Solo Queue at ASSIN: Mix of Traditional and Emerging Approaches

Nathan Siegle Hartmann
nathansh@icmc.usp.br

PROPOR 2016
July 13-15, 2016, Tomar, Portugal

Background

Measures of text similarity have been used for a long time in natural language processing applications and related research areas.

Text similarity has also been used for:

- Relevance feedback and text classification (Rocchio, 1971).
- Word sense disambiguation (Lesk, 1986; Schütze, 1998).
- Extractive summarization (Salton and Buckley, 1988).
- Machine translation (Papineni, 2001).
- Text summarization (Lin and Hovy, 2003).
- Text coherence (Lapata and Barzilay, 2005).

Background

There are different approaches to model the similarity of documents:

- Bag-of-words for lexical similarity.
- N-grams for semantics on sequence of words (Salton, 1989; Damashek, 1995).
- Latent Semantic Analysis (LSA) for semantics on a document (Deerwester et al., 1990; Landauer and Dumais, 1997).

Background

Other approaches to deal with text similarity use:

- Probability theory (Ponte and Croft, 1998).
- Lexical resources (Rada et al., 1989; Resnik, 1995).
- Both probability theory and lexical resources (Rodríguez and Egenhofer, 2003).

None of these works are appropriate to deal with sentence similarity because sentences pairs suffer of data sparsity.

Background

Recently several studies approached sentence similarity and its problem of data sparsity (Li et al., 2006; Liu et al., 2007):

- However, these works are dependent of corpora and/or lexical resources like wordnets.
- These dependencies limit the application of those approaches to other languages.
- Here, we are interested in language independent approaches.

Background

Word embedding has been used recently to measure similarity of sentences (Bjerva et al., 2014), paragraphs and documents (Kenter and de Rijke, 2015).

- The embedding approach is only dependent of a training corpus.
- Leads to low data sparsity if used with huge corpora.

Purpose

This work uses a classical feature (TF-IDF) and a more recent one (word embeddings) to propose a solution to the ASSIN Sentence Similarity shared-task.

- It is known that TF-IDF models well a document and it has been used for a long time to calculate similarity between documents.
- Word Embeddings model the context of a word and it can be useful when the context of a sentence matters.

Related work

SemEval-2014 Task 1 evaluated sentence similarity on english pairs of sentences.

A dataset called SICK was made available with 10,000 pairs of sentences: 5,000 pairs for training and 5,000 for testing.

Zhao et al. (2014):

- **Best results** at SICK: 0,828 Pearson Correlation (ρ) and 0,325 Mean Squared Error (MSE).
- Features: sentence length, cosine similarity, n-grams, etc.

Bjerva et al. (2014):

- Third best results at SICK: 0,827 ρ and 0,322 MSE.
- Features: sentence length, nouns and verbs shared between sentences, Wordnet synsets similarity, **embeddings**.

Algorithms

All our experiments were performed using:

- Linear Regression.
- SVR (linear).
- SVR (poly).
- SVR (RBF).

Because Linear Regression always had the best results, we only reported its results.

We used Pearson Correlation (ρ) and Mean Squared Error (MSE) to measure the performance of our systems.

We refer to Brazilian Portuguese as PT-BR and European Portuguese as PT-EU.

Baseline feature

Our baseline feature is the ratio of words shared between a pair of sentences.

It does not capture the semantics of a sentence.

Feature	PT-BR		PT-EU		Both	
	ρ	**MSE**	ρ	**MSE**	ρ	**MSE**
Baseline	0.57	0.50	0.60	0.49	0.59	0.50

Table 1: Evaluation of our baseline feature on ASSIN training dataset.

TF-IDF model

To model a TF-IDF representation of ASSIN training dataset, we had to investigate if a preprocessing step was necessary.

We tried three preprocessing methods and evaluated them on the training dataset using 10-fold cross-validation.

1. Tokens without stopwords and punctuation.
2. Stems without stopwords and punctuation.
3. Entities recognized by the parser Palavras (Bick, 2000) without stopwords and punctuation.

Feature	PT-BR		PT-EU		Both	
	ρ	MSE	ρ	MSE	ρ	MSE
Baseline	0.57	0.50	0.60	0.49	0.59	0.50
1	0.62	0.46	0.64	0.45	0.63	0.46
2	0.66	0.42	0.70	0.40	0.67	0.41
3	0.60	0.48	0.64	0.45	0.62	0.47

Table 2: Evaluation of tokens, stems and PALAVRAS entities to TF-IDF model on training data of ASSIN.

TF-IDF model

TF-IDF suffers with data sparsity because sentences are short and their size is a problem for TF-IDF model.

- We expanded our set of stems to better represent our sentences.
- We searched for synonyms on TEP thesaurus for BP (Maziero and Pardo, 2008).

We recursively searched for synonyms for different sets of words.

- We decided to expand synonyms of words that only have one synonym on TEP.
- Over-expansion of sentence stems made our TF-IDF model generic.

TF-IDF model

Recursion	Words	PT-BR ρ	PT-BR MSE	PT-EU ρ	PT-EU MSE	Both ρ	Both MSE
	Original stems	0.66	0.42	0.70	0.40	0.67	0.41
1	all	0.37	0.64	0.45	0.61	0.41	0.63
2	all	0.30	0.68	0.36	0.67	0.36	0.67
3	all	0.26	0.70	0.30	0.70	0.28	0.70
1	$\leq$ 10 syns	0.57	0.50	0.62	0.47	0.60	0.49
2	$\leq$ 10 syns	0.51	0.55	0.62	0.47	0.54	0.54
3	$\leq$ 10 syns	0.48	0.57	0.54	0.54	0.52	0.56
1	$\leq$ 4 syns	0.65	0.43	0.67	0.42	0.66	0.43
2	$\leq$ 4 syns	0.63	0.45	0.66	0.43	0.64	0.44
3	$\leq$ 4 syns	0.62	0.46	0.65	0.44	0.64	0.45
1	$\leq$ 2 syns	0.66	0.42	0.70	0.40	0.67	0.41
2	$\leq$ 2 syns	0.66	0.42	0.69	0.40	0.67	0.42
3	$\leq$ 2 syns	0.66	0.42	0.69	0.40	0.67	0.42
1	1 syn	0.67	0.41	0.70	0.39	0.68	0.41
2	1 syn	0.67	0.41	0.70	0.39	0.68	0.41
3	1 syn	0.67	0.41	0.70	0.39	0.68	0.41

Table 3: Evaluation of expansion recursive of stems to represent a sentence on a TF-IDF model (syns means synonyms).

TF-IDF model

Steps to get TF-IDF feature:

1. To remove stopwords and punctuation of a pair of sentences to reduce their TF-IDF matrices.
2. To use stems to reduce TF-IDF matrices.
3. To expand stems list by adding words synonyms. We only added synonym to a word that has just a synonym on TEP. It better describes rare words and better generalizes our TF-IDF model.
4. To calculate the cosine similarity between the two TF-IDF representations of a pair of sentences. This value is used as feature to our regression system.

Embedding Feature

An embedding representation can capture syntax and semantics of a word.

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

- We used word2vec package to create our embedding model.
- We used Skip-Ngram algorithm to model embeddings.
- We used a 600d array to embed a word (Mikolov et al., 2013).
- We used a Brazilian Portuguese corpus of 3B words compiled from website G1, Wikipédia and PLN-Br corpus (Bruckschen et al., 2008).
- All words were mapped to lowercase. We mapped words that occur once in the corpus to a token UNK. New words that are not found in our corpus are also mapped to UNK.

Embedding Feature

1. We used the embedding representation for each word of a pair of sentences.
2. The sentence composition is a summation of embeddings of component words.
3. We calculated the cosine similarity between the two embeddings which represent a pair of sentences. This value is used as feature to our regression system.

Feature	PT-BR		PT-EU		Both	
	ρ	**MSE**	ρ	**MSE**	ρ	**MSE**
Baseline	0.57	0.50	0.60	0.49	0.59	0.50
Embeddings	0.56	0.51	0.63	0.46	0.60	0.49

Table 4: Evaluation of the Embeddings model on ASSIN training data.

Evaluating on training dataset

Although we have not been sure if the Embeddings feature performed better than the Baseline, the first performs better when combined with TF-IDF than Baseline does.

Feature	PT-BR		PT-EU		Both	
	ρ	MSE	ρ	MSE	ρ	MSE
Baseline	0.57	0.50	0.60	0.49	0.59	0.50
Embeddings	0.56	0.51	0.63	0.46	0.60	0.49
TF-IDF	0.67	0.41	0.70	0.39	0.68	0.41
Baseline + TF-IDF	0.68	0.41	0.71	0.38	**0.71**	**0.38**
Embedding + TF-IDF	**0.69**	**0.39**	**0.73**	**0.36**	**0.71**	**0.38**

Table 5: Evaluation of our systems on ASSIN training data comparing to our Baseline system.

Evaluating on ASSIN testing dataset

Evaluating of testing set (results submitted to ASSIN):

- Embeddings feature did not outperform the Baseline feature for PT-EU testing corpus.
- TF-IDF is the best standalone feature.
- Embeddings feature improved the model that uses TF-IDF.
- The best results were obtained using both proposed features.

Feature	PT-BR		PT-EU	
	ρ	**MSE**	ρ	**MSE**
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

Table 6: Evaluation of our features on ASSIN test dataset.

Conclusion

- We obtained the best results for PT-BR sentence similarity and second best for PT-EU.
- The state of art for this task in English is 0,82 ρ and 0,32 MSE (SICK dataset).
- As we have tried a simple approach to solve the sentence similarity task, we believe that more improvements can be made to achieve state of art.

Future work

We believe that the summation of embeddings is not the best way to model a sentence.

- A LSTM network keep the order of the words. It can generate a better representation of a sentence.

Our embeddings model was trained only using a Brazilian Portuguese corpora.

- The embeddings model not always has PT-EU words in its vocabulary. Also, PT-EU syntactic constructions sometimes are different than PT-BR and our embeddings model is not able to deal with this.
- It explains why we achieved a higher MSE in PT-EU dataset than in PT-BR.

We only expanded words based on TEP (BP thesaurus).

References I

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press Aarhus.

Bjerva, J., J. Bos, R. van der Goot, and M. Nissim (2014). The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *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, and S. Aluísio (2008). Anotação Lingüística em XML do Corpus PLN-BR. NILC–TR–09–08. Technical report, University of São Paulo, Brazil.

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

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American society for information science 41*(6), 391.

Kenter, T. and M. de Rijke (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1411–1420. ACM.

Landauer, T. K. and S. 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.

Lapata, M. and R. Barzilay (2005). Automatic evaluation of text coherence: Models and representations. In *IJCAI*, Volume 5, pp. 1085–1090.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pp. 24–26. ACM.

Li, Y., D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett (2006). Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on 18*(8), 1138–1150.

References II

Lin, C.-Y. and E. Hovy (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 71–78. Association for Computational Linguistics.

Liu, X., Y. Zhou, and R. Zheng (2007). Sentence similarity based on dynamic time warping. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pp. 250–256. IEEE.

Maziero, E. and T. Pardo (2008). Interface de Acesso ao TeP 2.0 - Thesaurus para o português do Brasil. Technical report, University of São Paulo, Brazil.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Papineni, K. (2001). Why inverse document frequency? In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pp. 1–8. Association for Computational Linguistics.

Ponte, J. M. and W. B. Croft (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275–281. ACM.

Rada, R., H. Mili, E. Bicknell, and M. Blettner (1989). Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on 19*(1), 17–30.

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

Rocchio, J. J. (1971). Relevance feedback in information retrieval.

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

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

References III

Salton, G. and C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management 24*(5), 513–523.

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics 24*(1), 97–123.

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

Obrigado
Thank you

Perguntas
Questions