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.