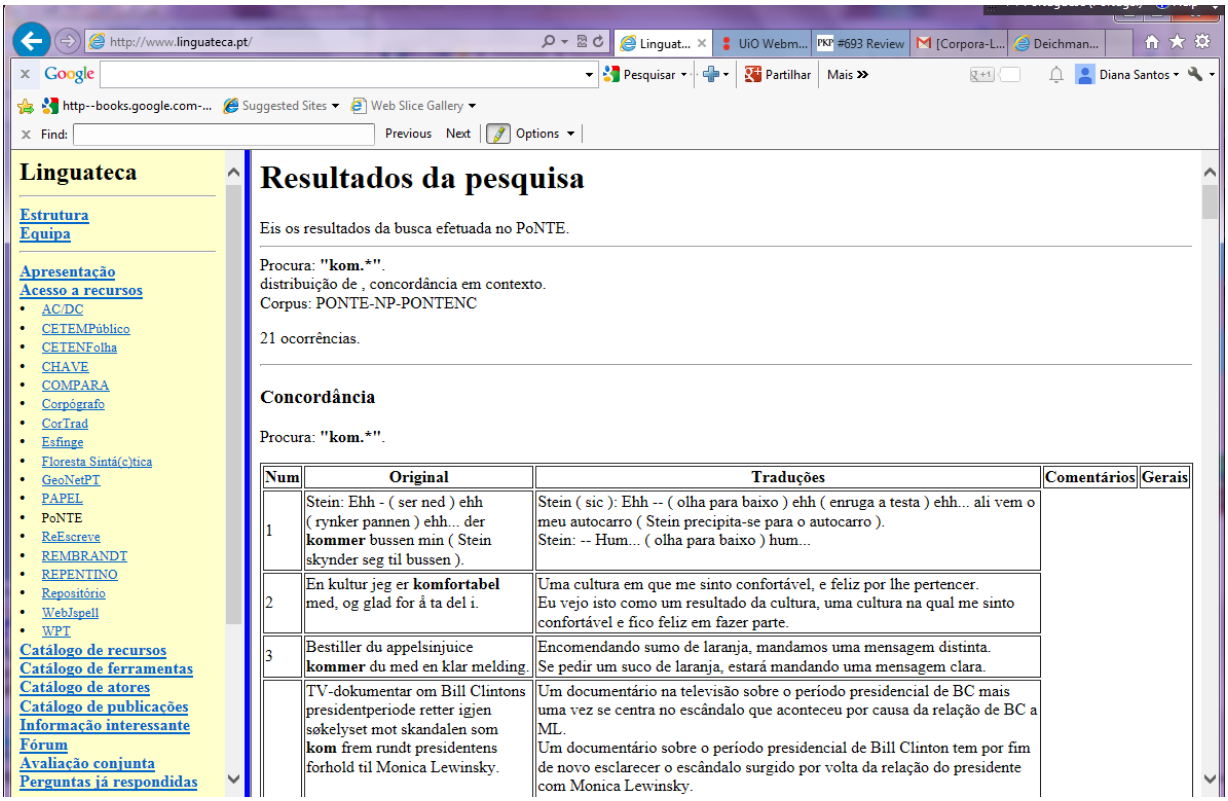


Figura 4: Exemplo de resultado do PoNTE compacto, mostrando ao mesmo tempo todas as traduções de uma frase original.

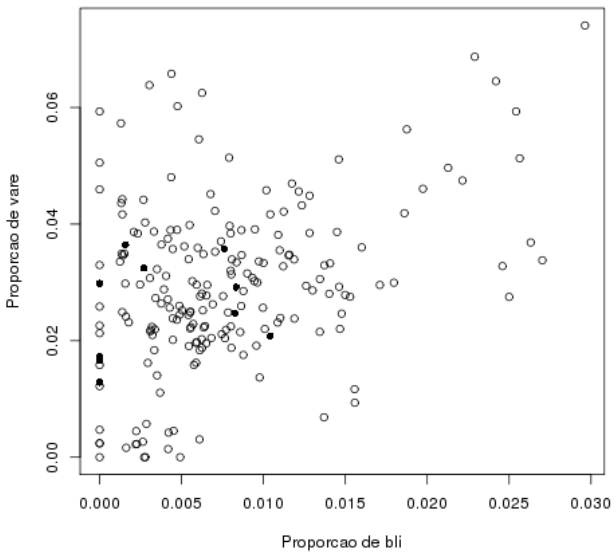


Figura 6: Distribuição da proporção de bli vs. vare em textos originais (bola preta) e traduzidos (bola branca)

Do ponto de vista de onde marcar, parece-me natural marcar associado à unidade de tradução como um todo os casos em que houve dificuldades generalizadas na tradução, enquanto que apenas faz sentido marcar em cada tradução os casos específicos em que houve um erro ou problema.

Ainda que com o advento de mais traduções de um mesmo texto possa haver casos problemáticos que deixem de o ser (por exemplo, casos em que todas as traduções estavam mal num dado ano podem passar a ter algumas traduções sem problemas ao adicionar as traduções dos anos subsequentes), parece-me relevante distinguir os casos em que a probabilidade de haver problemas (medida pelo número de más traduções) seja maior do que a existência de problemas pontuais, e que esta marcação, por corresponder a mais problemas, tenha prioridade.

Em segundo lugar, qual a "unidade" anotada? A frase inteira? Um sintagma, uma oração? De facto, não existe sempre uma resposta simples sobre o nível (linguístico) em que um determinado erro ou problema deve ser atacado, como aliás o demonstrou cabalmente a tentativa de produzir uma "ontologia" de erros de tradução com o TrAva (Santos, Maia e Sarmento, 2004; Sarmento et al., 2007). Por isso, desde logo consideramos a possibilidade, e vantagem, de um sistema aberto em que um mesmo caso possa ter múltiplas anotações.

Uma coisa é, contudo, clara: a existência de traduções com unidades (frases) relativamente

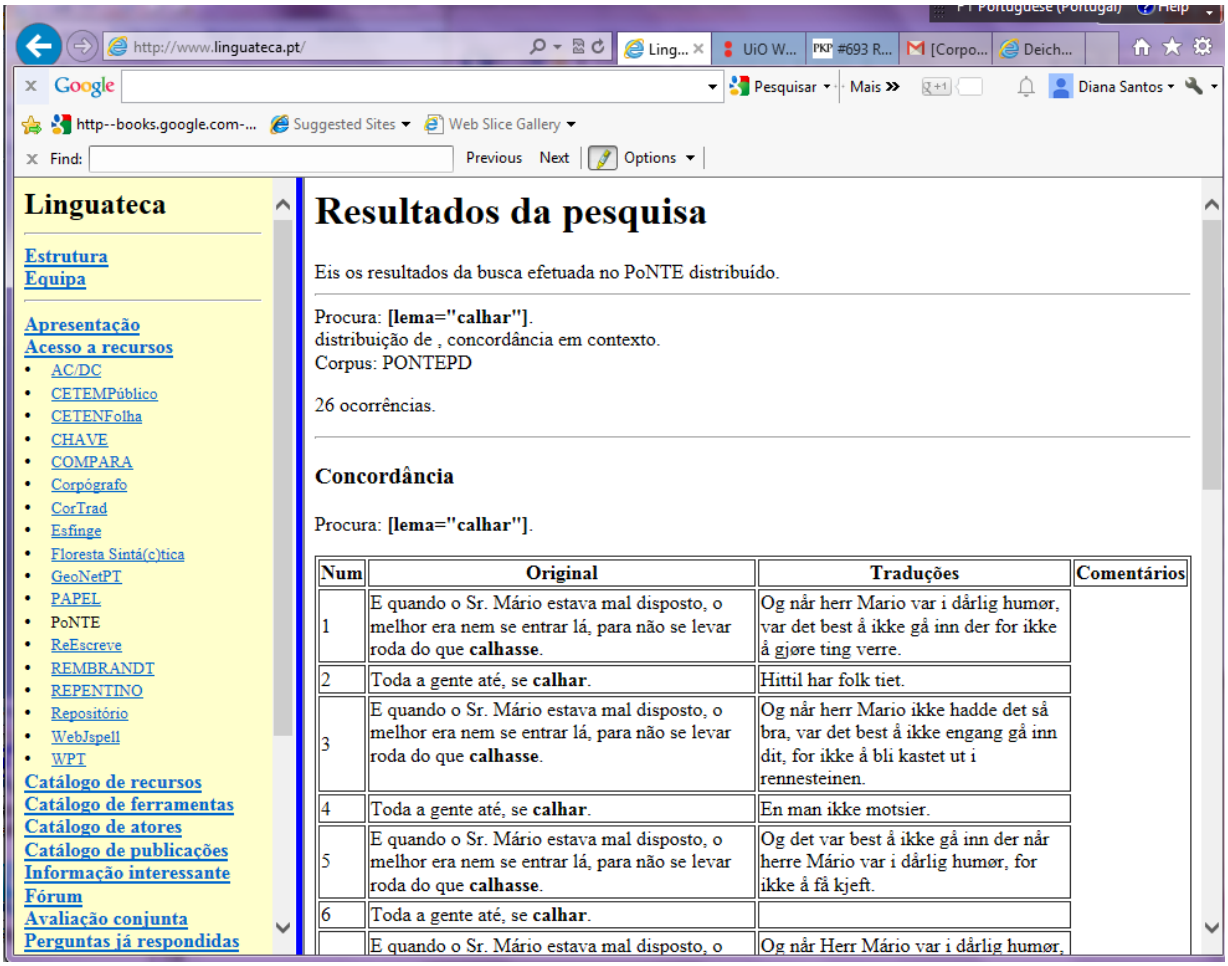


Figura 5: Exemplo de resultado do PoNTE distribuído, mostrando cada tradução separadamente.

extensas leva a que não seja apropriado marcar toda a unidade de tradução quando há um problema que se refere a um vigésimo da mesma. Além disso, essa mesma tradução pode ser digna de elogio em relação a um fenómeno e deficiente em relação a outro.

Para concretizar, veja-se o seguinte exemplo:

(2) A construção de mais 3.000 unidades, pelo programa «Minha Casa Minha Vida», dará continuidade ao Plano Municipal de Habitação Social no Cidade Aracy e na região do Zavaglia, além da construção de unidades para atender aos servidores públicos municipais.
Bygging av *mer enn 3000 enheter*, gjennom programmet «Mitt hus mitt hjem», fortsetter den kommunale planen for sosiale boliger i byen Cidade Aracy og i regionen Zavaglia, i tillegg til **bygging av enheter for kommunalt ansatte**.
A construção de mais de 3000 unidades (...) construção de unidades para empregados da câmara.

Enquanto o uso de um sintagma preposicional simples (*para empregados da Câmara*), em vez de ser fiel à perífrase original, é extremamente comendável, o aprendiz caiu na esparrela de confundir a adição (*mais 3000 unidades*) com a comparação ou com a aproximação (*mais de 3000 unidades*).

Ou seja, na tradução desta frase encontramos um caso exemplar, e um caso de erro, ambos devendo ser analisados e marcados.[6]

Como fazê-lo, porém, levanta a questão das unidades que devem ser o objeto da anotação, e a questão do contexto que deve ser mostrado numa procura (tanto de um lado como do outro), além de como executar essa mesma procura.

Vemos imediatamente que um alinhamento por frase é demasiado grosseiro, e que um alinhamento à palavra também muitas vezes é insuficiente. A própria palavra *alinhamento* é,

[6]Poder-se-ia argumentar que apenas os problemas mereceriam relevo, mas se fosse esse o caso não poderíamos nem usar o PoNTE para dar bons exemplos, nem para fazer estudos quantitativos, porque paar esses é preciso anotar tudo (no que se refere ao fenómeno em questão.)

aliás, pouco feliz[7], visto que queremos pôr em correspondência mas não alinhar.

A última pergunta, mais filosófica, posta pela Belinda Maia aquando da revisão deste artigo, é como escolher os problemas a debater ou focar na aula. Segundo a sua longa experiência, esse é um dos casos em que há menos consenso (no ensino de inglês como língua estrangeira – mas penso que é óbvio que tal problema deve ser extensivo ao ensino de qualquer língua estrangeira). A resposta é que tal é puramente subjetivo, e que depende de cada professor (e sua classe) os casos em que pega na sala de aula. É também relevante distinguir entre ensino de tradução propriamente dito ou ensino da língua. Mas aí penso que o PoNTE poderá ser usado de maneira diferente nos dois tipos de aula, embora não o vá demonstrar em seguida.

O resto desta secção descreve o tipo de comentários que estas traduções suscitam, tentando dar uma ideia dos fenómenos considerados interessantes e usados pedagogicamente.

Divido a discussão entre os casos identificados em cada uma das direções, notando que correspondem a problemas diferentes: enquanto a tradução para norueguês é geralmente feita para a língua materna, a tradução para português corresponde a uma retroversão, e foi sempre corrigida por mim no que se refere a problemas gramaticais, ortográficos ou morfológicos.

### 5.1 Na direção português-norueguês

Vejam-se alguns problemas que têm mostrado ser geralmente complicados:

Na frase seguinte, em doze traduções, onze estavam erradas no que se refere à estrutura argumental (mostro apenas uma).

(3) **Defendia** uma vez o meu amigo Nuno, num comentário que aqui deixou, que «as nossas memórias do que foi são muito seletivas».
Jeg forsvarte en gang min venn Nuno i en kommentar som han la ut her, at våre minner om det som var, er høyst selektive.
Eu defendi uma vez o meu amigo Nuno ...

O problema tem a ver com a ordem livre das palavras do português, que – quando comparada com uma língua com uma ordem fixa, como o norueguês – é dificilmente aceite/compreendida pelos alunos, que também aparentemente têm problemas com o facto de a completiva (o objeto) vir tão longe do verbo.

(4) – quando você terminar o mestrado, a não ser que consiga emprego como pesquisador em empresas privadas ( que são *pouquíssimos* ), você terá necessariamente que fazer um doutorado?

O problema aqui é que em português é claríssimo, devido ao masculino, que são os empregos que são pouquíssimos, e não as empresas privadas, como a totalidade das traduções permite erradamente inferir – porque não há género nem número de adjetivos em norueguês.

(5) – então, *com 3 anos de formado*, você terá que concorrer a bolsas de R$ 2.000 mensais para fazer doutorado?

Nesta frase, *com 3 anos de formado* significa que já passaram 3 anos desde que se formou... mas esta construção não era conhecida dos alunos, e por isso ninguém a conseguiu interpretar (e consequentemente traduzir) corretamente.

Outro problema bem distinto:

(6) Encararam e puseram o problema de fundir, adoptar soluções de compromisso ou separar radicalmente culturas por vezes altamente complexas (a Indiana, a Chinesa, as Africanas, a Brasileira) e religiões (Budismo, Bramanismo).

Possivelmente assoberbados com tanta informação numa frase, em (6) nenhum aluno conseguiu compreender que havia uma escolha entre três formas diferentes de proceder (nomeadamente (i) fundir, ou (ii) adotar soluções de compromisso, ou (iii) separar radicalmente), tendo portanto traduzido por algo muito pouco claro.

Outros casos são lexicais e têm a ver com diferentes culturas. No caso seguinte, é preciso compreender que as baratas não têm o mesmo papel de representantes do submundo em Moçambique e na Noruega:

(7) Mas o que dói saber é que, por extensão e acumulação, o não-funcionamento de tudo acabaria por servir aqueles que, como os ratos e as baratas, se movem melhor no caos e na podridão.

ou compreender o que significa "despachar":

(8) Despachar um documento preso nas burocracias;
gjøre kort prosess med et dokument som er fanget i byråkratiet;
Å sende et fast dokkument innenfor byrokra-

[7] Como já argumentado em Santos e Simões (2008).

tiet.
Avsende et dokument som var fastklemt i byråkratiet
Ekspedere et dokument som er fanget i byråkratiet;
avlevere et document som står fast i byråkratiet

É muito interessante constatar que quatro tradutores interpretaram *despachar* como enviar pelo correio, e um como liquidar ou contornar.

De qualquer forma, o caso mais complicado e a que costumo dar mais atenção tem a ver com a estruturação do discurso. Muitas vezes essa não é compreendida, e daí faltar coesão, por vezes mesmo qualquer lógica, à tradução resultante. De seguida, apresento alguns exemplos diversos de casos que tendem a ser ignorados:

(9) **E se** os Portugueses foram ajudados por inúmera gente de muitos países e tradições, não resta dúvida de que o esforço de aquisição foi seu, como sua foi a consciência primeira do novo mundo e o desafio àquele que existia.

(10) As mercearias são **o menos**.

(11) A contribuição de Portugal para o Renascimento, **todavia**, não se deu tanto no capítulo das Artes ou das Humanidades como no da Ciência.

(12) **Já** a solução do problema da dívida deve ser combinada com o crescimento econômico.

(13) Feitas as ressalvas, **vamos então** à minha campanha de anti-propaganda sobre a ciência no Brasil!

Outros comentários são mais subjetivos, e dependem naturalmente do estilo e do gosto, mas também do "espírito da língua". No caso seguinte, é natural comentar como "Desnecessário" traduzir uma oração participial (que não existe em norueguês) por uma relativa:

(14) A mesma comida, a mesma decoração, o mesmo programa na televisão **pendurada na parede**, ouvindo o atendente pronunciar a mesma palavra-chave: ≪hambúrguer≫.
Den samme maten, den samme dekorasjonen og det samme TV-programmet på TV-en **som er hengt på veggen**, og kelneren som sier det samme nøkkelordet: ≪hamburger≫.
'que está pendurada na parede'

Os casos mais interessantes – mas certamente mais difíceis de anotar – são aqueles em que mais de uma diferença contrastiva se conjuga para complicar a expressão do sentido original, veja-se por exemplo:

(15) Embora não haja um número expressivo de dados sobre os quais apoiar as diferenças, **qualquer observador mais atento** é capaz de verificar que elas existem e talvez até permitam identificar a que região da cidade um carioca pertence.

Neste caso, *qualquer observador mais atento* não foi bem traduzido/compreendido por nenhum dos 12 tradutores, possivelmente pela dificuldade conjugada (i) da palavra *observador* (o que observa e não necessariamente um papel/profissão), (ii) do uso da comparação sem termo de comparação, e (iii) do emprego do quantificador *qualquer*, que por ser intensional é sempre difícil de exprimir numa língua germânica.

(16) Queremos – e **podemos** – ajudar, enquanto há tempo, os países onde a crise já é aguda.

Neste caso, é preciso atentar à variada polissemia de *poder*, e escolher a mais adequada e não a mais frequente. No exemplo acima, *podemos* está empregue no sentido de "temos meios/capacidade para" e não no sentido de mera possibilidade.

E ainda mais difícil é quando é preciso traduzir formas mais criativas[8]:

(17) Eis que, de modo impensado, o subdesenvolvimento cria o seu próprio vocabulário, **as suas formas de se dizer**.

*As suas formas de se dizer* é certamente uma expressão que merece, mesmo para falantes nativos, alguma reflexão (basta ser de um escritor tão extraordinário como Mia Couto). Mas textos que valham a pena ser traduzidos são precisamente textos que provoquem reflexão.

Estes exemplos demonstram, espero, que a variedade de problemas ou de questões merecedoras de comentário, entrando apenas em conta com o texto fonte, é vastíssima, mas inescapável.

### 5.2 Na direção norueguês-português

Tendo em conta o objetivo destas traduções, ou retroversões, na direção norueguês-português,

[8]Note-se que não me estou a referir ao caso, também frequente, mas mais fácil de detetar, de formas elas próprias inventadas e portanto novas, como *brincriar* no texto em questão.

o tipo de comentários e de problemas é naturalmente outro, e está centrado no mantra do "não é assim que se exprime isso em português", no espírito por exemplo de (Bennett, 2010), ou então, no alertar para ter cuidado com outras interpretações que a tradução erroneamente adicionou.

Por exemplo, a questão da posição e forma do movimento, e dos modais, são críticas: na maior parte das vezes a primeira deve omitir-se, e a segunda traduzir-se pelo tempo e modo correspondentes.

Veja-se um primeiro exemplo em que a posição seria completamente irrelevante em português:[9]

(18) Her **sitter** jeg og faar ikke sove fordi jeg er saa fortvila.
Estou sentado aqui e não consigo dormir porque estou desesperado.
Agora estou tão desesperada que não consigo dormir,
Estou aqui, sem poder dormir, porque me sinto muito desesperada.
Aqui estou eu, tão preocupada que não consigo dormir.
Estou sentada e não consigo dormir por causa do desespero.
Estou sentado aqui sem conseguir dormir porque eu estou muito desesperado.

Um caso trivial em que o verbo modal nunca deveria ser traduzido literalmente é o seguinte.

(19) Hvorfor sauer ikke **kan** svømme
Porque as ovelhas não se põem a nadar.
Porque ovelhas não sabem nadar
Porque os carneiros não nadam
Porque os carneiros não sabem nadar?
Por que as ovelhas não nadam
Porque é que as ovelhas não sabem nadar
Porque é que os carneiros não nadam?
Porque ovelhas não sabem nadar
Porque as ovelhas não sabem nadar
Por que é que as ovelhas não sabem nadar.
Porque os carneiros não sabem nadar
Porque carneiros não sabem nadar
Porque as ovelhas não são capazes de nadar.

Repare-se que a variação entre *carneiros* e *ovelhas* é completamente irrelevante, neste caso, mas não a distinção entre *porquê* e *porque* (*hvorfor* e *fordi* em norueguês), em que quase nenhum tradutor reparou, e que fundamentaria, em minha opinião, algo tão diferente como *O porquê dos carneiros não nadarem...*

Noutros casos é simplesmente o tom, ou o tipo de língua, que não é conseguido pelas várias traduções. O exemplo seguinte é também interessante pela transformação de *der* (onde) para *quando* ou *enquanto*, necessária em português.

(20) Siss hadde mange tankar **der** ho gjekk, innballa for frosten.
Muitas coisas ocupavam a cabeça de Siss, que estava andando no frio ...
Siss tinha muitos pensamentos andando por ali, embalada contra o frio.
Siss tinha muitos pensamentos enquanto caminhava, coberta contra o frio
Siss tinha vários pensamentos, envolvida devido à geada.
Siss está a pensar, enrodilhada por causa do frio:
Siss estava pensando muito quando caminhava pela floresta, embrulhada em roupas contra o frio.

As questões discursivas são igualmente relevantes nesta direção da tradução, claro, e embora no que se segue eu apresente todas as traduções para dar uma ideia do conteúdo e da necessidade ou falta da sua tradução, essa ideia resulta bastante pálida porque a maior parte destes marcadores não foi bem traduzida:

(21) Jo mer overraskende resultatet er, jo bedre er forskningen, **på et vis**.
Quanto mais surpreendente o resultado, melhor a pesquisa, tipo.
Quanto mais estranhos os resultados, melhor a investigação... de uma certa forma.
Quanto mais surpreendente o resultado, melhor a investigação, por um lado.
Quanto mais surpreendente o resultado, melhor a pesquisa, de alguma forma.
De certa forma, mais surpreendente o resultado melhor a pesquisa.
O fato é que as pesquisas mais surpreendentes são as melhores, mas às vezes uma pesquisa torna-se mais surpreendente de que originalmente era para ser.
Duma maneira, quanto mais surpreendente for o resultado, melhor é a pesquisa.
De um jeito, quanto mais surpreendente for o resultado, tanto melhor para a investigação.
Tanto mais surpreendente o resultado, tanto melhor a investigação.
Quanto mais surpreendente o resultado, melhor a pesquisa, de alguma maneira.

[9] Na direção do norueguês para o português apresento muitas traduções, para ilustrar o nível do português dos alunos e a divergência de traduções no PoNTE.

E quanto mais surpreendente o resultado, tanto melhor a pesquisa, de algum modo.

(22) Noen ganger høres **nemlig** ting tullete ut fordi de er tullete.
Às vezes as coisas parecem ridículas exatamente porque são ridículas.
Às vezes as notícias parecem disparatadas, porque são disparatadas.
Algumas vezes as coisas parecem tolas porque o são.
Às vezes as coisas parecem tolas porque o são.
Às vezes , coisas que parecem absurdas, realmente são absurdas.
Algumas vezes as coisas soam disparatadas porque são disparatadas.
Algumas vezes coisas parecem desatinadas porque realmente são.
De vez em quando as coisas parecem asneiras porque é a característica correta.
Às vezes mesmo coisas parecem brincadeiras porque são brincadeiras.
**É que**, às vezes , uma coisa parece tola porque é tola.

(23) Men jeg er **ganske sikker** på at de regnet med at jeg drakk som dem. Når jeg **faktisk** lurte dem.
Mas tenho quase a certeza de que pensavam que eu estava a beber também, quando **realmente** os estava a enganar.
Mas tenho certeza que eles estavam contando com que eu estivesse bebendo álcool assim como eles, quando **na verdade** eu os enganei.

(24) Nogle af potterne er desværre knækket, men det kan man **næppe** undgå efter en forsendelse fra Sydamerika.
Infelizmente, alguns dos potes estão quebrados, mas isto não é possível evitar quando se trata de um envio da América do Sul.
Infelizmente alguns dos jarros estão partidos, mas isto é quase impossível de evitar num envio da América do Sul.
Alguns dos vasos estão infelizmente quebrados, mas isto é quase inevitável depois de uma viagem proveniente da América do Sul.

### 5.3 Pilotos de anotação

Para avançar na determinação da forma mais prática de adicionar a anotação, desenvolvi um programa auxiliar que permitia anotar informação sobre as traduções de dados verbos (qual verbo/lema era um parâmetro do programa), e criei um novo atributo posicional (coluna) no PoNTE distribuído, de forma a poder ter uma ideia da gama (e distribuição) dos diferentes comentários.

Anotei depois todas as ocorrências dos verbos *haver*, *ser*, *estar*, *ficar* e *ter*, que passaram a ser procuráveis pela interface do PoNTE distribuído, pedindo a distribuição de traduções, como ilustrado na figura 7.[10]

Ao fazer isso, logo várias questões surgiram, como os leitores pela simples leitura do exemplo podem apreciar:

1. Que tipo de valores associar?
2. Como padronizar os comentários?
3. A que nível de detalhe descer?
4. Como distinguir as duas situações diversas: (i) não traduzido porque o aluno simplesmente não fez de (ii) não traduzido porque o tradutor escolheu outra forma?
5. Será que associar as traduções às palavras é uma boa ideia, ou deveria ser a uma unidade maior?
6. E: Faz sentido anotar todas as palavras, ou apenas aquelas que nos suscitarem comentários?

E, talvez o problema maior de todos, imediatamente ilustrado pela nota de rodapé relativa aos "0", como garantir uma anotação atualizada, quando a adição de novas traduções transforma os dados quantitativos anteriormente adicionados em valores incorretos?

Todas estas perguntas correspondem a decisões que terão de ser tomadas, possivelmente com experimentação de vários caminhos e perguntas a vários utilizadores.

Outra anotação foi também efetuada no âmbito do estudo da tradução específica de sentimentos e sensações, agora classificando simplesmente os verbos (em português) em várias categorias sintático-semânticas (veja-se Santos (2014e) para mais pormenores). Aí ficou outra vez patente a dificuldade de refazer os corpos com mais informação, dado existirem diferentes fontes de anotação que é preciso harmonizar para cada nova versão. Ou seja, se da versão 3.0 para a 4.0 se passa a adicionar anotação automática de partes do corpo humano, como reintroduzir a anotação humana das sensações aqui mencionada, se ficou guardada em versões anteriores sem partes do corpo?

[10] Os casos de 0 correspondem a casos de ocorrências desses verbos que não foram anotadas, porque pertencem a traduções que foram incorporadas depois do piloto ter tido lugar.

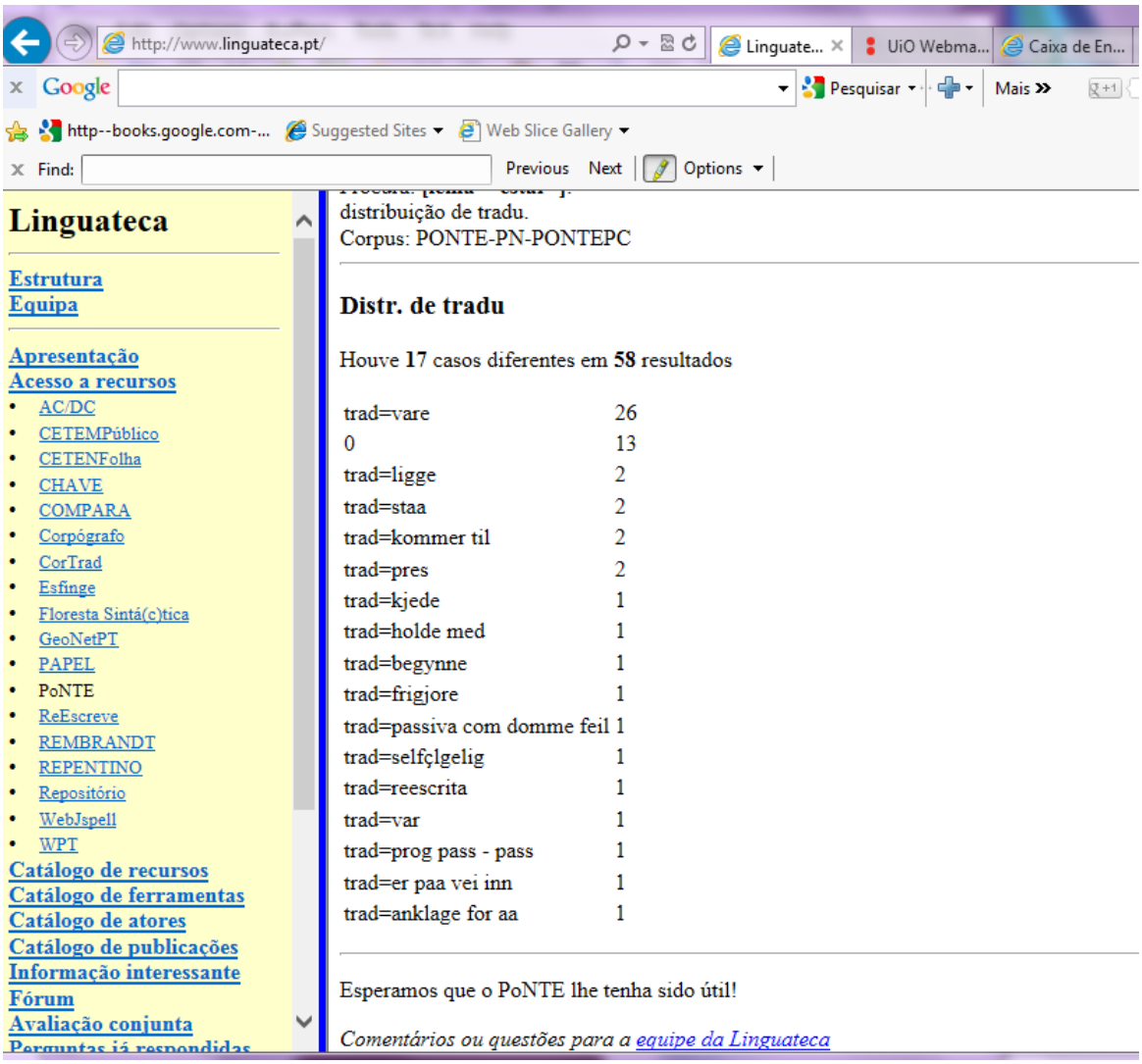


Figura 7: Um exemplo de comentários às traduções.

Estas questões levam a que seja mais natural tentar manter essa informação separadamente, para poder ser reposta numa nova versão, mas também isso não é necessariamente à prova de problemas entre versões, visto que pode haver diferenças de atomização ou outras, que inviabilizem uma mesma identificação... como tivemos experiência disso no caso de dois corpos que incorporámos no AC/DC mas que já vinham com anotação criada por métodos diferentes: a CDHAREM (Rocha e Santos, 2007) e o ReLi (Freitas et al., 2012).

Uma possibilidade é, naturalmente, transformarmos o processo de anotação em algo automático que é repetido para cada nova versão do corpo, como foi feito no caso do piloto do tipo de emoções, usando o `corte-e-costura` (Mota e Santos, 2009; Santos e Mota, 2010). Isso implica criar regras específicas, e usá-las de novo sempre que se recria o corpo. O problema é que se o corpo aumentou (incluindo novos textos, por exemplo), as regras muito provavelmente têm de ser alargadas para cobrir o novo material – ou então podem resultar mais prejudiciais do que úteis, se quisermos fazer fé nos dados quantitativos.

Parece que não podemos senão concluir que versões diferentes de um corpo em constante desenvolvimento terão de incluir questões diferentes, e que, para ter um produto acabado, não podemos melhorar noutros campos, o que é, de facto, na minha opinião um dos dilemas maiores dos compiladores de corpos:

- Dizemos que o corpo está pronto, estável, e não se mexe mais?
- Ou, pelo contrário, podemos/devemos ir melhorando e adicionando mais informação e mais textos?

Um sistema de versões para corpos é demasiado pesado, e a maior parte dos utilizadores nem sequer cita a versão (nem a data de acesso),

quanto mais compara com as anteriores... É a esse respeito instrutivo comparar o que escrevi em 2000 sobre o CETEMPúblico (Santos, 2000) e o que concluí mais tarde, nomeadamente que essas expetativas eram completamente desajustadas da realidade (Santos, 2014a).

Não tenho respostas definitivas para estas perguntas, mas devo mencionar que é pelo menos lícito fazê-las... e que estamos a voltar à carga neste assunto (a reutilização de anotações) com a iniciativa da Gramateca[11], lançada em janeiro do presente ano de 2014.

## 6 Comentários finais

Espero ter demonstrado que a avaliação de traduções na aprendizagem de uma língua é um fenómeno complexo que merece o desenvolvimento de ferramentas apropriadas ao seu estudo e ao seu reuso, e que um corpo como o PoNTE, dotado das ferramentas idealizadas acima, pode ser útil a dois níveis:

- a nível pedagógico, para ensinar este tipo de aprendentes e para formar futuros professores na área (e no par de línguas em questão);
- a nível gramatical, eventualmente para estudar outros tipos de fenómenos gramaticais que sejam suficientemente frequentes para permitir a obtenção de vários exemplos, bem e mal traduzidos.

Contudo, a implementação de um sistema que permita de facto procurar e codificar o tipo de informação desejada não é obviamente uma tarefa simples, e ainda nos encontramos nos primórdios de tal realização.

Que eu conheça, apenas Oliveira (2012) publicou um estudo que usa traduções de alunos num corpo paralelo que inclua o português. Kutuzov et al. (2012), por outro lado, é um projeto semelhante ao PoNTE em termos de abrangência e objetivos (para o par russo e inglês), mas com diferentes soluções técnicas e aparentemente ainda sem uso no próprio ensino.

Embora com espírito semelhante, nem Bernardini (2002) nem Abekawa e Kageura (2008) ou Bojar et al. (2008), que relatam trabalhos ou ideias interessantes associadas à anotação de corpos paralelos, têm um recurso do mesmo tipo que o PoNTE.

Embora existam alguns artigos relacionados com o ensino da tradução e/ou o estudo da atividade dos aprendizes de tradutores envolvendo o português como uma das línguas, não consegui identificar mais nenhum sobre um corpo de trabalhos produzidos por este tipo de aprendentes, o que me leva a insistir que este tipo de corpo é escasso e inovador.

Pela minha própria pesquisa de referências a trabalhos parecidos, é certamente Popescu-Belis, King e Bentanar (2002) aquele que é mais consonante com o descrito aqui, visto que os autores descrevem um corpo de traduções corrigidas, no âmbito do ensino (e avaliação) da tradução. Mas as semelhanças acabam aí: O formato escolhido é XML, e usam a nota atribuída a cada tradução pelos examinadores como (uma das) formas de avaliar cada tradução globalmente. Além disso, parece-me ser mais a correção da língua de chegada que está em questão na correção das traduções e não tanto a propria tradução. Finalmente, o par é francês para inglês; o conteúdo, em 2002, correspondia a 50 traduções de dois textos, e um dos objetivos da criação do corpo era para servir de treino na avaliação de tradução automática.

Não quero contudo dar a entender com as afirmações anteriores que não exista qualquer trabalho que estude aprendentes ou aprendizes de língua portuguesa, ou a aprender português:

- existem vários chamados "corpora de aprendiz de língua estrangeira", por exemplo Shepherd (2009), Dutra e Silero (2012) e Tagnin e Fromm (2008) tratam de corpos de alunos brasileiros a aprender inglês (no caso do COMAprend, também francês, alemão, italiano e espanhol), e Gamallo et al. (2013) refere-se a alunos portugueses a aprender galego;
- enquanto que Evers e Wilkens (2012) e os corpos PEAPL2[12] da Universidade de Coimbra e *Recolha de Dados de Aprendizagem de Português Língua Estrangeira*[13], do Centro de Linguística da Universidade de Lisboa, contêm textos de aprendizes de português como língua estrangeira.

A maior diferença é que nenhum destes corpos tem origem em traduções, mas sim em atividades de redação ou interação na língua a que se referem. Além disso, e pelo que me foi dado apreciar, a maioria também não tem

[11] Ver `http://www.linguateca.pt/Gramateca`, veja-se uma primeira apresentação do projeto em Santos (2014b; Santos (2014c).

[12] `http://www.uc.pt/fluc/rcpl2/`

[13] `http://www.clul.ul.pt/pt/recursos/314-corpora-of-ple`

associado qualquer sistema de procura específico, ao contrário do PoNTE.

Mas é inegável que a combinação de todos estas peças de um mesmo quebra-cabeças é interessante para estudar as dificuldades da aprendizagem do português e as características do português que diferem de outras línguas, e que um trabalho futuro pertinente seria desenhar estudos que usassem os três tipos de materiais.

Este artigo é, contudo, apenas dedicado ao PoNTE. Embora seja trabalho em progresso, já nos parece merecedor de publicação, tanto para congregar futuros utilizzadores como para pôr à consideração e discussão do público em geral possíveis caminhos a seguir num futuro próximo, nomeadamente:

- a criação de um novo "corpo" com apenas anotações, alinhado com os corpos de texto a que se refere, em vez de incluir as anotações como atributos dos corpos;
- a expansão do Ensinador (Simões e Santos, 2011) para corpos paralelos, em que ao aluno são apresentados alternativas relevantes para a própria tradução (desde que essa seja considerada correta), semelhante ao que é feito no RuN (Grønn e Marijanovic, 2010);
- o desenvolvimento de um programa que permita identificar automaticamente os casos de problemas, por exemplo comparando as várias traduções, o que é algo também relevante para estudos mais profundos do processo de tradução como os que pretendemos fazer no âmbito do CorTrad.

Gostaria contudo de terminar este texto com duas notas negativas, propostas pela Belinda Maia, e que me parecem importantes para diminuir um exagerado otimismo que possa ficar na imaginação dos leitores.

Em primeiro lugar, não é garantido que a crítica de traduções a alunos de língua não seja contraproducente. Repare-se que os alunos não são alunos de tradução, mas sim de língua. Uma coisa é achar que lhes torna as aulas mais interessantes – e a mim também, a outra é realmente conseguir demonstrar que aprendem melhor a língua portuguesa (e/ou que eu a ensino melhor).

Em segundo lugar, não é garantido que o trabalho – relativamente grande – de criar este corpo seja rentável em termos práticos (comparando com outra tarefas que eu poderia fazer para a melhoria do ensino), sobretudo devido à falta de utilizadores neste par de línguas. Só o futuro, realmente, o dirá. Futuro esse que, nos tempos mais próximos, pode ir sendo auscultado pelo ritmo de mudança e melhoria do corpo.

## Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN; de 2009 até 31 de dezembro de 2011 pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN), e a partir dessa data apenas apoiada pelas Universidades em que os membros trabalham.

Especificamente o projeto PoNTE foi apoiado pela Universidade de Oslo através da atribuição, no Outono de 2011, de uma mini-bolsa de investigação a Marcin Wlodek, a quem estou grata pelos comentários relativos às traduções desse semestre.

Agradeço também a Joacyr Oliveira as discussões sobre o uso de corpos de aprendentes de tradução para ensinar a língua e a tradução, e a Belinda Maia os comentários pertinentes na sua recensão.

## Referências

Abekawa, Takeshi e Kyo Kageura. 2008. Constructing a corpus that indicates patterns of modification between draft and final translations by human translators. Em *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA), 28-30 Maio, 2008.

Bakkejord, Kaja Rindal. 2008. Técnicas de substituição e supressão dos clíticos no português do Brasil. Tese de Mestrado, Universidade de Oslo.

Bennett, Karen. 2010. Academic discourse in portugal: A whole different ballgame? *Journal of English for Academic Purposes*, 9(1):21–32.

Bernardini, Silvia. 2002. Educating translators for the challenges of the new millenium: The potential of parallel bidirectional corpora. Em Belinda Maia, Johann Haller, e Margherita Ulrych, editores, *Training the language services provider for the New Millenium, Proceedings of the III Encontros de Tradução*. Astra-FLUP, FLUP, Porto, pp. 173–186.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de doutoramento, Aarhus University, Aarhus, Denmark, November, 2000.

Bojar, Ondrej, Miroslav Janicek, Zdenek Zabokrtsky, Pavel Ceska, e Peter Bena. 2008. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. Em *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA), 28-30 de maio, 2008.

Borin, Lars, Lauri Carlson, e Diana Santos. 2001. Corpus based language technology for computer-assisted learning of Nordic languages: Squirrel. Em Henrik Holmboe, editor, *Nordisk sprogteknolog (Nordic language technology), Aarbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Museum Tusculanumus Forlag, Københavns Universitet, Copenhaga, pp. 257–270, Setembro, 2001.

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.

Dutra, Deise Prina e R. P. Silero. 2012. O uso de for: uma análise de itens linguísticos em corpus de aprendizes brasileiros. Em Tania Shepherd, Tony Berber Sardinha, e Marcia Veirano Pinto, editores, *Caminhos da linguística de corpus*, pp. 325–341. Mercado de Letras.

Evers, Aline e Rodrigo Wilkens. 2012. Classificação de proficiência em língua adicional no português um estudo para a determinação de índices diferenciadores. Em *IX Encontro Nacional de Inteligência Artificial, ENIA 2012*, 20-25 de outubro, 2012.

Frankenberg-Garcia, Ana e Diana Santos. 2002. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*, IX(1):61–79.

Freitas, Cláudia, Eduardo Motta, Ruy Luiz Milidiú, e Juliana César. 2012. Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros. Em *XI Encontro de Linguística de Corpus - ELC 2012*, 13-15 de setembro, 2012.

Gamallo, Pablo, Marcos García, I. González, Muñoz. M., e I. Del Río. 2013. An evaluation of Avalingua based on learner corpora. Em *ICAME34 Workshop Learner Corpora and their Application in Language Testing and Assessment, May 22, Santiago de Compostela, Spain, 2013*, pp. 52–53.

Granger, Sylviane, Gaëtanelle Gilquin, e Fanny Meunier. 2013. *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead. Proceedings of the First Learner Corpus Research Conference (LCR 2011)*. Presses universitaires de Louvain.

Grønn, Atle e Irena Marijanovic, 2010. *Russian in contrast: form, meaning and parallel corpora*, pp. 1–24. Oslo Studies in Language 2(1).

Kutuzov, A. B., M. A. Kunilovskaya, A. Y. Oschepkov, e A. Y. Chepurkova. 2012. Russian-learner parallel corpus as a tool for translation studies. Em *Proceedings of Dialog 2012*.

Maia, Belinda e Diana Santos. 2012. Who is afraid of ... what? - In English and in Portuguese. Em Signe Oksefjell Ebeling, Jarle Ebeling, e Hilde Hasselgård, editores, *Aspects of corpus linguistics: compilation, annotation, analysis*, number 12 in Studies in Variation, Contact and Change in English, Dezembro, 2012.

Mota, Cristina e Diana Santos. 2009. Corte e costura no AC/DC: auxiliando a melhoria da anotação nos corpos, Setembro, 2009. `http://www.linguateca.pt/acesso/corte-e-costura.pdf`.

Nilsson, Kåre. 1997. A lusofonia vista por um lusitanista escandinavo, 11 de novembro, 1997. 1.o encontro de Professores de Português - Língua Estrangeira, organizado pelo Centro de Línguas, FFLCH, USP.

Oliveira, Joacyr. 2012. A linguística de corpus na formação de tradutores: compilação e análise de um corpus de aprendizes de tradução. Em *Trabalho em andamento - Anais do XI Encontro de Linguística de Corpus (ELC 2012)*.

Popescu-Belis, Andrei, Margaret King, e Houcine Bentanar. 2002. Towards a corpus of corrected human translations. Em Margaret King, editor, *Machine Translation Evaluation – Human Evaluators Meet Automated Metrics, Workshop Proceedings, LREC2002*, pp. 17–21.

Rocha, Paulo e Diana Santos. 2007. Disponibilizando a <OBRA>

Colecção Dourada </OBRA> do <ACONTECIMENTO > HAREM </ACONTECIMENTO> através do projecto <LOCAL|ORGANIZACAO| ABSTRACCAO> AC/DC </LOCAL|ORGANIZACAO| ABSTRACCAO>. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, pp. 307–326. Linguateca, 12 de novembro, 2007.

Rosa, Alexandra Assis. 2006. Does translation have a say in the history of our contemporary linguacultures? Some figures on translation in Portugal. *Polifonia*, 9:77–94.

Santos, Diana. 2000. O projecto Processamento Computacional do Português: Balanço e perspectivas. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pp. 105–113, São Paulo, 19-22 de novembro, 2000. ICMC/USP.

Santos, Diana. 2002. DISPARA, a system for distributing parallel corpora on the Web. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing (PorTAL 2002)*, Lecture Notes in Artificial Intelligence, pp. 209–218, Berlin/Heidelberg, 23-26 de junho, 2002. Springer-Verlag.

Santos, Diana. 2003. Timber! Issues in treebank building and use. Em Jorge Baptista, Isabel Trancoso, Maria das Graças Volpe Nunes, e Nuno J. Mamede, editores, *Computational Processing of the Portuguese Language: 6th In ternational Workshop, PROPOR 2003. Faro, Portugal, June 2003*, pp. 151–158, Berlin/Heidelberg. Springer Verlag.

Santos, Diana. 2012. Os possessivos estão-me a complicar o ensino :-), 26 de outubro, 2012. `http://www.linguateca.pt/Diana/download/posterAPL2012.pdf`.

Santos, Diana. 2013. Ser, estar, ficar, haver and ter against ha, bli and vare: who said it was easy to describe feelings and sensations?, 15 de maio, 2013. `http://www.linguateca.pt/Diana/download/KKPoNTE.pdf`.

Santos, Diana. 2014a. Corpora at Linguateca: vision and roads taken. Em Tony Berber Sardinha e Telma São Bento Ferreira, editores, *Working with Portuguese corpora*, pp. 219–236. Bloomsbury.

Santos, Diana. 2014b. First steps of Gramateca: a corpus-based grammar initiative for Portuguese, driven by Linguateca, 20 de fevereiro, 2014. `http://www.linguateca.pt/Diana/download/GramatecaOslo.pdf`.

Santos, Diana. 2014c. Gramateca: corpus-based grammar of Portuguese. Em Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A.S. Pardo, e Maria das Graças Volpe Nunes, editores, *PROPOR 2014*, pp. 214–219. Springer Verlag, outubro, 2014.

Santos, Diana. 2014d. Os possessivos estão-me a complicar o ensino :-) Um estudo do dativo possessivo baseado em corpos. Em apreciação, versão preliminar, `http://www.linguateca.pt/Diana/download/PossAprec.pdf`.

Santos, Diana. 2014e. Ser, estar, ficar, haver e ter vs. ha, bli e være: quem disse que era fácil traduzir sentimentos e sensações? Em Signe Oksefjell Ebeling, Atle Grønn, Kjetil Rå Hauge, e Diana Santos, editores, *Corpus-based Studies in Contrastive Linguistics*, pp. 271–288. Oslo Studies in Language 6(1).

Santos, Diana, Belinda Maia, e Luís Sarmento. 2004. Gathering empirical data to evaluate MT from English to Portuguese. Em Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Gudrun Magnusdottir, Anna Samiotou, e Khalid Choukri, editores, *Proceedings of LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*, pp. 14–17, 25 de maio, 2004.

Santos, Diana e Cristina Mota. 2010. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. Em Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 1437–1444. European Language Resources Association, 17-23 de maio, 2010.

Santos, Diana e Alberto Simões. 2008. Portuguese-English word alignment: some experiments. Em *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA), 28-30 de maio, 2008.

Sarmento, Luís, Anabela Barreiro, Belinda Maia, e Diana Santos. 2007. Avaliação

de Tradução Automática: alguns conceitos e reflexoes. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, pp. 181–190, Lisboa, Portugal, 20 de março, 2007. IST Press.

Shepherd, Tania. 2009. Corpora de aprendiz de língua estrangeira: um estudo contrastivo de n-gramas. *Veredas*, 11(2):100–116.

Simões, Alberto e Diana Santos. 2011. Ensinador: corpus-based Portuguese grammar exercises. *Procesamiento del Lenguaje Natural*, 47:301–309, Setembro, 2011.

Tagnin, Stella Esther Ortweiler e Guilherme Fromm. 2008. COMAprend – a experiência da construção de um corpus de aprendizes para estudos. *Domínios de Linguagem*, 2(2).

Tagnin, Stella O. E., Elisa Duarte Teixeira, e Diana Santos. 2009. CorTrad: a multiversion translation corpus for the Portuguese-English pair. *Arena Romanistica*, 4:314–323.

Teixeira, Elisa D., Diana Santos, e Stella E. O. Tagnin. 2011. CorTrad: um novo corpus paralelo multiversão para o par de línguas português-inglês. Em Tania Shepherd, Tony Berber Sardinha, e Marcia Veirano Pinto, editores, *Caminhos na Linguística de Corpus*. Mercado de Letras, pp. 151–176.

Volume 6, Número 1- Julho 2014

linguaMATICA

ISSN: 1647-0818

## Simpósio de Tecnologia da Informação e Linguagem Humana



## Artigos de Investigação



http://linguamatica.com