Solo Queue at ASSIN: Mix of a Traditional and an Emerging Approaches

Solo Queue at ASSIN: Purpose and Results

Nathan Siegle Hartmann
Universidade de São Paulo
nathansh@icmc.usp.br

Resumo

No presente artigo apresentamos uma proposta para atribuição automática da similaridade entre duas sentenças, tarefa definida na avaliação conjunta ASSIN 2016. Nossa proposta consiste no uso de uma *feature* clássica da classe *bag-of-words*, a TF-IDF; e uma *feature* emergente, capturada poe meio de *word embeddings*. Sabe-se que a medida TF-IDF pode ser utilizada para relacionar documentos que contém os mesmos elementos e, portanto, pode ser utilizada para documentos que compartilham palavras. *Word embeddings* são conhecidas por modelar a sintaxe e semântica das palavras e, segundo Mikolov et al. (2013a), podem ser utilizadas para modelar a *embedding* de uma sentença. Ao considerar ambas as *features*, ponderamos as palavras contidas nas sentenças e a semântica compartilhada entre elas. Como o rótulo de similaridade para o problema em questão é um valor real na escala entre 1 e 5, aplicamos uma técnica de regressão, a Regressão Linear. Os resultados obtidos mostraram que, apesar da *feature* de *embeddings* ter obtido resultados similares ao sistema *baseline*, ao ser combinada à *feature* TF-IDF, apresentou resultados levemente superiores aos obtidos ao ser usada somente a segunda *feature*. Os resultados obtidos nesse trabalho obtiveram a primeira colocação no ASSIN 2016 entre os trabalhos que participaram da tarefa de similaridade textual para português do Brasil e segunda colocação para português de Portugal.

Palavras chave

Similaridade Sentencial, *word embeddings*, Aprendizagem de Máquina

Abstract

In this paper we present a proposal to automatically label the similarity between a pair of sentences and the results obtained on ASSIN 2016 sentence similarity shared-task. Our proposal consists of using a classical feature of bag-of-words, the TF-IDF model; and an emergent feature, obtained from processing word embeddings. The TF-IDF is used to relate texts which share words. Word embeddings are known by capture the syntax and semantics of a word. Following Mikolov et al. (2013a), the sum of embedding vectors can model the meaning of a sentence. Using both features, we are able to capture the words shared between sentences and their semantics. We use linear regression to solve this problem, once the dataset is labeled as real numbers between 1 and 5. Our results are promising. Although the usage of embeddings has not overcome our baseline system, when we combined it with TF-IDF, our system achieved better results than only using TF-IDF. Our results achieved the first collocation of ASSIN 2016 for sentence similarity shared-task applied on brazilian portuguese sentences and second collocation when applying to Portugal portuguese sentences.

Keywords

Sentence Similarity, word embeddings, Machine Learning

1 Introdução

Pesquisas sobre similaridade entre documentos se iniciaram com foco na área de Recuperação de Informação em que, dada uma *query*, retorna os documentos mais similares a ela. A literatura apresenta diferentes abordagens para modelar a similaridade entre documentos. Podemos citar: abordagens por palavras (*bag-of-words*) que calculam a similariade lexical, ou n-grams (Salton, 1989) (Damashek, 1995) que conseguem capturar a semântica contida nas sequências de n palavras; e também abordagens mais complexas como *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990) (Landauer e Dumais, 1997), que visa calcular a similaridade semântica de todo o documento, e não apenas a lexical.

Entre os trabalhos clássicos da literatura de similaridade de documentos, podemos citar trabalhos que calcularam a similaridade textual de uma perspectiva matemática, utilizando estatística ou teoria de probabilidade (Ponte

e Croft, 1998), trabalhos que utilizam recursos léxicos para calcular a semântica em um parágrafo ou no documento (Rada et al., 1989) (Resnik, 1995) e outros trabalhos que combinam todas essas ideias (Rodríguez e Egenhofer, 2003). Esses métodos, no entanto, possuem dificuldades em lidar com a esparsidade de dados, que não proporciona frequência suficiente para métodos probabilísticos nem ocorrência de algumas palavras contidas em recursos lexicais. Portanto, nenhum desses trabalhos é apropriado para lidar com a similaridade sentencial.

Trabalhos subsequentes foram desenvolvidos com o propósito de lidar com a esparsidade de dados na similaridade sentencial (Li et al., 2006) (Liu, Zhou e Zheng, 2007). No entanto, esses trabalhos possuem a deficiência de serem dependentes de córpus ou *wordnet*. Essa dependência restringe um método, muitas vezes, a ser aplicado apenas a uma determinada língua devido à características únicas dessa língua, buscadas em um recurso compilado.

Trabalhos recentes utilizam o conceito de *embeddings* (Mikolov et al., 2013b) para calcular a similaridade entre sentenças, parágrafos e documentos. A vantagem dessa abordagem é, além da baixa esparsidade de dados, a independência de recursos léxicos, sintáticos e semânticos. Um modelo de *embeddings* necessita unicamente de um grande córpus de treinamento que, se for apropriado para a tarefa alvo, modelará bem o contexto das palavras e não acarretará na esparsidade de dados. Podemos citar o trabalho de Kenter e de Rijke (2015) que utilizou *word embeddings* para calcular a similaridade semântica entre textos curtos. Os autores treinaram um modelo de *embeddings* utilizando um córpus de 100 bilhões de palavras obtidas do *website* Google News. O gênero jornalístico é comumente utilizado para treinamento de *embeddings* por ser um gênero genérico, o que não limita o modelo treinado à um determinado cenário ou aplicação.

Esse trabalho apresenta uma proposta simples para cálculo da similaridade sentencial. Utilizamos uma *feature* clássica, a TF-IDF (*term frequency–inverse document frequency*), e também uma *feature* emergente, obtida por meio de *word embeddings*. As próximas seções seguem do seguinte modo: na Seção 2, são apresentadas as duas *features* propostas nesse trabalho e também a *baseline*, desenvolvida para validar a eficácia das *features* propostas; na Seção 3, são apresentados os resultados obtidos e uma breve discussão sobre eles; na Seção 4, são descritos alguns trabalhos relacionados, recuperados da SemEval-2014 Task 1, cujo objetivo também foi o cálculo da similaridade sentencial e; na Seção 5, são listadas as conclusões desse trabalho.

2 Features

Nesse trabalho, propomos o uso de duas *features*: uma relacionada com *word embeddings* e outra com o modelo TF-IDF. Também propomos uma *feature baseline* para validar a eficácia das *features* propostas. Nas subseções a seguir, apresentamos as *features* utilizadas nesse trabalho e a motivação para seu uso: na Subseção 2.1, detalhamos a *feature* obtido por meio de *word embeddings*; na Subseção 2.2, detalhamos a *feature* obtida por TF-IDF e, na Subseção 2.3, apresentamos a *feature baseline*.

2.1 Word Embeddings

A abordagem para modelagem de palavras no espaço vetorial utilizada nesse trabalho foi a Skip-Ngram, proposta por Mikolov et al. (2013b). Essa abordagem se baseou nos tradicionais modelos de língua, no entanto, ao invés de utilizar uma sequência de n palavras para predizer a palavra no instante $n+1$, ela utiliza uma única palavra i para predizer a janela j de palavras ao seu redor. Dessa forma, a *embedding* de uma palavra representa o contexto no qual ela ocorre, capturando relações sintáticas e semânticas. Um exemplo clássico da literatura para a língua inglesa mostra que ao subtrair o vetor da *embedding* de *homem* do vetor da *embeddings* de *rei* e somar o vetor da *embeddings* de *mulher*, chega-se a um *embedding* cujo vetor é muito similar ao de *rainha*. Com esse exemplo percebemos que a troca do gênero muda o substantivo em si, mas mantém a semântica correta, a versão feminina de *rei*.

Utilizamos o sistema word2vec[1] para a modelagem das *embeddings* por contér o algoritmo de treinamento Skip-Ngram. O córpus utilizado para treinamento contém 3 bilhões de tokens em português brasileiro, composto por textos do *website* G1, da Wikipédia e do córpus PLN-Br (Bruckschen et al., 2008). Definimos que cada *embedding* seria composta por um vetor de 600 dimensões, tamanho considerado suficiente nos experimentos realizados por Mikolov et al. (2013a). Todas as palavras foram mapeadas para caixa baixa a fim de reduzir esparsidade de dados no córpus. Também definiu-se um mapeamento das palavras com apenas uma ocorrência no córpus para um token genérico *UNK*. Toda

[1]Disponível em `https://code.google.com/archive/p/word2vec/`.

nova palavra não encontrada no vocabulário do córpus de treinamento também é mapeada para a *embedding* de *UNK*. É interessante observar que foi possível replicar o exemplo *rei-rainha*, clássico na literatura de *embeddings* da língua inglesa, para o nosso modelo treinado com textos em português brasileiro. Isso reforça que a abordagem de *embeddings* é independente de língua, dependendo apenas do córpus de treinamento.

Para calcularmos a similaridade entre os pares de sentenças, utilizamos o modelo treinado de *word embeddings* para representar as sentenças. O trabalho de Mikolov et al. (2013b) mostra que ao somar os vetores das *embeddings* das palavras de uma sentença, temos como resultado uma *embedding* que representa a sentença. Apesar de não terem sido encontrados trabalhos na literatura que avaliem a qualidade com que a composição de *embeddings* representa uma sentença, intuitivamente percebemos que, se a *embedding* de uma palavra representa o contexto em que ela ocorre, a soma das *embeddings* dessas palavras compõe a soma dos seus contextos. Uma abordagem similar para a tarefa de similaridade textual foi abordada por Bjerva et al. (2014) na SemEval-2014 Task 1. Os autores utilizaram, entre outras *features*, a similaridade do cosseno entre as somas das *embeddings* das sentenças. O sistema desenvolvido pelos autores obteve a terceira melhor colocação na tarefa de similaridade textual da SemEval-2014 Task 1.

O uso das *embeddings* como *feature* é dado pela similaridade do cosseno entre as *embeddings* dos pares de sentenças. O valor da similaridade entre os dois vetores de *embeddings* é utilizamo como uma *feature* para o sistema de regressão.

2.2 TF-IDF

A fim de utilizar uma abordagem clássica da área de PLN (Processamento de Linguagem Natural) para representação sentencial, realizamos uma modelagem TF-IDF das sentenças do córpus. Sabendo que a modelagem TF-IDF sofre com a esparsidade de dados, utilizamos apenas os *stems* das palavras de conteúdo das sentenças para representá-las, conseguindo dessa forma uma matriz TF-IDF reduzida. Além disso, sabemos que as sentenças a serem avaliadas são curtas e que não necessariamente contém as mesmas palavras. Assim, expandimos o vocabulário das sentenças buscando sinônimos para cada palavra de conteúdo no TEP (Thesaurus para o português do Brasil) (Maziero e Pardo, 2008). Verificamos que, ao expandir os sinônimos para todas as palavras de conteúdo de uma sentença, os vetores TF-IDF das sentenças se tornam muito similares, de forma a não conseguirmos distinguir sentenças similares das distintas. Portanto, empiricamente, limitamos a expansão de sinônimos para palavras de conteúdo que possuem até 2 sinônimos no TEP. Essa decisão foi tomada com base em experimentos no conjunto de treinamento disponibilizado pela comissão organizadora do ASSIN.

O uso do TF-IDF como *feature* é dado pela distância do cosseno entre os vetores TF-IDF dos pares de sentenças. Utilizamos esse valor como uma *feature* para o sistema de regressão.

2.3 Baseline

A fim de avaliar a eficácia das *features* propostas nesse trabalho, elaboramos um *baseline* para avaliação. A *feature baseline* consiste na proporção de palavras compartilhadas entre as duas sentenças. Essa *feature* não captura a semântica latente das sentenças. Por exemplo, mesmo que duas sentenças compartilhem uma quantidade substancal de palavras, um sinal de negação contido em uma dessas sentenças pode inverter o seu significado em relação a outra sentença. Assim, as *features* propostas são eficazes se capturarem informações latentes sobre as sentenças, de forma a proporcionar uma melhor performance ao sistema que automatiza a similaridade sentencial.

3 Experimentos

Nós treinamos 2 sistemas de Regressão Linear com os conjuntos de treinamento compostos por pares de sentença em português do Brasil (PTBR) e em português de Portugal (PTPT) disponibilizados pela comissão organizadora do ASSIN. Ambos os conjuntos contém 3,000 pares de sentenças cada. Cada sistema foi treinado com variação de *features*: utilizando a *feature baseline*; utilizando apenas *embeddings*; utilizando apenas *TF-IDF*; e uma versão utilizando *embeddings* e TF-IDF. Avaliamos as versões PTBR do nosso sistema sobre o conjunto de teste disponibilizado na *shared-task*, composto por 2,000 pares de sentenças em PTBR. Analogamente, avaliamos as versões PTPT do nosso sistema sobre o conjunto de testes PTPT da *shared-task*. Utilizamos as medidas Correlação de Pearson (CP) e Erro Quadrado Médio (EQM) para avaliar a qualidade das *features* propostas na tarefa de similaridade sentencial via método de regressão.

Verificando os resultados apresentados na Tabela 1, percebemos que o uso apenas da *feature* obtida das *word embeddings* não resultou em uma boa performance da Regressão Linear.

Feature	PT-BR		PT-PT	
	CP	EQM	CP	EQM
Baseline	0,57	0,50	0,60	0,49
Embeddings	0,58	0,50	0,55	0,83
TF-IDF	0,68	0,41	0,70	**0,39**
Embeddings + TF-IDF	**0,70**	**0,38**	**0,70**	0,66

Tabela 1: Avaliação das *features* propostas para cálculo de similaridade sentencial, utilizando Regressão Linear, nos conjuntos de teste da ASSIN *shared-task*.

Entendemos que, apesar da literatura apontar que a soma das *embeddings* de uma sequência de palavras representar a sintaxe-semântica dessa sequência, essa representação se torna genérica, não representando de fato a informação ali contida. Também devemos ponderar que, como o modelo de *embeddings* foi gerado sobre textos em PTBR, ele não está calibrado para lidar com a variante da língua PTPT – o que justifica o aumento de EQM na avaliação sobre o conjunto PTPT ao adicionar a *feature Embeddings* à TF-IDF. Além disso, a soma das *embeddings* pode não ser a melhor forma de manipular essa informação. O trabalho de Gabrilovich e Markovitch (2007) propõe o ponderamento das *embeddings* das palavras de um documento pela frequência com que essas palavras aparecem na língua. O trabalho de Yuan et al. (2016) mostra que o uso dessa modelagem melhora a performance da tarefa de desambiguização lexical de sentidos ao utilizar redes neurais.

Os resultados também nos mostram que o uso da *feature* TF-IDF resultou em uma performance significativa da Regressão Linear em relação ao uso da *feature baseline*. É interessante observarmos que a representação TF-IDF segue o modelo *bag-of-words*, o que implica na perda da ordem das palavras e na semântica latente. Não podemos afirmar que o resultado final do nosso sistema, que utiliza ambas as *features*, é superior ao do sistema que utiliza apenas TF-IDF, devido a falta de um teste de significância estatística. No entanto, especulamos que o uso das *embeddings* contribui para que o sistema capture a semântica da sentença em casos em que o significado do contexto importa – cenário em que o TF-IDF é insuficiente.

Os resultados obtidos pelo sistema desenvolvido nesse trabalho obtiveram primeiro lugar entre os competidores ao aplicar o sistema no córpus PTBR e segundo lugar ao aplicar o sistema no córpus PTPT. No caso geral, ao unir os córpus PTBR e PTPT, nós fomos os melhores colocados, com **0,68** de CP e **0,52** de EQM.

4 Trabalhos Relacionados

O SemEval 2014 disponibilizou uma *shared-task* (SemEval-2014 Task 1)[2], cujo um dos objetivos foi calcular a similaridade sentencial de um par de sentenças. Foi disponibilizado um dataset, o SICK, que contém 10,000 pares de sentenças, sendo 5,000 pares para treinamento e 5,000 pares para teste. Essa *shared-task* inspirou a organização da ASSIN, competição com propósito similar cujo foco voltou-se para a língua portuguesa. Nessa seção serão listados três trabalhos do SemEval-2014 Task 1 que trataram de similaridade sentencial.

O trabalho de Zhao, Zhu e Lan (2014) considerou um vasto conjunto de *features*. Entre as *features* utilizadas, podemos citar: tamanho de sentenças, similaridade superficial (distância do cosseno), similaridade semântica, *ngrams* com base em córpus de referência, entre outras. Esse trabalho foi o primeiro colocado para a tarefa de similaridade sentencial, obtendo 0,828 de CP e 0,325 de EQM.

O trabalho de Bjerva (2014) utilizou uma variedade de *features*, das quais podemos citar: tamanho das sentenças, substantivos e verbos compartilhados entre as sentenças, diferenças entre os conceitos Wordnet das palavras das sentenças e distância do cosseno das *word embeddings* das sentenças. Esse trabalho foi o terceiro colocado para a tarefa de similaridade sentencial, obtendo 0,827 de CP e 0,322 de EQM.

O trabalho de Lai e Hockenmaier (2014) utiliza *features* que consideram a proporção de palavras compartilhadas entre as sentenças, alinhamento entre as sentenças, presença de negação e a similaridade semântica entre o conjunto de palavras não compartilhado entre as sentenças. Esse trabalho foi o quinto colocado para a tarefa de similaridade sentencial, com 0,799 de CP e 0,369 de EQM.

5 Conclusão

Esse artigo apresentou os resultados obtidos pela equipe *Solo Queue* na tarefa de similaridade textual da ASSIN *shared-task*. Nossa proposta consiste no uso de uma *feature* clássica da classe *bag-of-words*, a TF-IDF; e uma *feature* emergente, obtida por meio de *word embeddings*. Sabemos que a medida TF-IDF pode ser utilizada para relacionar documentos que compartilham palavras e, portanto, pode ser utilizada para relacionar sentenças. *Word embeddings* são conhecidas

[2]Anais disponíveis em `http://www.aclweb.org/anthology/S/S14/S14-2.pdf#page=349`.

por modelar o contexto das palavras e podem ser utilizadas para modelar o contexto de uma sentença. Nossa equipe obteve os melhores resultados da *shared-task* ao avaliar o sistema desenvolvido sobre o conjunto de teste de pares de sentença em português do Brasil e segundo lugar ao avaliar sobre o conjunto de teste de pares de sentença em português de Portugal. No caso geral de avaliação, em que juntou-se os córpus, nosso grupo foi o melhor colocado. Acreditamos que melhores resultados podem ser obtidos ao investigar-se uma melhor ponderação das *embeddings* das palavras para modelar a *embedding* de sua sentença, como apresentado por (Gabrilovich e Markovitch, 2007) (Yuan et al., 2016). Ainda assim, a composição das *embeddings* de uma sequência de palavras não mantém a ordem delas, perdendo parte da semântica contida na sentença. Para resolver esse problema, vale avaliar o uso de uma rede LSTM para modelar a *embedding* de uma sentença a partir das *embeddings* das palavras dessa sentença. Redes LSTM são conhecidas por manterem a ordem de entrada dos elementos (1997). Também sabemos que o fato do nosso conjunto de *embeddings* ter sido treinado apenas sobre textos em Português do Brasil desafiou o sistema a lidar com textos em Português de Portugal. Assim, o treinamento de um modelo de *embeddings* que contemple ambas as línguas é o ideal pois, apesar das línguas compartilharem muitas características, suas nuances geram desafios particulares que merecem atenção.

Agradecimentos

Agradecemos ao aporte financeiro da FAPESP (p. 2016/00500-1) que financia esse projeto de pesquisa.

Referências

Bjerva, Johannes, Johan Bos, Rob van der Goot, e Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. Em *SemEval 2014: International Workshop on Semantic Evaluation*, pp. 642–646.

Bruckschen, M., F. Muniz, J. Souza, J. Fuchs, K. Infante, M. Muniz, P. Gonçalves, R. Vieira, e S. Aluísio. 2008. Anotação Lingüística em XML do Corpus PLN-BR. NILC–TR–09–08. Relatório técnico, University of São Paulo, Brazil.

Damashek, Marc. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, e Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Gabrilovich, Evgeniy e Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. Em *IJCAI*, volume 7, pp. 1606–1611.

Hochreiter, Sepp e Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kenter, Tom e Maarten de Rijke. 2015. Short text similarity with word embeddings. Em *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1411–1420. ACM.

Lai, Alice e Julia Hockenmaier. 2014. Illinois-lh: A denotational and distributional approach to semantics. *Proc. SemEval*.

Landauer, Thomas K e Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Li, Yuhua, David McLean, Zuhair A Bandar, James D O'shea, e Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on*, 18(8):1138–1150.

Liu, Xiaoying, Yiming Zhou, e Ruoshi Zheng. 2007. Sentence similarity based on dynamic time warping. Em *Semantic Computing, 2007. ICSC 2007. International Conference on*, pp. 250–256. IEEE.

Maziero, Erick e Thiago Pardo. 2008. Interface de Acesso ao TeP 2.0 - Thesaurus para o português do Brasil. Relatório técnico, University of São Paulo, Brazil.

Mikolov, Tomas, Kai Chen, Greg Corrado, e Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, e Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. Em *Advances in neural information processing systems*, pp. 3111–3119.

Ponte, Jay M e W Bruce Croft. 1998. A language modeling approach to information retrieval. Em *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275–281. ACM.

Rada, Roy, Hafedh Mili, Ellen Bicknell, e Maria Blettner. 1989. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30.

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

Rodríguez, M Andrea e Max J Egenhofer. 2003. Determining semantic similarity among entity classes from different ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, 15(2):442–456.

Salton, Gerard. 1989. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*.

Yuan, Dayu, Ryan Doherty, Julian Richardson, Colin Evans, e Eric Altendorf. 2016. Word sense disambiguation with neural language models. *arXiv preprint arXiv:1603.07012*.

Zhao, Jiang, Tian Tian Zhu, e Man Lan. 2014. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. Em *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics and Dublin City University*, pp. 271–277.