Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 9, Número 1- Julho 2017

ISSN: 1647-0818

Volume 9, Número 1 – Julho 2017

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Editorial

Este é o nono ano em que a Linguamática é editada. Todos os anos temos tido novidades, o que tem levado a nossa/vossa revista cada vez mais longe. Desde a publicação regular, até à indexação nos mais relevantes índices de indexação científica, a Linguamática tem-se superado. E isto só é possível graças aos nossos autores, que continuam a apostar na publicação nas línguas ibéricas, e aos nossos revisores, que avaliam os artigos mas também dão sugestões construtivas no sentido de melhorar todos os trabalhos publicados.

O nosso trabalho, como editores, tem sido a preparação das edições, mas também a contínua vontade de recompensar os nossos autores e revisores.

Nesse sentido, nos últimos meses todas as revisões que foram feitas sobre artigos publicados, foram registadas na plataforma Publons. O objetivo desta plataforma é registar oficialmente todo o trabalho de revisão que habitualmente é feito pro bono. Durante as próximas edições esse registo continuará a ser realizado, facilitando aos revisores o processo de registo desta tarefa tão valiosa.

Finalmente, mas não menos importante, a Linguamática tem, a partir de agora, através da Universidade do Minho, a possibilidade de atribuir aos artigos publicados um Document Object Identifier *(DOI). Assim, nesta edição, na folha de rosto de cada artigo, estará presente o número de DOI, bem como um QR-Code que permite aceder diretamente ao objeto respetivo. Durante os próximos meses serão adicionados registos para todos os trabalhos publicados na Linguamática, desde a sua primeira edição, a 5 de junho de 2009.*

A todos, o nosso obrigado!

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

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,
Université de Caen Basse-Normandie

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 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

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Marcos Garcia,
Universidade de Santiago de Compostela

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

Artigos de Investigação

Proposta recebida em Agosto 2016 e aceite para publicação em Junho 2017.

CORP: Uma Abordagem Baseada em Regras e Conhecimento Semântico para a Resolução de Correferências

CORP: A Rule Based Approach with Semantic Knowledge for Coreference Resolution

Evandro Fonseca
PUCRS
evandro.fonseca@acad.pucrs.br

Vinicius Sesti
PUCRS
vinicius.sesti@acad.pucrs.br

André Antonitsch
PUCRS
andre.antonitsch@acad.pucrs.br

Aline Vanin
UFCSPA
aline.vanin@ymail.com

Renata Vieira
PUCRS
renata.vieira@pucrs.br

Resumo

Neste trabalho propomos o uso de conhecimento lexical, sintático e semântico na tarefa de resolução de correferência. Para isso, realizamos experimentos envolvendo diferentes combinações de heurísticas. Como fruto deste estudo, geramos um sistema prático que resolve correferência em textos da língua portuguesa. Além disso, por meio do conhecimento semântico, introduzido pelo Onto.PT, foi possível obtermos um aumento significativo nos níveis de abrangência do nosso modelo.

Palavras chave

Resolução de Correferência, Conhecimento Semântico

Abstract

In this paper we propose the use of lexical, syntactic and semantic knowledge for coreference resolution. We conducted several experiments involving different heuristics. As a result of this study, we generated a practical system that solves coreference in Portuguese texts. In addition, it was possible to increase our recall through semantic knowledge provided by Onto.PT.

Keywords

Coreference Resolution, Semantic Knowledge

1 Introdução

A Resolução de correferências é um processo que consiste em identificar as diversas menções feitas a uma mesma entidade em um texto.

Encontramos diversas iniciativas para a língua portuguesa na literatura que abordam esse problema, geralmente separados entre a resolução de anáforas (Vieira et al., 2005; Bick, 2010; Rocha, 2000; Ferradeira, 1993; Basso, 2009) e o estudo da correferência nominal (Freitas et al., 2009; Fonseca, 2014; Fonseca et al., 2014, 2016a,b). Este último é o foco deste trabalho.

De forma geral, para este tipo de problema, muitos trabalhos adotam técnicas de aprendizado de máquina. Soon et al. (2001) são dos pioneiros nesse tipo de abordagem. Para o aprendizado, a obtenção de bons resultados depende da qualidade dos recursos utilizados. A língua portuguesa ainda possui uma carência por corpora com anotações de correferência suficientes para treinar modelos mais robustos. E, quando envolvemos o uso da semântica, a carência é ainda maior, dado que a quantidade de amostras é significativamente menor. Se compararmos os dois principais corpora para o Inglês e para o Português, temos, respectivamente, 34290 cadeias para o corpus Ontonotes (Pradhan et al., 2011) e 560 cadeias para o corpus Summ-it (Collovini et al., 2007). Dessa forma, em idiomas com carência de tais bases anotadas, uma abordagem baseada em regras linguísticas pode prover resultados mais significativos. Por outro lado, tem crescido a disponibilidade de recursos semânticos para o Português que podem ser utilizados para auxiliar em problemas relacionados a essa tarefa. Portanto, apresentamos neste artigo um sistema baseado em regras e conhecimento semântico para a resolução de correferências.

As principais contribuições deste trabalho são:

- a análise individual e conjunta das regras empregadas na solução do problema;
- um modelo para a resolução de correferências em Português, que faz uso de conhecimento semântico e, com isso, amplia a abrangência nos resultados.

DOI: 10.21814/lm.9.1.241

Este artigo está estruturado da seguinte forma: na Seção 2 é dada uma contextualização referente à tarefa de resolução de correferências e seus desafios, bem como é explorado o papel da semântica nesse processo; na Seção 3 são descritos os principais trabalhos relacionados, bem como os níveis de semântica e recursos utilizados por cada um; na Seção 4 são abordados os principais recursos utilizados na concepção de nosso modelo, que é descrito na Seção 5; na Seção 6 descrevemos os experimentos conduzidos, as métricas utilizadas na avaliação do modelo e a análise dos resultados; na Seção 7 é dada uma breve descrição do CORP, a ferramenta construída com base no modelo; na Seção 8 efetuamos uma análise de erros; e, por fim, na Seção 9 temos as conclusões e trabalhos futuros.

2 Semântica aplicada à Resolução de Correferência

A Resolução de correferências é um processo que consiste em identificar as diversas formas em que uma mesma entidade é evocada em um determinado texto. Em outras palavras, esse processo consiste em identificar as menções (expressões textuais) associadas a entidades ou eventos do mundo real. Em um discurso, menções que referem a uma mesma entidade são chamadas menções correferentes e formam um conjunto de menções, definido como cadeia de correferência (Poesio et al., 2016). Na sentença "A opinião é de Miguel Guerra, da Universidade de Santa Catarina (UFSC). Guerra participou...", podemos dizer que [Guerra] é uma correferência de [Miguel Guerra].

Existem casos em que estabelecer uma relação de correferência pode parecer uma tarefa simples, como em [Miguel Guerra] e [Guerra], dado que ambos os sintagmas compartilham o termo "Guerra". No entanto, ainda que estejamos lidando com a tarefa em nível lexical, existem situações mais complexas, que necessitam de tratamento distinto. Considere os seguinte exemplos:

(1) a. [**o sul** do Brasil], [**o sul** da África]
b. [**Universidade** do Paraná], [**Universidade** de São Paulo]

(2) [O **Brasil**], [a região sul do **Brasil**]

(3) [Adalberto **Portugal**], [**Portugal**]

(4) a. [a abelha], [o inseto]
b. [os ossos], [o fóssil]

Nos exemplos em 1 temos núcleos idênticos, mas os complementos indicam que os referentes são diferenciados. Em 2 temos o termo "Brasil" em ambos os sintagmas; no entanto, o primeiro refere-se ao país "Brasil" e o segundo a "a região sul do Brasil". Em 3, temos uma situação um pouco mais complexa, pois ambas as expressões possuem o termo "Portugal". Nesse caso, a palavra pode referir-se a uma entidade do tipo "Pessoa" ou "Local". Há casos, também, em que dois sintagmas podem discordar em gênero e (ou) número, mas ainda assim serem correferentes, como em 4. Em casos como esse, precisamos recorrer à semântica. Por meio dela, é possível identificar relações que vão além do reconhecimento de características lexicais.

Não é novidade que a semântica pode prover ganhos à resolução de correferência (Coreixas, 2010; Rahman & Ng, 2011; Ponzetto & Strube, 2006; Haghighi & Klein, 2009; Durrett & Klein, 2014; Fonseca et al., 2016b). Nesta Seção, citamos os principais recursos semânticos, utilizados na resolução de correferência, disponíveis para o Inglês e para o Português: para o Inglês, temos recursos bem conhecidos e consolidados, como a WordNet (Miller, 1995), um banco de dados lexical que possui informações sobre substantivos, verbos, adjetivos e advérbios. Todas essas classes de palavras são agrupadas em conjuntos de sinônimos, denominados synsets. Cada synset expressa um conceito distinto, que está interligado por meio de relações semânticas e lexicais. Temos também o FrameNet (Baker et al., 1998), contendo a similaridade semântica entre os verbos (caminhar, andar), e Yago (Suchanek et al., 2007), uma ontologia que contém relações semânticas como *Means* (significa) e *Type* (tipo de), análogas a, respectivamente, sinonímia e hiponímia.

Para o Português, temos algumas alternativas, como WordNet.PT. WordNet.BR, MultiWordNet.PT (Gonçalo Oliveira et al., 2015); FrameNetBR (Salomão, 2009), contendo relações semânticas entre verbos, com foco no domínio "Futebol". TEP2.0 (Maziero et al., 2008), um *thesaurus* contendo relações de sinonímia e antonímia; e, mais recentemente, foi criada a Onto.PT (Gonçalo Oliveira, 2012), uma ontologia semântica para o Português, sobre a qual são dados mais detalhes na Seção 4. Na Seção 3 detalham-se as características de cada recurso semântico que foram utilizadas na concepção de modelos de correferência.

3 Trabalhos Relacionados

Na literatura, encontramos muitos trabalhos voltados à resolução de correferências. Em sua grande maioria, esses trabalhos fazem um uso mais restrito da semântica, focando em categorias de entidades nomeadas e deixando de lado relações importantes, que poderiam trazer ganhos à tarefa. Nesta Seção, relatamos os principais trabalhos voltados à resolução de correferências para os idiomas Português e Inglês. Veremos que os níveis de semântica utilizados variam de acordo com o escopo e idioma de cada trabalho.

O trabalho de Lee et al. (2013), para a língua inglesa, faz uso de semântica para identificar menções que remetem a entidades do tipo "Pessoa", objetivando resolver correferências pronominais. Isto é, os autores utilizam semântica de forma mais simples, fazendo uso de apenas uma categoria de entidade, sem explorar quaisquer outras possíveis relações semânticas. Existem trabalhos que fazem um uso mais elaborado da semântica, como o de Rahman & Ng (2011), em que avaliaram a utilidade do conhecimento de mundo usando duas bases de conhecimento: Yago (Suchanek et al., 2007) e FrameNet (Baker et al., 1998). Utilizando os recursos citados, os autores fazem a identificação de relações semânticas como: "*Means*" (significa) e "*Type*" (tipo de). Cada relação semântica é representada por uma tripla *(AlbertEinsten, Type, physicist)*. Essa instância denota o fato de que Albert Einstein é um físico. A relação *"Means"*, análoga à sinonímia, provê as diferentes formas de expressar uma entidade. Portanto, permite tratar casos ambíguos, como: *(Einstein, Means, AlbertEinstein)* e *(Einstein, Means, AlfredEinstein)*, pois denotam o fato de que "Einstein" pode referir-se ao físico Albert Einstein e ao músico Alfred Einstein. Do FrameNet foram utilizados os papéis semânticos dos verbos, como por exemplo:

> Peter Anthony condena o programa de negociação, limitando o jogo para alguns, mas ele não tem certeza se quer denunciá-lo, porque...

Note que o papel semântico pode ajudar a estabelecer um link de correferência entre "programa negociação" e o pronome pessoal oblíquo "lo", uma vez que com o FrameNet é possível recuperar a relação entre "condena" e "denuncia", pelo fato dessas duas palavras aparecerem no mesmo *frame* e os dois sintagmas possuírem o mesmo papel semântico. Como resultado, os autores constataram que a semântica pode prover pequenos ganhos para a tarefa de resolução de correferências e, mesmo que pequenos, se acumulados, podem tornar-se algo substancial.

Hou et al. (2014) propôs um modelo baseado em regras, para a resolução de anáforas diretas e indiretas (*bridging*). A resolução de anáforas indiretas, consiste em reconhecer e criar um elo entre duas menções por meio de uma relação de "não identidade". Um bom exemplo de tal relação é a meronímia (parte de), como em: "a casa" e "a chaminé". Para identificar tais relações, os autores utilizaram o WordNet (Miller, 1995).

Para a língua portuguesa, Silva (2011) propôs um modelo para a resolução de correferências utilizando o conjunto de etiquetas semânticas providas pelo corpus do HAREM (Freitas et al., 2010). Para detectar tais categorias, Silva utilizou o parser PALAVRAS (Bick, 2000) e o reconhecedor de entidades nomeadas Rembrandt (Cardoso, 2012). Como base de conhecimento semântico, o autor utilizou o TEP2.0 (Maziero et al., 2008), um *thesaurus* contendo relações de sinonímia e antonímia para a língua portuguesa.

Ainda considerando o Português, Coreixas (2010) propôs a resolução de correferências, focando-se nas categorias "Pessoa", "Local", "Organização", "Acontecimento", "Obra", "Coisa" e "Outro". Como recursos, foram utilizados o corpus do HAREM, o parser Palavras e o corpus Summ-it. De forma a demonstrar que o uso de categorias semânticas pode auxiliar na tarefa de resolução de correferências, o autor compara duas versões de seu sistema: a primeira, sem fazer o uso de categorias semânticas; e a segunda, fazendo uso dessas categorias. Como resultado, Coreixas (2010) mostrou que o uso de categorias pode prover melhorias significativas, dado que o uso de categorias pode auxiliar a determinar se dado par de menções é correferente ou não. O autor também mostrou a importância do conhecimento de mundo para esta linha de pesquisa.

Garcia & Gamallo (2014a), propõem um modelo baseado em regras (semelhante ao de Lee et al. (2013), mas para múltiplos idiomas (Português, Espanhol e Galego). Em seu trabalho, os autores focam apenas na categoria semântica "Pessoa".

Em trabalhos anteriores (Fonseca et al., 2014) propusemos uma abordagem baseada em aprendizado de máquina, com foco em nomes próprios e nas categorias de entidades "Pessoa", "Local" e "Organização". Para detectar as entidades, utilizamos o Repentino (Sarmento et al., 2006) e NERP-CRF (do Amaral, 2013). Adicionalmente,

para casos mais genéricos de entidades, utilizamos listas, contendo substantivos comuns, que remetem a determinadas entidades, tais como: [advogado, agrônomo, juiz] para a categoria "Pessoa", e [avenida, rua, praça, cidade] para "Local".

Como podemos ver, existem muitos trabalhos propondo o uso de semântica, no entanto os níveis dessas regras variam de acordo com o escopo e quantidade de recursos disponíveis. Nosso modelo atual teve como objetivo avançar no estado da arte no que diz respeito à tarefa de resolução de correferências para o Português, utilizando recursos semânticos mais recentes, disponíveis para o português.

4 Recursos

Nesta Seção, apresentamos quatro recursos fundamentais para a concepção de nosso trabalho: o CoGrOO (Silva, 2013), um corretor gramatical com diversas funcionalidades para o português; o Onto.PT (Gonçalo Oliveira, 2012), ontologia utilizada para obtenção de relações semânticas (hiponímia e sinonímia); e CoNLL Scorer (Pradhan et al., 2014) e Summ-it++ (Antonitsch et al., 2016), utilizados na avaliação de nosso modelo.

CoGrOO

CoGrOO é um corretor gramatical de código aberto, capaz de prover anotação sintática. Tendo como principal funcionalidade a correção gramatical, o CoGrOO é capaz de identificar erros como: colocação pronominal, concordância nominal, concordância sujeito-verbo, uso da crase, concordância nominal e verbal e outros erros comuns de escrita em português do Brasil. Para tal, o CoGrOO realiza uma análise híbrida: inicialmente, o texto é anotado usando técnicas estatísticas de Processamento de Linguagens Naturais e, em seguida, um sistema baseado em regras é responsável por identificar os possíveis erros gramaticais. Além das funcionalidades já descritas, o CoGrOO possui, da mesma forma que o OGMA (Maia, 2008) e o PALAVRAS, a anotação de sintagmas nominais. Além disso, conta também com análise morfológica e com lematização.

Onto.PT

Construído de forma automática por meio de dicionários e de *thesaurus* da língua portuguesa, o Onto.PT é considerado uma ontologia de base para o português. Similar ao Wordnet (Miller, 1995), o Onto.PT possui uma estrutura baseada em *synsets*[1] e relações semânticas conectando esses *synsets*, como: hiperonímia, hiponímia, sinonímia, meronímia, entre outras. Na Tabela 1, podemos visualizar os tipos de relações semânticas consideradas por nosso modelo e suas quantidades, presentes na ontologia.

Para extrair as relações semânticas do Onto.PT, utilizamos uma API[2] que, para um dado par de palavras, retorna suas relações semânticas, conforme podemos visualizar na Tabela 2.

Relação	Tipo	Quantidade
Sinônimo_De	substantivo	84.015
	verbo	37.068
	adjetivo	45.149
	advérbio	2.626
Hipônimo_De	substantivo	91.466
Total	—	260.324

Tabela 1: Quantidade de relações no Onto.PT.

Par	Relação
estudo, pesquisa	sinonimoDe
abelha, inseto	hiponimoDe
animal, cachorro	hiperonimoDe

Tabela 2: Onto.PT: Exemplos de relações semânticas para um dado par de palavras.

Summ-it++

Concebido a partir do corpus Summ-it, o Summ-it++ consiste em uma nova versão do Summ-it portada para o formato SemEval (Recasens et al., 2010) e enriquecida com duas novas camadas de anotação semântica: Relação entre entidades nomeadas (Collovini et al., 2014); e Categorias de Entidades Nomeadas (do Amaral, 2013). O Summ-it++, assim como o Summ-it, possui 5033 menções, 3022 links, 560 cadeias de correferência. Adicionalmente, possui 1086 entidades nomeadas classificadas e 37 descritores de relação entre essas entidades. Para nossa avaliação, o corpus Summ-it++ mostrou-se o mais indicado, dado que possui anotação de correferência em nível de sintagmas nominais. Outros corpora para o Português, como o HAREM ou o de Garcia & Gamallo (2014b) possuem anotação de correferência apenas para categorias de entidades nomeadas. Na Tabela 3, podemos visualizar como são dis-

[1] Grupos de palavras que possuem um mesmo significado ex: [moço, menino, filho, garoto, rapaz].

[2] http://github.com/rikarudo/OntPORT

postas as informações do corpus. Essas são importantes, dado que para efetuar nossa avaliação, a saída de nosso modelo também teve de ser convertida para este formato. Na Tabela 3, cada coluna representa respectivamente:

ID: identificador de cada palavra na ordem em que elas aparecem na sentença;

Token: palavra ou multi-palavra;

Lemma: lema;

POS: análise morfológica (*part-of-speech*) de cada palavra;

Feat: gênero e número (*features*) de cada palavra;

Head: denota se a palavra é um núcleo (*head*) de sintagma nominal (caso sim, o campo recebe o valor '0');

NE: representa a categoria semântica das entidades nomeadas;

Rel: representa o descritor que expressa a relação entre um par de entidades nomeadas. Quando essa relação existe, ambas as entidades nomeadas envolvidas recebem o ID das palavras que compõem o descritor de relação.

Coref: contém o identificador da cadeia, sendo que o início de um sintagma é marcado por "(", e o seu final, por ")". Basicamente, menções correferentes recebem o mesmo ID.

CoNLL Scorer

Desenvolvido com o intuito de atender as necessidades da CoNLL shared task (Pradhan et al., 2011, 2012), o CoNLL Scorer (Pradhan et al., 2014) consiste em uma API cujo objetivo é avaliar modelos de resolução de correferência. Seu objetivo principal é prover uma forma automatizada e justa de avaliar tais modelos. Isso porque, como descrito por Pradhan et al. (2014), cada métrica favorece uma característica específica entre os links de menções. Dados os fatos, o recurso utiliza a média entre as três principais métricas, para determinar uma pontuação única.

Basicamente, tendo como entrada dois arquivos (ambos necessitam estar no formato SemEval (Recasens et al., 2010), um formato muito conhecido e utilizado pela maioria dos corpora): o primeiro, contendo as anotações que são o padrão de referência, e o segundo contendo as anotações, providas automaticamente pelo modelo a ser avaliado, o CoNLL Scorer calcula uma pontuação. Além disso, o recurso fornece também os resultados de todas as métricas conhecidas (MUC, B^3, Ceaf e BLANC) (Vilain et al., 1995; Bagga & Baldwin, 1998; Luo, 2005; Recasens & Hovy, 2011).

5 Descrição do Modelo

Nosso modelo segue o padrão de uma arquitetura multi-passos, baseada em regras linguísticas, assim como o modelo de Lee et al. (2013). Em uma arquitetura multi-passos, cada etapa consiste em aplicar determinada regra, objetivando agrupar duas menções m_x e m_y, caso suas restrições sejam satisfeitas. Diferente de Lee et al. (2013), nosso modelo é aplicado para o Português, e introduz o uso de conhecimento semântico provido pelo Onto.PT.

Nossas regras formam um conjunto facilmente encontrado em trabalhos realizados para o Inglês (Lee et al., 2013; Rahman & Ng, 2011; Soon et al., 2001). Contudo, nosso trabalho tem como diferencial o idioma para o qual é voltado e sua combinação específica de regras. Além disso, poucos trabalhos, mesmo para o Inglês, abordam o uso de regras semânticas, como Hiponímia e Sinonímia, para a resolução de correferências. Muitas de nossas regras foram adaptadas da literatura, considerando o padrão linguístico do Português e as limitações dos recursos disponíveis para o nosso idioma.

Inicialmente, realizamos a detecção de menções, por meio do parser CoGrOO (Silva, 2013); seguido de um pré-processamento, o qual removemos menções que: iniciem com entidades numéricas como percentual, dinheiro, cardinais e quantificadores (9%, $10,000, Dez, Mil, 100 metros). Apesar de existir correferência numérica, esta é responsável pela maioria das ligações incorretas. Portanto, optamos por não tratá-los. Após as etapas de detecção de menções e pré-processamento são aplicadas 13 regras (11 lexicais e 2 semânticas).

Regras Básicas

Casamento de Padrões Exato (Regra 1)
Considera como correferentes duas menções, cujos sintagmas nominais sejam exatamente iguais, incluindo seus modificadores e determinantes.

(5) a. [o Brasil], [o Brasil]

b. [a Amazônia], [a Amazônia]

Esta regra não agrupa pronomes e, para realizar o agrupamento, os sintagmas não podem pertencer

ID	Token	Lemma	PoS	Feat	Head	NE	Rel	Coref
1	A	o	art	F=S	_	_	_	_
2	opinião	opinião	n	F=S	0	_	_	_
3	é	ser	v-fin	PR=3S=IND	_	_	_	_
4	de	de	prp	_	_	_	_	_
5	o	o	art	M=S	_	_	_	(2
6	agrônomo	agrônomo	n	M=S	0	_	_	_
7	Miguel_Guerra	_	prop	M=S	0	PES	(9)	_
8			_	_	_	_	_	_
9	de	de	prp	_	_	_	_	_
10	a	o	art	F=S	_	_	_	_
11	UFSC	_	prop	F=S	0	ORG	(9)	(3)
12	(	(	(	_	_	_	_	_
13	Universidade_de_Santa_Catarina	_	prop	F=S	0	ORG	_	(3) | 2)
14	)	)	)	_	_	_	_	_
15	.	.	.	_	_	_	_	_
1	Guerra	_	prop	M=S	0	PES	_	(2)
2	participou	participar	v-fin	PS=3S=IND	_	_	_	_
...								

Tabela 3: Esquema de anotação Summ-it++.

a uma construção de aposto especificativo (regra 4); caso eles pertençam, seus sintagmas ligeiramente anteriores devem ser iguais. Com essa restrição evitamos links como:

(6) [[o telescópio] [**Gemini**]],
[[o projeto] [**Gemini**]]

Note que os sintagmas "Gemini" são exatamente iguais, no entanto são sub-sintagmas (adjuntos) de "o telescópio" e "o projeto". Em poucas palavras, após o processo de chunking[3], temos os seguintes sintagmas nominais: [o telescópio], [Gemini],[o projeto] e [Gemini]. Logo, mesmo esses sintagmas nominais possuindo um casamento exato não necessariamente significa que existe uma uma relação de correferência, dado que estes são adjuntos adnominais.

Casamento Parcial pelo Núcleo (Regra 2)
Considera como correferentes duas menções, cujo casamento obtido por meio do truncamento de seus sintagmas seja igual num mesmo contexto. O truncamento das menções é realizado levando em consideração seus núcleos, como nos exemplos abaixo:

(7) a. [o piloto americano], [o piloto]
b. [o ministro da justiça], [o ministro]

[3]Nem sempre o CoGrOO efetua a separação dos adjuntos adnominais. No entanto, para ambos os casos esta restrição é válida e previne links incorretos, aumentando a precisão do modelo

Assim como na regra Casamento de Padrões Exatos, pronomes e menções que estejam em uma construção de Aposto Especificativo não são agrupados por esta regra.

Aposto Explicativo (Regra 3)
Agrupa duas menções caso essas estejam em uma construção de aposto (Cadore & Ledur, 2013; Bechara, 1972). Essa regra consiste em buscar por marcações padrões que ajudam a identificar o aposto, como parênteses e menções entre vírgulas.

(8) a. [A Embrapa] ([Empresa Brasileira de Pesquisa Agropecuária])
b. [A ministra da justiça do país], [Elisabete Guigou], ...

Aposto Especificativo (Regra 4)
Consiste em verificar se duas menções vizinhas, m_i e m_{i+1}, estão em uma construção de aposto especificativo[4] (Cadore & Ledur, 2013; Bechara, 1972). Basicamente, se satisfazem as seguintes restrições:

- menção m_{i+1} é um nome próprio;
- menção m_i é um substantivo comum;
- menção m_i deve possuir um artigo definido;

[4]Diferente de Lee et al. (2013), aplicamos esta regra a todos os sintagmas nominais, não apenas a categoria pessoa.

- menção m_{i+1} não pode possuir um determinante;
- menções m_i e m_{i+1} devem estar na mesma sentença e serem adjacentes no texto (não pode haver outras palavras entre elas).
- caso o determinante de m_i esteja no plural, agrupa todas as menções subsequentes que:
  - sejam nomes próprios;
  - estejam na mesma sentença;
  - estejam separados por vírgula (ou "e" após as vírgulas).

(9) a. [o arqueólogo português], [Francisco Alves]

b. [o galeão], [Nossa Senhora dos Mártires]

c. [os brasileiros], [Gilson Rambelli, Paulo Bava de Camargo e Flávio Rizzi].

Acrônimo (Regra 5)
Agrupa duas menções se uma menção m_i é sigla de m_j.

(10) [Organização das Nações Unidas], [a ONU]

Predicado Nominativo (Regra 6)
Tem como objetivo identificar predicados nominativos e agrupá-los com suas respectivas referências. Para isso, buscamos por uma sequência que possua um verbo de ligação seguido de um determinante/artigo, como, por exemplo, (é um, é uma, foi o, foram os...); encontrada a sequência (verbo de ligação + determinante), agrupamos as menções adjacentes, como em:

(11) [A França] **é** [**o** único país que se recusa a aceitar a determinação europeia]

Nessa regra, consideramos apenas o verbo "ser", conjugado no passado, presente e futuro do singular e do plural. Outros verbos de ligação não foram considerados, pois geralmente associam-se a adjetivos, e não a substantivos, como por exemplo:

- Cláudia **anda** nervosa.
- Diana **continua** feliz.
- Nicole **ficou** triste.
- João **está** feliz.

Pronome Relativo (Regra 7)
Busca por menções que possuam/sejam pronomes relativos. Identificado um pronome relativo m_{i+1}, este é agrupado com a menção anterior adjacente m_i:

(12) [Wilkinson Microwave Anisotropy Probe], [cujos] primeiros dados.

Casamento Restrito pelo Núcleo (Regras 8 e 9)
Consiste em agrupar (por meio de um casamento ingênuo) duas menções, caso seus núcleos sejam iguais. Esse casamento, ao considerar apenas o núcleo dos sintagmas, muitas vezes pode causar um agrupamento incorreto, já que não considera que possam existir modificadores incompatíveis, como, por exemplo: Universidade de São Paulo e Universidade de Brasília. Note que os núcleos desses sintagmas são iguais, no entanto referem-se a entidades distintas. Para evitar esse tipo de agrupamento incorreto, esta regra implementa algumas cláusulas restritivas, que devem ser combinadas de modo a produzirem um link.

- **Casamento entre Núcleos:** O núcleo da menção atual m_j precisa ser o mesmo do antecedente m_i.

  (13) [Universidade Federal de São Paulo] ... [a Universidade] ...

- **Palavra Modificadora:** Todas as palavras de dada menção m_j, não consideradas como *stopwords* (substantivos comuns, próprios, verbos, adjetivos e advérbios) são incluídas em uma lista e comparadas com a menção antecedente m_i. Dessa forma, é possível verificar se existe alguma palavra que modifica o núcleo do antecedente. Essa cláusula explora a propriedade de discurso que nos diz que é incomum introduzirmos novas informações em novas menções a uma mesma entidade. Basicamente, menções subsequentes a uma mesma entidade possuem a tendência de serem menos explicativas.

  (14) [A menina que caiu e se machucou], [A menina que está feliz]

  Note que as palavras "está" e "feliz", existentes na menção atual, não são *stopwords*, então verificamos se essas duas palavras modificam o antecedente. Como o antecedente não possui as palavras "está e feliz", elas naturalmente o modificarão. Portanto, o agrupamento das menções não é realizado.

(15) [A estrada de Minas Gerais que ficará pronta], [A estrada que talvez esteja pronta]

As menções contidas no exemplo acima também não seriam agrupadas, dado que o advérbio "talvez" e o verbo "esteja" (contidos em "A estrada que talvez esteja pronta") modificariam o antecedente.

- **Modificadores Compatíveis:** Os modificadores de uma menção m_j atual são todos incluídos na lista de modificadores do candidato antecedente m_i. Essa cláusula é semelhante à "Palavra Modificadora", com o diferencial de que considera apenas modificadores que são substantivos e adjetivos. Em outras palavras, essa regra verifica se os modificadores do tipo adjetivos e substantivos, quando existem na menção, são iguais aos da menção anterior. Note que essa heurística realizaria o mesmo agrupamento que a regra "Palavra Modificadora" para o exemplo 14, porém teria um resultado diferente para o exemplo 15. Ou seja, o fato de haver um modificador — advérbio (talvez) e um verbo (esteja), por exemplo — não afeta o fato de serem correferentes, altera apenas o sentido do enunciado. Logo, a clásula "Modificadores Compatíveis" agruparia as duas menções do exemplo 15, pois as palavras da menção atual, m_j, (A estrada que talvez esteja pronta), consideradas não stopwords são: "Estrada" e "pronta", palavras que não modificariam o antecedente.

- **Encapsulamento de Menções** Esta cláusula nos diz que duas menções, para serem correferentes, uma menção não pode ser parte constituinte da outra. De forma a reconhecer este tipo de dependência, utilizamos o reconhecimento de preposições, como: "de" (e suas variações "do", "da", "dos", "das") e "em" (e suas variações "no", "na", "nos" e "nas"). No exemplo 16, [o menino] não pode fazer referência a [o pijama listrado] justamente porque a regra faz com que a preposição torne-se parte indispensável para haver correferência. Desse modo, a preposição "de" torna o sintagma [o pijama listrado] expressão adjunta de [o menino].

(16) [O menino de pijama listrado], [o pijama listrado].

É importante mencionar que a Regra "Casamento Restrito pelo Núcleo" consiste de duas etapas. A primeira (8) realiza o agrupamento das menções levando em consideração (Casamento entre Núcleos $\wedge$ Palavra Modificadora $\wedge$ Encapsulamento de Menções). A segunda (9) busca menções em que (Casamento entre Núcleos $\wedge$ Modificadores Compatíveis $\wedge$ Encapsulamento de Menções) sejam satisfeitas. Essas duas variações foram propostas por Lee et al. (2013) e mostraram uma melhoria de 0.9% na medida-f, quando utilizadas linearmente.

Casamento entre Nomes Próprios (Regra 10)
Agrupa duas menções caso as seguintes condições sejam satisfeitas:

- ambas as menções devem conter nomes próprios;
- os nomes próprios precisam ser iguais lexicalmente;
- as duas menções não devem estar encapsuladas, ou seja, devem respeitar a cláusula "Encapsulamento de Menções".

(17) [Califórnia],[a região sul da Califórnia].

No exemplo acima, temos a violação da terceira condição. Note que ambos os sintagmas nominais possuem o mesmo nome próprio, mas violam a cláusula "Encapsulamento de Menções", de modo semelhante ao exemplo 16. Neste caso, [Califórnia] e [da Califórnia] não podem ser correferentes pelo fato de a segunda menção estar ligada a uma preposição, tornando-a adjunto adverbial de lugar. Portanto, há uma especificação, em que não se está referindo a toda a Califórnia, mas somente à região sul desse estado.

Casamento Parcial entre Nomes Próprios (Regra 11)

Semelhante à regra "Casamento entre Nomes Próprios", mas permite que o núcleo da menção atual m_j combine com qualquer palavra existente na menção anterior m_i. Como em: [o agrônomo da UFSC, Miguel Guerra] e [Guerra]. Para realizar o agrupamento, algumas cláusulas devem ser respeitadas:

- ambas as menções devem conter nomes próprios;
- pelo menos uma palavra de m_j deve ser igual à m_i;
- o agrupamento deve respeitar a cláusula "Palavra Modificadora"

Regras Semânticas

Hiponímia (Regra 12)
Agrupa duas menções (m_i e m_j) se os lemas, provenientes dos núcleos de m_i e m_j, são hipônimos. Para encontrar tais relações, utilizamos o Onto.PT (Gonçalo Oliveira, 2012). Esta regra ajuda a agrupar menções como as do exemplo abaixo:

(18) Já se perguntou como as abelhas fabricam mel? Os insetos saem em busca de...

Para evitar o agrupamento incorreto de menções (exemplo 18), foram combinadas técnicas de pré e pós modificadores. Nesse exemplo, se extrairmos o lema do núcleo das menções e efetuarmos uma busca pela existência de relações semânticas entre "quebra-cabeça" e "problema", veremos que "quebra-cabeça" possui uma relação de hiponímia com "problema", mas note que as menções "o quebra-cabeça genético" e "problema ambiental" não são correferentes. Para evitar tal agrupamento, adicionamos a cláusula "Palavra Modificadora[5]". Dessa forma, o termo "ambiental" torna-se um modificador e o agrupamento das menções não é realizado.

(19) Foi o tempo em que decifrar o genoma ...o **quebra-cabeça** genético...
Isso é um **problema** ambiental...

Nesse sentido, para ocorrer o agrupamento de duas menções, duas condições precisam ser satisfeitas:

- o lema do núcleo das menções m_i e m_j necessita possuir uma relação de hiponímia;
- não podem haver palavras que modifiquem as menções (cláusula Palavra Modificadora).

Nós consideramos apenas a relação de hiponímia entre um referente e seu antecedente (não utilizamos hiperonímia), dado que no Português é mais comum introduzirmos uma entidade de forma mais específica e, em suas próximas menções, utilizarmos termos mais gerais para referir à mesma entidade, conforme o exemplo 19. Além disso, testes realizados com a regra Hiperonímia foram realizados, no entanto, a regra acabou gerando muitos links incorretos entre as menções. Contudo, não descartamos totalmente o uso de hiperônimos, estamos buscando apoio em Aprendizado de Máquina, objetivando descobrir a eficácia da regra Hiperonímia quando combinada com outras restrições e regras (Fonseca et al., 2016b).

[5]Nas regras de Hiponímia e Sinonímia os núcleos não são considerados palavras modificadoras.

Sinonímia (Regra 13)
Semelhante à regra Hiponímia, a regra Sinonímia agrupa duas menções quando há uma relação de sinonímia entre elas, respeitando as seguintes restrições:

- o lema do núcleo das menções m_i e m_j necessitam possuir uma relação de sinonímia;
- não podem haver palavras que modifiquem as menções;
- cada nova menção a ser agrupada a dada cadeia de correferência, por esta regra, necessita possuir uma relação de sinonímia com todas as menções desta cadeia. Respeitando esta restrição, evitamos agrupar menções como em:

(20) A Terra é um astro do sistema solar.
Esse planeta orbita a uma distância de 149.600.000 km do Sol.

6 Experimentos

De forma a avaliar nosso modelo, usamos seis métricas amplamente utilizadas pela literatura (descritas em 6.1). Cada uma delas objetiva avaliar um aspecto específico no modelo e calcular seu desempenho. Em nossos experimentos, efetuamos dois tipos de avaliação: na primeira (Tabela 4), avaliamos os ganhos que cada regra pode prover ao modelo, de forma independente; na segunda (Tabela 5), avaliamos os ganhos que cada regra agrega ao modelo, de forma cumulativa.

Note que no corpus Summ-it++, o aposto e sua menção referente formam apenas uma menção. Dessa forma, sintagmas que aparecem na forma de aposto são considerados como uma única menção, como em: "o Instituto Nacional de Pesquisas Espaciais (INPE)...". No corpus de referência temos apenas um sintagma [o Instituto Nacional de Pesquisas Espaciais (INPE)]. Já nosso modelo identifica como duas menções e as agrupa, formando uma cadeia: [o Instituto Nacional de Pesquisas Espaciais], [Inpe]. Dessa forma, na nossa avaliação, consideramos como acerto a criação de um link nesses casos.

Métricas de Avaliação

- MUC (Vilain et al., 1995): baseada em cadeias, mede quantos agrupamentos de menções são necessários para cobrir as cadeias padrão. O cálculo da métrica MUC é dado por meio das seguintes fórmulas:

$$Abrangência = \frac{\sum_{i=1}^{N_k} (\|K_i\| - \|p(K_i)\|)}{\sum_{i=1}^{N_k} (\|K_i\| - 1)}$$

$$Precisão = \frac{\sum_{i=1}^{N_r} (\|R_i\| - \|p'(R_i)\|)}{\sum_{i=1}^{N_r} (\|R_i\| - 1)}$$

Onde: K_i é i-ésima *key entity* (padrão) e $p(K_i)$ é o grupo de partições criado por meio da intersecção de K_i e os links preditos pelo modelo; R_i é a i-ésima *Response entity* (entidade predita pelo modelo) e $p'(R_i)$ é o conjunto de partições criadas por meio da intersecção de R_i e K_i. N_k e N_r representam a quantidade de menções padrão e resposta, respectivamente.

- B^3 (Bagga & Baldwin, 1998): baseada em menções, gera resultados tendo como foco as menções de cada entidade. Sua abrangência e precisão são obtidas por:

$$Abrangência = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \frac{\|K_i \cap R_j\|^2}{K_i}}{\sum_{i=1}^{N_k} K_i}$$

$$Precisão = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \frac{\|K_i \cap R_j\|^2}{R_j}}{\sum_{i=1}^{N_k} R_j}$$

Onde K representa o conjunto das *key entities* (menções padrão) e R o conjunto de menções preditas pelo modelo.

- CEAF (Luo, 2005): baseada no alinhamento de menções e entidades, possui duas variações: $CEAF_m$ (Φ_3) e $CEAF_e$ (Φ_4).

$$\Phi_3(K, R) = \|K \cap R\|$$

$$\Phi_4(K, R) = \frac{2\|K \cap R\|}{\|K\| + \|R\|}$$

$$Abrangência = \frac{\Phi_x}{\sum_{i=1} \|K_i\|}$$

$$Precisão = \frac{\Phi_x}{\sum_{i=1} \|R_i\|}$$

- BLANC (BiLateral Assessment of NounPhrase Coreference) (Recasens & Hovy, 2011): avalia tanto os links de correferência quanto os não correferentes. Temos, então, C_K e C_R respectivamente como: links de correferência padrão e preditos automaticamente e; N_K e N_R como grupo dos links de não correferência padrão e preditos automaticamente; $Abrangência_C$ e $Precisão_C$ remetem ao cálculo de abrangência e precisão dos links de correferencia, e $Abrangência_N$ e $Precisão_N$, aos links de não correferência.

$$Abrangência_C = \frac{\|C_k \cap C_r\|}{C_k}$$

$$Precisão_C = \frac{\|C_k \cap C_r\|}{C_r}$$

$$Abrangência_N = \frac{\|N_k \cap N_r\|}{N_k}$$

$$Precisão_N = \frac{\|N_k \cap N_r\|}{N_r}$$

- CoNLL (Pradhan et al., 2014): amplamente utilizada para avaliar modelos de resolução de correferência, a métrica CoNLL calcula um score único, baseando-se no cálculo da medida-f das métricas MUC, B^3 e $CEAF_e$:

$$\text{CoNLL} = \frac{(F(\text{MUC}) + F(B^3) + F(\text{CEAF}_e))}{3}$$

Análise dos Resultados

Analisando a Tabela[6] 4, podemos notar que as regras que lidam com o casamento de padrões entre palavras obtiveram precisões acima de 60%, tendo como destaque as regras 8 e 9 (Casamento Restrito pelo Núcleo), cujos resultados ultrapassaram 46% de score para a métrica CoNLL. Podemos notar também que a regra 3 (Aposto Explicativo) possui uma alta precisão, no entanto ocorre com pouca frequência no corpus utilizado para teste. Referente às regras semânticas Hiponímia e Sinonímia (12 e 13), notamos que sinonímia apresenta melhores resultados do que hiponímia. Apesar de individualmente não apresentarem os melhores resultados, quando utilizadas em conjunto com outras regras, podemos ver ganhos na abrangência.

[6] Nas Tabelas 4, 5 e 6 "P","A" e "F" representam respectivamente: Precisão, Abrangência e Medida-F.

	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			CoNLL
	P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	F
Regra 1	66.4	22.8	34.0	68.0	19.1	29.8	64.5	26.5	37.6	50.5	28.1	36.1	83.2	64.5	68.4	33.3
Regra 2	61.9	30.7	41.1	63.3	25.8	36.7	58.9	34.6	43.6	47.3	37.0	41.5	80.6	59.9	62.1	39.8
Regra 3	74.8	5.9	10.9	78.7	6.9	12.6	80.4	8.6	15.5	70.2	11.8	20.2	92.4	92.4	92.4	14.6
Regra 4	11.1	0.4	0.7	22.3	0.7	1.4	32.6	1.4	2.8	26.9	1.8	3.5	57.5	57.3	57.3	1.9
Regra 5	58.8	0.7	1.4	65.5	0.7	1.5	75.9	1.1	2.2	66.7	1.2	2.5	65.1	63.9	63.6	1.8
Regra 6	18.2	0.1	0.3	34.1	0.1	0.3	50.0	0.5	1.1	26.5	0.4	0.9	47.7	48.2	44.4	0.5
Regra 7	0.0	0.0	0.0	11.8	0.1	0.3	21.0	0.4	0.8	17.7	0.5	1.0	47.2	46.9	46.4	0.4
Regra 8	61.2	39.4	48.0	60.6	34.2	43.7	61.1	43.4	50.7	52.3	44.5	48.1	76.8	59.7	61.9	46.6
Regra 9	61.1	39.8	48.2	60.5	34.6	44.0	61.3	43.8	51.1	52.4	44.9	48.4	76.7	59.7	61.9	46.9
Regra 10	70.2	7.8	14.0	73.0	6.7	12.3	78.6	10.1	17.9	62.4	10.4	17.8	85.9	85.9	85.9	14.7
Regra 11	66.7	8.1	14.4	69.7	7.3	13.3	77.4	10.6	18.7	64.3	11.0	18.8	81.7	85.2	83.3	15.5
Regra 12	6.0	1.2	2.1	15.9	3.1	5.2	23.5	5.5	8.9	21.0	6.1	9.4	52.5	51.4	45.0	5.6
Regra 13	28.5	13.7	18.5	24.3	12.8	16.8	34.1	16.1	21.9	28.5	12.9	17.8	57.5	53.6	50.0	17.7

Tabela 4: Regras individuais.

	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			CoNLL
	P	A	F	P	A	F	P	A	F	P	A	F	P	A	F	F
Regra 1	66.4	22.8	34.0	68.0	19.1	29.8	64.5	26.5	37.6	50.5	28.1	36.1	83.2	64.5	68.4	33.3
+Regra 2	61.8	30.8	41.1	63.1	25.9	36.7	58.8	34.7	43.6	47.2	37.1	41.5	80.2	59.8	62.0	39.8
+Regra 3	63.3	36.4	46.3	64.8	32.8	43.6	61.2	41.5	49.5	51.7	46.5	49.0	81.5	60.4	63.2	46.3
+Regra 4	60.6	36.8	45.8	61.9	33.3	43.3	58.9	42.0	49.0	49.6	46.6	48.1	80.2	59.4	61.7	45.7
+Regra 5	60.4	37.0	45.9	61.7	33.5	43.4	58.7	42.2	49.1	49.6	46.8	48.1	79.9	59.3	61.6	45.8
+Regra 6	59.9	37.2	45.9	61.1	33.6	43.4	58.2	42.4	49.1	49.1	46.9	48.0	79.6	59.0	61.1	45.7
+Regra 7	58.3	36.9	45.2	59.7	33.5	42.9	56.8	42.2	48.4	47.7	46.5	47.1	78.9	58.4	59.9	45.1
+Regra 8	57.4	48.3	52.5	56.2	44.6	49.7	57.8	53.2	55.4	51.5	55.9	53.6	75.0	57.7	59.0	51.9
+Regra 9	57.4	48.6	52.6	56.2	44.8	49.8	57.9	53.4	55.6	51.6	56.2	53.8	75.0	57.7	59.0	52.1
+Regra 10	57.4	48.9	52.8	56.2	45.1	50.0	57.9	53.8	55.8	51.8	56.5	54.0	75.0	57.7	58.9	52.3
+Regra 11	57.0	48.7	52.5	55.4	45.1	49.7	57.9	53.5	55.6	52.0	55.7	53.8	74.1	57.8	59.1	52.0
+Regra 12	47.1	49.8	48.4	44.6	46.9	45.7	49.9	53.3	51.6	48.9	53.4	51.1	65.2	55.7	55.5	48.4
+Regra 13	42.3	53.6	47.3	38.7	50.8	43.9	45.2	55.6	49.9	45.6	52.8	48.9	62.9	54.6	53.3	46.7

Tabela 5: Regras cumulativas.

Por meio de nossas regras semânticas, foi possível identificar links como:

- [fungos], [pequenos cogumelos];
- [cientistas], [pesquisadores];
- [universo], [o cosmo].

Na Tabela 5, podemos inferir que a cada nova regra adicionada o modelo perde precisão, mas ganha em abrangência, aumentando, na maioria dos casos, sua medida-f. Adicionalmente, quando acrescentamos semântica ao modelo, há uma redução na medida-f. Contudo, há um aumento significativo em sua abrangência.

Na Tabela 6, temos os resultados dos principais trabalhos encontrados na literatura, avaliados utilizando as métricas da conferência CoNLL. Infelizmente, não é possível compararmos o nosso e os demais modelos, dado que cada modelo possui idioma e/ou escopos distintos. O trabalho de Garcia & Gamallo (2014a), por exemplo, resolve correferências para o Português, mas possui escopo limitado à categoria de entidade nomeada "Pessoa".

7 CORP

Como resultado da implementação do modelo de regras, o CORP (Coreference Resolution for Portuguese) é um sistema de resolução de correferências para o Português, disponível em duas versões: Desktop[7] e Web[8].

Ambas as versões produzem dois tipos de saída: a primeira, em HTML, objetiva facilitar a visualização da informação; e a segunda, em XML, que garante facilidade de processamento e reutilização da informação anotada.

Na Seção 8 são exibidas amostras de saídas em HTML, geradas pelo CORP. Menções correferentes entre si possuem o mesmo id e coloração. Contudo, existem casos em que algumas menções são parte constituinte de outras, como em: "[Claiton Campanhola, diretor de [a Embrapa[46]][35]]" (Figura 1). Em casos como esse, suas "sub-menções" recebem a mesma coloração da menção principal. Seus delimitadores e id recebem a cor correspondente à sua cadeia.

[7] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/corp-coreference-resolution-for-portuguese/

[8] http://ontolp.inf.pucrs.br/corref/

		MUC			B^3			$Ceaf_e$			CoNLL
Modelo	Idioma	P	A	F	P	A	F	P	A	F	F
Martschat et al., 2015	IN	76.8	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Fernandes et al., 2014	IN	75.9	65.8	70.5	77.7	65.8	71.2	43.2	55.0	48.4	63.4
	CH	71.5	59.2	64.8	80.5	67.2	73.2	45.2	57.5	50.6	62.9
	AR	49.7	43.6	46.5	72.2	62.7	67.1	46.1	52.5	49.1	54.2
Lee et al., 2013	IN	60.9	59.6	60.3	73.3	68.6	70.9	46.2	47.5	46.9	59.4
Garcia et al., 2014	ES	94.1	84.1	88.8	84.8	62.9	72.2	71.0	83.4	76.7	79.2
	GL	94.6	89.0	91.7	88.4	72.9	79.9	76.6	87.6	81.7	84.4
	PT	92.7	82.7	87.4	84.5	65.8	74.0	67.9	84.4	75.2	78.9
Nosso	PT	42.3	53.6	47.3	38.7	50.8	43.9	45.6	52.8	48.9	46.7

Tabela 6: Resultados não comparativos dos principais modelos da literatura.

8 Análise de Erros

Nesta Seção, apresentamos uma análise detalhada de erros do modelo. Para efetuar a análise, selecionamos três textos, pertencentes a dois corpora (Summ-it++ e CST-News (Maziero et al., 2010)). Podemos notar que os tipos mais comuns de erros ocorrem por meio do casamento parcial entre menções, agrupamento de duas ou mais cadeias de correferência, regra de aposto e regras semânticas.

Texto 1

O ministro **[Roberto_Rodrigues [66]]** (**[Agricultura [66]]**) anunciou ontem **[o nascimento de a bezerra Vitoriosa [73]]** . **[O animal [78]]** é **[um clone [78]]** gerado a_partir_de um clone a **[vaca [82]] [Vitória [82]]** , que havia sido clonada em 2001 . Para **[Rodrigues [66]]** , **[a cria [22]]** coloca a genética de o país em destaque em o cenário mundial . **[Clayton_Campanhola , diretor-presidente de [a Embrapa [46]] , [35]]** afirma que **[o método [40]]** ajudará em a multiplicação de **[animais [22]]** de elevado valor genético ou em **[a reprodução [33]]** de os ameaçados de extinção . " Se há um animal de **[boa qualidade genética [34]]** , a gente consegue manter isso em um filho (clonado) de o animal , mesmo que esteja velho . É **[a reprodução de [a qualidade [34]] . " [33]]** . Segundo **[Campanhola [35]]** , **[a técnica [40]]** pode ser aplicada imediatamente em a produção de carne e de leite . " **[A técnica [40]]** existe , pode ser utilizada e já foi testada . Agora é uma questão de aplicar e de divulgar melhor esse conhecimento . " . **[Vitoriosa [73]]** é o resultado de um experimento realizado por **[a Embrapa [46]]** (**[Empresa_Brasileira_de_Pesquisa_Agropecuária [46]]**) . Ela surgiu a_partir_de células isoladas de um pedaço de pele retirado de a orelha de **[a vaca [82]] [Vitória [82]]** , que foi **[o primeiro clone bovino de a América_Latina , nascida [78]]** em 2001 . " **[O clone de o clone [78]]** coloca o Brasil em a vanguarda científica de esse assunto , como já está em **[o seqüenciamento [63]]** (**[soletração [63]]**) de genoma " , afirmou **[Rodrigues [66]]** . em esse experimento , foram produzidos 35 embriões em seguida transferidos para 17 receptoras , as chamadas mães de aluguel . **[Vitoriosa , que [73]]** tem 15 dias , é a terceira tentativa de o órgão de criar **[um clone [78]]** a_partir_de outro . em o ano passado , duas cópias de **[Vitória [82]]** morreram , uma em o oitavo mês de gestação e outra 36 horas depois de **[o nascimento [73]]** .

Figura 1: Texto 1.

Cadeias Extraídas:

22. [a cria], [animais];
33. [a reprodução], [a reprodução da qualidade];
34. [elevado valor genético], [boa qualidade genética], [a qualidade];
35. [Clayton Campanhola , diretor-presidente da Embrapa], [Campanhola];
40. [a técnica], [A técnica];
46. [a Embrapa], [a Embrapa],[Empresa Brasileira de Pesquisa Agropecuária];
66. [Roberto Rodrigues], [Agricultura], [Rodrigues], [Rodrigues]
73. [o nascimento da bezerra Vitoriosa], [Vitoriosa], [Vitoriosa , que] , [o nascimento];
78. [O animal], [um clone], [o primeiro clone bovino da América Latina , nascida], [O clone do clone], [um clone];
82. [vaca], [Vitória], [a vaca], [Vitória], [Vitória];

Análise:
Na cadeia 22, podemos notar que o modelo agrupou incorretamente "a cria" e "animais". Note que "a cria" refere-se aos sintagmas "bezerra Vitoriosa, o animal e o clone". No entanto, como utilizamos o lema dos núcleos para as consultas semânticas, para a menção "animais", buscou-se por uma relação entre os sintagmas: "a cria" e "animal", a qual retornou uma relação de Hiponímia, que remete para o sintagma "animais". podemos notar o agrupamento de menções incorreto. Na primeira, trata-se da reprodução de animais ameaçados de extinção; a segunda, remete à reprodução da qualidade genética do animal gerado a partir da técnica.

Em 66, podemos ver que o sintagma "Agricultura" foi unido à cadeia "[Roberto Rodrigues], [Rodrigues], [Rodrigues]". Isso ocorre pelo fato do sintagma "Agricultura" estar entre parênteses após o nome "Roberto Rodrigues". Em 73 podemos notar a união de duas cadeias: " [Vitoriosa], [Vitoriosa , que]" e "[o nascimento da bezerra Vitoriosa], [o nascimento]". Este agrupamento incorreto deu-se por meio do casamento parcial entre os sintagmas "o nascimento da bezerra Vitoriosa" e "Vitoriosa".

Podemos notar, também, que a cadeia 78'ficou separada do sintagma "Vitoriosa". Isso porque dentro das regras implementadas não foi

possível criar um link entre as menções "Vitoriosa" e " O animal". Além disso, podemos notar que a última menção do sintagma [um clone] (...a terceira tentativa de criar um clone...) não faz referência a [o primeiro clone bovino da América_Latina], haja vista que o artigo indefinido gera uma expressão genérica, em que se pode fazer referência a qualquer clone no mundo real.

Texto 2

Após o anúncio de **[o sequenciamento [26]]** de **[o genoma [18]]** , em a semana passada , **[a França [34]]** resiste como **[único país de [a União_Européia [72]] a [34]]** não permitir **[patenteamento de genes [22]] [26]]** . **[A UE [72]]** adota , desde junho de 1998 , **[diretiva favorável [39]]** a **[o patenteamento [22]]** de **[genes [26]]** . O texto , redigido por o Parlamento_Europeu , Comissão_Européia_e_Conselho_de_Ministros , utiliza **[o princípio de que [39]]** " **[o genoma [18]]** não é patenteável , mas **[a sequência de [um gene [52]] [26]]** pode ser " . em o entanto , há restrições . **[O patenteamento [22]]** só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o funcionamento de **[o gene [26]]** é detalhado . **[A França [34]]** é **[o único país [34]]** que se recusa a aceitar **[a determinação européia [39]]** . **[A ministra de [a Justiça [64]] de [o país [34]] , [50]] [Elisabeth_Guigou [50]]** , disse que **[a norma [39]]** é incompatível com as leis francesas de bioética . em **[o início [39]]** de o mês , **[o CCNE ([69]] [Comitê_Consultivo_Nacional_de_Ética [69]]**) , órgão que orienta o governo francês sobre aspectos éticos de a biotecnologia , reforçou a posição de **[a ministra [50]]** , alegando que " o conhecimento de **[a sequência [52]]** de **[um gene [26]]** não pode ser assimilado como produto patenteado e , portanto , não é patenteável " . " Bem comum de a humanidade , (**[o sequenciamento de genes) [26]]** não pode ser limitado por patentes que pretendem , em nome de **[o direito [64]]** de propriedade industrial , proteger a exclusividade de esse conhecimento " , diz parecer de **[o CCNE [69]]** . O assunto deve ser debatido durante a presidência francesa de **[a UE [72]]** , em o segundo semestre .

Figura 2: Texto 2.

Cadeias Extraídas:

18. [o genoma], [o genoma];
22. [patenteamento de genes], [o patenteamento], [O patenteamento];
26. [o sequenciamento], [genes], [genes], [um gene], [um gene], [o gene], [um gene], [o sequenciamento de genes)];
34. [a França], [único país da União Européia a], [A França], [o único país], [o país];
39. [diretiva favorável], [o princípio de que], [a determinação européia], [a norma], [o início];
50. [A ministra da Justiça do país], [Elisabeth Guigou], [a ministra];
52. [a sequência de um gene], [a sequência];
64. [a Justiça], [o direito];
69. [o CCNE (], [Comitê Consultivo Nacional de Ética], [o CCNE];
72. [a União Européia], [A UE], [a UE];

Análise:

Analisando cadeias do texto 2, podemos notar que alguns dos erros encontrados foram decorrentes das regras semânticas Hiponímia e Sinonímia: na cadeia 39 alguns dos termos agrupados pelo sistema não são correferentes ('início' e 'diretiva') mas apresentam relações semânticas no Onto.PT ('início' SinonimoDe 'princípio' e 'diretiva' HipônimoDe 'norma'). Um problema semelhante ocorre na cadeia 64, dado que os termos 'justiça' e 'direito' apresentam relação de sinonímia, mas referem-se a menções distintas.

Texto 3

[A pista principal de [o Aeroporto_Internacional_de_São_Paulo [1]] ([40]] [Cumbica [1]]) , em Guarulhos , será totalmente reformada em março de 2008 , segundo **[informações [24]]** de o Ministério_da_Defesa anunciadas em esta segunda-feira , 6 . Com isso , **[a reforma emergencial [42]]** , que começaria em_breve , foi descartada . O ministro de a Defesa , Nelson_Jobim , anunciou **[a reforma [42]]** que , segundo estudos de **[a Empresa_Brasileira_de_Infra-Estrutura_Aeroportuária [16]]** (**[Infraero [16]]**) , **[a reforma [42]]** poderá ser feita sem que **[a pista [40]]** seja interditada . Apesar_da a definição , o cronograma de a obra não foi divulgado . De_acordo com **[informações [24]]** de a Defesa , a primeira etapa de **[a reforma [42]]** será feita com a reforma de um terço de **[a pista [40]]** , em uma de as cabeceiras . Com isso , as outras duas partes ficam disponíveis para pousos e decolagens . em **[a segunda parte [43]]** , a outra cabeceira será reformada e , em a terceira etapa , o centro de **[a pista [40]]** será reformado . em **[a terceira parte [43]]** de **[a reforma [42]]** , **[parte [43]]** de os voos de Cumbica serão transferidos para o Aeroporto_de_Viracopos , em Campinas .

Figura 3: Texto 3.

Cadeias Extraídas:

1. [o Aeroporto Internacional de São Paulo], [Cumbica];
16. [a Empresa Brasileira de Infra-Estrutura Aeroportuária], [Infraero];
24. [informações], [informações];
40. [A pista principal do Aeroporto Internacional de São Paulo], [a pista], [a pista], [a pista];
42. [a reforma emergencial], [a reforma], [a reforma], [a reforma], [a reforma];
43. [a segunda parte], [a terceira parte], [parte];

Análise:

Na cadeia 43 podemos notar que o modelo agrupou os sintagmas [a segunda parte], [a terceira parte] e [parte]. Note que a regra Palavra Modificadora serve justamente para evitar este tipo de agrupamento. No entanto, os sintagmas "[terceira parte]" e "[segunda parte]", foram ligados

por meio do sintagma "[parte]". Note que os sintagmas "[a segunda parte] e [a terceira parte]" remetem às etapas da reforma na pista do aeroporto. Embora o sintagma "[parte]" remete ao sintagma "[parte dos voos de Cumbica]", isso não foi identificado no pré-processamento.

9 Conclusão

Neste artigo, foi proposto um modelo baseado em regras linguísticas para a resolução de correferências em Português que emprega conhecimento semântico. Avaliamos os impactos de cada regra de forma individual e cumulativa. Mostramos também que modelos baseados em regras podem ser uma boa alternativa, quando há carência de corpora ricos em anotação, necessários para treinar modelos eficientes. Notamos que nossas regras semânticas obtiveram um impacto positivo na abrangência, com pequena queda na precisão. Contudo, mesmo com uma medida-F final um pouco menor, consideramos que o aumento significativo na abrangência é importante para esse tipo de tarefa. Em outras palavras, por meio da aplicação de regras semânticas foi possível identificar relações que vão além da análise de similaridade lexical e de justaposição, como no caso da relação entre o par [as abelhas], [os insetos].

Como trabalho futuro, pretendemos buscar novas alternativas semânticas e estudar novas cláusulas restritivas, de forma a fazer com que nossas regras consigam atingir uma precisão mais elevada sem abrir mão da abrangência. Outro objetivo futuro será testar nosso modelo utilizando outros corpora, como o de Garcia & Gamallo (2014b), de forma a efetuar uma comparação entre diferentes modelos.

Como resultado deste trabalho desenvolvemos e disponibilizamos o CORP, uma ferramenta para a resolução de correferências em língua portuguesa que pode auxiliar em diversas tarefas de PLN.

Agradecimentos

Os autores agradecem o suporte financeiro do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

Referências

do Amaral, Daniela Oliveira Ferreira. 2013. *O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa*: Pontifícia Universidade Católica do Rio Grande do Sul. Tese de Mestrado.

Antonitsch, André, Anny Figueira, Daniela Amaral, Evandro Fonseca, Renata Vieira & Sandra Collovini. 2016. Summ-it++: an enriched version of the Summ-it corpus. Em *10th edition of the Language Resources and Evaluation Conference (LREC)*, 2047–2051.

Bagga, Amit & Breck Baldwin. 1998. Algorithms for scoring coreference chains. Em *1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, 563–566.

Baker, Collin F., Charles J. Fillmore & John B. Lowe. 1998. The Berkeley framenet project. Em *17th International Conference on Computational Linguistics*, 86–90.

Basso, Renato Miguel. 2009. *A semântica das relações anafóricas entre eventos*: Universidade Estadual de Campinas, SP. Tese de Doutoramento.

Bechara, Evanildo. 1972. *Lições de português, pela análise sintática*. Editora Fundo de Cultura.

Bick, Eckhard. 2000. *The parsing system PALAVRAS: Automatic grammatical analysis of Portuguese in a constraint grammar framework*: Aarhus University Press. Tese de Doutoramento.

Bick, Eckhard. 2010. A dependency-based approach to anaphora annotation. Em *9th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, publicado online.

Cadore, Luiz Agostinho & Paulo Flávio Ledur. 2013. *Análise sintática aplicada: fundamentos de concordância, regência, crase, colocação, pontuação e significado*. Editora AGE 4th edn.

Cardoso, Nuno. 2012. Rembrandt: a named-entity recognition framework. Em *Eighth International Conference on Language Resources and Evaluation (LREC)*, 1240–1243.

Collovini, Sandra, Thiago I. Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino & Renata Vieira. 2007. Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. Em *V*

Workshop em Tecnologia da Informação e da Linguagem Humana, 1605–1614.

Collovini, Sandra, Lucas Pugens, Aline A. Vanin & Renata Vieira. 2014. Extraction of relation descriptors for Portuguese using conditional random fields. Em *14th Ibero-American Conference on Advances in Artificial Intelligence*, 108–119.

Coreixas, Tatiane. 2010. *Resolução de correferência e categorias de entidades nomeadas*: Pontifícia Universidade Católica do Rio Grande do Sul. Tese de Mestrado.

Durrett, Greg & Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics* 2. 477–490.

Ferradeira, José Eduardo de Sousa. 1993. *Resolução de anáfora pronominal*: Universidade Nova de Lisboa. Tese de Mestrado.

Fonseca, Evandro, Renata Vieira & Aline Vanin. 2014. Coreference resolution in Portuguese: Detecting person, location and organization. *Learning and NonLinear Models* 12(2). 86–97.

Fonseca, Evandro, Renata Vieira & Aline Vanin. 2016a. Adapting an entity centric model for Portuguese coreference resolution. Em *10th Annual Conference on Language Resources and Evaluation (LREC)*, 150–154.

Fonseca, Evandro, Renata Vieira & Aline Vanin. 2016b. Improving coreference resolution with semantic knowledge. Em *12th International Conference on the Computational Processing of Portuguese (PROPOR)*, 213–224.

Fonseca, Evandro Brasil. 2014. *Resolução de correferências em língua portuguesa: pessoa, local e organização*: Pontifícia Universidade Católica do Rio Grande do Sul. Tese de Mestrado.

Freitas, Cláudia, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira & Paula Carvalho. 2010. Second HAREM: advancing the state of the art of named entity recognition in Portuguese. Em *International Conference on Language Resources and Evaluation (LREC)*, 3630–3637.

Freitas, Cláudia, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira & Paula Carvalho. 2009. Relation detection between named entities: report of a shared task. Em *Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 129–137.

Garcia, Marcos & Pablo Gamallo. 2014a. An entity-centric coreference resolution system for person entities with rich linguistic information. Em *25th International Conference on Computational Linguistics*, 741–752.

Garcia, Marcos & Pablo Gamallo. 2014b. Multilingual corpora with coreferential annotation of person entities. Em *9th edition of the Language Resources and Evaluation Conference (LREC)*, 3229–3233.

Gonçalo Oliveira, Hugo. 2012. *Onto.PT: Towards the automatic construction of a lexical ontology for Portuguese*: Universidade de Coimbra. Tese de Doutoramento.

Gonçalo Oliveira, Hugo, Valeria de Paiva, Cláudia Freitas, Alexandre Rademaker, Livy Real & Alberto Simões. 2015. As wordnets do Português. *Oslo Studies in Language* 7(1). 397–424.

Haghighi, Aria & Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1152–1161.

Hou, Yufang, Katja Markert & Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. Em *Conference on Empirical Methods in Natural Language Processing (EMNLPL)*, 2082–2093.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu & Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics* 39(4). 885–916.

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 25–32.

Maia, Luiz Cláudio Gomes. 2008. *Uso de sintagmas nominais na classificação automática de documentos eletrônicos*: Universidade Federal de Minas Gerais. Tese de Doutoramento.

Maziero, Erick, Maria Lucía Jorge & Thiago Pardo. 2010. Identifying multidocument relations. Em *7th International Workshop on Natural Language Processing and Cognitive Science*, 60–69.

Maziero, Erick G., Thiago Pardo, Ariani Di Felippo & Bento C. Dias-da Silva. 2008. A base de dados lexical e a interface web do TeP 2.0:

thesaurus eletrônico para o Português do Brasil. Em *XIV Brazilian Symposium on Multimedia and the Web*, 390–392.

Miller, George A. 1995. WordNet: a lexical database for english. *Communications of the ACM* 38(11). 39–41.

Poesio, Massimo, Roland Stuckardt & Yannick Versley. 2016. *Anaphora resolution: Algorithms, resources, and applications*. Springer.

Ponzetto, Simone Paolo & Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. Em *Human Language Technology Conference*, 192–199.

Pradhan, Sameer, Xiaoqiang Luo, Marta Recasens, Eduard H. Hovy, Vincent Ng & Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. Em *52nd Annual Meeting of the Association for Computational Linguistics*, 30–35.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina & Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. Em *Joint Conference on Empirical Methods in Natural Language Processing and Conference on Natural Language Learning - Shared Task*, 1–40.

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel & Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes. Em *Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 1–27.

Rahman, Altaf & Vincent Ng. 2011. Coreference resolution with world knowledge. Em *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 814–824.

Recasens, Marta & Eduard H. Hovy. 2011. BLANC: implementing the rand index for coreference evaluation. *Natural Language Engineering* 17(4). 485–510.

Recasens, Marta, Lluís Màrquez, Emili Sapena, M Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio & Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. Em *5th International Workshop on Semantic Evaluation*, 1–8.

Rocha, Marco. 2000. A corpus-based study of anaphora in English and Portuguese. Em S. Botley & A. M. Mcenery (eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*, 81–94. John Benjamins Publishing Company.

Salomão, Maria Margarida Martins. 2009. FrameNet Brasil: um trabalho em progresso. *Calidoscópio* 7(3). 171–182.

Sarmento, Luís, Ana Sofia Pinto & Luís Cabral. 2006. REPENTINO - a wide-scope gazetteer for entity recognition in Portuguese. Em *7th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 31–40.

Silva, Jefferson Fontinele da. 2011. *Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado*: Universidade de São Paulo. Tese de Mestrado.

Silva, William Daniel Colen. 2013. *Aprimorando o corretor gramatical CoGrOO*: Universidade de São Paulo. Tese de Mestrado.

Soon, Wee Meng, Hwee Tou Ng & Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* 27(4). 521–544.

Suchanek, Fabian M., Gjergji Kasneci & Gerhard Weikum. 2007. Yago: a core of semantic knowledge. Em *16th International Conference on World Wide Web*, 697–706.

Vieira, Renata, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang & Gabriel Othero. 2005. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. Em A. Branco, T. McEnery & R. Mitkov (eds.), *Anaphora Processing: linguistic, cognitive and computational modeling*, 385–403. John Benjamins Publishing Company.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. Em *6th Conference on Message understanding*, 45–52.

Proposta recebida em Março 2017 e aceite para publicação em Junho 2017.

LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação

LinguaKit: a multilingual tool for linguistic analysis and information extraction

Pablo Gamallo
Centro Singular de Investigação de Tecnologias da Informação (CiTIUS)
Universidade de Santiago de Compostela
pablo.gamallo@usc.es

Marcos Garcia
Grupo LyS, Departamento de Letras
Faculdade de Filologia, Universidade da Corunha
marcos.garcia.gonzalez@udc.gal

Resumo

Este artigo apresenta LinguaKit, uma *suite* multilingue de ferramentas de análise, extração, anotação e correção linguísticas. LinguaKit permite realizar tarefas tão diversas como a lematização, a etiquetagem morfossintática ou a análise sintática (entre outras), incluindo também aplicações para a análise de sentimentos (ou minaria de opiniões), a extração de termos multipalavra, ou a anotação concetual e ligação a recursos enciclopédicos tais como a DBpedia. A maior parte dos módulos funcionam para quatro variedades linguísticas: português, espanhol, inglês e galego. A linguagem de programação de LinguaKit é Perl, e o código está disponível sob a licença livre GPLv3.

Palavras chave

extração de informação, tecnologia linguística

Abstract

This paper presents LinguaKit, a multilingual *suite* of tools for analysis, extraction, annotation and linguistic correction. LinguaKit allows the user to perform different tasks such as lemmatization, PoS-tagging or syntactic parsing (among others), including applications for sentiment analysis (or opinion mining), extraction of multiword expressions or conceptual annotation and entity linking to DBpedia. Most part of the developed modules work in four linguistic varieties: Portuguese, Spanish, English, and Galician. The system is programmed in Perl, and it is freely available under a GPLv3 license.

Keywords

information extraction, linguistic technology

1 Introdução

Neste artigo apresentamos LinguaKit, um pacote de ferramentas multilingues para o Processamento da Linguagem Natural (PLN), que contém módulos de análise, extração, anotação e correção linguística. Os diferentes módulos que compõem LinguaKit são interdependentes entre si, e estão organizados mediante uma arquitectura de *pipeline*. Permite realizar um vasto conjunto de tarefas de PLN, entre as quais: (i) identificação de orações e tokenização, (ii) lematização, (iii) etiquetagem morfossintática, (iv) identificação e (v) reconhecimento de entidades mencionadas, (vi) análise sintática de dependências, (vii) resolução de correferência a nível de entidade, (viii) extração de termos e (ix) de relações semânticas, (x) análise de sentimentos (minaria de opiniões), (xi) anotação conceitual com ligação a recursos enciclopédicos, (xii) correção e avaliação de léxico e sintaxe, (xiii) conjugação verbal automática, (xiv) resumo automático (sumarização), (xv) identificação de língua, ou (xvi) visualização de concordâncias (palavras chave em contexto).

As ferramentas foram desenhadas e desenvolvidas utilizando diferentes estratégias de PLN, tanto de base simbólica como estatística, com aprendizagem supervisionada, não supervisionada e semi-supervisionada. A maior parte dos módulos de LinguaKit funcionam em português, galego,[1] espanhol e inglês.[2]

[1] Neste trabalho consideramos *português* a variedade escrita utilizando as diferentes ortografias da Academia Brasileira de Letras e da Academia das Ciências de Lisboa, e *galego* a que segue (com maior ou menor fidelidade) as normas publicadas em Real Academia Galega e Instituto da Lingua Galega (2004).

[2] Exceto o sistema de correção e avaliação linguística —

DOI: 10.21814/lm.9.1.243

LinguaKit foi programado em Perl. Está disponível como um serviço web[3] e é acessível via RESTful API.[4] O código fonte está publicado sob uma licença GPL.[5]

A tabela 1 mostra os módulos da *suite* organizados em quatro categorias: análise básica, análise profunda, sistemas de extração, e aplicações linguísticas.

Uma das principais contribuições desta nova suite em código aberto é a criação de um ecossistema de ferramentas com diferentes níveis de complexidade. No primeiro nível, situam-se os módulos básicos de análise, que são utilizados para construir aqueles com uma complexidade maior, nomeadamente módulos de análise profunda e de extração. E estes, por sua vez, servem para desenvolver aplicações cada vez mais complexas, como a ferramenta de correção/avaliação linguística ou o anotador semântico.

O objetivo do presente artigo é descrever a arquitetura de LinguaKit, mencionando as metodologias utilizadas na implementação de cada módulo, e apresentar aquelas ferramentas que ainda não tinham sido tratadas em trabalhos precedentes.

Para além desta introdução, o artigo está organizado da seguinte maneira. Na secção 2 incluímos uma breve revisão do trabalho relacionado, e a secção 3 mostra a arquitetura do sistema. A seguir, apresentamos diferentes avaliações —já publicadas— dos diferentes módulos (secção 4), uma descrição pormenorizada dos extratores de termos (secção 5), e as conclusões do presente trabalho (secção 6).

2 Trabalho relacionado

Dado que existem numerosas ferramentas de PLN para diversas línguas e em várias linguagens de programação, nesta secção apresentamos sucintamente algumas das mais conhecidas e utilizadas *suites* de PLN em código aberto, tendo em conta também as línguas que cada uma delas suporta.

O *software* de PLN mais conhecido é provavelmente Stanford CoreNLP (Manning et al., 2014), que inclui módulos de análise tais como tokenizadores, etiquetadores morfossintáticos, reconhecedores de entidades, analisadores sintáticos, sistemas para a resolução da correferência, etc. Está escrito em Java e foi desenvolvido principalmente para o inglês, embora recentemente se tenham publicado modelos para diversas línguas como o chinês, o espanhol ou o árabe, entre outras.

FreeLing (Padró, 2011) é uma outra *suite* de PLN (escrita em C++) que inclui uma lista semelhante à de Stanford CoreNLP, mas dispõe de ferramentas para outras tarefas como a transcrição fonética ou a desambiguação semântica. A maior parte dos módulos analisa os textos em catalão, espanhol, português, galego, inglês, francês, e recentemente, alemão ou russo (entre outras línguas).

Um outro sistema de PLN escrito em Java é OpenNLP,[6] que realiza tarefas de análise similares aos que já foram referidos, mas que inclui, por exemplo, um módulo de categorização de documentos. Existem modelos disponíveis para várias línguas, nomeadamente inglês, espanhol e alemão.

Também programada em Java, IXA pipes (Agerri et al., 2014) é uma *suite* modular que realiza as tarefas mais habituais de processamento linguístico: tokenização, etiquetagem morfossintática, reconhecimento de entidades e análise sintática. Este sistema permite processar as seguintes línguas (com variações em função do módulo escolhido): espanhol, inglês, eusquera, italiano e galego.

Com a popularização da iniciativa *Universal Dependencies*,[7] que promove a unificação das diretrizes de anotação em diversas línguas, têm vindo a ser desenvolvidas algumas ferramentas compatíveis, como UDPipe (Straka et al., 2016). UDPipe inclui módulos de aprendizagem automática para tokenização, etiquetagem morfossintática, lematização e análise sintática.

Como foi referido, existem mais sistemas que realizam tarefas de PLN —alguns com objetivos ligeiramente diferentes, ou escritos noutras linguagens de programação—, tais como NLTK: *Natural Language Toolkit* (Bird et al., 2009), amplamente utilizado no ensino de PLN, ou spaCy[8] (mais focado em uso industrial), ambos escritos em *python*.

Para além dos diferentes *softwares* apresentados, cabe mencionar também CitiusTools (Garcia & Gamallo, 2015), *suite* de PLN a partir da qual foram desenvolvidos alguns dos módulos de LinguaKit. À diferença dos sistemas mencionados, que oferecem fundamentalmente módulos de análise, LinguaKit possui também um amplo le-

desenvolvido principalmente para a análise do galego—, e o conjugador verbal — que não funciona para o inglês.

[3] https://www.linguakit.com

[4] https://market.mashape.com/linguakit/linguakit-natural-language-processing-in-the-cloud

[5] https://github.com/citiususc/Linguakit

[6] http://opennlp.apache.org/

[7] http://universaldependencies.org/

[8] https://spacy.io/

tipo de módulo	módulos
análise básica	conjugador verbal
	segmentador de orações
	tokenizador e *splitter*
análise profunda	lematizador
	PoS-tagger
	identificador de entidades (NER)
	classificador de entidades (NEC)
	identificador de correferência
	analisador sintático em dependências
extração	palavras chave
	expressões multipalavra
	análise de sentimento/opinião
	relações semânticas (open IE)
aplicações	sumarização
	anotação semântica (com EL)
	concordâncias (palavras chave em contexto)
	identificação de línguas
	correção/avaliação linguística (léxica e gramatical)

Tabela 1: Módulos de LinguaKit organizados em quatro categorias.

que de ferramentas de extração, bem como de aplicações mais complexas baseadas nesses sistemas de extração.

3 Arquitetura

A figura 1 mostra as dependências entre os diferentes módulos apresentados na tabela 1, sendo esta arquitetura comum às quatro línguas processadas pelo sistema.

A análise básica consiste na segmentação de um texto em orações, que são a entrada do processo de tokenização. Por sua vez, o texto tokenizado é melhorado com regras básicas de *splitting*, que separam os elementos que compõem contrações (e.g., "*do* → *de o*", em português e galego) ou sequências de verbo e pronome clítico (e.g., "*comelo* → *comer o*", em galego). Este último módulo é dependente da língua, enquanto os processos anteriores são realizados com uma ferramenta única (utilizando listas de abreviaturas também dependentes de cada variedade linguística).

O conjugador verbal é um módulo isolado que toma como entrada um verbo em infinitivo tanto em espanhol como em galego e português. Neste último caso, o sistema pode realizar até quatro modelos de conjugação verbal, em função quer da variedade (português de Portugal ou do Brasil), quer do sistema ortográfico utilizado (antes ou depois do Acordo Ortográfico de 1990).[9]

[9] https://pt.wikipedia.org/wiki/Acordo_Ortografico_de_1990

Com base nos módulos de análise básica, foram implementadas duas aplicações diferentes: um identificador de língua e um gerador de concordâncias (palavras chave em contexto). O identificador de língua é também utilizado internamente pelo sistema para fazer a escolha automática dos módulos de uma ou outra língua, permitindo que o utilizador possa analisar um texto sem ter de selecionar a língua desejada.

Os módulos de análise profunda tomam como entrada a saída da análise básica. O primeiro processo é a lematização, que atribui todos os lemas e todas as etiquetas possíveis a cada forma (já tokenizada) do texto de entrada. O lematizador baseia-se num léxico computacional disponível para cada língua. Antes do processo de desambiguação realizado pelo etiquetador morfossintático (*PoS-tagger*, na tabela 1), é possível identificar as entidades mencionadas ou nomes próprios (NER). As entidades identificadas pelo NER serão classificadas após a etiquetagem morfossintática mediante um sistema de classificação semântica: o classificador de entidades mencionadas (NEC). O último módulo de análise é o *parsing* sintático em dependências, que toma como entrada o etiquetador morfossintático (com ou sem aplicação dos módulos de NER e NEC).

Várias ferramentas utilizam a saída dos módulos de análise profunda para extrair informação dos textos: extratores de opiniões (também conhecidos como analisadores de sentimento), de palavras chave, de expressões multipalavra, e de relações semânticas. Todos estes extratores tomam como entrada a saída do módulo

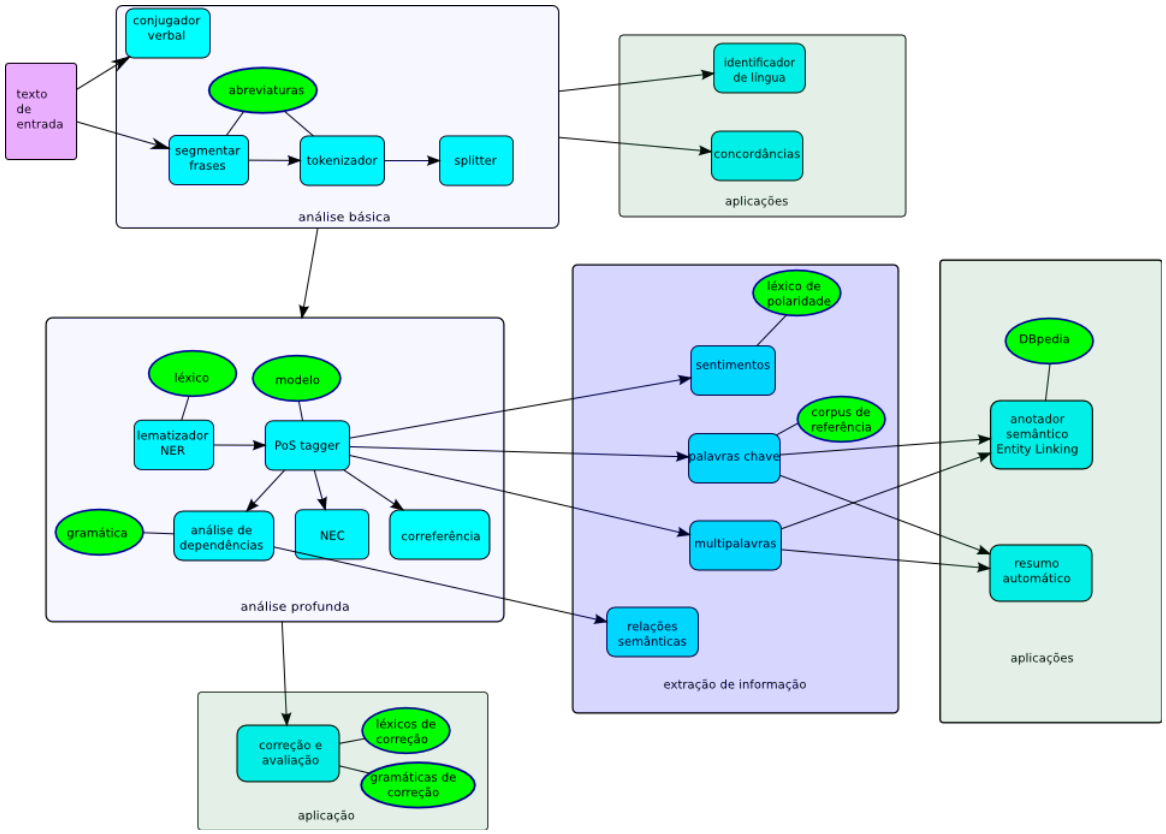

Figura 1: Arquitetura de LinguaKit.

de etiquetagem morfossintática. Para além disso, foi desenvolvida uma aplicação de correção lexical e gramatical que utiliza a saída do analisador sintático.

Finalmente, duas aplicações foram criadas a partir dos extratores de termos relevantes (isto é, palavras chave e expressões multipalavra): um gerador automático de resumos e um anotador semântico, que liga os termos extraídos a conceitos enciclopédicos armazenados em bases de conhecimento externas (por exemplo, a DBpedia).[10]

4 Módulos

Os principais módulos de LinguaKit foram desenhados e implementados nos últimos cinco anos, sendo a maior parte deles descritos em diferentes publicações. Assim, esta secção tem como objetivo pôr em conjunto as técnicas e metodologias empregadas em cada um dos principais módulos, bem como um breve resumo das avaliações realizadas.

[10] http://wiki.dbpedia.org/

Pré-processamento

Como foi referido, os primeiros módulos realizam um pré-processamento do texto que permite aplicar com maior precisão as ferramentas subsequentes: estes módulos realizam identificação de fronteiras de oração (com base em máquinas de estados finitas e em listas de abreviaturas que terminam com pontuação), de tokenização e *splitting* (processos pelos quais são separados os diferentes tokens de cada oração), e de lematização (que atribui um —ou mais— lemas possíveis a cada um dos tokens). Descrições mais pormenorizadas destes módulos podem encontrar-se em (Garcia & Gamallo, 2010) ou em (Garcia & Gamallo, 2015).

Etiquetagem morfossintática

Este módulo desambigua as etiquetas morfossintáticas[11] previamente atribuídos a cada token mediante um classificador *bayesiano* baseado em bigramas de tokens. Foi avaliado para três

[11] E também alguns lemas cuja atribuição varia em função da categoria morfossintática à que pertença o token. Por exemplo, as formas galegas/portuguesas *cala* ou *calas* podem ter como lema *calar* —se forem verbos—, ou *cala* —se forem nomes.

línguas: inglês, português e espanhol, com resultados próximos ao estado da arte: ≈ 96 para português e espanhol, e ligeiramente mais baixos (≈ 94%) para inglês (Gamallo et al., 2015b; Garcia & Gamallo, 2015).

Identificação e classificação de entidades mencionadas

O primeiro destes módulos identifica expressões *numex* (de base numérica) e *enamex* (nomes próprios) mediante máquinas de estados finitas, que têm em conta tanto as formas ortográficas (uso de maiúsculas) como palavras funcionais que possam conter (***U**niversidade **de** **S**antiago **de** **C**ompostela*). Uma vez identificadas as entidades, o módulo de classificação aplica um método de supervisão distante que lhe permite classificar as entidades em quatro classes: *pessoa*, *organização*, *local* ou *miscelânea*. O sistema emprega listas de entidades já conhecidas (*gazetteers*) e um conjunto de regras que permitem desambiguar as entidades que aparecem em mais de uma lista (que podem ser, por exemplo, *pessoa* ou *local*). Os *gazetteers* foram extraídos automaticamente de fontes externas com conhecimento enciclopédico.

Este módulo foi avaliado para as quatro línguas analisadas (inglês, português, espanhol e galego), utilizando diversos corpora e sendo comparando com sistemas supervisionados (Gamallo & Garcia, 2011; Garcia et al., 2012; Garcia & Gamallo, 2015). Os resultados obtidos —apesar de que não são sempre diretamente comparáveis— foram próximos aos atingidos por FreeLing e Stanford CoreNLP, superando nitidamente os modelos disponibilizados para OpenNLP.

Resolução de correferência a nível de entidade

Um outro módulo de análise linguística incluído em LinguaKit é o de resolução de correferência a nível de entidade. Este módulo utiliza como entrada um texto com as entidades mencionadas classificadas semanticamente, e aplica uma estratégia determinística baseada em filtros mediante os quais atribui um identificador numérico a cada uma das ocorrências (menções) das entidades previamente analisadas. Idealmente, este identificador será igual para cada uma das menções que refiram a mesma entidade do discurso (e.g., "António_Variações$_{\text{PESSOA_1}}$", "John$_{\text{PESSOA_2}}$", "John_Lennon$_{\text{PESSOA_2}}$", "António$_{\text{PESSOA_1}}$", "Lennon$_{\text{PESSOA_2}}$", ...). Este módulo é uma versão simplificada do apresentado em (Garcia & Gamallo, 2014).

Para além disso, este sistema inclui uma saída alternativa que aproveita a resolução de correferência para tentar corrigir erros prévios da classificação semântica. Assim, se a citada forma "Lennon" tivesse sido anteriormente classificada como *local*, mas identificada como menção da mesma entidade que "John_Lennon", a etiqueta semântica da primeira seria corrigida para *pessoa* (Garcia, 2016).

Analisador em dependências

O módulo de análise sintática, chamado DepPattern, baseia-se em regras formais de dependências e num algoritmo de *parsing* com técnicas de estados finitos. Foi avaliado para português e espanhol e comparado com MaltParser (Nivre et al., 2007), um *parser* determinístico de transições baseado em aprendizagem supervisionada. Os resultados obtidos por DepPattern com corpora de teste construído a partir de textos de diferentes domínios foram semelhantes aos obtidos por MaltParser: ≈ 82% de F-score (Gamallo, 2015).

Em Gamallo & González (2011) descrevem-se as características principais da gramática formal na qual se baseia o conhecimento linguístico de DepPattern. Um compilador transforma as regras formais, escritas com os princípios da gramática de dependências, em *scripts* Perl que representam os *parsers* de estados finitos.

Análise de sentimentos

O sistema de análise de sentimentos (tarefa também conhecida como minaria de opiniões) classifica uma oração como tendo uma opinião positiva, negativa ou neutra. O núcleo deste módulo é um classificador *bayesiano* treinado com texto previamente anotado com as opiniões referidas, que também utiliza um léxico de polaridade e regras sintáticas para a identificação de marcadores linguísticos que intensificam ou mudam a polaridade das palavras. Foi avaliado para inglês e espanhol, e participou em duas competições focadas na análise de opiniões em redes sociais: TASS 2013 (Gamallo et al., 2013a) para espanhol, e SemEval-2014 (Gamallo & Garcia, 2014) para inglês, mostrando um desempenho competitivo em ambas as línguas.

Extrator de relações

Este módulo consiste num sistema de extração de informação não supervisionado cujo obje-

tivo é obter um conjunto aberto de relações entre dous objetos. As relações (ou tripletas: *obj1,relação,obj2*) selecionadas por um sistema de extração de informação aberta (*Open Information Extraction*, OIE) representam as proposições básicas do texto de entrada. O nosso sistema, argOE (Gamallo & Garcia, 2015), está baseado em regras e toma como entrada um texto analisado em dependências em formato CoNLL-X. Foi avaliado em inglês, português e espanhol, e comparado com sistemas de OIE focados na extração numa única língua. O módulo incluído em LinguaKit melhora os resultados de muitos dos sistemas com os quais foi comparado, como ReVerb (Etzioni et al., 2011), embora os resultados sejam mais baixos do que um outro sistema baseado em regras, ClausIE (Corro & Gemulla, 2013).

Anotação e ligação semântica

Este módulo identifica os termos relevantes do texto que podem ser ligados a conceitos presentes em bases de dados externas, tais como a DBpedia. Esta tarefa, que consiste em relacionar os termos mencionados no texto e os conceitos de uma base ontológica e enciclopédica, é normalmente conhecido como *ligação de entidades* (*entity linking*, EL). O nosso sistema utiliza como recursos externos algumas relações da DBpedia e uma nova base construída mediante similaridade distribucional a partir das entradas textuais da Wikipedia. Foram avaliadas as versões portuguesa e inglesa (Gamallo & Garcia, 2016), com resultados similares a outros sistemas EL de referência, como *DBpedia Spotlight* (Mendes et al., 2011).

Corretor linguístico

O sistema de correção linguística de LinguaKit está, por enquanto, só disponível como módulo experimental na versão web.[12]

Esta ferramenta foi desenvolvida principalmente para galego, variedade na qual foi avaliada e comparada com revisões manuais de textos por parte de docentes profissionais (Gamallo et al., 2015a). O sistema contém diversos módulos que identificam e classificam diferentes tipos de erros habituais em aprendizes de galego, tanto de tipo léxico (castelhanismos, hipercorreções, etc.), como gramatical (concordância de género e número, posição dos pronomes átonos, etc.).

Existem, contudo, versões básicas para português e espanhol, mas precisam de um maior desenvolvimento no que diz respeito a recursos linguísticos tais como listas de tipologias de erros, ou regras sintáticas para a identificação e classificação de erros.

[12] https://linguakit.com/es/supercorrector

Outras ferramentas

Para além das ferramentas referidas (e das aplicações de extração mostradas na secção 5), LinguaKit também inclui as seguintes aplicações: (i) um gerador automático de resumos (sumarizador), (ii) um visualizador de palavras chave em contexto (concordâncias), e (iii) conjugadores verbais automáticos.

O sumarizador extrai as frases ou orações mais relevantes do texto de entrada. Utiliza a segmentação de orações, a análise morfossintática, e os extratores de palavras e multipalavras para ponderar as orações em graus de relevância. A partir da lista ponderada de orações, o usuário escolhe a percentagem de texto que quer extrair para construir o resumo.

O visualizador de concordâncias, também conhecido como *key word in context*, é uma ferramenta útil para estudos em linguística de corpus que procura no texto selecionado a palavra escolhida pelo utilizador, obtendo o seu contexto anterior e posterior em cada uma das suas ocorrências.

O módulo de conjugação verbal permite obter de modo automático a conjugação completa de um verbo a partir da sua forma em infinitivo. O sistema contém as regras de conjugação verbal do espanhol peninsular, do galego e de quatro normas do português: duas variedades diatópicas: português europeu e brasileiro; e duas variantes ortográficas para cada uma das anteriores: antes e depois do Acordo Ortográfico de 1990. Uma vez que o conjugador funciona aplicando diferentes regras em função do paradigma verbal, este pode gerar as formas conjugadas de verbos desconhecidos, tais como neologismos. Para além disso, identifica se o verbo é conhecido, com base em listas de verbos obtidos de recursos académicos para cada uma das línguas (Gamallo et al., 2013b).

Usabilidade

Para executar qualquer módulo em linha de comandos, disponibilizamos de um *script*, *linguakit*, que requer três argumentos: língua, nome do módulo e ficheiro TXT a ser processado. Por exemplo, o comando que faz a chamada básica do módulo de etiquetagem morfossintática em português é o seguinte:

```
./linguakit pt tagger input.txt
```

Com este comando, o utilizador não precisa de conhecer quais os módulos que dependem da etiquetagem (segmentação, tokenização, etc). De facto, o código executado por *linguakit* é um *pipeline* de *scripts*, cada um deles representando um módulo da suite. No caso da etiquetagem morfossintática para um texto em português, o *pipeline* invocado é o seguinte:

```
cat input.txt
|./tagger/pt/sentences-pt_exe.perl
|./tagger/pt/tokens-pt_exe.perl
|./tagger/pt/splitter-pt_exe.perl
|./tagger/pt/lemmas-pt_exe.perl
|./tagger/pt/tagger-pt_exe.perl
```

Na próxima versão de LinguaKit, os módulos poderão ser invocados também mediante funções Perl.

5 Extratores de termos

Uma vez apresentados os módulos e aplicações que já tinham sido avaliadas em diferentes publicações, nesta secção mostramos duas ferramentas de extração, que têm como objetivo identificar e selecionar os termos chave e relevantes de um texto. Consideram-se termos relevantes aquelas expressões mais importantes de um texto que são utilizadas como índices para —entre outras aplicações— a deteção imediata do tema ou tópico, para o etiquetado textual automático, ou bem para a classificação de documentos. Estes dous módulos de extração diferenciam-se no tipo de termos relevantes que extraem: (i) unidades monolexicais e nomes próprios (*termos básicos*), e (ii) unidades plurilexicais (*termos multipalavra*).

Termos básicos

Chamamos termos básicos àquelas unidades lexicais relevantes para um texto que se codificam como nomes comuns, nomes próprios (simples ou compostos), adjetivos e verbos. Exceto os nomes próprios, que podem ser expressões compostas por várias palavras (por exemplo, "Nova Iorque", "Universidade Nova de Lisboa", etc), os termos básicos são palavras simples monolexicais. O método de extração leva-se a cabo em duas fases: seleção de candidatos e ordenação por relevância.

Na primeira fase, o sistema identifica todos os candidatos a serem termos básicos mediante o etiquetador morfossintático. Deste modo, selecionam-se como candidatos todas as unidades lexicais que foram etiquetadas como nomes (comuns e próprios), adjetivos e verbos.

Na segunda fase, os termos ordenam-se por relevância e escolhem-se os N primeiros, sendo N um valor numérico parametrizável. Para calcular a relevância dos termos básicos recorremos à noção de *termhood*, é dizer, ao grau com que a unidade linguística está relacionada com conceitos específicos do domínio do texto (Kageura & Umino, 1996). Esta noção de *termhood* pode ver-se também como a probabilidade de um termo formar parte do domínio. O *termhood* não é, portanto, uma medida discreta, mas contínua. Em consequência, medimos a relevância de um termo básico (*termhood*) mediante um peso estatístico que é calculado contrastando as frequências dos candidatos no texto de entrada (dados observados) com um corpus de referência (dados esperados). Mais precisamente, o peso de um termo é o valor qui-quadrado que mede a divergência entre os dados observados e os esperados. Estes últimos são os dados obtidos a partir de um corpus de referência com um tamanho médio de 100M de tokens por língua, compilado pelo grupo ProLNat@GE, e que é composto por textos de vários géneros e domínios: jornalístico, técnico, literário, de redes sociais, etc. Finalmente, os termos são organizados em função do seu peso, de maior a menor, e o usuário escolhe os N mais relevantes em função do tamanho do texto e das necessidades de análise.

Termos multipalavra

Os termos multipalavra são expressões relevantes codificadas como unidades plurilexicais que instanciam padrões específicos de etiquetas morfossintáticas. Por exemplo, *língua natural*, *processamento da língua*, *tecnologias da língua* ou *analisador sintático* podem ser unidades multipalavra relevantes dentro de um texto de domínio científico focado em questões de PLN. Como no caso dos termos básicos, o processo de extração de multipalavras divide-se em duas fases: seleção de candidatos e ordenação dos mesmos por relevância. Porém, tanto a seleção de candidatos como a ordenação realizam-se mediante estratégias diferentes às utilizadas para a extração dos termos básicos.

Para a primeira fase utilizamos um conjunto de padrões de etiquetas (tabela 2) para identificar todas aquelas expressões plurilexicais que os instanciam (os artigos e determinantes das expressões não se tomam em conta na instanciação). O conjunto foi desenhado para a identi-

$nome - adj$	$adj - nome$
$nome - nome$	$nome - prep - nome$
$nome - prep - adj - nome$	$nome - prep - nome - adj$
$adj - nome - prep - nome$	$nome - adj - prep - nome$
$adj - nome - prep - nome - adj$	$nome-adj-prep-nome-adj$
$adj - nome - prep - adj - nome$	$nome-adj-prep-adj-nome$

Tabela 2: Conjunto de padrões de etiquetas utilizado para a identificação de candidatos a termos multipalavra (*adj* é adjetivo e *prep* é preposição).

peso	**multipalavra**	**padrão de etiquetas**
9,95	dación en pago	nome-prep-nome
7,94	viviendas vacías	nome-adj
7,27	renta básica	nome-adj
5,24	iniciativas legislativas	nome-adj
2,99	reuniones de representantes	nome-prep-nome

Tabela 3: As cinco multipalavras mais relevantes (*unithood*) extraídas do programa eleitoral do partido político espanhol *Podemos* para as eleições do 20D/2015.

ficação de multipalavras nas quatro línguas tratadas. Este método é semelhante ao descrito noutros trabalhos sobre extração terminológica (Vivaldi & Rodríguez, 2001; Sánchez & Moreno, 2006). Os padrões foram selecionados a partir da revisão manual de uma lista de n-gramas de etiquetas ordenadas por frequência em corpora de diferentes línguas.

Na segunda fase, a ordenação por relevância, utilizamos uma estratégia diferente à empregada na ordenação por termos básicos. Enquanto estes se ordenam em função da noção de *termhood*, a relevância das expressões multipalavra define-se mediante o conceito de *unithood*. Esta noção faz referência à associação das sequências de palavras com unidades lexicais estáveis. Mais concretamente, *unithood* refere-se ao grau de força e coesão entre as unidades lexicais que constituem os sintagmas e colocações (Kageura & Umino, 1996). A *unithood* só se aplica, portanto, a unidades plurilexicais com alguma coesão interna e não a unidades monolexicais.

O grau de coesão, ou *unithood*, pode calcular-se com diferentes medidas de associação lexical. O módulo de LinguaKit permite escolher entre 5 medidas para ordenar os candidatos a multipalavra: (a) qui-quadrado, (b) função de verosimilhança (*loglikehood*), (c) informação mutua (*mi*), (d) probabilidade condicional simétrica (*scp*), e (e) simples co-ocorrência. As medidas de associação aplicam-se para verificar se os constituintes co-ocorrem num sintagma aleatoriamente ou por atração. Assim, os valores observados equivalem à frequência da expressão multipalavra no texto de entrada, e os valores esperados calculam-se a partir das frequências dos constituintes por separado.

É importante sublinhar que estas estratégias básicas de extração são de propósito geral pois não estão adaptadas a um domínio específico. São aplicáveis portanto a qualquer domínio. No entanto, para serem mais eficientes, precisavam de incluir novos sub-módulos que permitissem uma fácil adaptação a domínios de especialidade. Na atualidade, a extração só permite selecionar e identificar candidatos a termo em geral, e não unidades terminológicas de um domínio previamente identificado.

Como exemplo de utilização, as tabelas 3 e 4 mostram as expressões multipalavra mais relevantes (usando qui-quadrado como peso para a ordenação) extraídas de dous programas de partidos políticos, *Podemos* e o *Partido Popular*, para as eleições ao parlamento espanhol de 20 de dezembro de 2015. Assim, este exemplo mostra como o extrator permite identificar as prioridades programáticas dos partidos políticos com uma simples vista de olhos sobre os termos mais relevantes.

Mesmo se a eficiência da extração de termos não foi avaliada quantitativamente, podemos encontrar alguns elementos que demonstram a sua usabilidade desde um ponto de vista qualitativo. Por um lado, os dous extratores de termos (básicos e multipalavra) foram inseridos no módulo mais complexo de anotação e ligação semântica, o qual sim foi avaliado quantitativamente e comparado com outros sistemas de anotação. Por outro lado, estes módulos foram utilizados por utentes muito variados com dife-

peso	multipalavra	padrão de etiquetas
20,37	inversores extranjeros	nome-adj
11,44	creación de empleo	nome-prep-nome
9,75	competitividad de economía	nome-prep-nome
7,73	reducción de impuestos	nome-prep-nome
2,93	ciudadanos españoles	nome-adj

Tabela 4: As cinco multipalavras mais relevantes (*unithood*) extraídas do programa eleitoral do partido político espanhol *Partido Popular* para as eleições do 20D/2015.

rentes aplicações e objetivos, tais como análises dos programas de partidos políticos feitas por jornalistas.[13]

6 Conclusões e trabalho futuro

Este artigo apresentou LinguaKit, um pacote linguístico que permite os utilizadores ter um acesso fácil e unificado a módulos de análise linguística muito diversos.

O conjunto de ferramentas disponível, mesmo se amplo e variado, fica ainda longe de cobrir todos as necessidades dos profissionais e utilizadores da língua. A este respeito, como trabalho futuro pretendemos, por um lado, continuar a melhorar o desempenho de alguns dos módulos de análise, e por outro lado ampliar o número de módulos com sistemas de transcrição fonética e fonológica. Além disso, está prevista a adaptação dos módulos de análise morfossintática e sintática para a sua compatibilidade com as diretrizes de anotação das *dependências universais*.

Para além de novos módulos, o sistema pode enriquecer-se com funcionalidades simples mas úteis para linguistas e investigadores. Por exemplo, um buscador de contextos léxico-sintáticos que utilize o analisador sintático para permitir procurar que nomes funcionam como sujeitos de um verbo específico, adjetivos que modifiquem um dado nome, etc. Em relação às novas funcionalidades, será preciso identificar os principais objetivos dos utilizadores para tentar que o sistema cubra as suas necessidades.

Agradecimentos

Este trabalho foi realizado graças ao financiamento da *Ayuda da Fundación BBVA para Investigadores y Creadores Culturales*, do projeto TELEPARES (MINECO, ref:FFI2014-51978-C2-1-R), da *Consellería de Cultura, Educación e Ordenación Universitaria* (2016-2019, ED431G/08), do *European Regional Development Fund (ERDF)*, e de um contrato *Juan de la Cierva-formación*, com referência FJCI-2014-22853.

[13] http://www.galiciaconfidencial.com/noticia/27170-son-galiza-galicia-marea

Referências

Agerri, Rodrigo, Josu Bermudez & German Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. Em *9th International Conference on Language Resources and Evaluation (LREC)*, 3823–3828.

Bird, Steven, Edward Loper & Ewan Klein. 2009. *Natural language processing with Python*. O'Reilly Media Inc.

Corro, Luciano Del & Rainer Gemulla. 2013. ClausIE: Clause-based open information extraction. Em *The World Wide Web Conference*, 355–366.

Etzioni, Oren, Anthony Fader, Janara Christensen, Stephen Soderland & Mausam. 2011. Open information extraction: the second generation. Em *International Joint Conference on Artificial Intelligence (IJCAI)*, 3–10.

Gamallo, Pablo. 2015. Dependency parsing with compression rules. Em *International Workshop on Parsing Technology (IWPT)*, 107–117.

Gamallo, Pablo & Marcos Garcia. 2011. A resource-based method for named entity extraction and classification. Em *Portuguese Conference on Artificial Intelligence (EPIA 2011)*, 610–623.

Gamallo, Pablo & Marcos Garcia. 2014. Citius: a naive-bayes strategy for sentiment analysis on English tweets. Em *8th International Workshop on Semantic Evaluation (SemEval)*, 171–175.

Gamallo, Pablo & Marcos Garcia. 2015. Multilingual open information extraction. Em *17th Portuguese Conference on Artificial Intelligence (EPIA)*, 711–722.

Gamallo, Pablo & Marcos Garcia. 2016. Entity linking with distributional semantics. Em *International Conference on the Computational Processing of the Portuguese Language (PROPOR)*, 177–188.

Gamallo, Pablo, Marcos Garcia & Santiago Fernández-Lanza. 2013a. TASS: a naive-bayes strategy for sentiment analysis on Spanish tweets. Em *Workshop on Sentiment Analysis (TASS@SEPLN)*, 126–132.

Gamallo, Pablo, Marcos Garcia, Isaac González, Marta Mu noz & Iria del Río. 2013b. Learning verb inflection using Cilenis conjugators. *The Eurocall Review* 21(1). 12–19.

Gamallo, Pablo, Marcos Garcia, Iria del Río & Isaac González López. 2015a. Avalingua: Natural language processing for automatic error detection. Em *Learner Corpora in Language Testing and Assessment*, vol. 70 Studies in Corpus Linguistics, 35–58. John Benjamins Publishing Company.

Gamallo, Pablo & Isaac González. 2011. A grammatical formalism based on patterns of part-of-speech tags. *International Journal of Corpus Linguistics* 16(1). 45–71.

Gamallo, Pablo, Juan Carlos Pichel, Marcos Garcia, José Manuel Abuín & Tomás Fernández-Pena. 2015b. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno big data. *Procesamiento del Lenguaje Natural* 53. 17–24.

Garcia, Marcos. 2016. Incorporating lexico-semantic heuristics into coreference resolution sieves for named entity recognition at document-level. Em *10th edition of the Language Resources and Evaluation Conference (LREC)*, 3357–3361.

Garcia, Marcos & Pablo Gamallo. 2010. Análise morfossintáctica para português europeu e galego: Problemas, soluções e avaliação. *Linguamática* 2(2). 59–67.

Garcia, Marcos & Pablo Gamallo. 2014. An entity-centric coreference resolution system for person entities with rich linguistic information. Em *25th International Conference on Computational Linguistics: Technical Papers (COLING)*, 741–752.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. Em *Symposium on Languages, Applications and Technologies (SLATE)*, 65–75.

Garcia, Marcos, Isaac González & Iria del Río. 2012. Identificação e classificação de entidades mencionadas em Galego. *Estudos de Linguística Galega* 4. 13–25.

Kageura, Kyo & Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology* 3(1). 259–289.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. Em *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Mendes, Pablo N., Max Jakob, Andrés García-Silva & Christian Bizer. 2011. DBpedia spotlight: Shedding light on the web of documents. Em *7th International Conference on Semantic Systems*, 1–8.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(2). 115–135.

Padró, Lluís. 2011. Analizadores multilingües en FreeLing. *Linguamática* 3(2). 13–20.

Real Academia Galega e Instituto da Lingua Galega. 2004. *Normas ortográficas e morfolóxicas do idioma galego*. Editorial Galaxia.

Sánchez, David & Antonio Moreno. 2006. A methodology for knowledge acquisition from the web. *Journal of Knowledge-Based and Intelligent Engineering Systems* 10(6). 453–475.

Straka, Milan, Jan Hajič & Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. Em *10th International Conference on Language Resources and Evaluation (LREC)*, 4290–4297.

Vivaldi, Jordi & Horacio Rodríguez. 2001. Improving term extraction by combining different techniques. *Terminology* 7(1). 31–47.

Projetos, Apresentam-se!

Proposta recebida em Março 2017 e aceite para publicação em Junho 2017.

Geração Automática de Sentenças em Língua Natural para Sequências de Pictogramas como Apoio à Comunicação Alternativa e Ampliada

Automatic generation of natural language sentences for pictogram sequences in support of Augmentative and Alternative Communication

Rafael Pereira
Universidade Federal de Sergipe
rafaelps@dcomp.ufs.br

Hendrik Macedo
Universidade Federal de Sergipe
hendrik@dcomp.ufs.br

Rosana Givigi
Universidade Federal de Sergipe
rosanagivigi@uol.com.br

Marco Túlio Chella
Universidade Federal de Sergipe
marco@dcomp.ufs.br

Resumo

A Comunicação Alternativa e Ampliada (CAA) é uma área de prática clínica educacional para fonoaudiólogos cujo objetivo é auxiliar indivíduos que possuam deficiência na oralidade. Os símbolos de comunicação pictórica constituem um dos sistemas da CAA que podem complementar ou mesmo substituir a linguagem falada desses indivíduos. É possível utilizar a habilidade já adquirida em comunicação pictórica por parte de crianças com deficiência para promover sua alfabetização. Infelizmente, a literatura relacionada parece não indicar solução prática para tal questão. Neste artigo, propomos um método para geração automática de sentenças naturais em língua portuguesa do Brasil que corresponda a uma dada sequência de símbolos pictóricos apresentados. Este método foi implementado em uma ferramenta visual de apoio ao profissional educador e atualmente faz parte de um dos recursos de CAA do Laboratório de CAA da Universidade Federal de Sergipe. Um conjunto de validação fornecido pelo Laboratório mostrou a corretude das sentenças geradas pela ferramenta.

Palavras chave

Geração de Linguagem Natural, Comunicação Alternativa e Ampliada, Símbolos Pictóricos

Abstract

The Augmentative and Alternative Communication (AAC) is an area of clinical educational practice for speech therapists whose goal is to assist individuals who are orally deficient. The pictorial communication symbols are one of the AAC systems that can complement or even replace the spoken language of these individuals. It is possible to use the ability already acquired in pictorial communication by children with disabilities to promote their literacy. Unfortunately, the related literature does not seem to indicate a practical solution to this question. In this paper, we propose a method for automatic generation of natural sentences in the Brazilian Portuguese language in regards to a given sequence of pictorial symbols presented. This method has been implemented in a visual tool to support professional educators and is currently part of one of the AAC tools of the AAC Laboratory at the Federal University of Sergipe, Brazil. A validation set provided by the Laboratory has shown the correctness of the sentences generated by the tool.

Keywords

Natural Language Generation, Augmented Alternative Communication, Pictograph Symbols

1 Introdução

Tecnologia Assistiva é o termo empregado a todo conjunto de dispositivos utilizados para auxiliar indivíduos com algum tipo de limitação intelectual, motora, visual ou auditiva a realizar atividades a que normalmente não estariam completamente aptos (Bharucha et al., 2009; Brodwin, 2010).

Um uso particular das tecnologias assistivas é feito pela chamada Comunicação Alternativa e Ampliada (CAA) (Beukelman & Mirenda, 2005; Alant & Bornman, 1994; Light, 1989). A CAA é uma área de prática clínica de pesquisa e educacional para fonoaudiólogos que visa auxiliar indivíduos que demonstrem prejuízos nos modos de comunicação gestual, oral e/ou escrita.

DOI: 10.21814/lm.9.1.242

Os sistemas de CAA dividem-se em picturais e linguísticos. Dentre os picturais destacam-se o Picture Communication Symbols (PCS), o Pictogram-Ideogram Communication (PIC), o Picsyms, o Rebus e o ARASAAC[1]

O sistema de símbolos de comunicação pictórica pode substituir ou complementar a linguagem falada e, desta forma, contribuir para o aumento da interação comunicativa dos indivíduos com deficiência na oralidade, suprindo as necessidades de recepção, compreensão e expressão da linguagem. Quando se utiliza o computador para CAA, o sistema de símbolos pictóricos associado a um mecanismo de entrada apresenta símbolos representativos que são selecionados pelo usuário, compondo uma mensagem que pode ser estruturada em um texto com apresentação na tela, sintetizado em voz ou a combinação de ambos.

Ainda são escassas as soluções de software e hardware para CAA para uso em computadores convencionais. Grande parte das propostas estão relacionadas à confecção de hardware específicos, tais como teclados e mouses especiais, que possuem alto custo, grandes dimensões e exigem grande treinamento para que todo potencial seja usufruído (Stephanick et al., 2010; Salsman et al., 2010). Estas características dificultam sobremaneira sua disseminação e uso por parte de laboratórios de informática de escolas convencionais. Uma iniciativa acadêmica recente alinha a confecção de um dispositivo de entrada do tipo mordedor com dois diferentes softwares: um para promover aceleração e corretude linguística da redação através da previsão inteligente de palavras e orações futuras e outro para gerar a sequência correspondente de símbolos pictóricos para uma dada sentença redigida em português do Brasil (Santos et al., 2015).

Ainda não existe, entretanto, solução para uma demanda essencial e que corresponde exatamente ao oposto da citada: como gerar automaticamente uma sentença em linguagem natural a partir de uma dada sequência de símbolos pictóricos? Solução apropriada para esta questão seria uma importante ferramenta de apoio à alfabetização de crianças com paralisia cerebral. Além disso, esta mesma solução poderia ser instrumento de comunicação efetivo para crianças já familiarizadas com a comunicação via simbolos pictóricos,

Dois trabalhos são parcialmente relacionados à problemática. Em YAG (McRoy et al., 2000), a solução combina a abordagem baseada em *template* (*template-based*) para representar a estrutura de texto com método para representação de conhecimento. Um template é uma forma predefinida contendo slots que são então preenchidos com informações especificadas por usuários. O texto gerado pelo YAG pode advir de diferentes tipos de entradas, como uma sequência de proposições em linguagem lógica ou uma estrutura de caractérisicas junto com o nome do *template*. A aplicação desenvolvida por Ramos-Soto et al. (2015) gera pequenos termos de previsões meteorológicas a partir de desenhos relacionados, como chuva, sol, nuvens, representadas em forma de dados numéricos. A solução consiste na combinação de técnicas de percepção, computação com palavras (Zadeh, 2002, 1996) e estratégias para descrição linguistica de dados.

Neste artigo, propomos um método para solução do problema, que consiste fundamentalmente em um modelo baseado em *templates* para Geração de Linguagem Natural, similar ao proposto por McRoy et al. (2000, 2003). Este método foi implementado sob a forma de um software de apoio ao profissional da área de fonoaudiologia ou educação especializada no trabalho de alfabetização de crianças que se utilizam desses símbolos pictóricos para comunicação.

O método para geração automática de sentenças representativas das sequências de símbolos pictóricos é apresentado na seção 2 deste artigo. A ferramenta desenvolvida a partir desta proposta é apresentada na seção 3, onde fazemos preliminarmente a análise de corretude de sentenças geradas para um conjunto de validação fornecido. A seção 4 traz a conclusão do artigo.

2 Método

O método proposto segue um pipeline de ações para geração de texto em linguagem natural linguisticamente correto e que traduza fielmente a semântica da sequência de símbolos pictóricos apresentada como entrada.

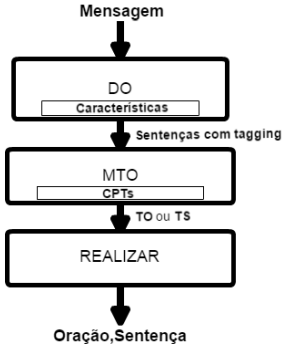

Figura 1: Componentes do método.

[1]Clik Tecnologia Assistiva, disponível em `http://www.clik.com.br/clik_01.html`.

Figura 2: Sequência de símbolos pictóricos (ou mensagem **M**) de entrada do pipeline de geração.

A figura 2 ilustra um exemplo de sequência de símbolos pictóricos para a qual o método deve produzir como saída de processamento a seguinte sentença, composta de duas orações: *Eu quero beber água de coco, mas em casa nós queremos brincar de bola com nossos amigos.*

O método para solução deste problema é fundamentado na fusão das ideias da gramática gerativa (GG) (Chomsky, 1965) e na representação de conhecimento (RC) através de *templates* (McRoy et al., 2000, 2003; Reiter, 1995). Da GG, a relevância para este problema reside na base categorial que faz parte do componente sintático, na qual uma oração é formada pelo $SN + SV$, onde SN é um sintagma nominal e SV um sintagma verbal (para este método o SV não inclui o verbo). Da RC, utilizamos, em nível de abstração, o conceito dos sintagmas nominal e verbal que foram representados na forma de *Proposição de Templete* (PT). Uma *proposição* é uma parte de uma oração, podendo ser um sujeito ou predicado da mesma. Dessa forma, uma PT é um micro-template que pode representar o sintagma SN ou SV de uma oração. Com isso, qualquer que seja o nível de granularidade de um *template*, ele deve possuir a estrutura síntatica característica da língua portuguesa do Brasil.

Para gerar orações ou sentenças por meio de *templates*, o método deve realizar as seguintes tarefas: (i) compor *Template de Oração* ($TO \sim SN + SV$) através de combinação de PTs com alguma Att (que é o verbo da oração); (2) compor *Template de Sentença* (TS) através de combinação de TOs, caso a mensagem de entrada necessite; e (3) realizar linguisticamente TOs ou TSs. Tanto TO quanto TS devem estar em conformidade com a mensagem (M) que foi dada como entrada para o CTO. Todas essas tarefas são, respectivamente, atribuídas aos módulos DO, CTO e Realizar, que são apresentados na figura 1.

A figura 1 apresenta a ligação entre os módulos que compõem este método. Os módulos são: *Detector de Oração (DO)*, que por sua vez contém um conjunto de caracterícas que identificam uma oração em uma sentença, *Construtor de Template de Oração (CTO)*, que contém um conjunto de PTs, e *Realizar*, cuja função é fazer a realização línguistica de um template. Mais detalhes sobre esses módulos serão descritos nas próximas subseções.

Na subseção 2.1, apresentamos o procedimento para aquisição do conhecimento e sua representação na base de *templates*. A subseção 2.2 descreve a técnica para detectar e separar as orações de uma sentença. A subseção 2.3 descreve o planejamento de documento e microplanejamento. Por fim, a subseção 2.4 descreve como é feita a realização linguística dos *templates*.

Aquisição e Representação do Conhecimento

A criação e validação do *corpus* linguístico, o qual foi utilizado para extrair o conhecimento necessário para este método, foram realizadas com supervisão de pesquisadora-chefe e estudantes de fonoaudiologia do Departamento de Fonoaudiologia da Universidade Federal de Sergipe.

As orações e sentenças que fazem parte desse *corpus*, quando relacionadas com as sequências de símbolos que as representam, possuem os seguintes níveis cognitivos: iconicidade, sintaxe e memorização. A iconicidade consiste na compreensão e percepção, ao selecionar pictogramas que represente alguma oração ou sentença. A sintaxe consiste no uso de preposição, pronome, advérbio e pontuação. Quanto à oração que se deseja gerar, ela pode ser simples ou composta. Já a memorização está relacionada à quantidade de símbolos pictóricos que podem ser representos por uma oração ou sentença. Além disso, os níveis cognitivos tercem o domínio deste método, juntamente com a necessidade de produzir orações que expressem noção de ação ou estado. Estes níveis são também aplicados às sentenças apresentadas na tabela 2.

Assim, a partir da análise desse *corpus*, foram confeccionadas 128 PTs que compoẽm a base de conhecimento para a solução do problema. Uma PT é então representada por *slots* (que são indicados por `<SLT>`, `<SLV>` ou `<ATPN>`), palavras da língua portuguesa e pela *Att*. A *Att* é in-

dicada pelo símbolo <ATPN>, onde A significa a atitude, T o tempo verbal, P a pessoa relacionada ao verbo (que é o sujeito da oração) e N o numeral (que pode ser singular ou plural) e está relacionado ao verbo. Dessa forma, cada PT deve pertencer a um *Conjunto de Proposição de Template* (CPT). Um CPT consiste no agrupamento de PTs por sua *chave*, a qual referencia o mesmo. Isso só é possível porque cada pictograma possui um ou mais nomes que o representam.

Segue alguns exemplos de PTs, CPTs, TO e TS que foram montados a partir de PTs:

1. **PTs**:
   `&a <SLT_NN>; &ao <SLT_NN>; &com &muita <SLV_VB> &de <SLT_NN>; &neste <SLT_NN> &de <SLT_NN>`
2. **CPTs**:
   (a) `Chave = IN_NN_PRP`
   (i) `em <SLT_NN> <SLT_PRP>; anzóis=[em]`
   (b) `Chave= VB_IN_NN_IN_PRP$_NN`
   (i) `<SLV_VB> de <SLT_NN> com nossas <SLT_NN>; anzóis=[de, com, nossas]`
   (ii) `<SLV_VB> de <SLT_NN> com nossos <SLT_NN>; anzóis=[de, com, nossos]`
3. **TOs**:
   `<ATPN> &com &muita <SLV_VB> &de <SLT_NN>; &a <SLT_NN> <AP3S> &ao <SLT_NN>`
4. **TS**:
   `<ATPN> &com &muita <SLV_VB> &de <SLT_NN>, mas &neste <SLT_NN> &de <SLT_NN> <AP1S> <SLV_VB>`

Note que nos *templates* que foram apresentados, introduzimos o conceito de "anzóis", indicados pelo símbolo &, que são as palavras classificadas morfologicamente como artigos, pronomes (exceto os pessoais), conjunções e preposições. Foi dessa forma, que representamos todo o conhecimento necessário para este método.

Detecção de Oração (DO)

Quando uma sequência de pictogramas representa uma mensagem que é composta por mais de uma oração, faz-se necessário identificar e extrair as orações. Isto acontece porque a estratégia é montar um determinado *template* para representar a estrutura de uma única oração ou sentença. Para tal tarefa, o algoritmo 1 é aplicado sobre uma mensagem m:

Considere uma mensagem m, como a do exemplo da figura 2, para o algoritmo 1. No passo 2 deste algoritmo, um pré-processamento sobre os nomes dos símbolos é realizado para que sejam retiradas as extensões e quaisquer caracteres que

Algorítmo 1 Detector de oração.

```
1: procedure DETECTAORACAO(m)
2:     m ← PREPROCESSAMENTO(m)
3:     sentencas ← DETECTESENTENCA(m)
4:     for all sentencas do
5:         tagSentencas ← POSTAG(sentencas)
6:     for all tagSentencas do
7:         oracoes ← REGRAS(sentencas, tagSentencas)
8:     return oracoes
```

não componham uma palavra válida na língua portuguesa (do Brasil). Então temos m igual a:

```
eu querer beber água de coco,
mas em casa nós querer brincar
de bola com nossos amigos.
```

No passo 3, detecta-se as sentenças que foram armazenadas em m. Em seguida, a função *POS-Tag* (Jurafsky & Martin, 2000) é aplicada sobre todas as sentenças (*sentencas*) nos passos 4 e 5. Ambas funções utilizadas, nos passos 3 e 5 do algoritmo 1, foram implementadas pela biblioteca OpenNLP.[2]

Já nos passos 6 e 7, são extraidas as orações atráves da aplicação das características apresentadas na tabela 1. Dentro da função REGRAS, cada vetor de *tokens* e vetor de morfemas de cada sentença são então varridos a fim de encontrar alguma característica listada na tabela 1. Quando isso acontece, uma oração é detectada e, então, esta é atribuida à variável *oracoes*.

Por fim, no passo 8, as orações que foram detectadas são retornadas. Sendo assim, obtemos como resultado do Detector de Orações para a mensagem m:

i) `eu_NNP querer_VB beber_VB água_NN de_IN coco_NN ,_SYM`

ii) `em_IN casa_NN nós_PRP querer_VB brincar_VB de_IN bola_NN com_IN nossos_PRP$ amigos_NN`

A Tabela 1 apresenta oito tipos de características que podem ocorrer na estrutura sintática da língua portuguesa, conforme as sentenças apresentadas no *corpus* (veja a subseção 2.1). Essas características são notadas quando se varre um vetor de *token* que representa algum sentença (tokenizada por espaço em branco e sinal de pontuação). Além do mais, para que as características sejam válidas, elas devem atender à pré-condição de que ao menos um verbo

[2]Disponível em http://opennlp.apache.org/.

Id	Morfema do *Token* Corrente	Morfema do *Token* Anterior	Morfema do *Token* Posterior
1	Conjunção	—	—
2	Vírgula	—	—
3	Pronome demonstrativo	Preposição	—
4	Pronome pessoal	—	Verbo
5	Pronome pessoal	Verbo	Verbo
6	Verbo	Vírgula	Vírgula
7	Interrogação	—	—
8	Exclamação	—	—

Tabela 1: As características que identificam uma oração para o nosso escopo.

deve existir antes da posição $i-1$ do *token* corrente i, exeto a sexta característica.

O Construtor de Templates de Orações (CTO)

O módulo CTO para esta ferramenta foi baseado nos módulos de planejamento definidos por Reiter & Dale (2000). Ele tem a responsabilidade de montar a estrutura sintática de uma oração ou sentença através do TO que pode ser concatenado a fim de montar o *template da sentença* (TS) de acordo com a sequência de pictogramas.

Com o resultado (i) do DO para sequência de pictogramas (ou mensagem **M**) ilustrado na figura 2, temos que as *proposições* são **M1**="eu" e **M2**="beber água de coco", consequentemente, **SN**="PRP", **Att**="querer" e **SV**="VB_NN_IN_NN" (para este módulo da ferramenta consideramos o SN e SV à nível morfológico, respectivamente, das *proposições* **M1** e **M2**). Desta forma, as *chaves* que mapeiam o conjunto onde devem ser encontradas as PTs de **M1** e **M2** são *chave*(**M1**)="PRP" e *chave*(**M2**)="VB_NN_IN_NN". Os resultados dessas *chaves* foram obtidos via concatenação dos valores armazenados no array de morfemas (visto na subseção 2.2). Então, para este exemplo, temos que os CPTs são:

1. `Chave(M1)= PRP`
   (a) `<SLT_PRP> anzóis=[]`
2. `Chave(M2) = VB_NN_IN_NN`
   (a) `<SLV_VB> <SLT_NN> com <SLT_NN> anzóis=[com]`
   (b) `<SLV_VB> <SLT_NN> de <SLT_NN> anzóis=[de]`

O primeiro *template* do CPT é selecionado, identificando-se pela *chave*(**M1**), que representa a *proposição* **M1**="eu". Este CPT contém apenas PT formado por *slot*, ou seja, qualquer um deles pode ser selecionado. A questão é como selecionar o *template* que melhor representa a *proposição* **M2**="*beber água de coco*". Para isso, utiliza-se o possível "anzol" de **M2**, &*de*, de modo que o template escolhido do CPT da *chave*(**M2**) é o segundo elemento do conjunto. O *template* completo para a oração TO_i seria então:

<SLT_PRP> <Att> <SLV_VB> <SLT_NN> de <SLT_NN>

Seja a segunda saída (ii) do DO, *em_IN casa_NN nós_PRP querer_VB brincar_VB de_IN bola_NN com_IN nossos_PRP$ amigos_NN*. As *proposições* neste caso são **M1**="em casa nós", **M2**="brincar de bola com nossos amigos" e **Att**="querer". Assim, **SN**="IN_NN_PRP", **SV**="VB_IN_NN_IN_PRP$_NN". As *chaves* que mapeiam os CPTs, onde deve ser encontrada as PTs **M1** e **M2**, são *chave* (**M1**)="IN_NN_PRP" e *chave* (**M1**)="VB_IN_NN_IN_PRP$_NN". Estas *chaves* mapeiam para os seguintes CPTs:

1. `Chave(M1) = IN_NN_PRP`
   (a) `em <SLT_NN> <SLT_PRP> anzóis=[em]`
2. `Chave(M2)= VB_IN_NN_IN_PRP$_NN`
   (a) `<SLV_VB> de <SLT_NN> com nossas <SLT_NN> anzóis=[de, com, nossas]`
   (b) `<SLV_VB> de <SLT_NN> com nossos <SLT_NN> anzóis=[de, com, nossos]`

Do primeiro conjunto, é selecionado o único template que representa a *proposição* **M1**="em casa nós". A seguir, o segundo elemento do CPT identificado pela *chave*(**M2**) é selecionado, já que os possíveis anzóis são &*de*, &*com* &*nossos* e que o template escolhido está relacionado à *proposição* **M2**="brincar de bola com nossos amigos". O template completo para a oração TO_{ii} é:

em <SLT_NN> <SLT_PRP> <Att> <SLV_VB> de <SLT_NN> com nossos <SLT_NN>

Caso a *chave(M1)* ou *chave*(**M2**) não referenciem nenhum CPT deste módulo, será necessário

inserir na base do sistema *templates* (PTs) que representem a oração desejada. Para que ainda assim se tenha ao menos uma sentença como saída, faz-se necessário um procedimento particular: outra *chave* dentre as existentes deve ser selecionada, desde que seja semelhante à *chave*(**M1**) ou *chave* (**M2**). Esta semelhança será computada através da similaridade do cosseno (intervalo [0, 1]):

$$\cos(z_i) = \frac{\vec{u}.P_i}{\|\vec{u}\|\,\|P_i\|},$$

tal que $0 \leq i \leq 9$ e $0 \leq z_i \leq \frac{\pi}{2}$, onde a *chave* de uma determinada proposição é o vetor $\vec{p} = (m^1, m^2, ..., m^9)$ com m sendo o valor da enumeração morfológica (número de classes gramaticais consideradas) e $P = (\vec{a}, \vec{b}, ..., \vec{n})$ é o conjunto de vetores pertencentes ao módulo CTO. O vetor com maior valor de similaridade será selecionado.

Sendo assim, o **TS** da sentença ilustrada na figura 2 é montado ao concatenar TO_i, a conjunção "mas" e $TO_{(ii)}$, então temos que TS é igual a:

<SLT_PRP> <Att> <SLV_VB> <SLT_NN> de <SLT_NN>, mas em <SLT_NN> <SLT_PRP> <Att> <SLV_VB> de <SLT_NN> com nossos <SLT_NN>

Realização Linguística

O propósito da realização linguística é realizar os **TOs** ou **TSs** (veja a subseção 2.3), ou seja, preencher os slots com as palavras correspondentes, respeitando a concordância nominal, fazer a concordância verbal correta com o sujeito da oração e, finalmente, adicionar os sinais de pontuação.

Dois dicionários de palavras foram construídos, considerando apenas palavras relacionadas com os símbolos pictóricos presentes na base. O primeiro dicionário possui palavras que não são verbos e está organizado da seguinte forma: (i) a primeira palavra é a palavra-chave que identifica as demais e não está flexionada em gênero ou número, (ii) as próximas palavras são flexionadas por número e, depois, por gênero. O segundo dicionário possui apenas verbos. O primeiro verbo está na forma infinitiva e funciona como a palavra-chave. Os verbos seguintes estão flexionados nos tempos verbais Presente e Futuro para cada pronome. Com este diciónario, pode-se realizar um template nesses dois tempos verbais. Por padrão, o tempo e a pessoa verbal utilizado para realização de *template* é o presente do indicativo e terceira pessoa do singular. Para as demais palavras que não são verbos, o número é singular e o gênero, masculino por padrão.

Para o TO_i do primeiro exemplo, *<SLT_PRP> <Att> <SLV_VB> <SLT_NN> de <SLT_NN>*, a realização linguística inicia com PT(**M1**)=< $SLT\_PRP >$ preenchendo-se o slot com o pronome "eu"; isto resulta em PT(**M1**)'="eu". A seguir **Att**="querer" deve concordar com o sujeito da *proposição* **M1** e deve ser conjugado no presente do indicativo (porque nenhum tempo verbal foi informado pelo usuário) de forma a concordar com o pronome: **Att**'="quero". Por fim, preenche-se os slots da PT(**M2**)=<SLV_VB> <SLT_NN> de <SLT_NN> e, dessa forma, temos que PT(**M2**)'="beber água de coco". Ao concatenar PT(**M1**)', **Att**', PT(**M2**)' e realizar a pontução, temos a oração *"eu quero beber água de coco,"*.

Para o $TO_{(ii)}$ do segundo exemplo, *em <SLT_NN> <SLT_PRP> <Att> <SLV_VB> de <SLT_NN> com nossos <SLT_NN>*, o primeiro slot do template da PT(**M1**)=*em <SLT_NN> <SLT_PRP>* é relacionado com o nome do pictograma "casa" e o segundo slot com o pronome "nós": PT(**M1**)'=*"em casa nós"*. Em seguida, **Att**=*"querer"* deve concordar com o sujeito da *proposição* **M1**, que nesse caso é o pronome e deve ser conjugado no presente do indicativo (por *default*): **Att**'="queremos". Finalmente, preenche-se os *slots* da PT(**M2**)=<SLV_VB> de <SLT_NN> com nossos <SLT_NN> com as descrições dos pictogramas, respectivamente, "brincar", "bola" e "amigos": PT(**M2**)'="brincar de bola com nossos amigos". Ao concatenar PT(**M1**)', **Att**' e PT(**M2**)', a oração realizada é *"em casa nós queremos brincar de bola com nossos amigos"*.

Com isso, temos que TS ($= TO_i + TO_{ii}$) realidazado lingusticamente é igual a esta sentença *"eu quero beber água de coco, mas em casa nós queremos brincar de bola com nossos amigos"*.

Discussão

Tendo em vista viabilizar uma comunicação simples e autônoma aos pacientes que ainda estam se familiarizando com os símbolos pictóricos, se faz necessário passar como entrada para o método ao menos um símbolo que represente um verbo (atitude). Isso se dá porque é bastante comum o uso de orações que expressem noção de ação ou estado neste tipo de comunicação.

Pela mesma razão em que se exige uma atitude, o uso de vírgulas na entrada do DO (ver seção 2.2) não é obrigatório. Se a vírgula fosse omitida no exemplo da subseção 2.2 depois da palavra "coco", ainda assim seria possível detectar a oração, pois neste caso, a entrada casaria

Nível	Descrição das Sentenças
1	Vamos tomar sorvete comigo
1	A menina foi para o mercado, mas não tinha dinheiro
2	A cidade é muito fria, por isso, o homem precisou de dois casacos para não adoecer
2	A festa foi hoje, se não tivesse chovido, ganharia muitos presentes, pois convidei muitos amigos
3	O tempo está chuvoso, por isso, não esqueça de fechar as janelas da casa quando sair para não molhar os móveis
3	Neste fim de semana, fui para fazenda de vovô. Calvaguei, me banhei de rio, comi manga, bebi leite da vaca e brinquei com meus amigos que moram lá

Tabela 2: Exemplos de sentenças fornecidas pelo Departamento de Fonoaudiologia.

com a caractéristica 1 da tabela 1 e estaria em conformidade com esta regra da gramatica: usa-se vírgula antes das conjuções "mas", "porém", "pois", "embora", "contudo", "todavia", "portanto" e "logo".

De certo, na montagem de TO ou TS, existe a amarração de PTs aos *anzóis* quando os utilizam para selecionar um PT de CPT (veja a subseção 2.3), embora isso ocorra somente se for passado algum símbolo pictórico como entrada para o método, que é interpretado como um *anzol*.

3 A Ferramenta CA²JU ESCRITO

O método de geração proposto para conversão de sequência de pictogramas em texto natural foi aplicado no desenvolvimento de uma ferramenta de apoio ao profissional que lida, em particular, com crianças que fazem uso da CAA para se comunicarem.

A composição visual ordenada de símbolos pictóricos deve ser feita da seguinte forma: (1) o profissional seleciona os pictogramas que estão apresentados em um teclado virtual localizado na parte inferior da imagem (figura 3), (2) os símbolos selecionados são apresentados em ordem da seleção no campo acima do teclado e (3) o texto será gerado a partir de um click.

Os símbolos utilizados na ferramenta pertencem ao sistema ARASAAC, que fora desenvolvido pelo Portal Aragonês de CAA. Esta é uma obra de Sergio Palao para CATEDU,[3] que os publica sob a licença Creative Commons.

Esta ferramenta faz atualmente parte de um conjunto de recursos de CAA do Laboratório de CAA da Universidade Federal de Sergipe para ensaios clínicos com pacientes.

A escolha de mensagens (em forma de sequência de símbolos pictóricos) que compõe a base experimental para testes e validação com crianças é baseada em protocolos de avaliação bem definidos pelos profissionais do Laboratório de CAA. Estes protocolos visam a seleção do sistema de signos por meio da compreensão, da percepção visual (escolha dos símbolos, tamanhos, etc), da mobilidade (acesso aos sistemas de auxílio técnico: precisão, rapidez, agilidade, força, etc), do nível cognitivo (nível de iconicidade, memória, léxico), de aspectos linguísticos, das posições posturais (ex: sentado, deitado, etc). Por fim, os protocolos analisam as formas de indicação dos sinais, sendo possível: (i) indicação direta, (ii) direta com auxílio, (iii) codificada, (iv) varredura (ou exploração) dependente ou (v) varredura independente. Um protocolo de acompanhamento vem sendo desenvolvido para registro semanal dos dados que evidenciem como estão sendo atingidos os objetivos comunicativos para a ferramenta.

A corretude dos textos produzidos pela ferramenta foi avaliada comparando-se com um conjunto de validação fornecido pelo Laboratório (ver tabela 2). A distância de Levenshtein, utilizada como métrica neste experimento preliminar, apontou valor próximo de 0 (zero) para todo o conjunto. Isto significa que a similaridade léxico-sintática entre as sentenças geradas automaticamente pela ferramenta e as pertencentes ao conjunto de validação foi muito alta.

4 Conclusão

Este artigo propôs um método para geração automática de sentenças em linguagem natural a partir de sequência de símbolos pictóricos, bastante utilizados em suporte à Comunicação Alternativa e Ampliada (CAA). O método proposto é baseado na confecção de *templates* que permitem boa variabilidade linguística das construções.

O método descrito foi utilizado para criação de uma ferramenta de suporte ao profissional de CAA que lida com crianças com paralisia cerebral e com crianças com transtorno do espectro autístico. O propósito específico da ferramenta é propiciar um ambiente computacional para facilitar a alfabetização destas crianças. A literatura relacionada não mostra quaisquer iniciativas com

[3] http://catedu.es/arasaac/

Figura 3: Interface gráfica Ca2ju Escrito.

este propósito. A ferramenta possui interface visual adequada para composição de sequência de pictogramas por parte do profissional e posterior geração do texto natural correspondente.

Em experimentação preliminar com um conjunto de validação fornecido pelo Laboratório de CAA da Universidade Federal de Sergipe foi mostrado que a geração das sentenças por parte da ferramenta condiz perfeitamente com as sentenças do conjunto de validação. A ferramenta é atualmente integrante do conjunto de recursos de CAA do respectivo laboratório e faz parte dos ensaios clínicos com grupos de controle e experimental.

Trabalhos em andamento consistem no aumento do conjunto de validação e complexidade das sentenças-alvo, finalização do protocolo de acompanhamento e, principalmente, avaliação quantitativa da contribuição da ferramenta enquanto mecanismo da CAA para a alfabetização de crianças com deficiência a partir dos grupos citados anteriormente. Resultados destes estudos são previstos até fim de 2017.

Referências

Alant, Ema & Juan Bornman. 1994. Augmentative and alternative communication. *South African Family Practise* 15(5).

Beukelman, David & Pat Mirenda. 2005. *Augmentative and alternative communication.* Brookes Publishin.

Bharucha, Ashok J., Vivek Anand, Jodi Forlizzi, Mary Amanda Dew, Charles F. Reynolds, Scott Stevens & Howard Wactlar. 2009. Intelligent assistive technology, applications to dementia care: Current capabilities, limitations, and future challenges. *The American Journal of Geriatric Psychiatry* 17.

Brodwin, Martin. 2010. Assistive technology. Em Irving B. Weiner & W. Edward Craighead (eds.), *Corsini Encyclopedia of Psychology*, 1–2. John Wiley and Sons.

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* MIT Press.

Jurafsky, Daniel & James H. Martin. 2000. *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition* Prentice Hall Series in Artificial Intelligence. Prentice Hall.

Light, Janice. 1989. Toward a definition of communicative competence for individuals using augmentative and alternative communication systems. *Augmentative and Alternative Communication* 5(2). 137–144.

McRoy, Susan W., Songsak Channarukul & Syed S. Ali. 2000. YAG: a template-based generator for real-time systems. Em *1st International Conference on Natural Language Generation (INLG)*, vol. 14, 264–267.

McRoy, Susan Weber, Songsak Channarukul & Syed S. Ali. 2003. An augmented template-based approach to text realization. *Natural Language Engineering* 9(4). 381–420.

Ramos-Soto, Alejandro, Alberto Jose Bugarín, Senén Barro & Juan Taboada. 2015. Linguistic

descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems* 23(1). 44–57.

Reiter, Ehud. 1995. NLG vs. templates. Em *European Workshop on Natural Language Generation*, vol. 5, 95–106.

Reiter, Ehud & Robert Dale. 2000. *Building natural language generation systems* Natural Language Processing. Cambridge University Press.

Salsman, Kenneth, John Sweetser & Anders Grunnet-Jepsen. 2010. Electronic equipment for handheld vision based absolute pointing system. Patente 7796116. US Patent and Trademark Office.

Santos, Flávio, Carlos Junior, Hendrik Macedo, Marco Chela, Rosana Givigi & Luciano Barbosa. 2015. CA^{2}JU: an assistive tool for children with cerebral palsy. *Studies in Health Technology and Informatics* 216. 589–593.

Stephanick, James, Christina James, Ethan R. Bradford & Michael R. Longé. 2010. Selective input system based on tracking of motion parameters of an input device. Patente 7750891. US Patent and Trademark Office.

Zadeh, Lofti A. 1996. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems* 4(2). 103–111.

Zadeh, Lofti A. 2002. From computing with numbers to computing with words – from manipulation of measurements to manipulation. *International Journal of Applied Mathematics and Computer Science* 12(3). 307–324.

BrAgriNews: Um Corpus Temporal-Causal (Português-Brasileiro) para a Agricultura

BrAgriNews: A Temporal-Causal Brazilian-Portuguese Corpus for Agriculture

Brett Drury
Faculty of I.T.,National University of Ireland Galway,Ireland
brett.drury@gmail.com

Robson Fernandes
ICMC, University of Sao Paulo, Sao Carlos,Brazil
robs.fernandes@outlook.com

Alneu de Andrade Lopes
ICMC, University of Sao Paulo, Sao Carlos,Brazil
alneu@icmc.usp.br

Resumo

Recentemente tem havido um aumento no interesse, tanto no meio acadêmico quanto na indústria, em aplicações de aprendizagem de máquina e técnicas de inteligência artificial relacionadas com problemas agrícolas. Mineração de texto e técnicas relacionadas com o processamento da língua natural, raramente foram usadas para resolver problemas agrícolas, e muito menos para a língua portuguesa. É possível que um dos fatores que influenciam a escassez no uso técnicas de mineração de texto, para analisar textos em português e resolver problemas agrícolas, pode ser devido à falta de um corpus anotado livremente disponível. Para colmatar a falta de um corpus agrícola em língua portuguesa, estamos liberando um recurso em português-brasileiro voltado para agricultura, descrito neste artigo. O corpus abrange um período parcialmente contínuo de tempo entre 1996 e 2016, consistindo de notícias em português-brasileiro que foram anotadas com o seguinte tipo de informação: causal, sentimento, entidades nomeadas que incluem expressões temporais. O corpus tem recursos adicionais como: treebank, listas de termos frequentes (sem stop-words): unigramas, bigramas e trigramas, bem como palavras ou frases que foram identificados por jornalistas como de domínio específico. Espera-se que a liberação do corpus estimule a adoção da mineração de texto na agricultura na comunidade de pesquisa lusófona.

Keywords

Mineração de Texto, Agricultura, Relações causais

Abstract

There has been a recent sharp increase in interest in academia and industry in applying machine learning and artificial intelligence to agricultural problems. Text mining and related natural language processing techniques, have been rarely used to tackle agricultural problems, and at the time of writing there was a single project in the Portuguese language. It is possible that the failure of researchers to use text mining techniques to analyze Portuguese texts to resolve agricultural problems may be due to a lack of freely available corpora. To correct the lack of a Portuguese language agriculture centric corpus we are releasing a Brazilian-Portuguese agricultural language resource, which is described by this paper. The corpus is partially non-contiguous and spans a time period from 1996 to 2016. It consists of news stories that have been scraped from Brazilian News sites that have been annotated with the following information types: causal, sentiment, named entities that include temporal expressions. The corpus has additional resources such as a: treebank, lists of frequent: unigrams, bigrams and trigrams, as well words or phrases that have been identified by journalists as either: "important" or domain specific. It is hoped that the release of this corpus will stimulate the adoption of text mining in agriculture in the Lusophonic research community.

Keywords

Text Mining, Agriculture, Causal Relations

1 Introdução

Este artigo descreve um corpus em português-brasileiro, em que se pretende ser útil para incentivar a pesquisa em mineração de texto para a agricultura.

O *BrAgriNews* é um corpus parcialmente não contíguo que abrange o período de 1997 a 2016. O corpus anota as seguintes informações: sentimento, informações temporais, causais e entidades nomeadas em notícias agrícolas. O corpus contém: Um "treebank" e documentos com parte de etiquetas de fala, bem como: modelos de tópicos e representações vetoriais de termos. Também fornece recursos léxicos, tais como:

1. Palavras frequentes;
2. Bigramas frequentes;

DOI: 10.21814/lm.9.1.245

3. Trigramas frequentes;
4. Palavra/frases que são considerados "importantes" pelos jornalistas com a adição de delimitadores, como aspas.

O restante do artigo está organizado da seguinte forma: Seção 2: Trabalhos Relacionados; Seção 3: Aquisição do Corpus e Visão Geral; Seção 4: Metodologia de Anotação; Seção 5: Recursos Léxicos; Seção 6: Treebank; Seção 7: Recursos de Relações entre Palavras; Seção 8: Informações de Nível de Documento; Seção 9: Licenciamento; Seção 10: Trabalhos Futuros; Seção 11: Conclusão.

2 Trabalhos Relacionados

Este corpus contém uma variedade de fenômenos da linguagem, incluindo causalidade, expressões temporais, bem como sentimento. O trabalho relacionado, portanto, concentra-se nas seguintes áreas:

1. Causalidade na linguagem.
2. Representação temporal no texto.
3. Sentimento na linguagem.

Causalidade

Há uma série de definições de causalidade. Uma definição bem conhecida foi preferida pelo filósofo escocês David Hume que afirmou que a causalidade tem três propriedades específicas: "(i) contiguidade no tempo e no lugar; (ii) prioridade no tempo, e (iii) constante conjunção entre a causa e o efeito" (Khoo et al., 2002). A causalidade na linguagem é expressa como "relações causais." As relações causais são relações dependentes entre eventos, fatos ou objetos (Vendler, 1967; Altenberg, 1984), onde um evento, fato ou objeto é a causa de outro evento, fato ou objeto (Altenberg, 1984).

As relações causais no texto como explicado anteriormente são relações dependentes entre eventos, fatos ou objetos. Os objetos de causa (eventos, fatos ou objetos) são ligados através de uma ligação causal aos objetos de evento (eventos, fatos ou objetos). Uma ligação causal é uma palavra ou frase que contém propriedade causal. Ligações causais são tipicamente verbos causais (Shams-Eddien, 2002), nos quais a causa ou objetos de evento podem ser expressos como frases nominais. As relações causais podem, portanto, ser expressas como simples padrões de extração, como:

$$NP\ CV\ NP,$$

no qual NP = *Frases Nominais* e CV = *Verbo Causal* (Shams-Eddien, 2002). O fluxo de causalidade neste padrão é da esquerda para a direita, onde o lado esquerdo (LHS) NP é o objeto de causa e o lado direito (RHS) é o objeto de efeito. Em português esta ordem pode ser alterada por uma preposição, por exemplo a expressão "por causa de" inverterá a ordem de causalidade em uma relação causal. A maior parte da pesquisa sobre a causalidade na língua foi realizada em inglês, por exemplo por Khoo et al. (2002); Altenberg (1984); Thomson (1987); Shams-Eddien (2002), sendo que poucos foram os estudos conduzidos em Português (Drury & de Andrade Lopes, 2015).

Representação e Extração do Tempo

Uma característica dos corpora disponíveis são as anotações temporais. Uma suposição deste artigo é que a representação temporal no texto é uma maneira de descrever expressões multi-palavras que representam:

1. Duração;
2. Expressão do tempo.

Por exemplo: "21 de maio de 2001" é uma expressão do tempo e "12/04/75 – 12/05/76", é uma duração de tempo. O tempo pode cobrir: segundos, minutos, horas, dias, décadas, anos e assim por diante.

Expressões de tempo podem ser feitas em linguagem natural em uma série de maneiras diferentes, consequentemente houve um padrão desenvolvido que tenta ter uma maneira uniforme de expressar informação temporal e de evento. Este padrão é o *TimeML* (Pustejovsky et al., 2003a)[1]. O *TimeML* é um dialeto XML, que permite a expressão padrão de:

1. Marcação de tempo de eventos;
2. Ordem de eventos com relação a um outro;
3. Raciocínio com expressões temporais contextualmente sub-especificadas;
4. Raciocínio sobre a persistência de eventos.

Além da padronização das expressões temporais, o consórcio *TimeML* lançou uma série de ferramentas que podem ser usadas para anotar ou extrair expressões de tempo no texto. O site documenta a Ferramenta de anotação (TANGO) e o *Tarsqi Toolkit*.

[1] http://www.timeml.org

O *Tarsqi Toolkit* contém um conjunto de ferramentas que podem ser usadas para extrair expressões de tempo, bem como garantir a sua consistência.

A literatura de pesquisa contém uma série de estratégias para extrair expressões temporais. Essas estratégias podem ser agrupadas em duas categorias: 1. aprendizagem de máquina (Bethard, 2013; Kolya et al., 2013; Llorens et al., 2010; UzZaman & Allen, 2010) e 2. híbrida de aprendizagem de máquina e linguística (Laokulrat et al., 2013; Jung & Stent, 2013).

Uma abordagem comum de aprendizado de máquina na literatura de pesquisa é a aprendizagem supervisionada com campos aleatórios condicionais (conditional random fiels — CRF) (Kolya et al., 2013; Llorens et al., 2010; UzZaman & Allen, 2010). As abordagens híbridas usam características linguísticas de dados rotulados para gerar modelos em uma estratégia de aprendizagem supervisionada. As duas principais características linguísticas utilizadas nas técnicas híbridas são as estruturas de dependência (Laokulrat et al., 2013) e informação semântica (Jung & Stent, 2013).

Existem vários corpora que podem ser usados para avaliar estratégias de extração temporal. Os dois principais corpora para o Inglês são: TimeBank (Pustejovsky et al., 2003b) e o AQUAINT Corpus[2]. Esses corpora são relativamente pequenos, com 183 e 73 notícias, respectivamente. Existem corpora em línguas não-inglesas, tais como para o Francês (Bittar, 2010), Italiano (Caselli et al., 2011), Romeno (Forascu & Tufiș, 2012), Espanhol[3] e Catalão.[4] Para o Português temos o HAREM (Carvalho et al., 2008), com 129 notícias.

Análise de Sentimentos

A análise do sentimento, de acordo com Liu e Zhang, é o estudo computacional das opiniões, avaliações, atitudes e emoções das pessoas em relação a entidades, indivíduos, questões, eventos, tópicos e seus atributos (Liu & Zhang, 2012). O campo é vasto, consequentemente esta pesquisa será limitada à análise de sentimentos da língua portuguesa.

Existêm vários métodos para a análise do sentimento, porém a abordagem dominante para o português descoberta nessa revisão é a baseada em dicionário. A análise do sentimento baseada em dicionário utiliza-se de recursos lexicais que possuem palavras ou frases com uma orientação de sentimento pré-definida. Existem três dicionários principais: dois multilíngue: *SentiLex* (Silva et al., 2012), *Opinion Lexicon* (Souza et al., 2011) e *LIWC* (Balage Filho et al., 2013), que é parte de uma aplicação de software. Avaliou-se os três dicionários e os principais pontos constatados foram que o *Sentilex* foi superior para a classificação de sentimento de documentos e *LIWC* produziu os melhores resultados para a classificação de opinião de sentenças. A análise do sentimento baseado no dicionário para o português foi aplicada a uma série de áreas que incluíram hotéis (Chaves et al., 2012), finanças (Alvim et al., 2010), crítica de cinema (Freitas & Vieira, 2013) e política (Silva et al., 2009).

As estratégias supervisionadas de classificação do sentimento de aprendizado de máquina exigem dados de treinamento. Um possível impedimento para o uso dessas técnicas é a falta de corpos anotados na língua portuguesa. Esta revisão da literatura descobriu um pequeno número de recursos que continham relativamente poucos recursos: Petronews (1500 documentos) (Alvim et al., 2010), ReLi (2056 documentos) (Freitas et al., 2012) e o conjunto de dados de Drury & de Andrade Lopes (2014) (500 documentos).

3 Aquisição do Corpus e Visão Geral

O corpus, como já comentado, contém notícias relacionadas à agricultura escritas em português-brasileiro. O corpus foi construído a partir de recursos inéditos pré-existentes e com notícias coletadas na Internet. As notícias foram coletadas com um "scraper" de sites respeitáveis, como:

1. Revista Canavieiros (Sugarcane Magazine).
2. Jornal Cana (Sugarcane Newspaper).

O "scraper" rodava às 8 horas da manhã, antes do início da bolsa de São Paulo. Esta decisão foi tomada para garantir que todas as experiências de negociação que foram feitas com modelos derivados deste corpus seriam "justas". O "scraper" correu de 2014 a 2016. O corpus final contém 96.784 documentos.

Características da Linguagem

Coleções de documentos ou corpus têm características específicas de linguagem que são determinadas pelo assunto e estilo do autor. Uma

[2] https://tac.nist.gov//data/data_desc.html#AQUAINT

[3] Disponível em https://catalog.ldc.upenn.edu/docs/LDC2012T12/

[4] Disponível em https://catalog.ldc.upenn.edu/docs/LDC2012T10/

maneira de comparar a linguagem é comparar a frequência de:

1. Advérbios com adjetivos.
2. Substantivos com verbos.

Os rácios foram 0,52 e 2,24, respectivamente. Uma comparação com outros textos pode ser encontrada nas Figuras 1 e 2[5].

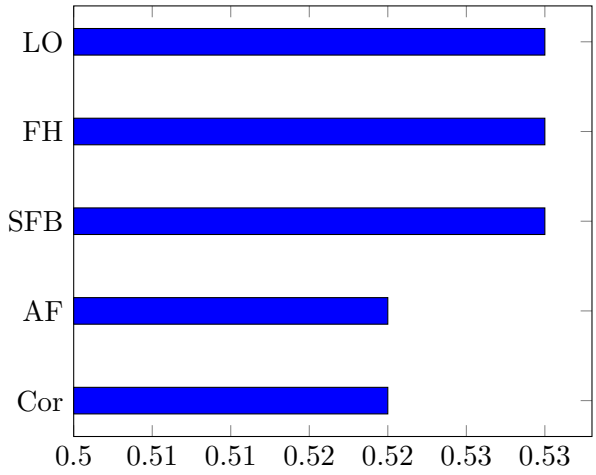

Figura 1: Relação entre advérbios e adjetivos, onde Cor = Corpus, AF = O Triunfo dos Porcos (Animal Farm), SFB = (Escândalo do Padre Brown) Scandal Of Father Brown, FH = História de Fanny Hill (Fanny Hill) e LO = Romance Lady Oracle (Lady Oracle).

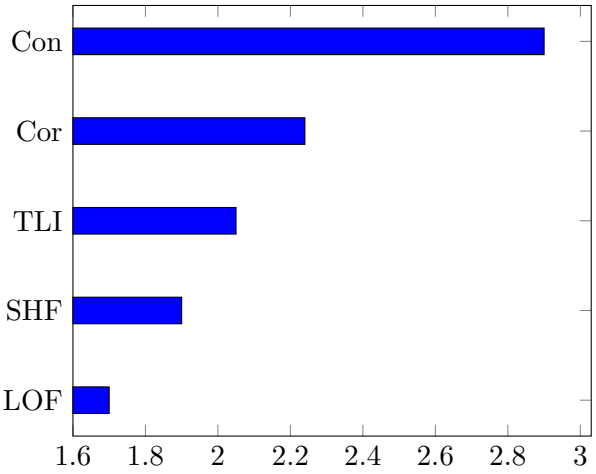

Figura 2: Relação entre Substantivo e Verbo, onde LOF = Vida de Johnson (Life Of Johnson), SHF = Forma das Coisas Por Vir (Shape Of Things To Come), TLI = O Instinto da Linguagem (The Language Instinct), Cor = Corpus (Corpus) e Con = Constituição (Constitution).

[5]Uma lista completa de rácios para textos alternativos para: substantivo/verbo e adjetivos/advérbios podem ser encontrados em: 1. https://goo.gl/1OZpNH e 2. https://goo.gl/6hzYPd.

Uma técnica de análise de linguagem complementar é listar as palavras mais frequentes no corpus. As palavras frequentes no corpus são boas indicadoras do assunto porque a frequência da palavra segue uma distribuição zipf, como demonstrado na Figura 3. A análise de palavras frequentes removeu *stop-words* (um, isto, o, etc), uma vez que elas não têm um significado específico de domínio, pois ocorrem com frequência relativa similar na maioria dos corpora ou coleções de textos. As palavras mais comuns neste corpus foram: Brasil; Milhões; Governo; Presidente; Mercado; Produção; Nacional; Acordo; Estado e Safra.

Uma representação visual da frequência de palavras na coleção de corpus é representada no diagrama de Nuvem de Palavra na Figura 4.

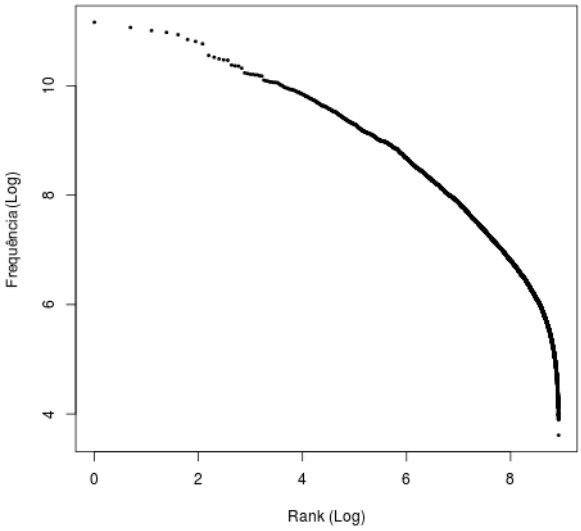

Figura 3: Relação entre a frequência das palavras e o seu rank.

Figura 4: Nuvem de Palavras de termos frequentes no corpus *BrAgriNews*.

A análise final considerou o tamanho do documento (número de palavras), frequência média das palavras e número total de palavras. Os valores foram: 1.127,14 palavras por documento,

frequência média de 1.617,82 palavras ±3504.12 e 12.305.150 de palavras no corpus *BrAgriNews*.

As técnicas de análise simples acima referidas forneceram uma visão geral das características linguísticas do corpus. A razão entre frequencias de substantivos e verbos indicam um corpus em liguagem objetiva, no qual a relação entre adjetivos e advérbios é similar a da literatura clássica. A contagem de frequência indica que os assuntos dominantes são: Estado; Comércio; e Agricultura. E que o comprimento médio do documento é relativamente pequeno.

Visão Geral do Corpus

O *BrAgriNews* está disponível em `https://goo.gl/lc0PzS`, e é distribuído como um arquivo compactado. A organização de pastas de nível superior é apresentado na Figura 5.

A pasta de nível superior contém: notícias, previsões meteorológicas e um *treebank*. As pastas *Weather Forecasts* e *Trees* contêm previsões meteorológicas e representação de árvore de dependência de sentenças aleatórias, respectivamente. A pasta *News Stories* tem um segundo nível de pastas que é demonstrado na Figura 5. O conteúdo das pastas será descrito posteriormente neste artigo.

Resumo da Etiqueta

A principal contribuição deste corpus é a anotação de notícias. A anotação delimita informações que podem ser úteis para categorização supervisionada ou técnicas de extração de relação. As notícias anotadas são armazenadas na pasta *Annotated Texts*. As anotações assumem a forma de marcações do tipo XML (etiquetas) que delimitam: uma única palavra ou uma sequência de palavras. As etiquetas anotam:

1. Sentimento.
2. Relações causais.
3. Porções de causa e efeito de relações causais.
4. Expressões de tempo.
5. Expressões de moeda.

Um resumo das etiquetas é descrito na Tabela 1.

Etiqueta	Explicação
Positive	Uma palavra que foi determinada como tendo uma orientação positiva
Negative	Uma palavra que foi determinada como tendo uma orientação negativa
Entity	Uma palavra ou n-grama que foi determinado como uma entidade nomeada
CRelation	Delimitação de uma relação causal
Effect	A parte de um efeito de uma relação causal
Cause	A parte de uma causa de uma relação causal
DOW	Dia da Semana
TOD	Hora do Dia
Season	Estação
Week	Expressão semanal
Date	Expressão diária
Currency	Expressão monetária
Quote	Discurso direto

Tabela 1: Resumo da Etiqueta.

4 Metodologia de Anotação

Esta seção discute as estratégias que foram usadas para anotar os documentos neste corpus. Haviam duas escolhas metodológicas possíveis para anotar este corpus:

1. Anotação manual.
2. Anotação automatizada.

A anotação manual é laboriosa e lenta, consequentemente seria impraticável usar esta técnica para este corpus e a anotação automatizada foi selecionada.

O resumo de etiqueta descrito na Tabela 1 revela que são 6 áreas de anotações principais:

1. Entidades nomeadas.
2. Anotação de sentimento.
3. Expressões de tempo.
4. Relações causais.
5. Discurso direto.
6. Parte da fala.

Entidades Nomeadas

As entidades nomeadas são palavras únicas ou expressões multi-palavras, que podem ser classificadas em uma categoria pré-existente, tais como: pessoa, empresa, organizações, e assim por diante. O suporte de entidade nomeada para o português-brasileiro é limitado e no momento da construção do corpus não havia nenhum classificador/extrator de entidade nomeada livremente disponível. Consequentemente, uma técnica baseada em regras foi desenvolvida para identificar candidatos de entidades nomeadas.

A técnica usou o seguinte procedimento:

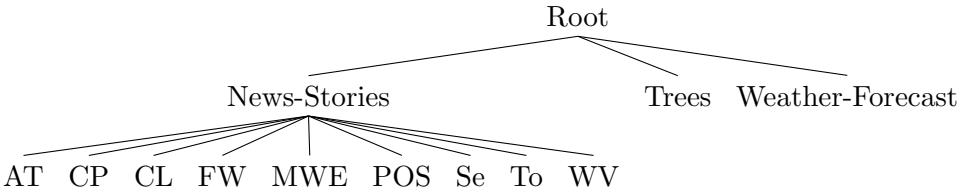

Figura 5: Organização das pastas, onde AT = Textos Anotados, CP = Frases de Causa, CL = Clusters, FW = Palavras Frequentes, MWE = Expressão Multi-Palavras, POS = Parte da Fala, Se = Sentimento, To = Topic and WV = Vetor de Palavras.

1. Identificar palavras maiúsculas que não iniciam sentenças.
2. Juntar os candidatos da Regra 1 se for separado por uma palavra de ligação.
3. Repetir a Regra 2 com entidades unidas geradas a partir dessa mesma regra.

O processo de união descrito nas Regras 2 e 3 pode ser ilustrada com o seguinte exemplo: uma entidade denominada candidata gerada por esta técnica é *Procuradoria-Geral da República*, que contém duas entidades candidatas denominadas *Procuradoria-Geral* e *República*, que é acompanhado por uma palavra de ligação *da*.

Uma pequena avaliação manual feita por um único especialista em domínios, onde 10 documentos foram escolhidos aleatoriamente, constatou que a técnica tinha uma precisão de 73.25%. A avaliação identificou manualmente as entidades em um documento, e verificou que a técnica as identificou corretamente. Correspondências parciais, bem como a falha ao identificar as entidades foram marcadas como incorretas.

Anotações de Sentimento

A anotação de sentimento foi alcançada usando um dicionário pré-compilado de sentimento: *Sentilex*. O dicionário contém palavras que têm uma orientação pré-determinada do sentimento. A estratégia divide as palavras em um documento e verifica a palavra contra a entrada no *Sentilex*. A estratégia aplica-se a uma das duas etiquetas: positiva ou negativa, as palavras com orientação de sentimento neutro são ignoradas. Por exemplo, <negative> ruim </negative>, ruim tem uma conotação de sentimento negativo e consequentemente é encapsulado com uma etiqueta negativa.

O dicionário *Sentilex* foi avaliado por Balage Filho et al. (2013), verificou-se que o Sentilex tem uma precisão de 44.17% no nível da sentença e 53.35% no nível do documento. *Sentilex* é um dos melhores dicionários de sentimentos para a língua portuguesa.

Expressões temporais

As expressões temporais para este corpus foram extraídas usando uma abordagem padrão baseada em regras. A expressão diária foi extraída com expressões regulares que identificaram sequências de números com separadores comuns. As expressões típicas captadas por esta abordagem foram “12/04/2016” e “12/04/16”.

As demais categorias de expressões de tempo foram capturadas usando listas codificadas de palavras. A lista de palavras foi compilada por um especialista em domínio.

As técnicas de anotação de expressão temporal baseadas na expressão regular, relataram exatidão muito alta, por exemplo, Strötgen & Gertz (2010) relataram que sua técnica de expressão regular registrou uma precisão de 85.00%.

Relações Causais

As anotações de causalidade seguem a noção de que as relações causais entre os eventos, e que a relação causal contém duas partes: (i) Evento de causa, e (ii) Evento de efeito.

Consequentemente, as anotações causais têm três anotações: (i) Toda a relação causal; (ii) Evento de causa; e (iii) Evento de efeito.

A estratégia de anotação causal foi uma estratégia de aprendizagem supervisionada descrita por Drury & de Andrade Lopes (2015). A estratégia utilizou uma visão local e global da causalidade no corpus. Dois separadores são criados a partir dessas duas visões. Os dois classificadores rotulam as relações causais no corpus e, quando os dois classificadores concordam com uma relação causal, uma anotação causal é feita. Exemplos das relações causais são demonstrados na Tabela 2.

Esta técnica foi avaliada por Drury & de Andrade Lopes (2015), verificou-se que tem uma precisão de 67.00% na anotação do nível da frase e 81.00% na classificação da relação causal.

Expressão Causal (Português)
preços gasolina alta aumentando demanda biocombustivel
politicas diminuído industria biocombustivel
consumo problemas logisticos causa destaca
surgiram oportunidades curto prazo exportacoes brasileiras biocombustivel

Tabela 2: Relações Causais relacionado com Biocombustíveis

Discurso Direto

Discurso direto para este artigo, é o discurso que foi citado diretamente no texto. Por exemplo, "Eu não estou em seu comitê de estratégia" Watson respondeu (https://goo.gl/VLeH18). O discurso é delimitado por marcas de fala, e seguido por uma entidade nomeada e um verbo.

A estratégia para anotar a fala direta foi outra técnica baseada em regras que identificou delimitadores de fala que foram as aspas, e as marcas de fala.

As palavras entre esses delimitadores foram assumidas como sendo de fala direta se a frase extraída tivesse uma contagem de palavras mínima de 6.

Uma pequena avaliação manual de 10 documentos que continham uma etiqueta de citação, realizadas por um único especialista de domínio, descobriu que as seqüências de texto que foram marcadas com aspas estavam corretas 86.66% do tempo. Uma citação correta foi assumida para ter um orador, como uma pessoa ou outra entidade, como uma empresa ou organização, bem como um elemento de fala. Marcação indevida ou obviamente incorreta foi marcada como um erro pelo anotador.

Marcação de Parte da Fala

A Marcação do papel morfo-sintático (part-of-speech tagging) aplica uma categoria de palavra como substantivo, adjetivo, advérbio, etc. a uma palavra. Para as marcações foi usado o *nlpnet* (Fonseca & Rosa, 2013) que é um rotulador baseado em rede neural. O rotulador foi treinado no corpus *mac-morpho* e tem: "97.33% a precisão de um token, 93.66% exatidão do token fora do vocabulário".

Um exemplo das anotações tipicamente encontradas no corpus pode ser encontrado na Tabela 3. A anotação é uma citação direta por "um consultor". A citação é encapsulada pela etiqueta "quote". A citação contém uma série de palavras sentimentais, em particular: "sofra" e "stress".

Essas palavras têm conotação negativa e, consequentemente, são encapsuladas por uma etiqueta "negativa". A citação contém uma relação causal: "o stress durante a pré-polinização pode resultar em produtividades menores.". Esta relação causal contém um evento de causa: "o stress durante a pré-polinização" e um evento de efeito: "produtividades menores". A citação também contém informações sobre o tempo: "Maio" e informação da entidade tal como: "Kansas" e "Iowa".

Exemplo Anotado

A agregação das anotações pode fornecer uma descrição detalhada dos dados. Um exemplo de anotações agregadas pode ser encontrado na Tabela 3.

Exemplo anotado
<Quote> "Minha preocupação é de que algum milho <Negative> sofra </Negative> com o <Negative> stress </Negative> hídrico durante a polinização, quando a planta está definindo o tamanho da orelha. Uma vez que este tamanho está definido, ele não pode ficar <Month> maio </Month> , assim sendo, <CRelation> <Cause> o <Negative> stress </Negative> durante a pré-polinização </Cause> pode resultar <Effect> em produtividades menores </Effect> </CRelation> . Eu acredito que isso possa já estar ocorrendo em alguns locais com o leste do <Entity> Kansas, </Entity> norte do <Entity> Missouri, </Entity> sul de <Entity> Iowa </Entity> e oeste de <Entity> Illinois, Indiana, Ohio </Entity> e <Entity> Michigan" </Quote> , </Entity> diz o consultor.

Tabela 3: Exemplo anotado

O exemplo de anotação demonstra claramente o esquema de anotação e como ele é usado dentro do corpus *BrAgriNews*, onde:

1. Etiqueta 'Quote' indica citação.
2. Etiqueta 'Negative' indica palavras com conotação negativa.
3. Etiqueta 'CRelation' indica citações que contém relação causal.
4. Etiqueta 'Month' indica citações que contém informações sobre o tempo.
5. Etiqueta 'Entity' indica informações sobre a entidade.

5 Recursos Léxicos

Há uma série de recursos léxicos que complementam o corpus principal. Os recursos léxicos estão localizados nas pastas *Multi-word Expressions* e *Frequent Words*.

Os recursos léxicos são: Palavras frequentes (não *stop-words*); Bigramas frequentes; e Trigramas frequentes.

Palavras Frequentes

As palavras frequentes, como descrito anteriormente, são palavras frequentes que não são *stop-words*. A técnica para identificar palavras frequentes eliminou qualquer palavra do corpus que estivesse em listas de stop-words comuns[6]. A frequência para o restante das palavras foi calculada. As 7499 palavras mais frequentes são armazenadas em um arquivo de texto e em formato "pickle" em Python (dicionário) e localizado na pasta *Frequent Words*.

Expressões Multi-palavras

Expressões multi-palavras são expressões que contêm 2 palavras ou mais. Existem várias estratégias para calcular expressões multi-palavras (MWE), e para os recursos MWE fornecidos com este corpus foram utilizadas três estratégias: Associação estatística; Co-ocorrência de palavras; Delimitadores de frases.

Associação estatística

É uma estratégia que identifica relações estatísticas entre palavras que aparecem em sequência (pares de palavras). Os pares de palavras que têm uma relação estatística significativa são susceptíveis de ser uma expressão de multipalavras (multi-word expression)(MWE) ou parte de um MWE. A técnica utilizada para calcular as MWEs foi *Pointwise Mutual Information* (PMI). O cálculo do PMI pode ser representado

$$PMI = \log\left(\frac{P(a,b)}{P(a)P(b)}\right)$$

onde "a" é a primeira palavra em uma sequência de duas palavras, "b" é a segunda palavra em uma sequência de duas palavras e "prob" é a probabilidade de uma palavra no corpus. Pares de palavras que tiveram um $PMI > 0$ foram considerados como bigramas. Os trigramas foram computados pelo cálculo da média PMI Para cada relação da sequência de 3 palavras. Esta técnica produziu: 6141 trigramas e 6491 bigramas. O bigrama e o trigrama estão localizados na pasta *Multi-Word Expressions* e estão disponíveis como: Arquivos de texto e formato de "pickle" em Python (Dicionários). Exemplos de MWEs extraídos com este método estão documentados na Tabela 4.

Bigramas
aparelhos celulares, principal adversário, laudo técnico, menor disponibilidade, tão difícil, investimento social, maior processadora, momento oportuno, agências internacionais, jogadas ofensivas, clubes participantes, primeira greve
Trigramas
contra a corrupção, dados foram divulgados, postos de combustíveis, investiga um esquema, abriu as portas, mês passado foi, plantio de mudas, área de educação, reduziu sua estimativa

Tabela 4: Amostra de MWE Extraído com Associação Estatística.

Co-ocorrência de palavras

Co-ocorrência é outra técnica a partir da qual os MWEs podem ser detectados. As palavras podem ser representadas como vetores, onde os valores no vetor são pesos que representam co-ocorrência com outras palavras. Esta representação combinada com skip-gramas pode ser usada para identificar frases (Mikolov et al., 2013) dentro de um fluxo de unigramas.

Este corpus vem com dois modelos que permitem a detecção de bigramas ou trigramas. Os modelos foram gerados a partir de Gensim[7]. Os modelos estão localizados na pasta *Word Vectors* e estão disponíveis como um formato Python "pickle".

Delimitadores de frases

Delimitador de frase é a pontuação que delimita palavras ou frases. Esta técnica identifica pares de marcas de citação ou sinais de pontuação que delimitam palavras, bigramas ou trigramas. Suponha-se que esses delimitadores fossem utilizados por jornalistas para indicar frases específicas de "domínio". Esta técnica identificou 1026 palavras, bigramas ou trigramas.

[6] Tais como https://snowballstem.org/algorithms/portuguese/stop.txt

[7] http://radimrehurek.com/gensim/models/phrases.html

6 Treebank

Uma árvore de dependência é uma forma de representação de dependências léxicas entre palavras e/ou frases. Uma coleção de árvores de dependência é conhecida como *treebank*. São relativamente poucos os *treebanks* portugueses quando comparados com o inglês. A mais conhecida *treebank* portuguesa é "Floresta" (Afonso et al., 2002).

Árvores de dependência têm sido usadas em tarefas comuns de processamento de língua natural (Qiu et al., 2009), tais como extração de relação causal (Khoo et al., 2000), área de pesquisa que a liberação deste corpus se destina a incentivar.

O *treebank* fornecido com este corpus consiste de 27931 sentenças que foram selecionadas aleatoriamente e analisadas com o analisador *LX-Dependency* (Rodrigues et al., 2014) cuja saída está em conformidade com a do analisador de *Stanford (Stanford Parser)*.

Em termos de avaliação do analisador *LX-Dependency*, o mesmo possui o UAS (*Unlabeled Attachment Score*) de 94,42 e a sua LAS (*Label Attachment Score*) é de 91,23 (Silva et al., 2010).

Uma saída típica do analisador é a seguinte:

```
(ROOT (S (NP (N' (N' (N Produção) (A
global)) (PP (P de) (NP (N açúcar)))))
(VP (V deve) (VP (V crescer) (PP (P
para) (NP (N' (N 165,1) (N' (N milhões)
(PP (P de) (NP (N toneladas)))))))))))
```

As dependências representadas por esta saída são apresentados na Figura 6.

7 Recursos de Relações entre Palavras

Este corpus contém modelos que podem ajudar na detecção de relações entre palavras ou frases. Os recursos liberados são métodos estatísticos, que são Vetores de palavras e Modelagem de tópicos; Estes modelos foram gerados com a biblioteca Gensim Python. Os recursos estão localizados nas pastas *Word Vector* e *Topic Resources*, respectivamente.

Vetores de Palavras

A representação de vetor de palavra é uma representação que trata palavras como vetores. Os vetores representam a co-ocorrência de uma determinada palavra com outras palavras no vocabulário. A frequência de co-ocorrência é representada como um peso. Os vetores são sistemas de coordenadas, portanto as semelhanças entre os vetores de palavras podem ser representadas como um ângulo. Isso permite o uso de medidas de similaridade como a similaridade de Cosseno para calcular a semelhança semântica entre as palavras.

O corpus tem um modelo de vetor de palavras que foi treinado a partir da informação no corpus. Para ilustrar a capacidade do modelo de vetor de palavras para identificar palavras relacionadas, um simples experimento foi conduzido para calcular o vizinho mais próximo com uma pequena seleção de palavras. As pontuações de similaridade foram computadas usando as chamadas de função Gensim[8]. A faixa de pontuação possível foi $0.0 \leq s \leq 1.00$, onde 1 com o maior índice de similaridade e 0 o menor. Os resultados são apresentados na Tabela 5. Os resultados mostram claramente que os pares de palavras com alta pontuação tinham similaridade, no entanto, os pares de palavras com as pontuações mais baixas não tinham relações óbvias. Os recursos de vetores de palavras estão localizados: */Data/News Stories/Topic Resources/Word Vectors/*.

Palavra	Palavras mais próximas	Palavras mais distantes
Etanol	Biocombustível (0.85), Álcool hidratado (0.84), Combustível (0.81), Álcool (0.87), Álcool anidro (0.81)	Vice-liderança(-0.47), Limão (-0.55), Sábado (-0.48), Rocher (-0.48)
Milho	Trigo (0.83), Soja (0.88), Grão de bico (0.84), Algodão (0.84)	Jogador Real (-0.46), Atenção (-0.45), Bonito (-0.44), Frutal, MG (-0.46)
Gasolina	Diesel (0.79), Combustível (0.81), Álcool (0.80)	Eroles (-0.46), PM (-0.42), Exultos (-0.42), Titã (-0.42)
Chuva	Tempestades (0.75), Sopros (0.78), Nuvens (0.74), Chuva (0.73), Isolado (0.74)	Discrepante (-0.39), Estradas (-0.39), T.M. (-0.36)

Tabela 5: Palavras com vizinhos mais próximos e mais distantes.

Os experimentos foram repetidos para verbos causais. Os verbos causais são verbos que descrevem uma relação causal entre eventos de causa e efeito. Os resultados para a experimento do verbo causal são demonstrados na Tabela 6. Os resultados mostram claramente que os vizinhos mais próximos têm propriedades causais. Isso tem implicações para a extração de relação causal, já que no momento da escrita não havia uma estratégia de extração de relação causal publicada que usasse vetores de palavras.

[8] https://radimrehurek.com/gensim/models/word2vec.html

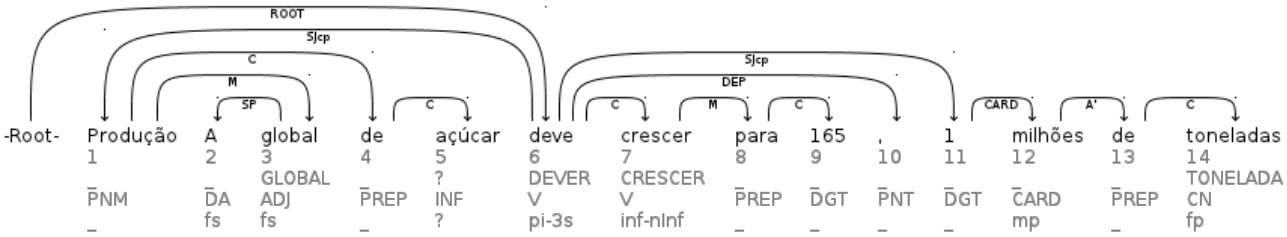

Figura 6: Dependências Léxicas

Verbo	Palavras mais próximas
causar	provocar (0.83), causam (0.67), sofrer (0.68), afetar (0.65), provoca (0.63)
afetar	prejudicar (0.85), comprometer (0.74), favorecer (0.73)
provocar	causar (0.82), gerar (0.71), sofrer (0.74)
causam	provoca (0.70), provocam (0.83)
provocam	causam (0.84), provoca (0.73)

Tabela 6: Verbos causais e seus vizinhos mais próximos.

Os vetores de palavras também podem ser usados para identificar frases semelhantes a uma frase de origem. A biblioteca Gensim fornece uma função de similaridade para n-gramas, que foi utilizada nos experimentos de bigramas e trigramas conduzidos neste artigo.

Os experimentos de bigramas usaram os seguintes bigramas de fonte selecionados aleatoriamente: Aparelhos celulares; Maior processadora; Dilma Rousseff; Receita bruta. A partir da qual foram calculados os bigramas mais próximos. Os resultados estão documentados na Tabela 7.

Bigramas	Mais próximos
aparelhos celulares	telefones móveis, canais eletrônicos, aparelhos eletrônicos, paredes celulares, equipamentos eletrônicos, caixas eletrônicos
maior processadora	maior importadora , maior produtora, maior produção, maior trading , maior exportadora, maior comercializadora, produção maior
Dilma Rousseff	Michel Temer, possível impeachment, eventual afastamento, recém-eleito presidente
receita bruta	captação líquida, dívida líquida, renda líquida, margem líquida

Tabela 7: Bigramas frequentes e seus vizinhos mais próximos.

O experimento do trigrama selecionou aleatoriamente trigramas e calculou seus vizinhos mais próximos. A técnica utilizada foi idêntica à utilizada para o experimento com bigrama. Os trigramas para esta experimento foram: Ministério da Cultura; Moagem de cana; Cultivares de soja. Os resultados estão descritos na Tabela 8.

Trigramas	Mais próximos
Ministério da Cultura	secretário-executivo do Ministério , Secretaria da Educação, ministro da Educação, Secretaria da Fazenda
moagem de cana	toneladas de cana-de-açúcar volume de moagem, safra de cana-de-açúcar, oferta de cana-de-açúcar, produção de cana-de-açúcar, capacidade de moagem
cultivares de soja	plantio de milho lavouras de milho lavouras de café

Tabela 8: Trigramas frequentes e seus vizinhos mais próximos.

Os experimentos com múltiplas palavras mostram que, embora os n-gramas mais próximos fossem compostos de sinônimos semelhantes semânticos, embora houvesse alguns erros óbvios. Exemplos de erros:

1. Aparelhos celulares e paredes celulares.
2. Aparelhos celulares e caixas eletrônicos.

Apesar dos erros, é claro que os experimentos retornam informações semânticas semelhantes nos n-gramas.

Modelagem de Tópicos

A modelagem de tópicos é um método não-supervisionado para agrupar palavras que co-ocorrem no mesmo tópico. A modelagem de tópicos pode ser usada para calcular semelhanças entre: frases e documentos.

Este corpus contém um número de modelos pré-treinados, bem como a distribuição de

tópicos pré-computados para cada documento no corpus. Os modelos pré-treinados têm uma série de variações de hiper parâmetros. As duas principais variáveis são: técnica de amostragem estatística *Latent Dirichlet Allocation* (LDA) ou *Latent Semantic Indexing* (LSI) (Blei et al., 2003) e 2. número de tópicos. Existem 5 modelos que usam LDA. Os modelos usam uma variedade de tópicos na faixa $500 \leq s \leq 2500$. O número de tópicos é incrementado em 500 para cada incremento do modelo. O modelo LSI tem um número de tópicos de 2000, o número de tópicos foi determinado pelo trabalho realizado por (Drury et al., 2015).

8 Informações de Nível de Documento

Informações de nível de documento no contexto deste artigo são aquelas que descrevem informações contidas em um documento individual. Existem 4 tipos de informações do documento: Distribuição do tópico; Orientação do sentimento; Número do grupo; e Frases de causa.

Os recursos estão localizados respectivamente nas pastas *Topic Resources*, *Sentiment*, *Clusters* e *Cause Phrases*.

Distribuição do Tópico

As informações do documento de distribuição de tópicos estão contidas em um arquivo de texto. Cada linha dentro do arquivo de texto representa um único documento. Cada linha contém o nome do documento e uma coleção de números de tópicos com uma probabilidade. O separador entre o número do tópico e sua probabilidade é um espaço, e o separador entre o número de tópicos e os pares de probabilidade é uma tabulação. A distribuição de probabilidade foi calculada com LDA e 2000 tópicos. Estes valores foram derivados do trabalho realizado por Drury et al. (2015).

Orientação do Sentimento

A orientação do sentimento para um documento foi alcançada contando o número de palavras com uma orientação sentimental. As palavras com uma orientação do sentimento neste caso são palavras com uma orientação positiva ou negativa do sentimento. As palavras com uma orientação neutra são ignoradas porque dominariam o documento. O cálculo pode ser representado:

$$S = freq(W_p) - freq(W_n),$$

onde $freq$ é a frequência de palavras com uma determinada orientação de sentimento,W_p são palavras com uma orientação positiva, W_n são palavras com orientação negativa e S é a orientação do sentimento. Documentos com uma pontuação de: 1. $S < 0$ recebem uma orientação negativa, 2. $S > 0$ recebem uma orientação positiva e 3. $S = 0$ recebem uma orientação neutra. O recurso é um arquivo de dicionário "pickled". O arquivo contém: a localização relativa de um documento, nome do arquivo e orientação de sentimento. Os valores das chaves são o local do arquivo e os valores são a orientação do sentimento.

Agrupamento

Documentos relacionados podem ser detectados por um processo de agrupamento. O processo de agrupamento para este corpus foi conseguido usando K-means, e a distribuição tópica acima mencionada. K foi ajustado para 200 usando Davies Bouldin Index (DBI) para calcular a "qualidade" de várias configurações de agrupamento. A medida de distância que foi usada para computar os agrupamentos foi a distribuição de tópicos de cada documento.

Os *clusters* e seus documentos componentes são fornecidos em um formato de dicionário "pickled". A chave é um número de cluster nominal e o valor são os documentos. Para ilustrar a semelhança de documentos que fazem parte do mesmo cluster são apresentados na Tabela 9.Os documentos contêm o mesmo tema da predição de colheita. O uso de tópicos em vez de semelhança de palavras produziu clusters que contêm o mesmo tema, ao invés da mesma palavra.

Documento 1	**Documento 2**
As usinas e destilarias do Centro-Sul do Brasil dão início nesta sexta, dia 1º de abril, a mais uma safra de cana-de-açúcar, com perspectivas favoráveis. A principal região produtora do país irá processar em 2016/2017 619,37 milhões de toneladas de cana (+2,3%).	A Organização Internacional do Café (OIC), em sua primeira estimava para a produção mundial no ano-safra 2015/2016, prevê colheita de 143,4 milhões de sacas de 60 kg, indicando um aumento modesto de 1,4% em relação ao ano-safra de 2014/2015 (141,4 milhões).....

Tabela 9: Fragmentos de texto dos documentos no mesmo grupo (*cluster*).

Relações Causais

Os documentos anotados fornecem uma relação de causa anotada, mas para extrair todas as relações de causa pode ser uma tarefa onerosa. O

corpus fornece uma lista de relações de causa pré-extraídas. A relação de causa é um arquivo delimitado por tabulação que representa a relação de causa como um triplo:

1. Evento de causa.
2. Ligação causal.
3. Evento de efeito.

Cada triplo tem um nome de documento que é o documento onde reside a relação causal. As palavras de parada *(stop-words)* foram removidas das relações causais. Uma amostra de relações causais pode ser encontrada na Tabela 10.

Relações Causais
governo aumente etanol anidro gasolina
clima seco produzidas milhoes toneladas acucar
taxa declinio diminuído levantando expectativas setor
chuvas últimos causa máquinas conseguem entrar lavoura

Tabela 10: Amostra de Relações Causais

9 Licenciamento

Este corpus é lançado sob a *Creative Commons License (4.0)* (https://wiki.creativecommons.org/wiki/Text).
É intenção dos autores que este corpus seja utilizado em sua amplitude, consequentemente esta licença foi escolhida porque permite o uso comercial e de redistribuição.

Este corpus se qualifica para a liberação de acordo com a legislação de uso justo[9] porque: é transformador, e nenhum ganho monetário será exigido para sua liberação.

10 Trabalhos Futuros

Pretende-se em trabalhos futuros considerar a avaliação de outras ferramentas que realizam detecção de entidades nomeadas, assim como outras formas de detecção de expressão multi-palavras, considerando o uso de opções como: OpenNLP, FreeLing, PALAVRAS e etc. Aplicar anotações baseadas em XML em relações causais que apresentam estruturas fracas. Além disso, vamos considerar alternativas abertas ao LX-Dependency, como por exemplo o UDPortugueseBR[10]

11 Conclusão

Este artigo descreve um corpus português-brasileiro que contém notícias relacionadas a agricultura. Essas notícias têm anotações causais e sentimentais relacionadas a informações temporais, bem como anotações de entidades nomeadas. O corpus contém recursos de linguagem, tais como: árvores de dependência, modelos de tópicos e modelos de vetor de palavras, bem como meta-informações, como distribuição de tópicos. Além disso, contém informações sobre o nível do documento, como distribuição de tópicos e informações sobre o sentimento.

Este recurso que acreditamos ser único e substancial, foi liberado para incentivar pesquisas de mineração de texto no campo da agricultura, bem como pesquisas em áreas relacionadas, como relação de causalidade e extração de conhecimento.

Agradecimentos

Esta pesquisa teve apoio financeiro das agências brasileiras: FAPESP (processos 15/14228-9 e 11/20451-1) e CNPq (processo 302645/2015-2). Somos gratos aos árbitros pelos comentários e sugestões no desenvolvimento deste trabalho.

Referências

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002. Floresta sintá (c) tica: A treebank for portuguese. Em *International Conference on Language Resources and Evaluation (LREC)*, 1698–1703.

Altenberg, Bengt. 1984. Causal linking in spoken and written English. *Studia Linguistica* 38(1). 20–69.

Alvim, Leandro, Paula Vilela, Eduardo Motta & Ruy Luiz Milidiú. 2010. Sentiment of financial news: a natural language processing approach. Em *1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology*, edição online.

Balage Filho, Pedro P., Thiago A. S. Pardo & Sandra M. Alusio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. Em *9th Brazilian Symposium*

[9] https://www.copyright.gov/fair-use/more-info.html

[10] https://github.com/UniversalDependencies/UD_Portuguese-BR

in Information and Human Language Technology (STIL), 215–219.

Bethard, Steven. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. Em *Second Joint Conference on Lexical and Computational Semantics (SEM)*, 10–14.

Bittar, André. 2010. *Building a TimeBank for French: a reference corpus annotated according to the ISO-TimeML standard*: Paris 7. Tese de Doutoramento.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3. 993–1022.

Carvalho, Paula, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas & Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 11–31. Linguateca.

Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta & Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. Em *5th Linguistic Annotation Workshop*, 143–151.

Chaves, Marcírio Silveira, Larissa A. de Freitas, Marlo Souza & Renata Vieira. 2012. Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. Em *International Conference on Application of Natural Language to Information Systems*, 296–301.

Drury, Brett & Alneu de Andrade Lopes. 2014. A comparison of the effect of feature selection and balancing strategies upon the sentiment classification of Portuguese news stories. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 413–417.

Drury, Brett & Alneu de Andrade Lopes. 2015. The identification of indicators of sentiment using a multi-view self-training algorithm. *Oslo Studies in Language* 7.

Drury, Brett, Jorge Carlos Valverde-Rebaza & Alneu de Andrade Lopes. 2015. Causation generalization through the identification of equivalent nodes in causal sparse graphs constructed from text using node similarity strategies. Em *International Symposium on Information Management and Big Data*, 58–65.

Fonseca, Erick R. & João Luís G. Rosa. 2013. A two-step convolutional neural network approach for semantic role labeling. Em *International Joint Conference on Neural Networks*, 2955–2961.

Forascu, Corina & Dan Tufiș. 2012. Romanian TimeBank: An annotated parallel corpus for temporal information. Em *Eight International Conference on Language Resources and Evaluation (LREC)*, 3762–3766.

Freitas, Cláudia, Eduardo Motta, R. Milidiú & Juliana César. 2012. Vampiro que brilha... rá! desafios na anotaçao de opinião em um corpus de resenhas de livros. Em *XI Encontro de Linguística de Corpus*, s/p.

Freitas, Larissa A. & Renata Vieira. 2013. Ontology based feature level opinion mining for Portuguese reviews. Em *22nd International Conference on World Wide Web (WWW)*, 367–370.

Jung, Hyuckchul & Amanda Stent. 2013. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. Em *Second Joint Conference on Lexical and Computational Semantics (SEM)*, 20–24.

Khoo, Christopher, Syin Chan & Yun Niu. 2002. The many facets of the cause-effect relation. Em Rebecca Green, Carol A. Bean & SungHyon Myaeng (eds.), *The Semantics of Relationships*, vol. 3 Information Science and Knowledge Management, 51–70. Springer.

Khoo, Christopher S. G., Syin Chan & Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. Em *38th Annual Meeting on Association for Computational Linguistics*, 336–343.

Kolya, Anup Kumar, Amitava Kundu, Rajdeep Gupta, Asif Ekbal & Sivaji Bandyopadhyay. 2013. JU_CSE: A CRF based approach to annotation of temporal expression, event and temporal relations. Em *Second Joint Conference on Lexical and Computational Semantics (SEM)*, 64–72.

Laokulrat, Natsuda, Makoto Miwa, Yoshimasa Tsuruoka & Takashi Chikayama. 2013. Uttime: Temporal relation classification using deep syntactic features. Em *Second Joint Conference on Lexical and Computational Semantics (SEM)*, 88–92.

Liu, Bing & Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. Em Charu C. Aggarwal (ed.), *Mining text data*, 415–463. Springer.

Llorens, Hector, Estela Saquete & Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. Em *5th International Workshop on Semantic Evaluation (SemEval)*, 284–291.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Advances in neural information processing systems*, 3111–3119.

Pustejovsky, James, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer & Graham Katz. 2003a. TimeML: robust specification of event and temporal expressions in text. Em Mark T. Maybury (ed.), *New directions in question answering*, 28–34. AAAI Press.

Pustejovsky, James, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro & Marcia Lazo. 2003b. The TIMEBANK corpus. Em *Corpus linguistics*, 647–656.

Qiu, Guang, Bing Liu, Jiajun Bu & Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. Em *International Joing Conference on Artificial Intelligence (IJCAI)*, vol. 9, 1199–1204.

Rodrigues, João, Francisco Costa, João Silva & António Branco. 2014. Automatic syllabification of portuguese. *Encontro Anual da Associação Portuguesa de Linguística* 715–720.

Shams-Eddien, Katrin. 2002. *Beth Levin's English verbs classes and alternations*. Free University of Berlin.

Silva, Joao, António Branco, Sérgio Castro & Ruben Reis. 2010. Out-of-the-box robust parsing of Portuguese. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 75–85.

Silva, Mário J., Paula Carvalho & Luís Sarmento. 2012. Building a sentiment lexicon for social judgment mining. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 218–228.

Silva, Mário J., Paula Carvalho, Luís Sarmento, Pedro Magalhães & Eugénio Oliveira. 2009. The design of OPTIMISM, an opinion mining system for Portuguese politics. Em *New trends in artificial intelligence: Proceedings of EPIA*, 12–15.

Souza, Marlo, Renata Vieira, Débora Busetti, Rove Chishman & Isa Mara Alves. 2011. Construction of a Portuguese opinion lexicon from multiple resources. Em *8th Brazilian Symposium in Information and Human Language Technology*, 59–66.

Strötgen, Jannik & Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. Em *5th International Workshop on Semantic Evaluation*, 321–324.

Thomson, Judith Jarvis. 1987. Verbs of action. *Synthese* 72(1). 103–122.

UzZaman, Naushad & James F. Allen. 2010. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. Em *5th International Workshop on Semantic Evaluation (SemEval)*, 276–283.

Vendler, Zeno. 1967. Causal relations. *The Journal of Philosophy* 64(21). 704–713.

http://www.linguamatica.com/

linguaMATICA