ResSim at SemEval-2017 Task 1: Multilingual Word Representations for Semantic Textual Similarity

Johannes Bjerva
Center for Language and Cognition Groningen
University of Groningen
The Netherlands
j.bjerva@rug.nl

Robert Östling
Department of Linguistics
Stockholm University
Sweden
robert@ling.su.se

Abstract

Shared Task 1 at SemEval-2017 deals with assessing the semantic similarity between sentences, either in the same or in different languages. In our system submission, we employ multilingual word representations, in which similar words in different languages are close to one another. Using such representations is advantageous, since the increasing amount of available parallel data allows for the application of such methods to many of the languages in the world. Hence, semantic similarity can be inferred even for languages for which no annotated data exists. Our system is trained and evaluated on all language pairs included in the shared task (English, Spanish, Arabic, and Turkish). Although development results are promising, our system does not yield high performance on the shared task test sets.

1 Introduction

Semantic Textual Similarity (STS) is the task of assessing the degree to which two sentences are semantically similar. Within the SemEval STS shared tasks, this is measured on a scale ranging from 0 (no semantic similarity) to 5 (complete semantic similarity) (Cer et al., 2017). Monolingual STS is an important task, for instance for evaluation of machine translation (MT) systems, where estimating the semantic similarity between a system's translation and the gold translation can aid both system evaluation and development. The task is already a challenging one in a monolingual setting, e.g., when estimating the similarity between two English sentences. In this paper, we tackle the more difficult case of cross-lingual STS, e.g., estimating the similarity between an English and an Arabic sentence.

Previous approaches to this problem have focussed on two main approaches. On the one hand, MT approaches have been attempted (e.g. Lo et al. (2016)), which allow for monolingual similarity assessment, but suffer from the fact that involving a fully-fledged MT system severely increases system complexity. Applying bilingual word representations, on the other hand, bypasses this issue without inducing such complexity (e.g. Aldarmaki and Diab (2016)). However, bilingual approaches do not allow for taking advantage of the increasing amount of STS data available for more than one language pair.

Currently, there are several methods available for obtaining high quality multilingual word representations. It is therefore interesting to investigate whether language can be ignored entirely in an STS system after mapping words to their respective representations. We investigate the utility of multilingual word representations in a cross-lingual STS setting. We approach this by combining multilingual word representations with a deep neural network, in which all parameters are shared, regardless of language combinations.

The contributions of this paper can be summed as follows: i) we show that multilingual input representations in some cases can be used to train an STS system without access to training data for a given language; ii) we show that access to data from other languages in some cases improves system performance for a given language.

2 Multilingual Word Representations

2.1 Multilingual Skip-gram

The skip-gram model has become one of the most popular manners of learning word representations in NLP (Mikolov et al., 2013). This is in part owed to its speed and simplicity, as well as the per-

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 154–158,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

formance gains observed when incorporating the resulting word embeddings into almost any NLP system. The model takes a word w as its input, and predicts the surrounding context c. Formally, the probability distribution of c given w is defined as

$$p(c|w;\theta) = \frac{\exp(\vec{c}^T\vec{w})}{\Sigma_{c\in V}\exp(\vec{c}^T\vec{w})}, \quad (1)$$

where V is the vocabulary, and θ the parameters of word emeddings ($\vec{w}$) and context embeddings ($\vec{c}$). The parameters of this model can then be learned by maximising the log-likelihood over (w,c) pairs in the dataset D,

$$J(\theta) = \sum_{(w,c)\in D} \log p(c|w;\theta). \quad (2)$$

Guo et al. (2016) provide a multilingual extension for the skip-gram model, by requiring the model to not only learn to predict English contexts, but also multilingual ones. This can be seen as a simple adaptation of Firth (1957, p.11), i.e., you shall judge a word by the *multilingual* company it keeps. Hence, the vectors for, e.g., *dog* and *perro* ought to be close to each other in such a model. This assumes access to multilingual parallel data, as word alignments are used in order to determine which words comprise the multilingual context of a word. Whereas Guo et al. (2016) only evaluate their approach on the relatively similar languages English, French and Spanish, we explore a more typological diverse case, as we apply this method to English, Spanish and Arabic. We use the same parameter settings as Guo et al. (2016).

2.2 Learning embeddings

We train 100-dimensional multilingual embeddings on the Europarl (Koehn, 2005) and UN corpora (Ziemski et al., 2016). Word alignment, which is required for the training of multilingual embeddings, is performed using the Efmaral word-alignment tool (Östling and Tiedemann, 2016). This allows us to extract a large amount of multilingual (w,c) pairs. We then use these pairs in order to learn multilingual embeddings, by applying the *word2vecf* tool (Levy and Goldberg, 2014).

3 Method

3.1 System architecture

We use a relatively simple neural network architecture, consisting of an input layer with pre-trained word embeddings and a network of fully connected layers. Given word representations for each word in our sentence, we take the simplistic approach of averaging the vectors across each sentence. The resulting sentence-level representations are then concatenated and passed through a single fully connected layer, prior to the output layer. In order to prevent any shift from occurring in the embeddings, we do not update these during training. The intuition here, is that we do not want the representation for, e.g., *dog* to be updated, which might push it further away from that of *perro*. We expect this to be especially important in cases where we train on a single language, and evaluate on another.

We apply dropout ($p = 0.5$) between each layer (Srivastava et al., 2014). All weights are initialised using the approach in Glorot and Bengio (2010). We use the Adam optimisation algorithm (Kingma and Ba, 2014), monitoring the categorical cross entropy of a one-hot representation of the (rounded) sentence similarity score, while sanity-checking against the scores obtained as measured with Pearson correlation. All systems are trained using a batch size of 40 sentence pairs, over a maximum of 50 epochs, using early stopping. Hyperparameters are kept constant in all conditions.

3.2 Data

We use all available data from previous editions of the SemEval shared tasks on (cross-lingual) STS. An overview of the available data is shown in Table 1.

Language pair	N sentences
English / English	3750
English / Spanish	1000
English / Arabic	2162
Spanish / Spanish	1620
Arabic / Arabic	1081

Table 1: Available data for (cross-lingual) STS from the SemEval shared task series.

4 Experiments and Results

We aim to investigate whether using a multilingual input representation and shared weights allow us to ignore languages in STS. We first train and evaluate single-source trained systems (i.e. on a single language pair), and evaluate this both using the same language pair as target, and on all other target language pairs.[1] Secondly, we investigate the effect of bundling training data together, investigating which language pairings are helpful for each other. We measure performance between gold similarities and system output using the Pearson correlation measure, as this is standard in the SemEval STS shared tasks. We first present results on the development sets, and finally the official shared task evaluation results.

4.1 Single-source training

Results when training on a single source corpus are shown in Table 2. Training on the target language pair generally yields the highest results, except for one case. When evaluating on Arabic/Arabic sentence pairs, training on English/Arabic texts yields comparable, or slightly better, performance than when training on Arabic/Arabic.

Train Test	en/en	en/es	en/ar	es/es	ar/ar
en/en	**0.69**	0.07	-0.04	0.64	0.54
en/es	0.19	**0.27**	0.00	0.18	-0.04
en/ar	-0.44	0.37	**0.73**	-0.10	0.62
es/es	0.61	0.07	0.12	**0.65**	0.50
ar/ar	0.59	0.52	**0.73**	0.59	0.71

Table 2: Single-source training results (Pearson correlations). Columns indicate training language pairs, and rows indicate testing language pairs. Bold numbers indicate best results per row.

4.2 Multi-source training

We combine training corpora in order to investigate how this affects evaluation performance on the language pairs in question. In the first condition, we copy the single-source setup, except for that we also add in the data belonging to the source-pair at hand, e.g., training on both English/Arabic and Arabic/Arabic when evaluating on Arabic/Arabic (see Table 3).

Train Test	en/en	en/es	en/ar	es/es	ar/ar
en/en	0.69	0.68	0.67	0.69	**0.71**
en/es	0.22	0.27	**0.30**	0.22	0.24
en/ar	0.72	0.72	**0.73**	0.71	0.72
es/es	0.63	0.60	0.63	0.65	**0.66**
ar/ar	0.71	0.72	**0.75**	0.70	0.71

Table 3: Training results with one source in addition to in-language data (Pearson correlations). Columns indicate added training language pairs, and rows indicate testing language pairs. Bold numbers indicate best results per row.

We observe that the monolingual language pairings (en/en, es/es, ar/ar) appear to be beneficial for one another. We therefore run an ablation experiment, in which we train on two out of three of these language pairs, and evaluate on all three. Not including any Spanish training data yields comparable performance to including it (Table 4).

Ablated Test	en/en	es/es	ar/ar	none
en/en	**0.60**	0.69	0.69	0.65
es/es	0.64	**0.64**	0.67	0.60
ar/ar	0.68	0.66	**0.58**	0.72

Table 4: Ablation results (Pearson correlations). Columns indicate ablated language pairs, and rows indicate testing language pairs. The *none* column indicates no ablation, i.e., training on all three monolingual pairs. Bold indicates results when not training on the language pair evaluated on.

4.3 Shared Task Test Results

The results from the official SemEval-2017 evaluation are shown in Table 5. Although our results for Spanish/Spanish and English/English are in line with our development results, the results for all other language pairs are far lower than expected. This might be explained by overfitting to the training/dev sets we use. After the official evaluation period ended, we also attempted to perform a sanity check. We allowed our model to tune on the gold data, which surprisingly did not increase performance particularly much. We therefore suspect that the poor system performance we observe, may be partially owed to two factors: i) overfitting on the tracks involving Arabic, as we did not apply any type of pre-processing, and our vector set was tuned on relatively little Arabic data; ii) discrepancies between the mix of training-data (and possibly annotators) from previous editions of the

[1] This setting can be seen as a sort of model transfer.

	Primary	ar/ar	ar/en	es/es	es/en	es/en (wmt)	en/en	en/tr
Single-source	0.3148	0.2892	0.1045	0.6613	0.2389	0.0305	0.6906	0.1884
Multi-source	0.2938	0.3120	0.1288	0.6920	0.1002	0.0162	0.6877	0.1195
Ablation	0.2145	0.0033	0.1098	0.5465	0.2262	0.0199	0.5057	0.0902

Table 5: Results on SemEval-2017 Shared Task Test sets.

shared task, and test data in this year's edition.

An interesting option to attempt to solve part of this problem, would be to frame this as a multi-task learning problem. This could be done by assigning each year's data set a separate output layer. Should annotator conventions differ, e.g., if a score of 2.5 in 2015 is equivalent to a score of 3.5 in 2016, the network should be able to learn this and compensate for such effects.

5 Discussion

In all cases, training on the target language pair is beneficial. We also observe that using multilingual embeddings is crucial for multilingual approaches, as monolingual embeddings naturally only yield on-par results in monolingual settings. This is due to the fact that using the shared language-agnostic input representation allows us to take advantage of linguistic regularities across languages, which we obtain solely from observing distributions between languages in parallel text. Using monolingual word representations, however, there is no similarity between, e.g., *dog* and *perro* to rely on to guide learning.

For the single-source training, we in one case observe somewhat better performance using other training sets than the in-language one: training on English/Arabic outperforms training on Arabic/Arabic, when evaluating on Arabic/Arabic. We expected this to be due to differing data set sizes (English/Arabic is about twice as big). Controlling for this does, indeed, bring the performance of training on English/Arabic close to training on Arabic/Arabic. However, combining these datasets increases performance further (Table 3).

In single-source training, we also observe that certain source languages do not offer any generalisation over certain target languages. Interestingly, certain combinations of training/testing language pairs yield very poor results. For instance, training on English/English yields very poor results when evaluating on English/Arabic, and vice versa. The same is observed for the combination Spanish/Spanish and English/Arabic. This may be explained by domain differences in training and evaluation data. A general trend appears to be that either monolingual training pairs and evaluation pairs, or cross-lingual pairs with some overlap (e.g. English/Arabic, Arabic/Arabic) is beneficial.

The positive results on pairings without any language overlap are particularly promising. Training on English/English yields results not too far from training on the source language pairs, for Spanish/Spanish and Arabic/Arabic. Similar results are observed when training on Spanish/Spanish and evaluating on English/English and Arabic/Arabic, as well as when training on Arabic/Arabic and evaluating on English/English and Spanish/Spanish. This indicates that we can estimate STS relatively reliably, even without assuming any existing STS data for a given language.

6 Conclusions and Future Work

Multilingual word representations allow us to leverage more available data for multilingual learning of semantic textual similarity. We have shown that relatively high STS performance can be achieved for languages without assuming existing STS annotation, and relying solely on parallel texts. An interesting direction for future work is to investigate how multilingual character-level representations can be included, perhaps learning morpheme-level representations and mappings between these across languages. Leveraging approaches to learning multilingual word representations from smaller data sets would also be interesting. For instance, learning such representations from only the New Testament, would allow for STS estimation for more than 1,000 languages.

Acknowledgments

We would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster.

References

Hanan Aldarmaki and Mona Diab. 2016. GWU NLP at SemEval-2016 Shared Task 1: Matrix factorization for crosslingual STS. In *Proceedings of SemEval 2016*. pages 663–667.

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 1–5. http://www.aclweb.org/anthology/S17-2001.

John R Firth. 1957. *A synopsis of linguistic theory, 1930-1955*. Blackwell.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*. volume 9, pages 249–256.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Proc. of AAAI*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit.*. Phuket, Thailand.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL*. pages 302–308.

Chi-kiu Lo, Cyril Goutte, and Michel Simard. 2016. Cnrc at semeval-2016 task 1: Experiments in crosslingual semantic textual similarity. *Proceedings of SemEval 2016* pages 668–673.

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

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics* 106(1):125–146.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Micha Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.