Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

Avaliando a similaridade semântica entre frases curtas através de uma abordagem híbrida

Allan de Barcelos Silva, Sandro José Rigo, Isa Mara Alves, Jorge L. V. Barbosa

[1]Programa de Pós-Graduação em Computação Aplicada –

Universidade do Vale do Rio dos Sinos

Caixa Postal 93.022-000 – 93.022-750 – São Leopoldo – RS – Brasil

allanbs@edu.unisinos.br, {rigo, ialves, barbosa}@unisinos.br

Abstract. *The task of evaluating textual semantic similarity is one of the challenges in the Natural Language Processing area. It is observed in the literature the experimentation with priority use of probabilistic resources, and linguistic aspects explored in an incipient way. This paper presents an experiment with a hybrid approach, in which both resources of distributed representation and also lexical and linguistic aspects are integrated for the evaluation of semantic similarity between short sentences in Brazilian Portuguese. The proposed technique was evaluated with a dataset known in the literature and obtained good results.*

Resumo. *A tarefa de avaliação da similaridade semântica textual é um dos desafios na área de Processamento de Linguagem Natural. A literatura descreve a experimentação com uso prioritário de recursos probabilísticos, sendo que aspectos linguísticos ainda são explorados de forma incipiente. O presente trabalho apresenta um experimento com uma abordagem híbrida, na qual tanto recursos de representação distribuída como aspectos léxicos e linguísticos são utilizados em conjunto para a avaliação de similaridade semântica entre frases curtas em português do Brasil. A técnica proposta foi avaliada com datasets conhecidos na literatura e obteve bons resultados.*

1. Introdução

Este artigo trata da análise de similaridade textual, tarefa que representa um desafio nas pesquisas relacionadas à área de Processamento de Linguagem Natural (PLN) [Kao and Poteet 2007] [Gomaa and Fahmy 2013] [Pradhan et al. 2015]. A identificação de similaridade entre frases e textos é uma parte fundamental para muitas tarefas em PLN [Gomaa and Fahmy 2013]. Observa-se que boa parte dos métodos atuais para esta tarefa são baseados prioritariamente na similaridade entre as palavras, representando as sentenças de modo simplificado, como um vetor de termos. Ainda, uma parte significativa dos trabalhos restringe a análise ao tratamento da informação léxica, utilizando-se pouco de outros recursos linguísticos. Ao adotar estas abordagens muitas vezes a ordem das palavras e o seu significados nas sentenças como um todo são desconsideradas [Ferreira et al. 2016]. Logo, podem ocorrer falhas quando as frases não possuem termos comuns devido à diversidade do vocabulário. Além disso, a dificuldade na identificação do contexto em

frases curtas é maior do que em documentos, pois estas possuem volume limitado de texto quando comparadas aos mesmos [Metzler et al. 2007].

Foram analisados estudos de similaridade semântica textual voltados para a língua portuguesa brasileira. Observou-se uma linha de desenvolvimento de trabalhos que incorporam prioritariamente características léxicas em suas técnicas [Fialho et al. 2016] e [Alves et al. 2016], valendo-se de materiais disponíveis em bases de dados abertas, tais como WordNet[1], FrameNet[2] ou VerbNet[3], entre outros, devido à qualidade das relações descritas nestes recursos. Em outra linha de trabalhos, os Modelos de Espaço Vetorial (MEV) são destacados [Barbosa et al. 2016] e [Freire et al. 2016] devido às possibilidades da sua abordagem probabilística, independência de domínio e capacidade em obtenção automática de relações semânticas dado um espaço de contextos. Ao mesmo tempo, trabalhos como [Ferreira et al. 2016] e [Alves et al. 2016] empregam recursos linguísticos tais como as relações de hiponímia, antonímia e sinonímia, obtendo resultados relevantes.

O trabalho descrito neste artigo consiste em uma abordagem híbrida na qual são integradas técnicas usando um conjunto de recursos linguísticos e probabilísticos. Através destes, foram definidos e analisados diversos conjuntos de atributos empregados na tarefa de avaliação da similaridade semântica entre sentenças curtas, através de sua combinação em um algoritmo para regressão linear. Para tanto, foram utilizados recursos como os Modelos de Espaços Vetoriais, bem como a exploração das relações semânticas de aspectos como hiponímia e antonímia [Cançado 2013], através das bases *Portuguese Unified Lexical Ontology* (PULO) [Simões and Guinovart 2014] e Thesaurus para o português do Brasil (TeP) [Maziero et al. 2008]. Como forma de realizar uma comparação dos resultados obtidos com o estado da arte na área, foi utilizado um conjunto de dados anotados disponibilizado no evento PROPOR 2016[4], junto ao workshop de Avaliação de Similaridade Semântica e Inferência Textual (ASSIN). Os resultados obtidos foram considerados competitivos e permitiram também a análise de impacto dos conjuntos de atributos empregados.

2. Trabalhos relacionados

Atualmente a tarefa de avaliação de similaridade textual vem recebendo bastante atenção [Ferreira et al. 2016], [Agirre et al. 2012], [Hartmann 2016], [Barbosa et al. 2016] e [Freire et al. 2016], o que também é observado em eventos como o SemEval[5] e PROPOR[6], os quais possuem tarefas para mensurar a similaridade semântica entre sentenças, tanto para a língua inglesa quanto para portuguesa.

Em seu trabalho, [Hartmann 2016] faz o mapeamento de todas as palavras não encontradas no vocabulário ou com apenas uma ocorrência no *corpus* para um *token* genérico UNK. Na sequência, o autor expandiu as sentenças utilizando recursos de sinonímia para palavras de conteúdo que possuíam até dois sinônimos no TeP e aplicou *Stemming* para obter somente o radical das palavras. Após, o autor calculou a similaridade do cosseno entre a soma dos *word embeddings* obtidos através do *word2vec* [Miko-

[1] `http://www.nilc.icmc.usp.br/wordnetbr/`.
[2] `http://www.ufjf.br/framenetbr/`.
[3] `http://www.nilc.icmc.usp.br/verbnetbr/`.
[4] http://www.propor2016.di.fc.ul.pt
[5] *International Workshop on Semantic Evaluation*
[6] *International Conference on the Computational Processing of Portuguese*

lov et al. 2013], em conjunto com a similaridade dos vetores TF-IDF de cada frase para estimar o quão similar são as sentenças através do algoritmo de regressão linear *Support Vector Machines* (SVM).

No trabalho de [Barbosa et al. 2016] são criadas métricas com *word embeddings* e *Inverse Document Frequency* (IDF) para utilização no algoritmo SVM e também em uma rede siamesa (*Siamese Networks*) de [Chopra et al. 2005]. Em [Freire et al. 2016] é um proposto um *framework* de três sistemas: STS_MachineLearning, STS_HAL e STS_WORDNET_HAL. O primeiro utiliza a similaridade entre palavras pelo coeficiente DICE e pela WordNet, enquanto que os demais utilizam a abordagem simbólica com o cálculo da similaridade de palavras através da *Latent Semantic Analysis* (LSA) e possuem uma variação que utiliza a WordNet no mesmo cálculo.

O trabalho de [Alves et al. 2016] utilizou em sua abordagem o cálculo de heurísticas sob um conjunto de nove redes semânticas (dentre elas PULO e TeP) para extrair relações entre as palavras e sentenças. O autor realizou contagens de relações léxicas, sintáticas e semânticas, além de empregar recursos como os tipos de entidades nomeadas e diversas outras medidas de similaridade/distância entre os nós da rede semântica. Ao final, todos os atributos gerados foram combinados em três técnicas de regressão linear para mensurar a similaridade entre as sentenças.

Com base no estudo realizado, observam-se trabalhos na literatura utilizando apenas recursos probabilísticos ou então heurísticos para avaliação da similaridade entre sentenças. Além disso, muitas das pesquisas que fazem uso de recursos linguísticos abordam de forma superficial a capacidade destes, pois utilizam apenas a existência ou não de relações para tratar o problema. Desta forma, o trabalho aqui apresentado é motivado pelo interesse na integração destas classes de recursos visando aproveitar de maneira mais efetiva e aprofundada cada uma de suas potencialidades. Além de propor uma abordagem aplicável para mensurar a similaridade semântica entre frases curtas, através da aplicação de contagem em relações de antonímia, penalização de diferença de tamanho entre sentenças, bem como do uso de relações de hiperonímia, hiponímia e sinonímia no apoio de modelos de espaço vetorial para redução e dimensionalidade e análise de similaridade.

3. Materiais e métodos

Para melhor compreensão da abordagem proposta a metodologia aplicada no presente trabalho foi dividida em sete passos, os quais tem como início (passo 1) a captura de textos em páginas de notícias através de um *Web Crawler*[7]. Na medida em que ocorre a coleta, a cada página visitada o software realiza a extração dos elementos textuais, a remoção de marcações *HTML*, e após grava o texto em um arquivo contendo um parágrafo por linha. Após a coleta de centenas de milhares de páginas (passo 2), é formado o *corpus* descrito na Subseção 3.2.

No passo 3 são aplicadas operações para preparação do *corpus* como entrada para o MEV, o qual obtém os *word embeddings* e armazena-os no formato *Comma Separated Values* (CSV) no servidor (passo 4). Uma vez que o recurso foi gerado, ocorre a etapa de pré processamento (passo 5) do conjunto de dados para reduzir a esparcidade da

[7]Software destinado a coleta e captura de textos na internet.

informação, através da remoção da pontuação, transformação do texto para caixa baixa e remoção de dados numéricos. No passo 6 são utilizados os recursos léxico-semânticos (PULO e TeP) para tratar as relações de hiperônimos, hipônimos e sinônimos.

Para melhor entendimento da abordagem, considerando as sentenças originais (números 1 e 2) descritas a seguir:

1. `A comissão apura denúncias de abuso e exploração sexual em meninas da comunidade quilombola.`
2. `O grupo apura denúncias de abusos e exploração sexual de crianças da Comunidade Quilombola.`

Após o passo de pré processamento (5 e 6), são obtidas as seguintes sentenças de exemplo:

1. `comissao apurar denunciar abusar exploracao sexual criancas comunidade quilombola`
2. `comissao apurar denunciar abusos exploracao sexual criancas comunidade quilombola`

Por fim, os dados resultantes do processo até este momento são utilizados como entrada para treinamento e teste dos algoritmos de aprendizagem de máquina, onde serão gerados os modelos classificadores de similaridade (passo 7).

3.1. Conjunto de dados

O presente trabalho utilizou como base para comparação de resultados o conjunto de dados disponibilizado pelo Workshop ASSIN, pertencente ao evento PROPOR/2016. O objetivo do *workshop* é a identificação da similaridade semântica e classificação entre pares de frases curtas disponibilizados no conjunto de dados. Segundo [Fonseca et al. 2016], o conjunto de dados disponibilizado foi anotado pelo total de 36 pessoas que participaram em diferentes quantidades, sendo que cada frase foi avaliada por 4 pessoas.

O conjunto de dados conta com 10.000 pares de sentenças coletadas através do Google News (divididos igualmente para o português do Brasil e de Portugal), destes 6.000 registros são dados para treinamento e os demais para teste, ambos os conjuntos contendo o valor de similaridade entre os pares de sentenças no intervalo $[1, 5]$. A avaliação dos trabalhos submetidos para a tarefa deu-se através da Correlação de *Pearson* (CP) e do Erro Médio Quadrado (EMQ), onde as técnicas deveriam possuir a maior CP e o menor EMQ possível [Fonseca et al. 2016].

3.2. *Corpus* para treinamento

Neste trabalho foi obtido um *corpus* em português para identificação das *word embeddings* através do algoritmo *GloVe*. Para tanto, foi desenvolvido um Web Crawler[8] para captura de textos em páginas de notícias como Google News e Wikipédia. No decorrer do processo de captura de textos, a cada página visitada o software realiza a extração dos elementos textuais e a remoção de marcações *HTML*. Após realizar a captura dos textos, foram removidos caracteres especiais diferentes de: $., ; ?! -$ nas sentenças [Manning and

[8] `http://www.projeto.unisinos.br/pipca_sts/web_service`.

Schütze 2000], bem como removidas as sentenças compostas somente com números ou que continham menos de cinco palavras. Na sequência, todo o texto foi transformado para minúsculo com o objetivo de reduzir a esparsidade dos dados e eliminar a redundância de palavras.

Foi disponibilizado o *corpus* utilizado (em sua forma original) e os *word embeddings* através do endereço `http://www.projeto.unisinos.br/pipca_sts`, pois tal ato contribui para o aumento da disponibilidade de recursos na área de PLN e possibilita que outras pesquisas possam utilizar os recursos no desenvolvimento de seus trabalhos.

3.3. Técnica

No presente trabalho foi utilizado o algoritmo *GloVe*[9] [Pennington et al. 2014] para modelagem do espaço de vetores e obtenção dos *word embeddings*, devido a disponibilidade da técnica *word2vec* para a linguagem R[10]. Apesar do modelo utilizado diferir do *word2vec*, pois o primeiro é baseado na contagem de elementos e o segundo é um modelo de linguagem neural, é possível observar nos experimentos de [Pennington et al. 2014], o desempenho do *Glove* em capturar a semântica das palavras.

O *corpus* elaborado foi utilizado para o treinamento do GloVe no servidor utilizado para o processamento, o qual conta com dois processadores *E5-2620* versão 4 2.1GHz, 128 gigabytes RDIMM (2400MT/s) e placa de vídeo *Matrox G200eR2* com 16 megabytes. O modelo foi treinado durante 10 épocas, com 6 elementos na janela de contexto, 100 co-ocorrências e taxa de aprendizagem de 0.15. Além disso, o tamanho dos vetores foi definido para 600 posições, pois notou-se nos testes realizados por [Pennington et al. 2014] o aumento da acurácia do algoritmo em capturar as semânticas das sentenças. Inicialmente foi realizada a composição de cada frase através dos *word embeddings* correspondentes a cada palavra e desta maneira foi obtida uma matriz de contextos com W palavras e 600 dimensões. Neste ponto, assim como nos trabalhos de [Hartmann 2016] e [Mikolov et al. 2013], criou-se um atributo através da similaridade do cosseno entre a soma da matriz de contextos de cada sentença. Contudo, [Hartmann 2016] comenta que a soma da matriz de *word embeddings* cria uma representação genérica da frase e acaba por não refletir seus contextos. Desta forma, aplicou-se a técnica *Principal Component Analysis* (PCA) para redução de dimensionalidade e calculou-se a distância euclidiana entre o primeiro componente de cada sentença, o qual contém os itens com maior variação na matriz de contextos.

Além dos atributos que fazem uso dos *word embeddings*, foram elaboradas mais 10 medidas através do processamento de outros recursos léxicos e semânticos das sentenças, os quais podem ser observados na Tabela 1. O atributo TF-IDF foi utilizado com as orações originais e também com uma variação onde através da base PULO e do TeP foram utilizados os atributos 9 e 10 (Tabela 1) para a substituição de sinônimos, hipônimos e hiperônimos. A utilização da variação do atributo TF-IDF como métrica para avaliação de similaridade ocorre como tentativa para redução da esparsidade dos dados, pois a abordagem TF-IDF utiliza em seu cálculo a contagem de palavras compartilhadas entre as sentenças. Logo, quanto mais elementos compartilhados entre os textos, maior

[9]Disponível em `https://nlp.stanford.edu/projects/glove/`
[10]Disponível em `https://www.r-project.org/`

será a similaridade entre ambos.

Tabela 1. Lista de atributos elaborados

Índice	Atributo
1	Similaridade do cosseno entre a soma dos *word embeddings*
2	Distância euclidiana entre o primeiro componente principal de cada sentença
3	Similaridade do cosseno entre os vetores TF-IDF de cada sentença
4	Coeficiente de penalização pelo tamanho das sentenças
5	Proporção de palavras em comum entre as sentenças
6	Proporção de *ngramas* em comum das sentenças
7	Proporção de palavras diferentes entre as sentenças
8	Contagem de antônimos nas sentenças
9	Substituição dos hipônimos e hiperônimos nas sentenças
10	Substituição de sinônimos

Utilizou-se a equação indicada por [Ferreira et al. 2016] para o cálculo da penalização de sentenças com tamanhos diferentes, porém o valor da similaridade usada na fórmula do autor foi substituído pela média aritmética das similaridades entre os *word embeddings* e TF-IDF. A adaptação da fórmula pode ser vista na Equação 1, onde T corresponde ao tamanho das sentenças e $Sim(frase)$ é o valor da média.

$$Penalizacao = \left\{ \begin{array}{cc} \frac{|T(frase_1)-T(frase_2)|\times Sim(frase)}{T(frase_1)} & se\ T(frase_1) > T(frase_2) \\ \frac{|T(frase_1)-T(frase_2)|\times Sim(frase)}{T(frase_2)} & caso\ contrario \end{array} \right\} \quad (1)$$

A medida da proporção de *ngramas* deu-se através da busca por bigramas ou trigramas em ambas as sentenças, utilizando as bibliotecas da ferramenta *Weka* de [Witten et al. 2016] para encontrar termos compostos e comuns com pelo menos uma ocorrência.

4. Resultados

Inicialmente foram realizados uma série de experimentos para avaliar a contribuição dos *word embeddings* na obtenção da similaridade semântica. Como se pode observar na Tabela 2, os resultados obtidos com os atributos isolados não foram suficientes para um bom desempenho do SVM, resultado também observado no trabalho de [Hartmann 2016]. Entende-se que a utilização de PCA ao invés de soma para obtenção da similaridade das embeddings mantém o desempenho não satisfatório porque a redução de dimensionalidade dos *word embeddings* pode levar a perda das nuances e peculiaridades das sentenças, ocasionando assim a perda do contexto.

Analisando os resultados da Tabela 2, observa-se que a maior Correlação de Pearson (CP) e o menor Erro Quadrado Médio (EQM) foram obtidos através dos experimentos com utilização de recursos linguísticos, tais como os antônimos e as relações de hiponímia. Entretanto, ao analisar a quantidade de antônimos por tuplas no conjunto de

Tabela 2. Experimentos e resultados

Atributos *	Correlação de Pearson	Erro Médio Quadrado
1	0.3165	0.6847
2	0.2641	0.7226
3	0.4448	0.6174
9	0.0355	0.7754
2,4	0.2672	0.7087
1,3,6,5	0.6364	0.4535
1,3,6,5,7	0.5782	0.5102
2,3,6,5	0.6357	0.4543
2,3,6,5,7	0.6343	0.4622
3,6,5	0.6160	0.4790
1,3,5,6,7,8,9,10	0.6394	0.4499
1,3,5,6,7,8,10	0.6370	0.4522
1,3,5,7,8,10	0.6408	0.4482
1,3,5,7,9,10	**0.6410**	**0.4479**

* A primeira coluna representa o índice dos atributos descritos na Tabela 1.

dados do PROPOR/ASSIN, notou-se que em raros casos foram identificadas uma ou mais relações de antonímia na mesma sentença, o que é justificado pelo baixo volume de registros da relação na base PULO. Tal fato dificultou a utilização das relações linguísticas e contribuiu para o desempenho da técnica no uso dos atributos de antônimos e hiponímia. Além disso, nota-se o baixo desempenho do atributo de penalização pela diferença de tamanho entre as sentenças. Após aplicada uma análise estatística, foi constatada a não existência de correlação com o valor esperado de similaridade ($p > 0.05$).

Na Tabela 3 são apresentados os melhores resultados no estado da arte para avaliação de similaridade semântica, os quais são comparados com o atual trabalho através do conjunto de dados do PROPOR/ASSIN (Seção 3.1). Apesar de ser possível observar na mesma tabela que este trabalho não obteve o melhor resultado para CP ou EQM, ressaltamos que o número de *tokens* no *corpus* usado para obtenção dos *word embeddings* foi extremamente reduzido.

Em [Hartmann 2016], o autor utiliza os *word embeddings* treinados em um *corpus* contendo cerca de três bilhões de tokens coletados dos websites G1 e Wikipédia, além da utilização do *corpus* PLN-Br de [Bruckschen et al. 2008]. Enquanto que foram utilizados apenas 1584492 *tokens* para o treinamento dos *word embeddings* no atual trabalho, o que corresponde cerca de $0,05\%$ do que foi usado por [Hartmann 2016]. Deste modo,

Tabela 3. Comparação com o estado da arte

	Abordagem	CP	EMQ
Técnica proposta	Embeddings com PCA	0,30	0,69
	Soma dos *word embeddings*	0,30	0,68
	TF-IDF	0,44	0,61
	Embeddings com PCA + TF-IDF	0.46	0.59
	Soma dos *word embeddings* + TF-IDF	0.55	0.52
	Melhor resultado da Tabela 2 *	0.64	0.44
[Hartmann 2016]	Soma dos *word embeddings*	0,58	0,50
	TF-IDF	0,68	0,41
	Soma dos *word embeddings* + TF-IDF	0,70	**0,38**
[Fialho et al. 2016]	Soft TF-IDF Similaridades entre palavras Sobreposição de *ngramas*	**0,73**	0,63
[Alves et al. 2016]	Métricas de similaridade, distância e contagens	0,65	0,44

* A linha com o título ”Melhor resultado da Tabela 2”corresponde à combinação dos atributos: Soma dos *word embeddings*, proporção de palavras em comum, TF-IDF, proporção de palavras diferentes, contagem de antônimos, substituição de sinônimos e relações de hiponímia.

é possível que a quantidade de *tokens* pode ser uma das causas para o desempenho dos experimentos com os atributos derivados dos *word embeddings*. Porém, os resultados obtidos foram superiores aos de [Fialho et al. 2016] para português do Brasil quando observado apenas o EQM e próximos aos de [Alves et al. 2016] mesmo sem uma análise sintática ou reconhecimento de entidades nomeadas.

Os melhores resultados obtidos pelo presente trabalho envolveram a substituição dos sinônimos e relações de hiponímia das sentenças. Tal recurso não afeta o sentido da frase e permite a comparação direta entre ocorrências de palavras comuns em ambas as sentenças. A abordagem descrita maximizou os resultados da técnica TF-IDF, agregando para esta um papel fundamental na obtenção da similaridade entre as frases. Entretanto, é visto que apesar da métrica dos antônimos não apresentar correlação com o valor esperado de similaridade ($p > 0.05$), este demonstrou bom desempenho quando utilizado em conjunto com outros atributos, tal como é possível observar na Tabela 2.

5. Conclusões

Neste trabalho foi apresentada uma abordagem híbrida para avaliar a similaridade semântica entre frases curtas. Para tanto, foram integrados recursos como Modelos de Espaço Vetorial e também as relações linguísticas de antonímia, hiperonímia, hiponímia e sinonímia. Através do emprego de recursos linguísticos, foram observados resultados próximos ao estado da arte apesar do uso de um *corpus* limitado para o treinamento do

MEV ($0,05\%$ da quantidade de *tokens* que são vistos na literatura). Além disso, os experimentos realizados demonstram que a utilização de relações de hiperonímia e hiponímia, por si só, não apresentam informações suficientes para uma melhor avaliação de similaridade. Porém a utilização destas como atributos, auxiliou na generalização dos termos das sentenças e consequentemente trouxe melhores resultados para técnicas como TF-IDF e *word embeddings*.

Como trabalhos futuros, apesar da alta exigência de hardware para as soluções que envolvem aprendizado profundo, é interessante a avaliação de desempenho do algoritmo SVM frente as redes neurais multicamadas e *Long-Short Term Memory Networks*, pois já são vistos em outros trabalhos como [Mueller 2016], a capacidade destas para tratar representações e modelagens semânticas complexas com o objetivo de mensurar a similaridade entre sentenças.

Referências

[Agirre et al. 2012] Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, number 3, pages 385–393.

[Alves et al. 2016] Alves, A. O., Rodrigues, R., and Oliveira, H. G. (2016). ASAPP: Alinhamento Semântico Automático de Palavras aplicado ao Português. In *PROPOR - International Conference on the Computational Processing of Portuguese*, Tomar, Portugal.

[Barbosa et al. 2016] Barbosa, L., Cavalin, P., Guimarães, V., and Kormaksson, M. (2016). Blue Man Group at ASSIN: Using Distributed Representations for Semantic Similarity and Entailment Recognition. In *PROPOR - International Conference on the Computational Processing of Portuguese*, Tomar, Portugal.

[Bruckschen et al. 2008] Bruckschen, M., Muniz, F., Guilherme, J., De Souza, C., Fuchs, J. T., Infante, K., Muniz, M., Gonçalves, P. N., Vieira, R., and Aluísio, S. (2008). Anotação Linguística em XML do Corpus PLN-BR. Technical report, Universidade de São Paulo, São Paulo.

[Cançado 2013] Cançado, M. (2013). *Manual de Semântica: Noções Básicas e Exercícios*. UFMG.

[Chopra et al. 2005] Chopra, S., Hadsell, R., and Y., L. (2005). Learning a similiarty metric discriminatively, with application to face verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 349–356.

[Ferreira et al. 2016] Ferreira, R., Lins, R. D., Simske, S. J., Freitas, F., and Riss, M. (2016). Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer Speech & Language*, 39:1–28.

[Fialho et al. 2016] Fialho, P., Marques, R., Martins, B., Coheur, L., and Quaresma, P. (2016). INESC-ID at ASSIN: medidor de similaridade semântica e classificador de inferência textual. In *PROPOR - International Conference on the Computational Processing of Portuguese*, Tomar, Portugal.

[Fonseca et al. 2016] Fonseca, E. R., Borges, L., Santos, D., Criscuolo, M., and Aluísio, S. M. (2016). ASSIN: Evaluation of Semantic Similarity and Textual Inference. In *PROPOR - International Conference on the Computational Processing of Portuguese*, Tomar, Portugal.

[Freire et al. 2016] Freire, J., Pinheiro, V., and Feitosa, D. (2016). LEC_UNIFOR no ASSIN: FlexSTS Um Framework para Similaridade Semântica Textual. In *PROPOR - International Conference on the Computational Processing of Portuguese*, Tomar, Portugal.

[Gomaa and Fahmy 2013] Gomaa, W. and Fahmy, A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18.

[Hartmann 2016] Hartmann, N. S. (2016). Solo Queue at ASSIN : Combinando Abordagens Tradicionais e Emergentes. In *PROPOR - International Conference on the Computational Processing of Portuguese*, page 6.

[Kao and Poteet 2007] Kao, A. and Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer London, London.

[Manning and Schütze 2000] Manning, C. D. and Schütze, H. (2000). Foundations of Natural Language Processing. *Reading*, page 678.

[Maziero et al. 2008] Maziero, E. G., Pardo, T. a. S., Di Felippo, A., and Dias-da Silva, B. C. (2008). A base de dados lexical e a interface web do TeP 2.0. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, page 390, New York, New York, USA. ACM Press.

[Metzler et al. 2007] Metzler, D., Dumais, S., and Meek, C. (2007). Similarity Measures for Short Segments of Text. In *Proceedings of the 29th European Conference on IR Research (ECIR 2007)*, volume 4425, pages 16–27.

[Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Interspeech*, (1):104–108.

[Mueller 2016] Mueller, J. (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. In *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*, number 2012, pages 2786–2792.

[Pennington et al. 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

[Pradhan et al. 2015] Pradhan, N., Manasi Gyanchandani, B., and Wadhvani, R. (2015). A Review on Text Similarity Technique used in IR and its Application. *International Journal of Computer Applications*, 120(9):975–8887.

[Simões and Guinovart 2014] Simões, A. and Guinovart, X. G. (2014). *Bootstrapping a Portuguese WordNet from Galician, Spanish and English Wordnets*, pages 239–248. Springer International Publishing, Cham.

[Witten et al. 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 4 edition.