SEF@UHH at SemEval-2017 Task 1: Unsupervised Knowledge-Free Semantic Textual Similarity via Paragraph Vector

Mirela-Stefania Duma and Wolfgang Menzel
University of Hamburg
Natural Language Systems Division
{mduma, menzel}@informatik.uni-hamburg.de

Abstract

This paper describes our unsupervised knowledge-free approach to the SemEval-2017 Task 1 Competition. The proposed method makes use of Paragraph Vector for assessing the semantic similarity between pairs of sentences. We experimented with various dimensions of the vector and three state-of-the-art similarity metrics. Given a cross-lingual task, we trained models corresponding to its two languages and combined the models by averaging the similarity scores. The results of our submitted runs are above the median scores for five out of seven test sets by means of Pearson Correlation. Moreover, one of our system runs performed best on the Spanish-English-WMT test set ranking first out of 53 runs submitted in total by all participants.

1 Introduction

Semantic Textual Similarity (STS) aims to assess the degree to which two snippets of text are related in meaning to each other. The SemEval annual competition offers a track on STS (Cer et al., 2017) where submitted STS systems are evaluated in terms of the Pearson correlation between machine assigned semantic similarity scores and human judgments.

We participated in both monolingual sub-tracks and cross-lingual sub-tracks. Given a sentence pair in the same language, the SemEval STS task is to assign a similarity score to it ranging from 0 to 5, with 0 implying that the semantics of the sentences are completely independent and 5 denoting semantic equivalence (Cer et al., 2017). The cross-lingual side of STS is similar to the initial task, but differs in the input sentences which come from two languages.

This year's shared task features six sub-tasks: Arabic-Arabic, Arabic-English, Spanish-Spanish, Spanish-English (two test sets), English-English and a surprise task (Turkish-English) for which no annotated data is offered.

For example, for the English monolingual STS track, the pair of sentences below had a score of 3 assigned by human annotators, meaning that the two sentences are roughly equivalent, but some essential information differs or is missing (Cer et al., 2017).

Bayes' theorem was named after Rev Thomas Bayes and is a method used in probability theory.

As an official theorem, Bayes' theorem is valid in all universal interpretations of probability.

We present an unsupervised, knowledge-free approach that utilizes Paragraph Vector (Le and Mikolov, 2014) to represent sentences by means of continuous distributed vectors. In addition to experimenting with feature spaces of different dimensionality, we also compare three state-of-the-art similarity metrics (Cosine, Bray-Curtis and Correlation) for calculating the STS scores. We do not make use of any lexical or semantic resources, nor hand-annotated labeled corpora in addition to the distributed representations trained on non-annotated text. The approach gives promising results on all sub-tasks, with our submitted systems ranking first out of 53 for one Spanish-English sub-track and above the median scores for five out of seven test sets.

We first shortly summarize related work in STS and describe Paragraph Vector in Section 2. Then we present our method in Section 3 along with the corpora we used in training the Paragraph Vector models. Section 4 contains an overview of the evaluation and the results.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 170–174,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

2 Related Work

2.1 Semantic Textual Similarity

We present in this subsection the state-of-the-art in STS-Task 1 using Paragraph Vector since it is the most relevant to our work. King et al. (2016), for instance make use of Paragraph Vectors as one approach in the English monolingual sub-task. Results are reported for a single vector size and the Cosine metric which is employed in obtaining the similarity score between sentences. Brychcín and Svoboda (2016) follow a similar approach but apply it also to the cross-lingual task.

We raise three research questions regarding the usage of Paragraph Vector in STS:

- To which degree does the vector size matter?
- What could be a better alternative to the traditional Cosine metric for measuring the similarity between two vectors (obtained with Doc2Vec[1])?
- Given a cross-lingual task, does averaging the similarity scores obtained using the Doc2Vec models trained on both language corpora result in an improvement over using only the scores from one model?

2.2 Paragraph Vector

In order to assess the semantic textual similarity of two sentences, methods of representing them are crucial. Le and Mikolov (2014) propose a continuous, distributed vector representation of phrases, sentences and documents, Paragraph Vectors. It is a continuation of the work in Mikolov et al. (2013a) where word vectors (embeddings) are introduced in order to semantically represent words.

The strength of capturing the semantics of words via word embeddings is visible not only when considering words with similar meaning like ”strong” and ”powerful” (Le and Mikolov, 2014), but also in learning relationships such as *male/female* where the vector representation for *King* - *Man* + *Woman* results in a vector very close to *Queen* (Mikolov et al., 2013b).

In the Paragraph Vector framework, the paragraph vectors are concatenated with the word vectors to form one vector. The paragraph vector acts as a memory of what is missing in the current context. The word vectors are shared across all paragraphs, while the paragraph vector is shared across all contexts generated from the same paragraph. The vectors are trained using stochastic gradient descent with backpropagation (Le and Mikolov, 2014).

Since the STS task requires assigning a similarity score between two sentences, we apply Paragraph Vector at the sentence level. The models are trained using the Gensim library (Řehůřek and Sojka, 2010).

3 Semantic Textual Similarity via Paragraph Vector

3.1 Corpora

For training the Doc2Vec models we used various corpora available for the different language pairs. Following the rationale from Lau and Baldwin (2016), we concatenated to the corpora the test set too as the Doc2Vec training is purely unsupervised. The corpora we used are made available by Opus (Tiedemann, 2012) (except Commoncrawl[2] and SNLI (Bowman et al., 2015)): Wikipedia (Wolk and Marasek, 2014), TED[3], MultiUN (Eisele and Chen, 2010), EUBookshop (Skadiņš et al., 2014), SETIMES[4], Tatoeba[5], WMT[6] and News Commentary[7]. The following table presents which corpora were used and how many sentences they consist of. The corpora marked with * were used only for the third run.

Track / Corpora	AR-AR	AR-EN	ES-ES	ES-EN	EN-EN	TR-EN
Commoncrawl	-	-	1.84M	-	2.39M	-
Wikipedia	151K	151K	-	1.81M	-	160K
TED	152K	152K	-	157K	-	137K
MultiUN	1M	1M	-	-	-	-
EUBookshop	-	-	-	-	-	23K
SETIMES	-	-	-	-	-	207K
Tatoeba	-	-	-	-	-	156K
SNLI*	-	150K	-	150K	150K	150K
WMT*	-	16K	-	16K	16K	16K
News Commentary*	-	238K	-	238K	238K	238K

Table 1: Corpora used in training Doc2Vec models

The SNLI, WMT and News Commentary corpora were used for run 3 in some sub-tasks where we aimed to assess whether using more data makes

[1] The terms Paragraph Vector and Doc2Vec are used interchangeably as follows.

[2] http://commoncrawl.org/

[3] http://www.casmacat.eu/corpus/ted2013.html

[4] http://nlp.ffzg.hr/resources/corpora/setimes/

[5] http://tatoeba.org/

[6] http://www.statmt.org/wmt14/

[7] http://www.casmacat.eu/corpus/news-commentary

a difference. For training the English models only the EN side of the ES-EN language pair was used.

3.2 Preprocessing

For the sub-tasks that included the Arabic language we utilized the Stanford Arabic Segmenter (Monroe et al., 2014) in order to reduce lexical sparsity. For all the other sub-tasks, we performed text normalization, tokenization and lowercasing using the scripts available in the Moses Machine Translation Toolkit (Koehn et al., 2007).

3.3 Methods

We assess the semantic similarity between two sentences based on their continuous vector representations obtained by means of various Paragraph Vector models. A similarity metric is applied afterwards in order to determine the proximity between the two vectors. This measure is directly used as the similarity score of the two sentences.

For all sub-tasks we experiment with the PV-DBOW training algorithm, various vector sizes (200, 300 and 400) and with various state-of-the-art similarity metrics (Cosine, Bray-Curtis, Correlation) defined as:

$$\text{Cosine: } 1 - \frac{u \cdot v}{||u||_2 ||v||_2}$$

$$\text{Bray-Curtis: } \frac{\sum |u_i - v_i|}{\sum |u_i + v_i|}$$

$$\text{Correlation: } 1 - \frac{(u-\bar{u}) \cdot (v-\bar{v})}{||(u-\bar{u})||_2 ||(v-\bar{v})||_2}$$

where u and v are the vector representations of the two sentences, $\bar{u}$ and $\bar{v}$ denote the mean value of the elements of u and and v, and $x \cdot y$ is the dot product of x and y.

The Cosine metric is directly available from the *Gensim* library, while the Bray-Curtis and Correlation metrics are part of the *spatial* library from *scipy*[8]. We need to invert the score produced by the *spatial* library as it provides dissimilarity scores instead of the required similarity measures.

Given a monolingual sub-task $L_1 - L_1$ and multiple bilingual corpora, the L_1 side of the corpora is used to train Doc2Vec models. For all cross-lingual sub-tasks $L_1 - L_2$ we used Google Translate to obtain the test set translation from L_1 to L_2 and vice versa. Then we trained the Doc2Vec models for the two languages separately and combined the similarity scores obtained by the two models by averaging. Since the scores are in the range $(0, 1]$ we multiply them by 5 in order to return a continuous valued similarity score on a scale from 0 to 5, as the competition requires.

We submitted three runs to the competition:

run1	Model(size=200), Cosine similarity EN-ES: Model_ES AR-EN: Model_AR TR-EN: Model_TR
run2	Model(size=400), Cosine similarity EN-ES: Model_ES AR-EN: Model_AR TR-EN: Model_TR
run3	Model(size=200), Bray-Curtis similarity, more training data EN-ES: Model_EN AR-EN: Model_EN TR-EN: Model_EN

Table 2: Submitted runs settings

4 Evaluation and Results

The similarity scores are evaluated by computing the Pearson Correlation between them and human judgments for the same sentence pairs. This section presents our results for all sub-tasks of the 2017 test sets and also for the STS Benchmark[9] (Cer et al., 2017).

4.1 STS 2017 Test Sets

When considering all 85 submitted runs (including the monolingual runs and the baseline), our best runs ranked 26 out of 49 for AR-AR, 21 out of 45 for AR-EN, 22 out of 48 for ES-ES, 28 out of 53 for ES-EN-a, 1 out of 53 for ES-EN-b, 35 out of 77 for EN-EN and 16 out of 48 for TR-EN (Cer et al., 2017).

Several experiments were conducted with size 200, 300 and 400 for the Doc2Vec vectors, training on both sides of the corpora for the cross-lingual tasks and applying Cosine, Bray-Curtis and Correlation similarity metrics. We detail in Table 3 the Pearson Correlation scores obtained.

The results indicate that the Bray-Curtis metric performs better than the other two in five out of seven test sets, with a tie on the EN-EN test set. Regarding the dimension of the Doc2Vec vectors, a conclusion cannot be simply drawn from these results, since size 200 leads to best results for ES-ES, ES-EN-a and EN-EN, size 300 gives best results for AR-AR, size 400 for AR-EN and ES-EN-b and a tie for TR-EN when using sizes 300 and 400. It is also important to note that the

[8] https://docs.scipy.org/doc/scipy-0.18.1/reference/spatial.html

[9] http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

Task	Cosine			Bray-Curtis			Correlation		
AR-AR									
200		0.5587			0.5790			0.5579	
300		0.5825			**0.5984**			0.58	
400		0.5773			0.5943			0.5767	
AR-EN	AR	EN	Mean	AR	EN	Mean	AR	EN	Mean
200	0.4789	0.4971	0.5221	0.755	0.503	0.5268	0.4779	0.4997	0.5227
300	0.4963	0.5141	0.5429	0.502	0.5085	0.5432	0.4963	0.5154	0.5437
400	0.4813	0.5266	0.5381	0.4949	0.5288	**0.5469**	0.4796	0.5275	0.5372
ES-ES									
200		**0.7455**			0.7423			0.7434	
300		0.7002			0.7054			0.6991	
400		0.6979			0.7072			0.6982	
ES-EN-a	ES	EN	Mean	ES	EN	Mean	ES	EN	Mean
200	0.5738	0.6021	0.6212	0.5852	0.6208	**0.6353**	0.5748	0.6041	0.6227
300	0.5676	0.6162	0.6219	0.5793	0.6253	0.6299	0.566	0.6171	0.6213
400	0.566	0.6092	0.6187	0.5767	0.6162	0.6253	0.5643	0.606	0.6163
ES-EN-b	ES	EN	Mean	ES	EN	Mean	ES	EN	Mean
200	0.3069	0.1933	0.3111	0.306	0.1686	0.2953	0.307	0.1919	0.31
300	0.3234	0.1784	0.3193	0.3187	0.1685	0.3099	0.323	0.1826	0.3222
400	0.3407	0.1873	0.3303	**0.3436**	0.1575	0.3113	0.342	0.1854	0.3284
EN-EN									
200		**0.7880**			**0.7880**			0.7871	
300		0.7237			0.7396			0.7249	
400		0.7185			0.7264			0.7178	
TR-EN	TR	EN	Mean	TR	EN	Mean	TR	EN	Mean
200	0.4990	0.5554	0.5804	0.5080	0.5577	0.5846	0.5052	0.5540	0.5837
300	0.4919	0.5718	0.5792	0.4869	**0.6001**	0.5879	0.4909	0.5705	0.5770
400	0.4878	0.5832	0.5775	0.5024	**0.6000**	0.5930	0.4857	0.5836	0.5772

Table 3: Pearson Correlation results for various parameters

Pearson correlation scores range from 0.1575 to 0.3436 for the ES-EN-b test set and from 0.7178 to 0.788 for the EN-EN test set which suggests that experimenting with various sizes of Doc2Vec vectors is worth investigating, contrary to the common practice of experimenting with just a single vector size.

Averaging the similarity scores for the source and the target language also seems to be a promising approach. This combination led to best Pearson correlation scores for two of the four cross-lingual test sets (AR-EN and ES-EN-a).

We report in Table 4 the Pearson correlation results of the runs we submitted to the competition. For the first two runs we used Cosine for computing the similarity between the sentence pairs and for the third run we used Bray-Curtis.

	average	AR-AR	AR-EN	ES-ES	ES-EN-a	ES-EN-b	EN-EN	TR-EN
run 1	0.5644	0.5588	0.4789	0.7456	0.5739	0.3069	0.7880	0.4990
run 2	0.5528	0.5774	0.4813	0.6979	0.5660	0.3407	0.7186	0.4878
run 3	**0.5676**	0.5790	0.5384	0.7423	0.5866	0.1802	0.7256	0.6211

Table 4: Results for the submitted runs

The non-English language side of the corpora was used for training the Doc2Vec models for the cross-lingual tasks in the first two runs, while for the third run we trained the Doc2Vec models on the English side of the corpora. In the third run we also included additional data (except for AR-AR and ES-ES) in order to assess how the size of the training corpus for the Doc2Vec models influences the results. For the AR-EN, ES-EN-b and TR-EN sub-tasks the scores improved when using more training data, but the differences were small.

4.2 STS Benchmark

The Semeval STS organizers made available the STS Benchmark for the EN-EN task with the purpose of creating state-of-the-art approaches and collecting their results on standard data sets. The benchmark data consist of a selection of previous data sets used in the competition between 2012 and 2017.

Since the methods we presented are unsupervised and knowledge-free, we did not make use of the annotated training data when computing the similarity scores for the development and test sets. We tested two approaches for obtaining similarity scores on the EN-EN sub-task: the first infers the vectors for the development and test set sentences from the already trained Doc2Vec models (**Post-training inference**) and the other one retrains from scratch new models by adding the development and test sets to the initial Doc2Vec training data (**New-Model**).

As it can be noted in Table 4, the best Pearson correlation result for EN-EN was obtained using the settings from our submitted run 1. These settings also gave the best results for the STS Benchmark test data (Table 5).

Approach	Development set	Test set
Post-training inference	0.6670	0.5915
New-Model	0.6158	0.5922

Table 5: Results for the STS Benchmark

5 Conclusions

We presented in this paper our unsupervised knowledge-free approach to the STS task. A wide range of experiments were carried out in order to assess the impact of the similarity metric if Paragraph Vector is used to represent sentences. Our results indicate that Bray-Curtis might be a good choice, because it outperformed the commonly used Cosine metric on five out of seven test sets. Moreover, training the Doc2Vec models on both sides of the language corpora and averaging their similarity scores seems to be a promising approach for the cross-lingual STS task.

The proposed method achieved encouraging results as we ranked first on the EN-ES-b sub-task and obtained Pearson correlation scores above the median score for five out of seven test sets.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tomáš Brychcín and Lukáš Svoboda. 2016. Uwb at semeval-2016 task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, pages 588–594.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 1–14. http://www.aclweb.org/anthology/S17-2001.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*.

Milton King, Waseem Gharbieh, SoHyun Park, and Paul Cook. 2016. Unbnlp at semeval-2016 task 1: Semantic textual similarity: A unified framework for semantic processing and evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. ACL, San Diego, California, pages 732–735.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL, Stroudsburg, PA, USA.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. pages 1188–1196.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*. pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*. pages 746–751.

Will Monroe, Spence Green, and Christopher D. Manning. 2014. Word segmentation of informal arabic with domain adaptation. In *Association for Computational Linguistics (ACL)*.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, pages 45–50.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Krzysztof Wolk and Krzysztof Marasek. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. In *Procedia Technology, 18*. Elsevier, pages 126 – 132.