ANOTAÇÃO DE SENTIDOS DE VERBOS EM NOTÍCIAS JORNALÍSTICAS EM PORTUGUÊS DO BRASIL

[1]Marco A. Sobrevilla Cabezudo, [1]Erick G. Maziero, [2]Jackson W. C. Souza,
[1]Márcio S. Dias, [1]Paula C. F. Cardoso, [1]Pedro P. Balage Filho,
[1]Verônica Agostini, [1]Fernando A. A. Nobrega,
[3]Cláudia D. Barros, [2]Ariani Di Felippo, [1]Thiago A. S. Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
[2]Departamento de Letras, Universidade Federal de São Carlos
[3]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo

1. Introdução

A ambiguidade lexical é um fenômeno relevante no campo da semântica e sua resolução é importante em várias aplicações de Processamento de Linguagem Natural (PLN). A seguir, apresentamos um exemplo de ambiguidade lexical: "o banco *quebrou*". Nesse exemplo, vemos que o verbo "quebrar" poderia ter o significado de "falir financeiramente" ou "despedaçar". Em PLN, a Desambiguação Lexical de Sentido (DLS) (Jurafsky e Martins, 2009) é a tarefa que se ocupa de tratar a ambiguidade lexical e escolher o sentido mais adequado (dentre um conjunto de possíveis sentidos, disponível no que se é chamado de "repositório de sentidos") para uma palavra dentro de um contexto.

Uma etapa importante durante o desenvolvimento de métodos de DLS é a anotação de sentidos em um córpus, já que, a partir dela, pode-se desenvolver e treinar sistemas de DLS, fazer avaliações, e prover um recurso útil para futuras pesquisas em DLS, servindo, eventualmente, como um *benchmark* para a área. Para o português, tem-se poucos trabalhos de DLS e, portanto, existem poucos córpus com anotação de sentidos. Specia (2007) usou um córpus paralelo para implementar um método de DLS focado na tradução automática entre o português e o inglês. Machado et al. (2011) apresentaram um córpus formado por notícias jornalísticas em português extraídas de internet para a implementação de um método para desambiguação geográfica. Travanca (2013) focou nos verbos do córpus PAROLE (Ribeiro, 2013), que inclui textos sobre livros, jornais, periódicos e outros. Nóbrega (2013) anotou os sentidos dos substantivos comuns mais frequentes no córpus CSTNews (Cardoso et al., 2011), composto por notícias jornalísticas em português provenientes de agências de notícias on-line.

Pelo que se sabe, o trabalho de Nóbrega (2013) foi o primeiro trabalho a lidar com a DLS para o português com propósito geral, isto é, sem visar uma aplicação específica. O autor usou como repositório de sentidos a WordNet de Princeton, em sua versão 3.0 (Fellbaum, 1998), em inglês. Com isto, fez-se necessário traduzir as palavras do português para suas versões adequadas em inglês antes de se identificar o sentido correto na wordnet. Para tanto, o dicionário bilíngue WordReference® foi utilizado como recurso de suporte. No presente artigo, relata-se o processo de anotação de sentidos para os verbos no córpus

CSTNews, dando-se continuidade ao trabalho de Nóbrega (2013) e visando-se investimentos em DLS de verbos para o português.

O artigo está estruturado da seguinte maneira: na Seção 2, apresenta-se a metodologia usada para a anotação de córpus; na Seção 3, apresentam-se os resultados e avaliação da anotação.

2. Metodologia de Anotação

A anotação focou em desambiguar as palavras identificadas como verbos, pois se sabe que o verbo é uma classe gramatical de grande relevância na estrutura de uma sentença (Fillmore, 1968).

A anotação de sentidos foi realizada no córpus multidocumento CSTNews (Cardoso et al., 2011), composto por 50 coleções de textos jornalísticos com 2 ou 3 textos cada uma (por isso o nome multidocumento). Os motivos que levaram a usar o CSTNews para a anotação foram: o córpus foi utilizado para a tarefa de anotação de sentidos de substantivos (Nóbrega, 2013); ao ser um córpus jornalístico, abrange muitos domínios, o que garante que poderemos utilizar uma grande quantidade de sentidos para as palavras a serem anotadas e subsidiar métodos de DLS de uso geral para os verbos.

Para este trabalho, seguindo-se o trabalho de Nóbrega, também foi usada a WordNet de Princeton como repositório de sentidos, sendo que os sentidos são representados pelos conjuntos de sinônimos da WordNet (chamados *synsets*). Os motivos que levaram ao seu uso foram os seguintes:

- é um dos recursos semânticos mais difundidos e usados na literatura;
- é considerada como uma ontologia linguística (Di Felippo, 2008), isto é, abrange o conhecimento geral do mundo, com conceitos representados em língua natural (no caso, o inglês);
- dá-se continuidade ao trabalho feito por Nóbrega.

Para auxiliar a anotação, foi desenvolvida a ferramenta NASP++, que apoia, de forma semiautomática, toda a metodologia explicitada a seguir. A metodologia de anotação utilizada foi a usada por Nóbrega. Esta é dividida em duas partes, uma metodologia geral e outra individual. A metodologia geral faz referência às etapas que todos os anotadores devem seguir para anotar uma coleção de textos. Os passos a executar foram os seguintes:

- escolher um texto da coleção para ser anotado;
- anotar todas as palavras indicadas como “verbo” nesse texto e, depois disso, anotar o texto seguinte da coleção;
- após anotar todos os textos, revisar e salvar os mesmos.

Além disso, para cada um dos textos anotados, foi seguida a metodologia individual apresentada na Figura 2.1.

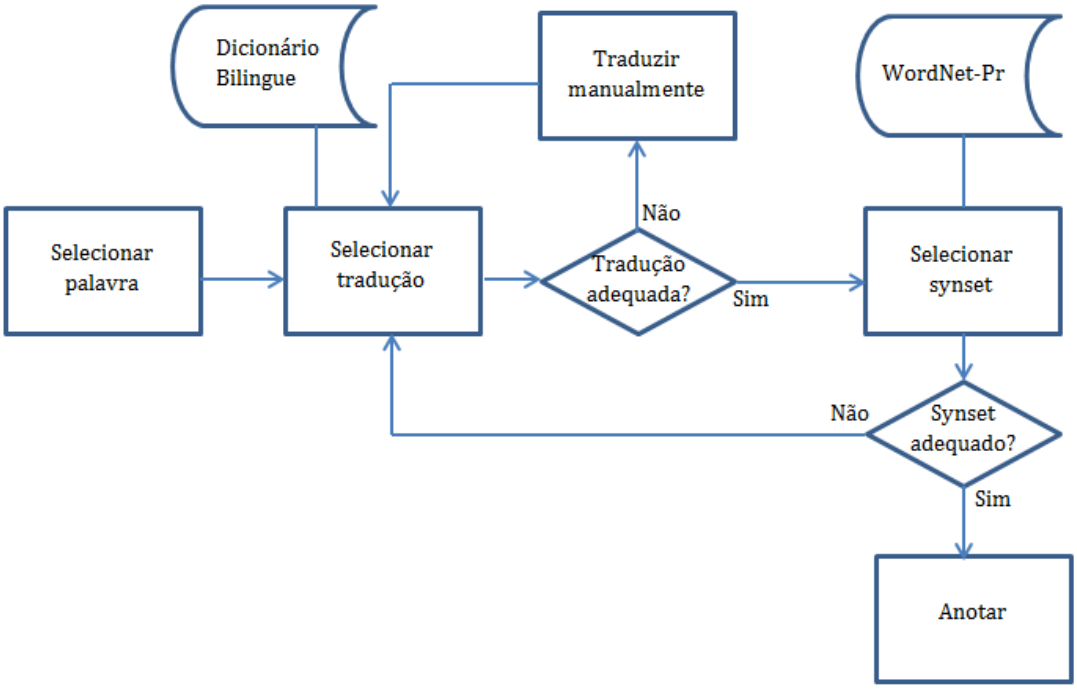

Figura 2.1. Metodologia de Anotação

Embora a anotação das palavras tenha seguido a metodologia mencionada, ocorreram exceções que foram tratadas de maneira particular. A seguir, apresentamos os desafios encontrados:

- verbos auxiliares: no caso de uma palavra a ser anotada ser um verbo auxiliar, optou-se por adicionar um comentário do tipo "Verbo auxiliar";
- verbos "independentes": por exemplo, na sentença "ele havia prometido retornar", os verbos "prometer" e "retornar" são considerados independentes, portanto, são anotados separadamente;
- predicados complexos: a maneira de anotar predicados complexos consistiu em identificar o verbo prinicipal do predicado complexo, adicionar um comentário do tipo "É predicado complexo" e, seguindo a metodologia individual, anotar o sentido referente ao predicado complexo e não só ao verbo;
- verbos no particípio e adjetivos: para efeitos da anotação, optou-se por distinguir uma palavra como sendo verbo no particípio se o anotador conseguir trocar da voz passiva à ativa na sentença em questão.

O uso de ferramentas de suporte foi essencial para a tarefa de anotação. Como parte da metodologia de anotação, em caso de dúvidas, os anotadores puderam usar o Google Translate (https://translate.google.com.br/) e o Linguee (http://www.linguee.com.br/) como dicionários bilíngues. No caso em que não conseguiram resolver as dúvidas, puderam fazer uma consulta a toda a equipe de anotação.

A anotação foi realizada por grupos de 2 a 3 anotadores, que era trocado a cada rodada de anotação, de forma a (i) se evitar vícios (*bias*) de anotação e (ii) permitir o compartilhamento de experiências de anotação. Em geral, cada grupo anotava uma coleção

de textos por dia, em horários diários pré-fixados. No total, a anotação durou 7,5 semanas, incluindo o treino dos anotadores.

3. Resultados da Anotação

Na Tabela 3.1, apresentam-se dados gerais obtidos da anotação:

	Total	Verbos principais	Predicados complexos	Verbos auxiliares	Erros de anotação
# palavras anotadas	6494	5082	146	949	317

Tabela 3.1. Estatísticas da anotação de instâncias do córpus CSTNews

Da Tabela 3.1, salienta-se que as 5082 instâncias de verbos principais que foram anotadas representam 844 verbos principais diferentes, e que, para estes, foram indicadas 787 traduções e anotados 1047 *synsets* diferentes. O item "Erros de anotação" refere-se às palavras erroneamente indicadas como verbos pela ferramenta NASP++ e que foram manualmente corrigidas.

Comparando as estatísticas apresentadas por Nóbrega (2013) e neste trabalho (apresentadas na Tabela 3.2), ressalta-se que os verbos possuem uma maior variação de sentidos, tanto em nível do córpus, quanto em coleções de textos. Além disso, o número de possíveis *synsets* para cada palavra e a porcentagem de palavras ambíguas neste trabalho é também maior ao apresentado por Nóbrega para os substantivos. Com estes resultados, infere-se que a tarefa de anotação para os verbos é mais difícil do que para os substantivos. Uma razão é porque os verbos são mais polissémicos (Miller et al., 1990).

	Substantivos (Nóbrega, 2013)	Verbos
Número máximo de *synsets* anotados por palavra no córpus	5	18
Número máximo de *synsets* anotados por palavra em uma coleção (e não no córpus todo)	3	4
Média do número de *synsets* possíveis (disponíveis para escolha na anotação) por palavra	6	12
Porcentagem de palavras ambíguas	77%	82.11%

Tabela 3.2. Comparação entre as estatísticas da anotação no trabalho de Nóbrega (2013) e no presente trabalho

Para a avaliação do nível de concordância na anotação, foi usada a medida Kappa (Carletta, 1996) e outras medidas mencionadas a seguir:

- Concordância Total: número de vezes em que todos os anotadores concordaram, em relação ao total de instâncias.
- Concordância Parcial: número de vezes em que a maioria dos anotadores concordou, em relação ao total de instâncias.
- Concordância Nula: número de vezes em que houve discordância total ou em que não se configurou uma maioria na anotação, em relação ao total de instâncias.

Devido a tarefa de anotação ter uma etapa de tradução para poder obter o sentido da WordNet, fez-se necessária a avaliação da concordância na (1) etapa de tradução, na (2) escolha do *synset* e na (3) seleção da tradução com seu respectivo *synset*.

Para a avaliação, foram escolhidas as mesmas 3 coleções de texto usadas no trabalho de Nóbrega (2013), visando comparações entre os resultados. Cada uma das coleções foi anotada por 4 grupos diferentes de anotadores. Na Tabela 3.3, apresentam-se os valores de concordância obtidos no trabalho de Nóbrega (2013) e neste trabalho.

	Substantivos (Nóbrega, 2013)				Verbos			
	Kappa	Total (%)	Parcial (%)	Nula (%)	Kappa	Total (%)	Parcial (%)	Nula (%)
Tradução	0.853	82.87	11.08	6.05	0.648	48.81	48.50	2.69
Synset	0.729	62.22	22.42	14.36	0.509	35.12	58.47	6.41
Tradução-*Synset*	0.697	61.21	24.43	14.36	0.474	31.73	61.29	6.98

Tabela 3.3. Valores de Concordância obtidos por Nóbrega (2013) e neste trabalho

Analisando os resultados da Tabela 3.3, nota-se que os valores de concordância para os substantivos são, na maioria, superiores aos verbos. Este resultado era esperado, devido a maior complexidade que os verbos apresentam e ao maior grau de polissemia presente nos verbos. Contudo, os valores obtidos na anotação de sentidos de verbos são aceitáveis no cenário da DLS. A seguir, apresentam-se dois fragmentos de sentenças, cujos verbos obtiveram concordância parcial e concordância nula, respectivamente.

1. Os advogados de mais de 500 pessoas que se *dizem* vítimas de abusos sexuais...

2. ...fontes da polícia moscovita *adiantaram* que ela teria acontecido provavelmente por causa da explosão acidental de um bujão de gás.

Algumas das dificuldades encontradas na anotação, que também ajudam a explicar o nível mais baixo de concordância, foram: a falta de traduções e/ou *synsets*, a ocorrência de lacunas lexicais (natural, em função das línguas diferentes envolvidas), a identificação de predicados complexos e verbos auxiliares, e a distinção de verbos no particípio de adjetivos.

Finalmente, o córpus anotado e a NASP++ estão disponíveis na página do projeto SUCINTO, em www.icmc.usp.br/pessoas/taspardo/sucinto/resources.html. Mais detalhes sobre a anotação e o próprio córpus podem ser encontradas em Sobrevilla-Cabezudo et al. (2014).

Agradecimentos

Nosso agradecimento à FAPESP, à CAPES e à Samsung pelo apoio a esta pesquisa.

Referências

Cardoso, Paula C. F.; Erick G. Maziero; Maria L. R. C. Jorge; Eloise M. R. Seno; Ariani Di Felippo; Lucia H. M. Rino; Maria das Graças V. Nunes; Thiago A. S. Pardo. 2011. "CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese." In *Proceedings of the 3rd RST Brazilian Meeting*. pp. 88–105, Cuiabá, MT, Brasil.

Carletta, Jean C. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22: 249-254

Di Felippo, Ariani. 2008. *"Delimitação e Alinhamento de Conceitos Lexicalizados no Inglês Norte-americano e no Português Brasileiro."* Tese de doutorado, Faculdade de Ciências e Letras, Universidade Estadual Paulista. São Paulo, Brasil.

Fellbaum, Christiane. 1998. *WordNet An Eletronic Lexical Database*. MIT Press

Fillmore, Charles J.1968. "The Case for Case." In *Universals in Linguistic Theory*. pp. 1-88, New York, USA.

Jurafsky, Daniel; James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.* 2nd edition. Prentice-Hall.

Machado, Ivre M.; Rafael O. de Alencar; Roberto de Oliveira Campos Junior; Clodoveu A. Davis. 2011. "An ontological gazetteer and its application for place name disambiguation in text." *Journal of the Brazilian Computer Society* 17: 267-279.

Miller, George A.; Richard Beckwith; Christiane Fellbaum, Derek Gross; Katherine J. Miller. 1990. Introduction to Wordnet: An on-line lexical database, *International Journal of Lexicography* 3: 235-244.

Nóbrega, Fernando A. A. 2013. *"Desambiguação Lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento."* Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Ribeiro, Ricardo. 2003. *"Anotação Morfossintáctica Desambiguada do Português."* Dissertação de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Specia, Lucia. 2007. "*Uma abordagem híbrida relacional para a desambiguação lexical de sentido na tradução automática*." Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, Brasil.

Sobrevilla-Cabezudo, Marco A.; Erick G. Maziero; Jackson W. C. Souza; Márcio S. Dias; Paula C. F. Cardoso; Pedro P. Balage Filho; Verônica Agostini; Fernando A. A. Nóbrega; Cláudia Dias de Barros; Ariani Di-Felippo; Thiago A. S. Pardo. 2014. *"Anotação de Sentidos de Verbos em Notícias Jornalísticos em Português do Brasil."* Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. NILC - TR - 14 - 05. São Carlos, SP. Em publicação.

Travanca, Tiago. 2013. *"Verb Sense Disambiguation."* Dissertação de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.