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

Análise de Medidas de Similaridade Semântica na Tarefa de Reconhecimento de Implicação Textual

David B. Feitosa[1], **Vládia C. Pinheiro**[1]

[1]Programa de Pós-Graduação em Informática Aplicada (PPGIA)
Universidade de Fortaleza (UNIFOR)
Caixa Postal 60.811-905 – Fortaleza – CE – Brasil

davidfeitosa@gmail.com, vladiacelia@unifor.br

Abstract. *In this work, we present a feature-based approach to the RTE (Recognizing Text Entailment) task that verifies the similarity between two sentences including syntactic and semantic aspects. The selected features come from the winning work of the RTE task of the workshop ASSIN (Semantic Similarity Evaluation and Textual Inference) with some changes and addition of other semantic feature. The evaluation methodology consisted in replicating the task with the database used in the workshop, analyzing the results with and without the semantic features. Besides the numerical approach, we mention a symbolic one with its characteristics and limitations.*

Resumo. *Neste trabalho, apresentamos uma abordagem baseada no uso de features para a tarefa de RTE (Recognizing Text Entailment) que verifica a similaridade entre duas frases incluindo aspectos sintáticos e semântico. As features selecionadas são oriundas do trabalho vencedor da tarefa de RTE do workshop ASSIN (Avaliação de Similaridade Semântica e Inferência Textual) com algumas alterações e adições de outra feature semântica por nós. A metodologia de avaliação consistiu em replicar a tarefa com a base de dados usada no workshop, analisando os resultados com e sem as features semânticas. Além da abordagem numérica, citamos uma simbólica com suas características e limitações.*

1. Introdução

No uso de linguagem natural, um fenômeno comum é a existência de várias maneiras de se expressar, de forma idêntica ou similar, um significado [Sha et al. 2015]. Para descobrir a equivalência ou relação entre textos ou sentenças, a tarefa de Reconhecimento de Implicação Textual (em ingles, *Recognizing Textual Entailment - RTE*) é proposta como uma forma de avaliar se o significado de um texto "H" pode ser inferido de outro texto "T" [Dagan et al. 2006]. Ela é mais relaxada que a tarefa de inferência lógica pura, pois podemos considerar que "T infere H" ($T \rightarrow H$) se, tipicamente, um ser humano que ler T puder inferir que H é uma verdade provável, e não que T é condição suficiente para H (ou seja, sempre que T é verdade então H é verdade). A relação é direcional porque mesmo que "T infere H" seja verdade, o reverso "H infere T" é bem menos provável. Como exemplo de implicação textual, temos o par: ("Edgar Freitas Gomes da Silva nasceu no Funchal a 25 de Setembro de 1962, tem 53 anos, é casado e com um filho.", "Edgar Silva

nasceu em 1962 no Funchal.") e, como contraexemplo, o par: ("Eram doentes que estavam internados e debilitados pelas suas patologias.", "Estão identificados 30 doentes com a bactéria.").

Muitas aplicações de Processamento de Linguagem Natural (PLN), tais como: Resposta automática a perguntas, Recuperação de Informação, Sumarização ou Tradução automática de textos, Classificação de textos, dentre outras [Fialho et al. 2016], necessitam de sistemas eficientes para reconhecimento de implicação textual.

Para língua inglesa, desde 2005 são propostas competições de RTE. Como exemplo, tem-se as tarefas do eventos SEMEVAL – Semantic Evaluation [1] e PASCAL Recognizing Textual Entailment (RTE) Challenges [2] [Dagan et al., 2006]. Ambos reuniram excelentes sistemas para RTE que, principalmente, empregam técnicas superficiais (*shallow techniques*) tais como: sobreposição de termos, analise morfossintática e análise de dependência sintática [Vanderwende et al. 2006, Jijkoun and Rijke 2005, Malakasiotis and Androutsopoulos 2007, Haghighi et al. 2005]. O vencedor do último SEMEVAL atingiu uma acurácia de 77% com o uso de um algoritmo de máxima entropia e features sintáticas. Uma abordagem híbrida proposta em - simbólica e probabilística - de [Sha et al. 2015] atingiu uma acurácia de 85.16%.

Para língua portuguesa, o primeiro evento que propôs a tarefa RTE foi o Workshop ASSIN 2016 [3] , o qual reuniu 6 sistemas participantes para as variações do português brasileiro (PT-br) e o português de Portugal (PT-pt). O sistema proposto em [Fialho et al. 2016] foi o vencedor da competição e usa aprendizagem automática supervisionada, algoritmo SVM [Malakasiotis and Androutsopoulos 2007], explorando propriedades lexicais como Maior Subsequência Comum, Distância de Edição, Comprimento etc das sentenças T e H. O resultado de tal sistema, em termos de f-measure, foi de 0.69 e representa o estado da arte desta tarefa para o Português.

Neste trabalho, apresentamos uma análise de medidas de similaridade semântica para a tarefa RTE em textos da língua portuguesa, usando como referência o corpus de sentenças do ASSIN 2016. Uma série de experimentos foram realizados, visando analisar a relevância das métricas para RTE e a influência de bases de conhecimento léxico-semanticas como WordNet [Miller 1995]. Ao final, verificamos a influência das métricas semânticas e discorremos sobre os resultados.

2. Fundamentação Teórica

A tarefa de RTE busca identificar se um dado texto pode ser inferido de outro. Como exemplo, dadas as sentenças "Taciana trabalha na UNIFOR que fica em Fortaleza." e "UNIFOR está em Fortaleza", podemos afirmar que a segunda pode ser inferida da primeira? Quais as relações entre as sentenças que indicam a inferência? Vários dos sistemas propostos, tanto para língua inglesa como para o português, se baseiam em técnicas superficiais que com suporte em características sintáticas e lexicais dos textos. Nas subseções seguintes, apresentamos as principais métricas usadas e o estado da arte da RTE.

[1] http://alt.qcri.org/semeval2014/
[2] http://u.cs.biu.ac.il/~nlp/RTE1/Proceedings/
[3] http://propor2016.di.fc.ul.pt/?page_id=381

2.1. Métricas de RTE

Dentre as funções para a tarefa de RTE, segundo [Fialho et al. 2016], as que obtiveram os melhores resultados foram:

1. Soft TF-IDF: mede a similaridade entre representações vetoriais das frases, mas considera a métrica Jaro-Winkler como métrica de similaridade interna para encontrar palavras equivalentes, com um limiar de 0.9. A métrica Jaro-Winkler atribui maior peso quando há um prefixo em comum [TeamCohen 2016];
2. Jaccard: distância entre os dois conjuntos como a razão entre o tamanho da interseção e o da união. Portanto, um valor 1 significa que as frases são iguais e 0, totalmente diferentes [TeamCohen 2016];
3. Comprimento: representa a diferença de comprimento absoluta (número de símbolos) entre o texto e a hipótese. Os comprimentos máximo e mínimo são também considerados (separadamente) como características;
4. LCS (Longest Common Subsequence): representa o tamanho da maior subsequência comum entre o texto e a hipótese. O valor é definido entre 0 e 1, dividindo-se o tamanho da LCS pelo tamanho da frase mais longa [Hirschberg 1977];
5. Numérica: consiste no resultado da multiplicação de duas similaridades de Jaccard. Uma entre os caracteres numéricos no par texto-hipótese, e outra entre as palavras em torno de tais caracteres numéricos. O resultado é um valor contínuo entre 0 e 1, com o valor 0 indicando que as frases são, possivelmente, contraditórias;
6. Sobreposição NE: mede a similaridade de Jaccard considerando apenas as entidades mencionadas (NE - Named Entities), ou seja, que contém letras maiúsculas;
7. ROUGE-N: representa a sobreposição de n-gramas com base em estatísticas de co-ocorrências [Lin and Och 2004];
8. ROUGE-L: representa uma variação da métrica ROUGE-N baseada em skip-bigrams [Lin and Och 2004].;
9. TER (Taxa de Erros de Tradução): consiste em uma extensão da Taxa de Erros em Palavras (ou Word Error Rate - WER), que é uma métrica simples baseada em programação dinâmica e que é definida como o número de alterações necessárias para transformar uma sequência em outra [Snover et al. 2006].

2.2. Trabalhos Relacionados

[Lai and Hockenmaier 2014] descreve o sistema vencedor da tarefa de RTE no SEMEVAL em 2014 para lingua inglesa. O sistema combina diferentes fontes semânticas para predizer a relação e implicação textual. Foram usadas features de similaridade distribuídas, similaridade denotacional e de alinhamento baseadas em estruturas sintáticas superficiais. Houve combinações de múltiplas métricas, dentre elas: negação, sobreposição de palavras, sinônimo, hiperônimo. Para a tarefa, foi usada MALLET [McCallum 2002] com um algoritmo de máxima entropia. O resultado final foi de uma acurácia de 77% sobre a base fornecida.

O sistema de [Sha et al. 2015] propõe uma abordagem híbrida, também para o inglês, que utiliza o sistema ReVerb [Fader et al. 2011] para extrair relações verbais na forma $relacao$ ($objeto_1$, $objeto_2$) de cada frase do par T/H. Com essas relações extraídas, ocorre um mapeamento com a base de conhecimento Yago [Suchanek et al. 2007]

de onde também são obtidas regras de relacionamento, atraves de um sistema de rule learner – AMIE [ref]. Um exemplo de regra aprendida pelo AMIE é "se uma pessoa mora na cidade A e trabalha na empresa B, a localização da empresa B é a cidade A". Esse conjunto de informações é a entrada para uma Rede Lógica de Markov [Richardson and Domingos 2006] que avalia a probabilidade do conjunto de relações com as regras da base de conhecimento resultar em uma situação de *entailment* entre T e H. Em testes para reproduzir o trabalho de [Sha et al. 2015],encontramos uma série de limitações como restrições às sentenças com pelo menos um verbo e dois objetos; esparsidade da base de dados que não permite o correto mapeamento da relação verbal e a extração das regras, ao extrair a relação da frase "The dog should sleep in the bed with you", o verbo *sleep* não existia na base para que a relação fosse mapeada.

No Workshop ASSIN, a tarefa de RTE foi proposta para as duas variantes da língua portuguesa. L2F/INESC-ID [Fialho et al. 2016] atingiu 0,70 em termos de f-measure de 0.7 e foi o vencedor para a variação do português europeu. O sistema é baseado no uso de features sintáticas e aprendizado supervisionado com SVM.

[Barbosa et al. 2016] apresenta o sistema da equipe Blue Man Group, que também é baseado em um classificador supervisionado (SVM). A diferença de abordagem é que este sistema usa como característica a similaridade semântica baseada nas palavras anteriores e posteriores relativas à palavra analisada na frase. A base usada para aprendizado foi a Wikipedia [Wikipedia 2014] em português com um total aproximado de 540.000 palavras distintas. A ferramenta *word2vec* [4] foi utilizada para o cálculo dos vetores. Embora outras técnicas tenham sido testadas,o algoritmo SVM foi o que apresentou melhor desempenho com f-measure de 0,52 para o português europeu e f-measure de 0,61 para o português brasileiro.

Como melhor resultado geral em acurácia, a abordagem de [Oliveira Alves et al. 2016], ASAPP, obteve um valor de 80.27% e f-measure de 0.54. Perdendo apenas na f-measure para Blue Man Group que obteve, no geral, acurácia de 79.62% e f-measure de 0.58. ASAPP [Oliveira Alves et al. 2016] é baseado em múltiplas features lexicais, sintáticas, semânticas - semelhança entre a vizinhança das palavras, estrutura das redes de palavras, presença e pertença em synsets difusos; aplicando o classificador Vote e AdditiveRegression . Apresentou desempenho com f-measure de 0,54 e um valor de 80.27% para acurácia.

Como estado da arte para a tarefa de RTE para o português, temos o sistema L2F/INESC-ID com f-measure de 0.7 (português europeu) e o sistema Blue Man Group com f-maesure de 0.52 (português brasileiro).

3. Similaridade Semântica Aplicada a RTE

A figura 1 ilustra o fluxo do processo de RTE com adição de features semânticas. Em resumo, inicialmente uma etapa de Pre-Processamento realiza a limpeza - remoção de stopwords e pontuação - ou formatação - transformando todos os termos em minúsculos ou na simbologia do Metaphone 3. Em seguida são calculadas as features sintáticas e semânticas, com uso de uma base de conhecimento (Knowledge Base – KB). Por fim, é gerado o conjunto de exemplos para treinamento e submetido a um algoritmo de aprendizagem supervisionada.

[4] `http://code.google.com/archive/p/word2vec/`

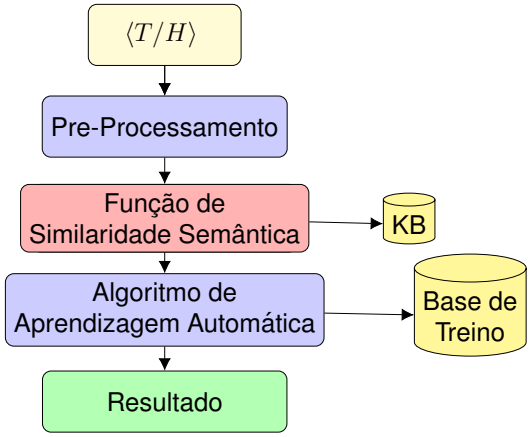

Figura 1. Fluxo para geração das features semânticas.

3.1. Similaridade Semântica Textual

Cada feature semântica é calculada a partir de uma função de similaridade semântica textual, variando-se a métrica de similaridade entre palavras.

A função de similaridade semântica textual STS implementa a fórmula 1, pela qual calcula-se a média aritmética da similaridade semântica entre as palavras ou termos de T e H. Assim, se $T = \{t_1, t_2, \ldots, t_n\}$ e $H = \{h_1, h_2, \ldots, h_m\}$, o produto cartesiano $T \times H$, $n \times m$, denotado por w, é o conjunto das combinações dos termos de T, denotado por $C_w = \{c_1 : (w_{t_1}, w_{h_1}), c_2 : (w_{t_1}, w_{h_2}), c_3 : (w_{t_1}, w_{h_3}), \ldots, c_w : (w_{t_n}, w_{h_m})\}$; Temos que $f(c_i)$ é uma função de similaridade entre palavras que representa uma das métricas descritas na seção a seguir.

$$STS(T, H) = \frac{\sum_{i=1}^{w} f(c_i)}{w} \quad (1)$$

3.2. Métricas de Similaridade Semântica entre Palavras

A seguir são descritas as métricas de similaridades entre palavras, calculadas na base léxico-semântica WordNet [Miller 1995]. A WordNet é um grande banco de dados léxico da língua inglesa. Substantivos, verbos, adjetivos e advérbios são agrupados em conjuntos de sinônimos cognitivos (synsets), cada um expressando um conceito distinto. Os synsets estão interligados por meio de relações conceituais semânticas e lexicais. Ela se assemelha superficialmente a um dicionário de ideias afins, na medida em que agrupa as palavras em conjunto com base em seus significados. No entanto, existem algumas distinções importantes. Primeiramente, a WordNet interliga não apenas palavras, mas sentidos específicos. Como resultado, palavras que são encontradas em estreita proximidade umas com as outras na rede estarão semanticamente próximas. Em segundo lugar, a WordNet rotula as relações semânticas entre palavras, enquanto que os agrupamentos em

um dicionario de sinônimos não seguem nenhum padrão explícito alem da similaridade de significado.

1. HirstStOnge: dois conceitos lexicalizados são semanticamente próximos se seus conjuntos de sinônimos (ou synsets) na WordNet são conectados por um caminho que não é muito longo e que "não muda de direção com muita freqüência" [Hirst et al. 1998].;
2. LeacockChodorow: esta medida baseia-se no comprimento do caminho mais curto entre dois conjuntos de synsets para sua medida de similaridade. Limitando-se às relações do tipo É-UM e escalando o comprimento do percurso pela profundidade total da taxonomia [Leacock and Chodorow 1998];
3. Lesk: em 1985, Lesk, propôs que a relação de duas palavras é proporcional à extensão das sobreposições de suas definições de dicionário [Banerjee and Pedersen 2002] estenderam essa noção para usar o WordNet como dicionário para as definições de palavras;
4. WuPalmer: a medida proposta por [Wu and Palmer 1994] calcula o relacionamento considerando as profundidades dos dois synsets nas taxonomias da WordNet, junto com a profundidade do LCS;
5. Resnik: [Resnik 1995] define a similaridade entre dois synsets para ser o conteúdo de informação de seu superordenado mais baixo;
6. JiangConrath: também usa a noção de conteúdo de informação (IC), mas sob a forma da probabilidade condicional de encontrar uma instância de um synset filho dado uma instância de um synset pai: $\frac{1}{jcn\_distance}$, onde $jcn\_distance$ é igual a $IC(synset_1) + IC(synset_2) - 2 * IC(lcs)$ [Jiang and Conrath 1997];
7. Lin: a equação matemática proposta por [Jiang and Conrath 1997] é modificada e o valor de relacionamento será maior ou igual a zero e menor ou igual a um [Lin et al. 1998];
8. Path: esta métrica assume o valor de -1 se não houver distância entre os synsets e, $\frac{1}{distancia}$, se a distância for maior que 0.

4. Avaliação Experimental

Neste trabalho, a hipótese de pesquisa é que a adição de conhecimento semântico melhora a performance da tarefa de RTE e os experimentos realizados visaram verificar esta hipótese.

4.1. Configuração e Metodologia de Avaliação

O procedimento adotado para a avaliação consistiu em:

1. Iniciamos com a verificação da base de dados fornecida pelo workshop ASSIN;
2. Usamos um algoritmo de balanceamento para equilibrar as classes fornecidas pela base;
3. Testamos nossa hipótese com o algoritmo de aprendizagem automática.

As features elaboradas por [Fialho et al. 2016] são uma combinação entre uma representação que pode alterar o estados das palavras da frase, como os termos estarem em minúsculas ou convertidos a uma forma fonética, com uma função. Em relação às representações citadas na seção 4, optamos por uma única alteração: no lugar do Double

Metaphone, usamos o Metaphone 3 (M3) [Philips 2010]. M3 foi projetado para retornar uma chave fonética "aproximada" - e uma chave fonética alternativa quando apropriado - que deve ser a mesma quando o idioma for o inglês.

Os cenários definidos para a avaliação foram:

1. Cenário 1 - Execução com as features sintáticas usadas em [Fialho et al. 2016];
2. Cenário 2 - Execução com todas as features sintáticas e todas as features semânticas calculadas para toda variação de métrica de similaridade entre palavras, definidas em 3.2;
3. Cenário 3 - Execução com todas as sintáticas e uma feature semântica por vez.

Os seguintes parâmetros e ferramentas foram adotados para realização dos experimentos:

- SMOTE: algoritmo de balanceamento entre as classes da base de teste [Chawla et al. 2002]. Único atributo alterado foi o *percentage* valorado em 181;
- Base de dados fornecida pelo workshop ASSIN. A base de testes que usamos consistia de 3000 entidades anotadas, mas a classe de "Entailment" representava 24% do total das instâncias. Para nossos experimentos, foram mantidas as classes "None" e "Entailment";
- SVM: o algoritmo de aprendizagem automática. Usamos a implementação fornecida no Weka com as configurações padrões, *cross validation folds* valorado em 10;
- As combinações entre funções e representações: alterações nas palavras das frases - que mais contribuíram na classificação foram: a) Soft TF-IDF, em símbolos originais, b) Jaccard, sobre Metaphone 3, c) Jaccard, sobre símbolos em minúsculas, d) Comprimento absoluto, em Metaphone 3, e) Maior subsequência comum (LCS), sobre símbolos em minúsculas, f) Numérica, em símbolos originais, g) Sobreposição NE, em Metaphone 3, h) ROUGE-N, em símbolos originais, i) ROUGE-L, sobre símbolos em minúsculas e j) TER, sobre símbolos em minúsculas.

A tabela 1 resume os resultados dos três cenários em termos de das medidas de precisão, recall e f-measure.

Conforme coletado, embora exista ganho, foi bem reduzido o acréscimo nas métricas observadas: para a precisão, não houve acréscimo em quaisquer casos; no recall, uma tênue variação; na f-measure, uma variação ligeiramente maior que as anteriores.

5. Análise dos Resultados

Pelos resultados obtidos, não houve confirmação forte da hipótese levantada, ou seja, adicionar informações semânticas não importa em acréscimo de performance significativo da tarefa de RTE. No Cenario 1 foi obtido f-measure de 0,706, e o melhor cenário com adição de feature semântica foi a variação do CENARIO 3 (Sint + PATH), que resultou f-measure de 0,710. Com o objetivo de identificar razões para a irrelevância de conhecimento semântico, foi analisado o corpus de referência usado no Workshop ASSIN.

Por observação através de análise gráfica no corpus, verificamos a distribuição de valores. A função projetada foi valor da variável PATH em relação ao número de

Cenário	Configuração	Precisão	Recall	F-Measure
1	Todas as Sintáticas (S)	0,710	0,707	0,706
2	S + Todas Semânticas	0,708	0,702	0,708
3	S + Hirststonge	0,705	0,706	0,701
3	S + Lesk	0,707	0,705	0,704
3	S + Leacockchodorow	0,709	0,709	0,706
3	S + Wupalmer	0,710	0,710	0,708
3	S + Jiangconrath	0,710	0,709	0,709
3	S + Lin	0,710	0,709	0,709
3	S + Resnik	0,710	0,710	0,709
3	S + Path	0,710	0,708	0,710

Tabela 1. Resultados dos cenários de avaliação, considerando features semânticas e sintáticas.

sobreposições dos termos em comum entre T e H. Verificamos uma concentração de valores próximos ao quartil inferior por toda a amostra e um incremento de valores com a redução da sobreposição, implicando em alto valor de relacionamento semântico.

6. Conclusão

Neste artigo, realizamos o levantamento da importância da tarefa de RTE para a área de PLN. Levantamos os melhores trabalhos desenvolvidos em workshops e competições tanto para língua inglesa, quanto para a portuguesa. Dentre os trabalhos para o português, citamos as abordagens dos vencedores do último workshop que abordou a tarefa. Implementamos a solução do vencedor desse workshop com as features sintáticas adicionando combinações de semânticas - nossa feature de teste para este trabalho. Coletamos os resultados e os apresentamos para análise.

Seguindo as hipóteses levantadas na seção 5, prosseguiremos os testes com o auxílio de bases maiores e, possivelmente, mais representativas para verificar se há evolução nos resultados por adição do conhecimento semântico da nossa feature ou se há necessidade de ajustes dessa para agregar valor ao processo de classificação.

Referências

Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 136–145. Springer.

Barbosa, L., Cavalin, P., Guimaraes, V., and Kormaksson, M. (2016). Blue man group no assin: Usando representações distribuídas para similaridade semântica e inferência textual. Linguamática, 8(2):15–22.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment, pages 177–190. Springer.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1535–1545. Association for Computational Linguistics.

Fialho, P., Marques, R., Martins, B., Coheur, L., and Quaresma, P. (2016). Inesc-id@ assin: Mediçao de similaridade semântica e reconhecimento de inferência textual. Linguamática, 8(2):33–42.

Haghighi, A. D., Ng, A. Y., and Manning, C. D. (2005). Robust textual inference via graph matching. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 387–394. Association for Computational Linguistics.

Hirschberg, D. S. (1977). Algorithms for the longest common subsequence problem. Journal of the ACM (JACM), 24(4):664–675.

Hirst, G., St-Onge, D., et al. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database, 305:305–332.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

Jijkoun, V. and Rijke, M. (2005). Recognizing textual entailment using lexical similarity. In Proceedings of the PASCAL Challenge Workshop on Recognising Textual Entailment, 2005, pages 73–76.

Lai, A. and Hockenmaier, J. (2014). Illinois-lh: A denotational and distributional approach to semantics. In SemEval 2014. Special Interest Group on the Lexicon of the Association for Computational Linguistics.

Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. WordNet: An electronic lexical database, 49(2):265–283.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the

42nd Annual Meeting on Association for Computational Linguistics, page 605. Association for Computational Linguistics.

Lin, D. et al. (1998). An information-theoretic definition of similarity. In ICML, volume 98, pages 296–304. Citeseer.

Malakasiotis, P. and Androutsopoulos, I. (2007). Learning textual entailment using svms and string similarity measures. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 42–47. Association for Computational Linguistics.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Miller, G. A. (1995). Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

Oliveira Alves, A., Rodrigues, R., and Gonçalo Oliveira, H. (2016). Asapp: alinhamento semântico automático de palavras aplicado ao português. Linguamática, 8(2):43–58.

Philips, L. (2010). Metaphone 3. "Disponível em `https://searchcode.com/codesearch/view/2366000/` versão 2.1.3".

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.

Richardson, M. and Domingos, P. (2006). Markov logic networks. Machine learning, 62(1):107–136.

Sha, L., Li, S., Chang, B., Sui, Z., and Jiang, T. (2015). Recognizing textual entailment using probabilistic inference. In EMNLP, pages 1620–1625.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006). Cambridge, Massachusetts.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web, pages 697–706. ACM.

TeamCohen (2016). secondstring. `https://github.com/TeamCohen/secondstring`.

Vanderwende, L., Menezes, A., and Snow, R. (2006). Microsoft research at rte-2: Syntactic contributions in the entailment task: an implementation. In Proceedings of the Second PASCAL Recognising Textual Entailment Challenge, pages 27–32.

Wikipedia (2014). Wikipedia, the free encyclopedia. [Acessada em 2014].

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133–138. Association for Computational Linguistics.