Proposta recebida em Novembro 2018 e aceite para publicação em Dezembro 2018.

Análise da capacidade de identificação de paráfrase em ferramentas de resolução de correferência

Analyzing paraphrase identification capabilities in coreference resolution tools

Bernardo Scapini Consoli
Pontifícia Universidade Católica do Rio Grande do Sul
bernardo.consoli@acad.pucrs.br

Joaquim Francisco dos Santos Neto
Pontifícia Universidade Católica do Rio Grande do Sul
joaquim.santos@acad.pucrs.br

Sandra Collovini de Abreu
Pontifícia Universidade Católica do Rio Grande do Sul
sandra.abreu@acad.pucrs.br

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

Resumo

Os fenômenos linguísticos de correferência e paráfrase compartilham certos aspectos. É comum, por exemplo, referir-se a uma mesma entidade de maneiras diferentes em um mesmo contexto, assim, a resolução de correferências pode auxiliar o processo de identificação de paráfrases. Este artigo apresenta uma análise das capacidades da ferramenta de resolução de correferência CORP, para Português, no contexto de identificação de paráfrases nos níveis de sentença e de sintagma.

Palavras chave

resolução de correferência, identificação de paráfrase

Abstract

The linguistic phenomena known as coreference and paraphrasing share certain aspects among themselves. It is common, for example, to refer to an entity in different ways within the same context, and as such the resolution of such coreferent mentions may be of aid to the process of identifying paraphrases. This paper presents an analysis of the capabilities of the coreference resolution tool CORP, created for use with the Portuguese language, within the context of paraphrase identification in the sentence and noun phrase levels.

Keywords

coreference resolution, paraphrase identification

1 Introdução

Aplicações de PLN que lidam com paráfrase e correferência tem o potencial de melhorar o entendimento e a geração de sistemas. Extração de paráfrases e resolução de correferência podem ser aplicadas nas tarefas de perguntas e respostas, extração de informação, tradução automática, entre outras (Recasens & Vila, 2010).

Paráfrase é definida como a relação entre duas expressões que possuem o mesmo sentido, enquanto correferência é definida como a relação entre duas expressões que possuem o mesmo referente no mesmo contexto (Jurafsky & Martin, 2009). Pares parafrásicos podem ser correferentes e vice-versa (Recasens & Vila, 2010). A relação entre correferência e paráfrase implica uma possibilidade da utilização de resolução de correferência para facilitar a identificação de paráfrases.

Contudo, correferência é um fenômeno linguístico muito mais dependente em contexto do que a paráfrase (Recasens & Vila, 2010). No exemplo: "Ana foi ao cardiologista entregar alguns exames. O médico pediu a ela para retornar na próxima semana". Podemos ver que [cardiologista] e [médico] são sintagmas correferentes e parafrásicos, enquanto o nome próprio [Ana] e o pronome pessoal [ela] são somente correferentes, pois possuem uma referência mas não um significado intrínseco (Recasens & Vila, 2010). Deste modo, é necessário distinguir quais menções de uma cadeia de correferências são parafrásicas.

DOI: 10.21814/lm.10.2.288

Este trabalho visa analisar a capacidade da ferramenta de resolução de correferência para o Português CORP descrita em (Fonseca et al., 2017) no contexto da área de identificação de paráfrases. Duas análises foram realizadas: a primeira estuda os tipos de sintagmas parafrásicos encontrados pelo CORP, comparando-os com resultados obtidos pela função de resolução de correferência para o Inglês da ferramenta Stanford CoreNLP apresentada em (Manning et al., 2014). Essa análise foi feita sobre um subconjunto de textos da revista Pesquisa FAPESP[1]; a segunda é uma análise da performance do CORP sobre um corpus de sentenças parafrásicas descrito em (Fonseca et al., 2016), visando identificar sintagmas nominais correferentes que possam servir como âncoras para um subsequente processo de identificação parafrásica.

O restante deste trabalho está organizado nas seguintes seções: a Seção 2 descreve o referencial teórico juntamente com trabalhos relacionados; a Seção 3 apresenta os recursos utilizados; a Seção 4 descreve a avaliação realizada; e por fim, na Seção 5 as considerações finais são apresentadas.

2 Correferência e Paráfrase

As ferramentas computacionais para resolução de correferência lidam com a tarefa de identificar as expressões textuais associadas a entidades ou eventos do mundo real. Recasens & Vila (2010) destacam que paráfrase e correferência são geralmente definidos como relações de similaridade, ou seja, dadas duas expressões que têm o mesmo significado, estas são parafrásicas; e dadas duas expressões referentes à mesma entidade em um discurso, estas são correferentes. Partindo desse princípio, quando usamos uma ferramenta de resolução de correferência e dela obtemos suas respectivas cadeias, podemos obter paráfrases nas cadeias identificadas.

Para um melhor entendimento das relações entre paráfrase, correferência e sistemas de resolução de correferência, considere o seguinte fragmento de texto[2], processado pelo sistema de resolução de correferência (CORP) (Fonseca et al., 2017):

"[...] Para **[Luiz_Eugênio_Mello [258]]**, vice-presidente de a Associação_Nacional_de_ Pesquisa_e_Desenvolvimento_das_Empresas_Inovadoras (Anpei) [...] Para **[Luiz_Mello [258]]**, há uma baixa intensidade de P&D mesmo entre empresas líderes [...]. Já as 10 mais em a Espanha, que foram HP, Airbus, Ericsson, CSIC, Fractus, Gamesa, Vodafone, Laboratórios_Dr._Esteve, Intel e Telefonica, depositaram 739 patentes em os Estados_Unidos, 88 % a mais, diz **[Mello [258]]**. [...] **[Luiz_Eugênio_Mello [258]]** também critica a dificuldade de trabalhar com prioridades. [...]"

Os termos destacados em negrito foram identificados como correferentes e categorizados com a categoria Pessoa, como ilustra a Tabela 1.

Cadeia Pessoa
[Luiz Eugênio Mello](2)
[Luiz Mello](1)
[Mello](1)

Tabela 1: Cadeia Pessoa do fragmento textual.

2.1 Trabalhos Relacionados

O estudo de identificação de paráfrase de Shinyama & Sekine (2003) utiliza Reconhecimento de Entidades Nomeadas para encontrar o que chamam de "âncoras", expressões que provavelmente não mudariam em frases parafrásicas (nomes de pessoas, datas, entre outros). De acordo com a definição de paráfrase dada previamente, sentenças parafrásicas devem possuir sintagmas nominais que se referem às mesmas entidades. Se identificadas as relações de correferência entre os sintagmas das sentenças, é possível removê-los do processo de identificação de paráfrase, substituindo-os por uma âncora. Um exemplo é a entidade nomeada [Presidente Temer] sendo mencionado em outras frases como [Michel Temer] ou até mesmo através de um pronome, no caso [ele].

O estudo de Regneri & Wang (2012) utilizou informações de estrutura e de contexto dos documentos analisados para auxiliar na coleta de sentenças parafrásicas. A resolução de correferência foi utilizada para adicionar mais informações de contexto à análise. Especificamente, a correferência incluiu a informação de quais partes das sentenças possuem o mesmo sentido no contexto analisado.

Nos estudos apresentados, por mais que os sintagmas em si não sejam parafrásicos, como no caso do pronome que por natureza não pode ser paráfrase de um nome, a relação de correferência entre um nome e um pronome ajuda a identificar sentenças parafrásicas.

Quanto a avaliação realizada no presente trabalho, a primeira análise é uma busca por sintagmas nominais parafrásicos encontrados pelo sistema CORP, com o objetivo de analisar os pa-

[1] http://revistapesquisa.fapesp.br/

[2] http://revistapesquisa.fapesp.br/2017/06/19/financiamento-em-crise/

drões em que estes sintagmas se encaixam, enquanto a segunda é um estudo de como o CORP poderia ser utilizado como um auxiliador para um sistema de identificação de paráfrases.

3 Recursos

Nesta Seção, nós descrevemos as ferramentas de correferência CORP e Stanford CoreNLP, que tratam as línguas Portuguesa e Inglesa, respectivamente, bem como o corpus ASSIN 2016 de sentenças parafrásicas.

CORP.
CORP é um recurso para a resolução de correferência em Português descrito por Fonseca et al. (2017). O CORP utiliza um conjunto de regras sintáticas e semânticas, propostas por Fonseca (2018), para decidir se dois sintagmas nominais (nomes próprios ou comuns) são correferentes, bem como as informações de *Part-of-Speech* (PoS) e sintáticas provindas da ferramenta Cogroo (Silva, 2013). Para um melhor entendimento, na Tabela 2 temos exemplos de cadeias de correferência de um texto da revista Pesquisa FAPESP[3]. A primeira coluna indica a categoria da cadeia (Sarmento et al., 2006) e na coluna seguinte as cadeias de correferência com a respectiva frequência de cada menção. Podemos notar que o primeiro exemplo indica a categoria Pessoa em que temos as diferentes menções que designam [Carlos Américo Pacheco]. Já para Organização/Local temos a cadeia referindo-se ao [Brasil], e por fim temos a cadeia com diferentes menções da [Universidade Estadual de Campinas] por meio do acrônimo [Unicamp].

Stanford CoreNLP.
Stanford CoreNLP (Manning et al., 2014) provê um conjunto de ferramentas de tecnologia de linguagem humana, incluindo Reconhecimento de Entidades Nomeadas, identificação de dependências sintáticas e resolução de correferências. Para este trabalho, utilizamos suas capacidades de resolução de correferência na versão determinística, que é baseada em regras sintáticas. O sistema determinístico foi utilizado para melhor comparar com o CORP, que também é um sistema determinístico. Para exemplificar, a Tabela 3 ilustra exemplos de cadeias de correferência de um texto da FAPESP[4]. A primeira coluna ilustra as categorias tratadas pelo Stanford CoreNLP, nas colunas seguintes temos as cadeias de correferência com a respectiva frequência de cada menção. Destaca-se as diferentes menções na cadeia de [Carlos Américo Pacheco] (Pessoa), a qual inclui o pronome pessoal [he].

Corpus ASSIN 2016.
O corpus ASSIN 2016[5] descrito por Fonseca et al. (2016) foi construído para as tarefas ASSIN da conferência PROPOR 2016. Este corpus possui 10.000 pares de sentenças parafrásicas anotados para grau de similaridade semântica e inferência textual. A Figura 1 demonstra um par do corpus ASSIN 2016, na linguagem XML. *Pair similarity* é a média da medida de similaridade semântica textual selecionada por 4 anotadores; o *ID* é o número identificador do par; *entailment* é a classe de inferência textual dada por anotadores; t e h são as sentenças e indicam qual delas é o texto e a hipótese para objetivos de inferência textual.

4 Avaliação

Duas análises foram realizadas sobre a capacidade do CORP de auxiliar na tarefa de identificação de paráfrase: a primeira trata de estudar a relação entre o fenômeno da correferência e o fenômeno da paráfrase no contexto de sintagmas nominais, subsequentemente realizando uma análise para discernir os padrões em que se encaixam os sintagmas nominais parafrásicos encontrados pelo CORP; a segunda trata de sua capacidade de encontrar relações de correferência que auxiliem em uma potencial tarefa de identificação de sentenças parafrásicas.

Análise 1.
Para a primeira análise foram utilizados 10 textos paralelos em Inglês e Português retirados da revista Pesquisa FAPESP. O método de identificação de padrões parafrásicos foi dividido em 2 passos:

1. O CORP e o Stanford CoreNLP são utilizados para extrair automaticamente as cadeias de correferência nos 10 textos paralelos em Inglês e Português da revista Pesquisa FAPESP.
2. Os sintagmas extraídos são manualmente analisados e classificados em padrões de acordo com características identificáveis.

Considerando que o Stanford CoreNLP é uma das melhores ferramentas disponível para a resolução de correferências, decidimos utilizar os

[3] http://revistapesquisa.fapesp.br/2017/06/19/financiamento-em-crise/
[4] http://revistapesquisa.fapesp.br/en/2017/12/10/funding-in-crisis/
[5] http://nilc.icmc.usp.br/assin/

Categorias	Menções (Frequência Individual)
Pessoa	[Carlos Américo Pacheco, professor de o Instituto de Economia de a Universidade Estadual de Campinas] (1) [Carlos Américo Pacheco] (2)
Organização/Local	[o Brasil] (11) [O Brasil] (3) [a Universidade Estadual de Campinas] (1) Unicamp (1)

Tabela 2: Cadeias de Correferência - CORP

Categorias	Menções (Frequência Individual)
Pessoa	[Carlos Américo Pacheco, a professor at the Institute of Economics of the University of Campinas (Unicamp)] (1) [Carlos Américo Pacheco] (4) [a professor at the Institute of Economics of the University of Campinas (Unicamp)] (1) [Pacheco, Chief Executive of the FAPESP Executive Board] (1) [Pacheco] (1) [he] (2)
País	[Brazil] (17) [Brazil,that filed the most patent applications in the U.S.] (1) [Brazil, which has become increasingly more complex in recent decades] (1)

Tabela 3: Cadeias de Correferência - Stanford

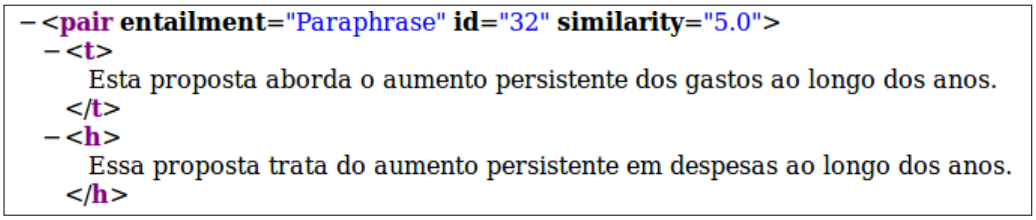

Figura 1: Amostra do corpus ASSIN 2016

resultados do toolkit como uma referência à qual podemos comparar os resultados encontrados pelo CORP. Cabe salientar que as ferramentas utilizadas possuem diferenças na identificação dos sintagmas nominais extraídos (menções) e no conjunto de categorias tratadas.

Como resultado da avaliação temos 3 padrões: Acrônimos, Nomes Próprios e Nomes, os quais são ilustrados nas Tabelas 4 e 5.

A análise dos sintagmas nominais parafrásicos identificados pelo Stanford CoreNLP mostrou que alguns padrões são melhor identificados do que outros. Exemplos de cadeias de correferência contendo menções parafrásicas de cada padrão são apresentados na Tabela 4. Podemos observar que a ferramenta conseguiu tratar poucos casos de sintagmas nominais parafrásicos envolvendo acrônimos, como por exemplo, [United States] - [US]. Em geral, o padrão de Nomes Próprios (ou Entidades Nomeadas) conseguiu identificar eficientemente as paráfrases a partir de cadeias bem completas, como por exemplo, na cadeia de Pessoa temos: [biochemist María Elena López of the Institute of Biological Sciences] - [López] - [María Elena López]. Podemos notar que na cadeia [O Rio de Janeiro] - [O Rio] a ferramenta não conseguiu identificar a categoria e classificou como Outro sendo a correta Local. O padrão de Nomes destaca-se por cadeias extensas contendo vários sintagmas parafrásicos, como no exemplo referindo-se a [gravitational wave] (Outro).

A avaliação do CORP com base nos 3 padrões identificados mostra que, em geral, o CORP consegue identificar eficientemente as paráfrases envolvendo acrônimos por meio das cadeias de correferência com diferentes menções para as categorias tratadas.

Padrão	Categoria	Menções (Frequência Individual)
Acrônimos	País	[United States] (1) [US] (1)
Nomes Próprios	Pessoa	[biochemist María Elena López of the Institute of Biological Sciences] (1) [López] (5) [María Elena López] (1)
	Outro	[o Rio de Janeiro] (2) [o Rio] (1)
Nomes	Outro	[another gravitational wave] (1) [this gravitational wave an identical instrument in Livingston, Louisiana] (1) [this gravitational wave] (1) [an identical instrument in Livingston , Louisiana] (1)

Tabela 4: Padrões de Paráfrases identificadas pelo Stanford CoreNLP.

Padrão	Categoria	Menções (Frequência Individual)
Acrônimos	Organização/Local	[o Instituto Brasileiro de Geografia e Estatística] (1) [IBGE] (1)
Nomes Próprios	Pessoa	[Gustavo Gomes] (1) [Gomes] (1)
	Organização/Local	[o Rio de Janeiro] (2) [o Rio] (1)
Nomes	Outro	[o debate] (1) [o desafio] (1)
	Comunicação	[um sinal sazonal , cujo pico máximo] (1) [O sinal] (1)

Tabela 5: Padrões de Paráfrases identificadas pelo CORP

Na Tabela 5 são ilustrados exemplos, como na cadeia [Instituto Brasileiro de Geografia e Estatística] - [IBGE] (Organização/Local). Um outro padrão são os nomes próprios (de Pessoas, Organização/Local, entre outros), como por exemplo, a cadeia [Gustavo Gomes] -[Gomes] (Pessoa) e a cadeia [o Rio de Janeiro] - [o Rio] (Organização/Local). Por fim, o padrão referente a Nomes refere-se aos sintagmas nominais que possuem o mesmo significado, como por exemplo, a cadeia de [o debate] - [o desafio] (Outro). O outro exemplo desse padrão tratou a categoria Comunicação em que as menções [um sinal sazonal, cujo pico máximo] e [O sinal] são parafrásicos. Nota-se que o primeiro sintagma nominal teve problemas na sua identificação por parte do Cogroo.

Análise 2.
Para a segunda análise foram utilizados 116 pares de sentenças parafrásicas do corpus ASSIN 2016. O método proposto de identificação dos sintagmas nominais para auxiliar na identificação de paráfrase entre sentenças foi dividido em 3 passos:

1. Pares de sentenças parafrásicas são anotados manualmente para correferência.

2. O CORP é utilizado para anotar automaticamente os pares de sentenças.

3. A anotação automática é comparada à manual.

Dada a falta de um corpus paralelo entre o Português e o Inglês anotado para correferência, o Stanford CoreNLP não foi utilizado para comparação, e julgamos que uma tradução automática geraria ruído demais nos dados. Desta forma, esta análise verifica o desempenho do CORP na identificação dos sintagmas nominais correferentes que possam servir como âncoras para um subsequente processo de identificação de paráfrases. Para isso, os sintagmas nominais parafrásicos contidos nos 116 pares de sentenças parafrásicas do corpus ASSIN 2016 foram anotados manualmente. Um total de 339 sintagmas foram anotados e serviram de referência para a avaliação da performance do CORP na tarefa proposta.

Sentenças	Menções - Referência	Menções - CORP
O tremor também deixou quase 100 mortos em a Índia e China.	[O tremor] - [O terremoto] [quase 100 mortos] - [cerca de 100 vítimas fatais]	[O tremor] - [O terremoto] —
O terremoto fez também cerca de 100 vítimas fatais em a Índia e em a China.	[Índia] - [Índia] [China] - [China]	[Índia] - [China] - [China] —
Esta proposta lida com o persistente aumento em os gastos a o longo de os anos. Esta proposta enfrenta o persistente aumento de despesas por anos.	[Esta proposta] - [Esta proposta] [os gastos] - [despesas] [os anos] - [anos]	— [os gastos] - [despesas] [os anos] - [anos]
Cerca de 5 mil pessoas trabalham usando a plataforma Uber hoje em o Brasil. Atualmente , cerca de 5 mil profissionais atuam usando a plataforma Uber em o país.	[Cerca de 5 mil pessoas] - [Cerca de 5 mil profissionais] [a plataforma Uber] - [a plataforma Uber] [o Brasil] - [o país]	— [a plataforma Uber] - [a plataforma Uber] [o Brasil] - [o país]

Tabela 6: Sintagmas correferentes identificados pelo CORP.

Como resultado temos 152 acertos; uma taxa de Precisão de 67%; Abrangência de 43% e F-measure de 53%. Na Tabela 6 são ilustrados exemplos de sintagmas correferentes identificados pelo CORP.

Na primeira coluna temos pares de sentenças parafrásicas, na segunda temos a anotação manual dos sintagmas nominais, e na última, os sintagmas nominais extraídos pelo CORP. Podemos observar que o CORP conseguiu identificar corretamente sintagmas parafrásicos, como por exemplo, nas cadeias [os gastos] - [as despesas]; [o tremor] - [o terremoto]; [o Brasil] - [o País]. Entretanto, como o CORP desconsidera sintagmas com dados numéricos, não foram identificados sintagmas contendo quantidades, como por exemplo, nas cadeias [quase 100 mortos] - [cerca de 100 vítimas fatais]; [cerca de 5000 pessoas] - [cerca de 5000 profissionais]. Além disso, ocorreram casos em que o CORP agrupou menções de cadeias diferentes em uma mesma cadeia, como no exemplo: [Índia] - [China] - [China].

5 Considerações Finais

Apresentamos neste trabalho as relações entre paráfrase e correferência, bem como uma avaliação da capacidade do CORP na identificação de paráfrases por meio de duas análises. A primeira análise resultou na identificação de três padrões de sintagmas correferentes que podem auxiliar na identificação de sintagmas parafrásicos. O CORP se mostrou capaz de identificar sintagmas correferentes e parafrásicos em textos do Português, destacando-se para os casos com acrônimos, como por exemplo, [o Estatuto de a Criança e de o Adolescente] - [ECA]. A segunda análise mostrou que o CORP auxilia na identificação de sintagmas nominais correferentes e parafrásicos em pares de sentenças parafrásicas. Uma das dificuldades da avaliação manual foi a delimitação dos sintagmas nominais, como por exemplo, no fragmento da sentença parafrásica "presidente Blatter não vai mais responder perguntas" o CORP identificou dois sintagmas: [presidente Blatter] e [Blatter]. Cabe ressaltar que a etapa de identificação dos sintagmas do CORP é provida pela ferramenta Cogroo.

Como trabalhos futuros, pretendemos tipificar/classificar os padrões para sintagmas parafrásicos propostos com auxilio de linguistas, e disponibilizar recursos para o Português anotados para correferência e paráfrases. Além disso, planejamos utilizar as cadeias de correferência extraídas pelo CORP para enriquecer uma análise de similaridade semântica deste mesmo corpus.

Agradecimentos

Agradecemos à PUCRS, CNPQ, CAPES e FAPERGS pelo seu apoio financeiro.

Referências

Fonseca, Erick, Leandro Borges dos Santos, Marcelo Criscuolo & Sandra Aluísio. 2016. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Fonseca, Evandro. 2018. *Resolução de correferência nominal usando semântica em língua portuguesa*. PUCRS: Programa de Pós-Graduação em Ciência da Computação, PUCRS. Tese de Doutoramento.

Fonseca, Evandro, Vinicius Sesti, André Antonitsch, Aline Vanin & Renata Vieira. 2017.

CORP: uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. *Linguamática* 9(1). 3–18.

Jurafsky, Daniel & James H. Martin. 2009. *Speech and language processing*. Prentice-Hall.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven 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.

Recasens, Marta & Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics* 36(4). 639–647.

Regneri, Michaela & Rui Wang. 2012. Using discourse information for paraphrase extraction. Em *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 916–927.

Sarmento, Luís, Ana Sofia Pinto & Luís Cabral. 2006. REPENTINO - a wide-scope gazetteer for entity recognition in Portuguese. Em *Computational Processing of the Portuguese Language*, 31–40.

Shinyama, Yusuke & Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. Em *Second International Workshop on Paraphrasing*, 65–71.

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