Desambiguação dos termos do Atlas Linguístico do Brasil através da OpenWordnet-PT-ALiB

Augusto Sampaio Barreto, Daniela Barreiro Claro

[1]FORMAS Research Group
Instituto de Computação, Universidade Federal da Bahia
Salvador - Bahia - Brazil

{augusto.barreto,dclaro}@ufba.br

Abstract. *This work describes the disambiguation of terms from the Linguistic Atlas of Brazil (ALiB) via OpenWN-PT-ALiB through a Twitter corpus. The study presents two main contributions: the incorporation of some ALiB terms in OpenWordNet-PT (OpenWN-PT) and the development of a disambiguation method using Word Embeddings and the Soft Cosine Measure (SCM). The proposed method uses Word Embeddings to represent the words in a vector space and calculates the SCM between the context of the tweets and the possible synsets of OpenWN-PT-ALiB for disambiguation. Results demonstrate the effectiveness of the method, with higher disambiguation rates even in the context of Twitter.*

Resumo. *Este trabalho descreve a desambiguação de termos do Atlas Linguístico do Brasil (ALiB) via OpenWN-PT-ALiB através de um corpus do Twitter. O estudo apresenta duas principais contribuições: a incorporação de alguns termos do ALiB na OpenWordNet-PT (OpenWN-PT) e o desenvolvimento de um método de desambiguação utilizando Word Embeddings e a Soft Cosine Measure (SCM). O método proposto utiliza Word Embeddings para representar as palavras em um espaço vetorial e calcula a SCM entre o contexto dos tweets e os possíveis synsets da OpenWN-PT-ALiB para a desambiguação. Os Resultados demonstram a eficácia do método, com taxas de desambiguação superiores mesmo no contexto do Twitter.*

1. Introdução

O ALiB começou em 2001, encerrando seus inquéritos na década passada. Os termos coletados ao longo de duas décadas de inquéritos suscitam a análise de sua presença nas redes sociais devido à disseminação dessas plataformas e à quantidade significativa de postagens em linguagem escrita. A difusão das redes sociais e o grande volume de postagens destacam a importância de investigar a presença dos termos coletados pelo ALiB nessas plataformas.

Analisar a vitalidade linguística dos termos do ALiB nas redes sociais exige automação devido ao volume de tweets. Muitos termos do ALiB são ambíguos e necessitam de desambiguação conforme o contexto do Atlas. Por exemplo, *bala* em tweets geralmente se refere a munição, enquanto no ALiB pode significar um tipo de doce.

Assim, o presente trabalho evidencia duas principais contribuições com o intuito de automatizar a desambiguação dos termos do ALiB. A primeira contribuição se refere

à OpenWordNet-PT (OpenWN-PT), à qual foram adicionados os synsets de alguns dos termos do ALiB. A segunda contribuição se refere ao método de desambiguação dos termos do ALiB para usufruir da OpenWordNet-PT-ALiB (OpenWN-PT-ALiB) que foi gerada. O método proposto se baseia no uso de *Word Embbedings* para encapsular as palavras no espaço vetorial, e no cálculo da *Soft Cosine Measure*(SCM) entre o contexto dos tweets com a palavra a ser desambiguada e os possíveis *synsets* existentes para essa palavra na OpenWordNet-PT-ALiB.

2. OpenWordNetPt-ALiB

A OpenWordnet-PT [de Paiva et al. 2012], abreviado como OpenWN-PT, surgiu com base no projeto da WordNet [Fellbaum 1998], de modo a atender à demanda da comunidade científica por ferramentas disponíveis em língua portuguesa de forma gratuita, acessível online e disponível para download e uso offline. Ela mapeia os *synsets* da língua inglesa para os correspondentes em português.

Com base no arquivo original em formato RDF da OpenWN-PT, alguns dos *synsets* do ALiB foram incorporados diretamente em seu arquivo fonte no formato RDF, sendo esta uma das contribuições científicas deste trabalho. A análise lexical incluída na nova versão, a OpenWN-PT-ALiB, foi baseada no questionário semântico-lexical obtido do Projeto ALiB. Variantes lexicais foram fornecidas por informantes brasileiros localizados em todo o território nacional. Esses dados foram publicados nos dois volumes do Atlas Linguístico do Brasil [Cardoso and Mota 2014]

Com o intuito de analisar se os resultados do ALiB seriam retornados dada uma consulta a um *synset*, o termo *Goleiro* foi utilizado.

A Figura 1 apresenta o resultado da consulta ao *synset* do termo *Goleiro* no arquivo original da OpenWN-PT.

```
WORD: goleiro
WORD_SENSES: ['goleiro', 'guarda-redes']
```

Figure 1. Consulta aos *synsets* da palavra "goleiro" na OpenWN-PT

Já a Figura 2 apresenta o resultado da consulta ao *synset* do termo *Goleiro* na OpenWN-PT-ALiB, após adição do novo sentido "sutiã"

```
WORD: goleiro
WORD_SENSES: ['sutiã', 'goleiro', 'guarda-redes']
```

Figure 2. Consulta aos *synsets* da palavra "goleiro" OpenWN-PT-ALiB

3. Modelos de Linguagem

Diversos modelos de linguagens foram propostos [Bengio et al. 2003], com o objetivo de representar o espaço semântico ideal de palavras em um espaço vetorial contínuo com valor real. As representações distribuídas dos termos em um espaço vetorial (*word embeddings*) ajudam os algoritmos de aprendizado a obter melhores desempenhos em tarefas de processamento de linguagem natural, agrupando palavras similares [Mikolov et al. 2013]

O NILC-Embeddings [Hartmann et al. 2017] é um repositório público que foi desenvolvido com o objetivo de compartilhar *word embeddings* gerados para a Língua Portuguesa. Ele contribui para tornar acessível recursos vetoriais a serem utilizados em tarefas de Processamento da Linguagem Natural e Aprendizado de Máquina para língua portuguesa.

Neste trabalho, os vetores de embeedings pré-treinados disponibilizados no proejto NILC-Embeddings foram utilizados para a tarefa de desambiguação de sentido de palavras, do inglês Word Sense Disambiguation (WSD).

4. Desambiguação automática dos termos do AliB

Em termos gerais, a tarefa de *word sense disambiguation* envolve a associação de uma determinada palavra em um texto com uma definição ou significado (sentido) que é distinguível de outros significados potencialmente atribuíveis a essa palavra. A tarefa, portanto, necessariamente envolve duas etapas: (1) a determinação de todos os diferentes sentidos para cada palavra relevantes (pelo menos) para o texto em consideração; e (2) um meio para atribuir cada ocorrência de uma palavra ao sentido apropriado [Ide and Véronis 1998].

O Soft Cosine Measure (SCM), uma extensão da medida de similaridade do cosseno, oferece a capacidade de avaliar a similaridade entre dois documentos, mesmo quando não compartilham palavras em comum. Inicialmente proposto por Mikolov et al. (2013), esse método emprega uma medida de similaridade entre palavras, que é obtida por meio de operações vetoriais entre as word embeddings individuais das palavras. No contexto desse estudo, a SCM foi empregada para calcular a similaridade entre o contexto de uma palavra sujeita a desambiguação e a sua definição na base OpenWN-PT/ALiB.

Com o intuito de desambiguar os termos do ALiB, um método foi proposto, como mostrado na Figura 3.

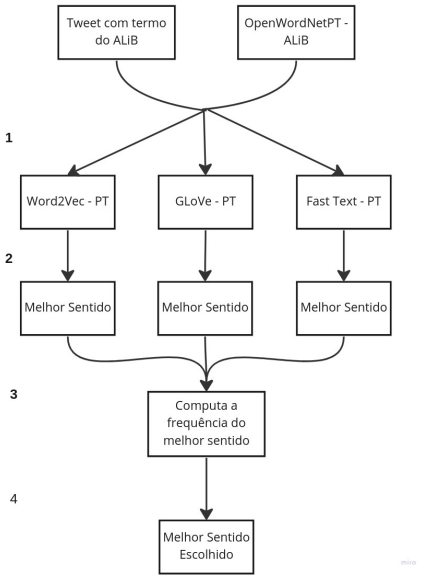

Figure 3. Representação arquitetural do método de desambiguação utilizado neste trabalho.

Nesta arquitetura, é importante evidenciar quatro macros etapas:

1. Cálculo da SCM pelos modelos de word embbedings CBOW de 600 dimensões, utilizando o contexto dos tweets minerados com a palavra a ser desambiguada, e os *synsets* desta palavra na OpenWN-PT/ALiB
2. Escolha do melhor sentido, com base na maior similaridade encontrada com base no cálculo da SCM.
3. Totaliza o melhor sentido escolhido por cada modelo.
4. Escolhe o melhor sentido para desambiguação, com base no que foi escolhido pela maioria dos modelos.

5. Experimentos e Resultados

O corpus, denominado TweetALiB/PT, composto de tweets coletados com o uso da biblioteca Tweepy foi utilizado.

Os experimentos foram realizados para identificar a vitalidade dos termos ALiB no Twitter e comparar os métodos de Word Sense Disambiguation (WSD) para a língua portuguesa em Tweets. Para comparar os resultados, foi escolhida uma amostra aleatória de 100 tweets para cada termo com sentido a ser desambiguado.

Como não existe um dataset rotulado com o sentido original dos tweets minerados para o ALiB, foi necessária realizar uma conferência manual na tarefa de desambiguação dos sentidos dos tweets. Para isso, 100 tweets aleatórios foram selecionados e o procedimento de desambiguação foi realizado manualmente, comparado com o sentido encontrado pelo método proposto.

Um importante resultado a se destacar é que, através do método proposto, a tarefa de desambiguação automática pode ser realizada em maior escala.Utilizando o SCM, a desambiguação nos experimentos atingiu valores superiores a 25 %, chegando a 55 % para a palavra “goleiro”.

Ressalta-se que o Twitter é uma plataforma em que a linguagem empregada é majoritariamente informal e que se resume a poucas palavras de contexto, dificultando a tarefa de desambiguação até mesmo para um humano.

6. Conclusão e Trabalhos Futuros

O presente trabalho descreveu o método referente à incorporação de alguns dos termos do ALiB na OpenWordnetPT-ALiB. Os resultados obtidos, evidenciam que ainda existem desafios na tarefa de desambiguação utilizando o método proposto em compreender as nuances semânticas e realizar associações relevantes em textos curtos e informais, característicos da plataforma. Como trabalhos futuros, pode-se explorar a investigação de técnicas de pré-processamento de texto específicas para lidar com a linguagem peculiar do Twitter.

Agradecimentos

O presente trabalho conta com o apoio da CAPES-Brasil - Código de Financiamento 001 e da FAPESB - Projeto TIC.

References

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

Cardoso, S. and Mota, J. (2014). *Atlas Linguístico do Brasil*. Addison-Wesley Longman Publishing Co., Inc.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, RS, Brasil. SBC.

Ide, N. and Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.