FBK HLT-MT at SemEval-2016 Task 1: Cross-lingual Semantic Similarity Measurement Using Quality Estimation Features and Compositional Bilingual Word Embeddings

Duygu Ataman[1,2], José G. C. de Souza[1,2], Marco Turchi[1], Matteo Negri[1]
[1] HLT - MT
Fondazione Bruno Kessler, Trento, Italy
[2]Department of Information Engineering and Computer Science
University of Trento, Italy
{ataman,desouza,turchi,negri}@fbk.eu

Abstract

This paper describes the system by FBK HLT-MT for cross-lingual semantic textual similarity measurement. Our approach is based on supervised regression with an ensemble decision tree. In order to assign a semantic similarity score to an input sentence pair, the model combines features collected by state-of-the-art methods in machine translation quality estimation and distance metrics between cross-lingual embeddings of the two sentences. In our analysis, we compare different techniques for composing sentence vectors, several distance features and ways to produce training data. The proposed system achieves a mean Pearson's correlation of 0.39533, ranking 7^{th} among all participants in the cross-lingual STS task organized within the SemEval 2016 evaluation campaign.

1 Introduction

Semantic textual similarity (STS) measures the degree of equivalence between the meanings of two text sequences (Agirre et al., 2015). The similarity of the text pair can be represented as a continuous or discrete-time value ranging from irrelevance to exact semantic equivalence (Agirre et al., 2015).

STS has been one of the official shared tasks in SemEval since 2013 and has attracted the participation of many researchers from the scientific community; enabling the evaluation of several different approaches in natural language processing with a common benchmark and the production of novel annotated data sets that can be used in future research. State-of-the-art monolingual STS methods make use of several approaches including word alignments and distributional semantics, which are typically employed in a machine learning scenario (Sultan et al., 2015; Hänig et al., 2015).

This is the first year in which SemEval has organized a cross-lingual STS (CL-STS) sub-task, for which a baseline system applicable to the problem has not been defined yet. Similar to the monolingual STS task, the cross-lingual task requires the interpretation of the semantic similarity of two cross-lingual sentences, one in English and another one in Spanish, with a score ranging from 0 to 5. CL-STS measurement could be extremely useful for achieving textual entailment, paraphrase identification, word-sense disambiguation or sentiment analysis at the cross-lingual level as well as providing new means for an adequacy-oriented evaluation of machine translation outputs.

A related task in natural language processing is quality estimation. Quality estimation (QE) is used for automatically predicting the quality of machine translation outputs with respect to the source sentences in the original language (Mehdad et al., 2012; Turchi et al., 2014; C. de Souza et al., 2014a; C. de Souza et al., 2014b; C. de Souza et al., 2015). One shortcoming of QE approaches is that the QE system may not capture all aspects of the semantic representations of sentences. For instance, from a QE perspective, under which the number of edit operations required to fix a translation is used as a proxy of quality, a fluent translation containing an unnecessary negation would likely be labelled as a "good" translation. Therefore, a better solution would be

Proceedings of SemEval-2016, pages 570–576,
San Diego, California, June 16-17, 2016.

geared to also capture the adequacy aspects of cross-lingual comparison of the sentences. In order to improve the quality of the comparison, the features used in a QE system can be improved using distributional semantics. Neural language models, such as CBOW or Skipgram (Mikolov et al., 2013a) have proved to be useful in the monolingual STS task before (Agirre et al., 2015). Recent studies have extended these models to create bilingual word embeddings such that the embeddings are mapped to a common cross-lingual vector space by using a parallel training corpus or a dictionary (Klementiev et al., 2012; Mikolov et al., 2013b; Luong et al., 2015).

In light of these considerations, our submission to the first SemEval CL-STS task combines features derived from QE with distance features obtained by applying cross-lingual word embeddings. These features are used to feed an Extremely Randomized Trees (ET) regressor (Geurts et al., 2006) trained to predict the similarity score of the two sentences.

The rest of this paper is organized as follows. Section 2 describes the components of our CL-STS system. The details of the experimental analysis carried out on different composition approaches, the characteristics of the system under the influence of each vector space feature and varying data distributions are presented in Section 3. The final ranking of our system can be found in Section 4 along with the conclusions of our study in Section 5.

2 System Description

This section describes the proposed CL-STS system. The data to be semantically compared is first pre-processed as described in Section 2.1. The embedding corresponding to each word is retrieved and composed to form a sentence embedding using one of the methods described in Section 2.2. The features to be used in the regression are extracted from the sentence embeddings using 8 different distance measures listed in Section 2.2. These features are combined with 79 more features obtained by QE (Section 2.3) to produce a final set of 87 features, which is used to predict the similarity score of the two sentences. The overall system is illustrated in Figure 1.

2.1 Pre-processing

The data used in training and testing the system are processed before feature extraction in the following way. First of all, a language identification package developed by Lui et al. (2012) is used to detect the order of English and Spanish sentences in each line of the data set. This step is meant as a first sanity check since some of the next processing steps are language-dependent and hence sensitive to the order in which the two sentences are presented. The data is then tokenized and lowercased before the extraction of features using the text processing system of Moses (Koehn et al., 2007).

2.2 Bilingual Embedding Features

To obtain word embeddings, we use the bilingual Skipgram model by Luong et al. (2015). The embeddings are trained using the default parameters described by the authors and with a dimension of 200. We constructed an English-Spanish parallel corpus from Europarl (Koehn, 2005), UN (Rafalovitch et al., 2009), data sets of the quality estimation shared task in WMT 2012 (Callison-Burch et al., 2012), as well as the training data of the monolingual STS task from previous years (See Subsection 3.1) and used this data to train our bilingual embeddings.

Sentence embeddings are then generated by averaging the word embeddings in each sentence. Averaging is a simple and powerful composition method for monolingual word embeddings which has not been outperformed yet by much more sophisticated schemes, such as the recurrent neural networks and long short-term memories (Blacoe and Lapata, 2012; Wieting et al., 2015). Moreover, the latter often requires language-specific syntactic parsers which are not available in all languages, thus are not generally suitable for cross-lingual applications.

In our system, we implement three different averaging strategies to see the influence of stop words or term frequencies in the final sentence embedding. Our approaches consist of:

1. Averaging with all the tokens in the sentence including punctuation;
2. Averaging after removing stop words and punctuation in the sentence;

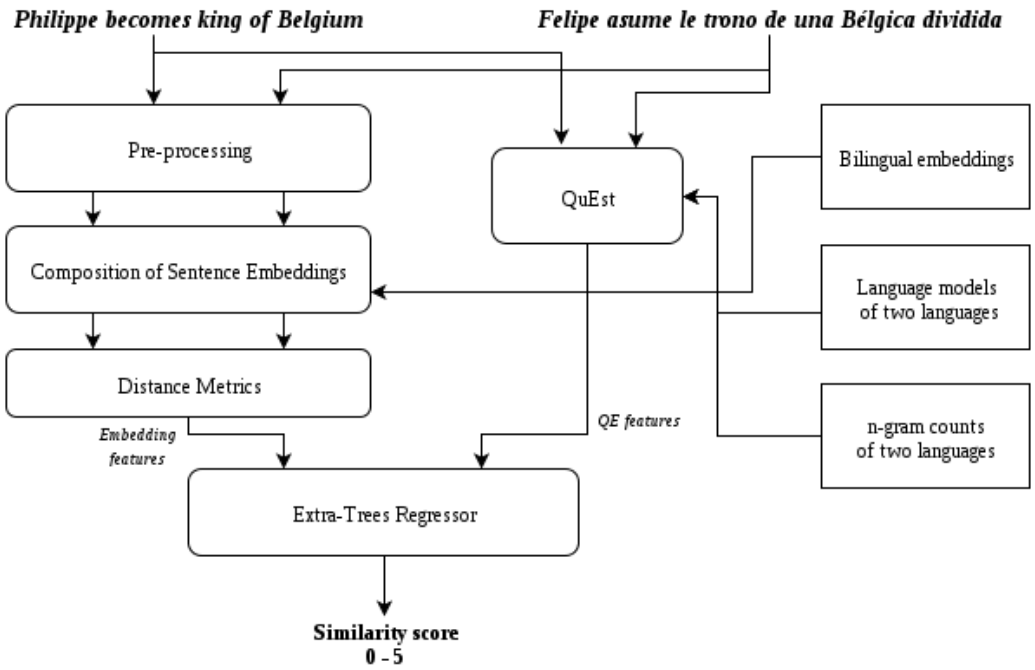

Figure 1: The schematic of the overall system for CL-STS

3. Averaging by weighting each word in the sentence by their inverse term frequencies.

During processing of the two sentences in English and Spanish, the words and punctuation in each corresponding sentence are averaged according to one of the methods above to generate one final embedding representing each sentence in the two languages.

In order to apply semantic comparison between the sentence embeddings, we select 8 distance measures that are defined in vector space. Given the two sentence vectors, the selected distance metrics include:

1. Cosine distance
2. Euclidean distance
3. Manhattan distance
4. Chebyshev distance
5. Canberra distance
6. Pearson's correlation
7. Ratio of number of words
8. Ratio of means

We further analyze the usefulness of these features through a set of experiments before using in the CL-STS task (Section 3).

2.3 Quality Estimation Features

The QE features used in our system are obtained by QuEst tool (Specia et al., 2013). The QuEst tool uses language models, POS-taggers or word aligners to extract many features that can represent complexity (*e.g.* language model probabilities or n-gram counts in the source segment) fluency (*e.g.* language model probabilities or number of tokens in the target segment), adequacy (*e.g.* word alignment features such as ratio of nouns/verbs/etc. in each sentence) or confidence (*e.g.* global scores or n-best lists) of a translation pair. For a more detailed description of each feature, we refer the reader to Specia et al. (2013). The language models used to extract the features are trained with the NY Times portion of English Gigaword (v.5) (Parker et al., 2011) and Spanish Gigaword (v.2) (Mendonca et al., 2009) and the same parallel data as described in the training of bilingual word embeddings (See Section 2.2)

Features	# Features	Pearson's correlation		
QE	79	0.5899		
Embeddings	8	0.4690		
QE + sentence embedding features				
		Composition method		
		Average	**Average w/o stop words**	**ITF-weighted Average**
+ Cosine	80	0.6106	0.6399	0.6017
+ Manhattan	81	0.6127	0.6429	0.5897
+ Euclidean	82	0.6152	**0.6445**^{run1}	0.5906
+ Canberra	83	0.6149	0.6435	0.5962
+ Pearson's correlation	84	0.6145	0.6361	0.6039
+ Chebyshev	85	0.6136	0.6364	0.5907
+ Ratio of NumWords	86	0.6138	0.6366	0.5915
+ Ratio of Means	87	0.6154	**0.6375**^{run2}	0.5996

Table 1: Pearson's correlation of system predictions using cross-validation. The first and second entries in the table indicate performance of the system using the two approaches separately. The second part indicates the system performance when the two approaches are combined, revealing the effect of each distance-based feature on the performance. Each column represents a different sentence composition method; including averaging, averaging after removing stop words and weighted averaging. Numbers in bold are the two runs, **[run1]** wth 82 and **[run2]** wth 87 features respectively, submitted to the SemEval 2016 - CL-STS shared task.

2.4 Ensemble Regression

An ET regressor is used as the learning method in the system. The ET regressor applies bagging to generate a number of random subsets of the training data and fits individual decision trees using different subsets of features and hyper parameters. The final prediction is produced by an ensemble average over all of the decision trees.

3 Experiments

3.1 Corpus

For this first round of the CL-STS task, training data was not released by the organizers. Therefore, the data for training the regressor was generated using those from the monolingual STS tasks organized in 2012, 2013, 2014 and 2015 (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015). The data was translated to Spanish using the MateCat tool (Federico et al., 2014) to create a cross-lingual data set where one of the sentences is in English and the other is in Spanish. In order to compensate for the different characteristics of data which occur after translation, we generate three different sets as follows:

1. The first sentence in the data set is selected as the English sentence, the second sentence is translated to Spanish, denoted as: s_1*(en)* - $\overline{s_2}$*(es)*.

2. The second sentence in the data set is selected as the English sentence, the first sentence is translated to Spanish, denoted as: s_2*(en)* - $\overline{s_1}$*(es)*.

3. The first two data sets are concatenated, denoted as the *merged* set.

The size of each set is given in Table 2. The three data sets are evaluated during our experiments in terms of the capability to represent the true data distribution and used in the test phase with the selected settings (Section 3.3).

Data set	# sent	# src	# tgt
s1(en) - $\overline{\text{s2}}$(es)	18,105	233,141	250,359
s2(en) - $\overline{\text{s1}}$(es)	18,091	230,300	253,115
merged	36196	463441	503474

Table 2: Sizes of the training sets: number of sentence pairs and number of words in source and target languages

3.2 Cross-validation on the merged set

The performance of our ET regressor is evaluated using 10-fold cross-validation on the merged set. The results are presented in Table 1.

We run a detailed analysis to evaluate each individual distance metric according to their effect on the performance. For evaluating the performance of the system we use Pearson's correlation. The experiments show that QE features and embedding features perform poorly when used as two different groups, but the performance can be improved when they are integrated together. The best correlation is achieved when cosine, manhattan and euclidean distances are combined with the QE features. All distance metrics show a consistent improvement over the system trained only on QE features, except for the Pearson's correlation (Table 1).

The experiments also reveal that averaging after removing the stop words and punctuation is the best sentence composition method among the three averaging strategies. Compared to this method, inverse term-frequency weighted averaging aims to decrease the effect of stop words in a less greedy way. However, we see that this approach results in even poorer performance than straightforward averaging; suggesting that we may lose important information from word semantics necessary in composing the sentence meaning when we weight each word with their inverse term frequencies. Thus, we select averaging without stop words as the composition method to use in our system to participate in the SemEval 2016 CL-STS shared task.

3.3 Performance of different training data sets on the evaluation data

In the second phase, we compare the performance of our ET regressor in terms of three different training sets, as explained in Section 3.1. 30% of the training set is used for validation and the remaining 70% is used for training purposes. The regressor then performs predictions for the same validation set using the different sets. Finally, we implement an ensemble model which is an average of the predictions of the three subsystems. The results of these experiments can be seen in Table 3.

In light of these experiments, we choose the feature sets composing of 82 and 87 features and the merged set for training purposes (See Table 1). These two systems are submitted as our **[run1]** and **[run2]** to the CL-STS shared task. We contribute with the ensemble average system as shown in Table 3 for **[run3]**.

Data set	Pearson's correlation
s1(en) - $\overline{\text{s2}}$(es)	0.5499
s2(en) - $\overline{\text{s1}}$(es)	0.5815
merged	0.6227
Average	$\mathbf{0.6131}^{run3}$

Table 3: Pearson's correlation of system (run2) predictions using 30-70 split in the (merged) data set. Performance is shown according to three different training data distributions. Average indicates the performance of ensemble of the three approaches, and is the system chosen to be submitted as **[run3]** to the SemEval 2016 - CL-STS shared task.

System	News	Multi-source	Mean
run3	0.25507	0.53892	0.39533
run1	0.24318	0.53465	0.3872
run2	0.24372	0.5142	0.37737

Table 4: Official results of SemEval 2016 - CL-STS shared task

4 Results

The performance of our system in the SemEval 2016 - CL-STS task is given in Table 4. The test set of the CL-STS task contains two parts with different characteristics. The first set contains 301 sentence pairs in the news domain and the second set consists of 2973 sentence pairs drawn from different domains. By using these data sets to test our system, we achieve a Pearson's correlation of 0.539 on the multi-source test set and 0.255 on the news set between our predictions and the true labels (Table 4).

We observe that using **[run3]**, which combines the three systems trained on three data sets with different distributions, is the most successful approach. The decreased values of Pearson's correlation are consistent with the fact that test data of the task and our training set are sampled from different distributions. Moreover, our system performs better in the general domain and worse in the specific domain of news. However, this performance can be improved by extending the training corpus to a similar content with the test corpus. Other aspects to improve are the quality of bilingual word embeddings, which should also be trained with more data and parameters that are more suitable for the specific task, and addition of feature selection quality to the system.

The results also show that performance of the system using only QE features and the three distance metrics consisting of cosine, manhattan and euclidean distances provide better results than the system using all features. Therefore, one can see that these three features can provide significant information when comparing two sentence embeddings and could be reliably used in future applications.

5 Conclusion

We have presented the CL-STS measurement system with which we participated in the SemEval 2016 CL-STS shared task. Our system used QE and distance features based on bilingual word embeddings to train an ET regressor that predicts the cross-lingual semantic similarity between a pair of sentences. We used an ensemble method to generate and use training data for the task and saw that this approach improved the performance of our system. Our best performance achieves a Pearson's correlation of 0.53892 while placing FBK HLT-MT as the 7^{th} out of 10 teams in the task.

Acknowledgments

This work has been partially supported by the EC-funded H2020 project QT21 (grant agreement no. 645452).

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Baneab, Claire Cardiec, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirrea, Weiwei Guof, Inigo Lopez-Gazpioa, Montse Maritxalara, Rada Mihalceab, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556. Association for Computational Linguistics.

José G. C. de Souza, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014a. FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 322–328.

José G. C. de Souza, Marco Turchi, and Matteo Negri. 2014b. Machine translation quality estimation across domains. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 409–420.

José G. C. de Souza, Matteo Negri, Elisa Ricci, and Marco Turchi. 2015. Online Multitask Learning for Machine Translation Quality Estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–228, Beijing, China, July.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, et al. 2014. The matecat tool. In *COLING (Demos)*, pages 129–132.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning*, 63(1):3–42.

Christian Hänig, Robert Remus, and Xose de la Puente. 2015. Exb themis: Extensive feature extraction from word alignments for semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 264–268, Denver, Colorado, June. Association for Computational Linguistics.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*, pages 171–180, Montréal, Canada, June.

Angelo Mendonca, David Andrew Graff, Denise DiPersio, Linguistic Data Consortium, et al. 2009. *Spanish gigaword second edition*. Linguistic Data Consortium.

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

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition, june. *Linguistic Data Consortium, LDC2011T07*.

Alexandre Rafalovitch, Robert Dale, et al. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299.

Lucia Specia, Kashif Shah, José G. C. de Souza, and Trevor Cohn. 2013. Questa translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 79–84.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. Dls@cu: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 148–153, Denver, Colorado, June. Association for Computational Linguistics.

Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Baltimore, Maryland, USA, June.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.