Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

Anotação de corpus com a OpenWordNet-PT: um exercício de desambiguação

Cláudia Freitas[1], **Livy Real**[3], **Alexandre Rademaker**[3,2]

[1]PUC-Rio, Brazil

[2]FGV/EMAp, Brazil

[3]IBM Research, Brazil

Abstract. *This paper presents the first effort towards a portuguese wordnet annotated corpus. We mannualy annotated 30 sentences, using the OpenWordNet-PT as a lexicon, and then compared the results with an automatic annotation. In addition to the system's evaluation, the results provided valuable insights about how to deal with this ambitious task.*

Resumo. *O presente trabalho apresenta o primeiro passo em direção à construção de um corpus alinhado com uma wordnet — especificamente, com a OpenWordNet-PT. Fizemos um exercício de anotação manual dos substantivos de 30 frases, e comparamos os resultados com os de uma anotação automática. Para além dos índices de acerto do sistema, este breve exercício foi capaz de apontar caminhos para a construção de um corpus alinhado com uma wordnet.*

1. Introdução

No atual contexto do processamento computacional das línguas, em que sistemas já não são protótipos, recursos capazes de lidar com o processamento de sentido estão no centro das atenções. Tais recursos podem assumir a forma de corpora semanticamente anotados ou de léxicos computacionais ou bases de dados lexicais. Para a língua inglesa, a WordNet de Princeton [Fellbaum 1998] [1] é o exemplo canônico de uma base lexical geral e robusta, amplamente utilizada por sistemas de PLN. Por outro lado, ainda são poucos os trabalhos relacionados à construção de corpora alinhados à wordnets. Para a língua portuguesa, com relação a recursos similares à WordNet [Oliveira et al. 2015], destacamos a OpenWordNet-PT [de Paiva et al. 2012] [2] (doravante OpenWN-PT), alinhada à WordNet de Princeton e que conta hoje com 47.702 synsets, dos quais 32.855 correspondem a substantivos, 5.060 a verbos, 8.753 a adjetivos e 1.034 a advérbios.

A OpenWN-PT foi escolhida pelos organizadores dos projetos FreeLing [Padró and Stanilovsky 2012], Open Multilingual Wordnet [Bond and Foster 2013] e ainda Google Translate [3] como a representante das wordnets abertas em português. No entanto, a OpenWN-PT ainda não dispõe de um corpus alinhado, e este trabalho relata o primeiro passo nesta direção.

[1]Usaremos "WordNet" para nos referirmos à WordNet de Princeton e "wordnet" como termo geral para a classe de recursos léxicos com estrutura similar à WordNet.

[2]Disponível para download em `http://github.com/own-pt/openWordnet-PT/` e para navegação online em `http://wnpt.brlcloud.com/wn/`.

[3]`http://translate.google.com/about/intl/en_ALL/license.html`.

Alinhar um corpus com uma wordnet ainda em construção também é uma maneira de avaliar e melhorar a própria wordnet: a verificação da cobertura leva à adição de sugestões, além de garantir que tais adições são palavras de uso comum na língua.

Existem mais de 60 wordnets disponíveis [4] e, segundo [Petrolito and Bond 2014], há pelo menos 20 corpora anotados semanticamente a partir de wordnets, para mais de 10 línguas. Diferentemente dos corpora alinhados a wordnets de que temos conhecimento, que foram feitos manualmente [Koeva et al. 2010] ou consistem da tradução automática de algo feito manualmente [Bentivogli and Pianta 2005], pretendemos realizar a anotação por meio do módulo de desambiguação de sentidos (WSD) da suíte Freeling [Padró and Stanilovsky 2012]. O Freeling disponibiliza um conjunto de ferramentas abertas para o processamento de diferentes línguas, e o módulo WSD dedicado à língua portuguesa já incorpora a OpenWN-PT. Uma primeira etapa, portanto, na criação do corpus anotado e alinhado à openWordnet-PT é avaliar a qualidade da ferramenta WSD, comparando-a com o desempenho humano. O presente trabalho relata os resultados de um breve exercício que teve como objetivo principal produzir essa avaliação.

2. Formas de avaliar wordnets e relações semânticas

Boa parte dos trabalhos em PLN utiliza como forma de avaliação as medidas de precisão e abrangência. Para que essas medidas sejam calculadas, é fundamental a existência de um gabarito. No entanto, para a avaliação de bases lexicais criadas automaticamente, tais medidas não são facilmente aplicáveis. O que significaria, nesse contexto, a noção de abrangência? A quantidade de conhecimento corretamente codificado, com relação a todo o conhecimento que deveria ser adquirido? O problema está em como definir "todo o conhecimento que deve ser adquirido", já que o mesmo conjunto de fatos pode levar a diferentes interpretações e, consequentemente, a diferentes tipos de "conhecimento".

Ainda que existam tentativas de avaliar wordnets ou recursos similares em português [Oliveira et al. 2015], tais avaliações são sempre comparações, e pouco nos informam quanto à qualidade intrínseca de cada recurso. Adicionalmente, concordamos com [Brewster et al. 2004] quando indicam que uma possibilidade para a avaliação de ontologias é direcioná-las aos dados (uma avaliação data-driven). Por isso, um alinhamento entre os synsets existentes e um corpus é uma boa maneira verificar a sua completude – ainda que saibamos que um corpus será sempre uma porção limitada da língua.

3. Descrição do experimento

A suíte Freeling dispõe de um módulo desambiguação de sentidos (WSD), que realiza um alinhamento entre as palavras do texto e a OpenWordNet-PT. Com o objetivo de verificar a precisão do sistema automático de desambiguação, criamos um experimento no qual diferentes anotadores deveriam selecionar o synset adequado para uma palavra em contexto. Em seguida, comparamos os resultados obtidos com os synsets sugeridos pelo módulo de WSD do Freeling [Agirre and Soroa 2009].

Foram selecionadas 30 frases da porção brasileira do corpus Bosque, a parte revista da Floresta Sintá(c)tica [Afonso et al. 2002]. A escolha pela variante brasileira teve como objetivo garantir segurança na atribuição dos sentidos, já que os anotadores eram

[4] `http://globalwordnet.org/wordnets-in-the-world/`.

brasileiros. Além disso, consideramos apenas os substantivos, e selecionamos frases com pelo menos 5 deles. A restrição aos substantivos se deve à reconhecida polissemia verbal, o que tornaria a tarefa mais difícil para os avaliadores. O número total de substantivos avaliados foi de 226, com 204 palavras distintas.

Cada avaliador recebeu um formulário com as 30 frases, e abaixo de cada frase listamos os substantivos alvo, que por sua vez direcionavam o avaliador para a página da OpenWN-PT com todos os synsets em que palavra analisada participava. O avaliador então deveria selecionar o synset adequado, indicando no campo do formulário o código do synset. Mais de um synset poderiam ser escolhidos, desde que ambos se adequassem igualmente ao contexto, segundo o avaliador. Os avaliadores foram instruídos a deixar o campo em branco caso não considerassem nenhum synset adequado, independentemente na natureza da inadequação.

Os anotadores não receberam nenhum treinamento especial que garantisse familiaridade com a OpenWN-PT. Participaram da anotação 9 alunos de graduação do curso de Letras-Tradução e 1 tradutor (anotadores "inexperientes"). Adicionalmente, duas das autoras do artigo também participaram da anotação (anotadoras "experientes").

4. Resultados

Usando o coeficiente *Kappa* [Carletta 1996], que mede o grau de concordância entre anotadores, fizemos dois tipos de avaliação da concordância: apenas a concordância entre humanos, e a concordância entre humanos e o módulo de desambiguação do Freeling.

Na concordância inter-anotadores, considerando apenas os anotadores “inexperientes"e apenas um synset por anotador [5], o índice de concordância foi de 0.67. Quando, no mesmo grupo de anotadores, consideramos todos os synsets escolhidos para uma mesma palavra, o índice de concordância cai para 0.55. Chama a atenção o baixo índice de concordância, mas é igualmente surpreendente que a concordância apenas entre as anotadoras experientes também seja de 0.67.

Especificamente quanto às anotadoras experientes, quando comparamos o módulo WSD do Freeling e a anotadora 1, a concordância é de 0.45; a concordância entre o módulo WSD e a anotadora 2 é de 0.52; e a concordância entre ambas as anotadoras e o módulo WSD é 0.56. Porque a concordância foi baixa mesmo entre as anotadoras experientes, a avaliação com o módulo WSD do Freeling é pouco informativa com relação à qualidade do sistema. Isto é, se entre humanos é difícil acordar sobre qual o synset adequado, que desempenho esperar do sistema?

5. Análise dos erros

Em cerca de 20% dos casos foi apontada a ausência de um synset adequado. Essa ausência, por sua vez, não significa necessariamente uma lacuna na OpenWN-PT, já que o alinhamento de palavras com synsets é precedido pelas etapas de tokenização e lematização. Quando há falha em alguma dessas etapas, falha também a atribuição de sentido. [6]

[5] Ao longo da avaliação, percebemos que haviam anotadores mais criteriosos, que sistematicamente optavam por listar todos os synsets considerados adequados, em oposição a anotadores mais econômicos, que listavam apenas o primeiro synset adequado que encontravam. A opção de avaliação de um synset por anotador buscou evitar que a divergência na quantidade dos synsets escolhidos influenciasse a discordância.

[6] Todas as etapas do processamento foram realizadas pela suíte do Freeling.

A seguir, detalhamos as situações em que isso ocorreu: (1) Erro na atribuição da classe gramatical: 6 casos, em que estava em jogo a flutuação entre N e ADJ; (2) Erro de lematização quanto ao número: há palavras que atribuem sentidos ligeiramente diferentes quando estão no singular ou no plural: *recursos* pode ser o plural de recurso mas, com o sentido de *bens, riquezas, recursos financeiros*, será usado sempre no plural; *vésperas* também tem um sentido menos preciso que *véspera*; (3) Erros de tokenização e unidades multipalavra: quando a tokenização é feita palavra por palavra, é difícil apontar para o synset adequado se ele for composto por uma unidade multipalavra, e isso aconteceu em cerca de 20% das palavras não alinhadas.

Sabemos que algumas dessas "falhas" não são exatamente erros, mas antes pontos não consensuais no PLN e que se refletem nas wordnets.

Outro ponto é a necessidade de um tratamento mais sistemático de prefixos e outros compostos com hífen. Em nosso exercício, não foi possível anotar *super-acordo*, ausente na OpenWN-PT, e não nos parece que deveria ser diferente. Por outro lado, gostaríamos que *social-democrata* estivesse em algum synset.

A existência de synsets relacionados à política norte-americana também traz desafios no que se refere à anotação de textos de uma outra cultura, e talvez seja preciso criar synsets relevantes para o mundo lusófono.

Por fim, não sabemos como lidar com efeitos de estilo, como o emprego da expressão *a ferro e fogo*, em "*Iti Fuji conquista clientela a ferro e fogo. Restaurante tem seu ponto forte no balcão de grelhados, que se sobrepõe aos prosaicos sushis e sashimis.*", que deve remeter à expressão *a ferro e fogo*, mas, simultaneamente, também ao ferro e ao fogo das grelhas.

A possibilidade de atribuição de mais de um synset a uma palavra também contribuiu para a baixa concordância. Apesar de cientes da granularidade talvez excessiva da WordNet, e da dificuldade inerente à tarefa lexicográfica de separação dos sentidos das palavras [Kilgarriff 1997], não foram raros os casos em que mais de um sentido era possível, e isso só foi verificado após a anotação, com uma análise caso a caso das divergências. Tomando por base uma das anotadoras experientes, em ao menos 8% das palavras anotadas mais de um synset seria aceitável.

6. Considerações finais e trabalhos futuros

O objetivo inicial deste exercício foi verificar a qualidade de um sistema de desambiguação com base na OpenWN-PT. Para isso, criamos uma tarefa de anotação semântica. Considerando a baixa concordância entre os anotadores, a proposta inicial de avaliação de um sistema automático de desambiguação deve ser vista com cautela, uma vez que não está claro o que esperar exatamente como desempenho de um sistema nesta tarefa. Por outro lado, o exercício nos permitiu um instantâneo da OpenWN-PT versão 1.0, fornecendo pistas relativas a pontos que devem ser tratados na construção de uma wordnet cada vez mais robusta.

O exercício também nos apontou caminhos para etapas futuras na criação de um corpus anotado alinhado com a OpenWN-PT. Pretendemos refazer o experimento com uma ferramenta de anotação específica para isso, e já contando com uma versão melhorada da OpenWN-PT.

Referências

Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá(c)tica: um treebank para o português. In Gonçalves, A. and Correia, C. N., editors, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)*, pages 533–545, Lisboa, Portugal. APL.

Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (ACL)*, volume 1, page 1352–1362.

Brewster, C., Alani, H., and Dasmahapatra, A. (2004). Data driven ontology evaluation. In *In Int. Conf. on Language Resources and Evaluation*.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An open brazilian wordnet for reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Kilgarriff, A. (1997). I dont believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Koeva, S., Leseva, S., Tarpomanova, E., Rizov, B., Dimitrova, T., and Kukova, H. (2010). Bulgarian sense annotated corpus - results and achievements. In *Proceedings of the 7th International Conference of Formal Approaches to South Slavic and Balkan Languages*, volume FASSBL-7, page 41–48, Dubrovnik, Croatia.

Oliveira, H. G., de Paiva, V., Freitas, C., Rademaker, A., Real, L., and Simões, A. (2015). *As Wordnets do Português*, volume 7, pages 397–424. OSLa, Oslo, Noruega.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 8th LREC*, page 2473–2479.

Petrolito, T. and Bond, F. (2014). A survey of wordnet annotated corpora. In *Proceedings of the Seventh Global WordNet Conference*, volume 1, pages 236–243, Tartu, Estonia.