Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity

Cristina España-Bonet
[1]University of Saarland
[2]DFKI, German Research Center for Artificial Intelligence
Saarbrücken, Germany
cristinae@dfki.de

Alberto Barrón-Cedeño
Qatar Computing Research Institute
HBKU, Qatar
albarron@hbku.edu.qa
albarron@gmail.com

Abstract

This is the Lump team participation at SemEval 2017 Task 1 on Semantic Textual Similarity. Our supervised model relies on features which are multilingual or interlingual in nature. We include lexical similarities, cross-language explicit semantic analysis, internal representations of multilingual neural networks and interlingual word embeddings. Our representations allow to use large datasets in language pairs with many instances to better classify instances in smaller language pairs avoiding the necessity of translating into a single language. Hence we can deal with all the languages in the task: Arabic, English, Spanish, and Turkish.

1 Introduction

The Semantic Textual Similarity (STS) task poses the following challenge. Let s and t be two text snippets. Determine the degree of equivalence $\alpha(s,t) \mid \alpha \in [0,5]$. Whereas 0 represents complete independence, 5 reflects semantic equivalence. The current edition (Cer et al., 2017) includes the monolingual ar–ar, en–en, and es–es, as well as the cross-language ar–en, es–en, and tr–enlanguage pairs. We use the two-letter ISO 639-1 codes: ar=Arabic, en=English, es=Spanish, and tr=Turkish.

Multilinguality is the premise of the Lump approach: we use representations which lie towards language-independence as we aim to be able to approach similar tasks on other languages, paying the least possible effort. Our regression model relies on different kinds of features, from simple length-based and lexical similarities to more sophisticated embeddings and deep neural net internal representations.

2 Features Description

The main algorithm used in this work is the support vector regressor from LibSVM (Chang and Lin, 2011). We use an RBF kernel and greedily select the best parameters by 5-fold cross-validation. In addition, we experiment with a different machine learning component built with gradient boosting algorithms as implemented by the XGBoost package.[1]

We describe the features in growing level of complexity: from language flags up to embeddings derived from neural machine translation.

2.1 Language-Identification Flags (6 feats.)

The novelty of the cross-language tasks causes a noticeable language imbalance in the amount of data (cf. Table 1). To deal with this issue, one of our systems learns on the instances in all the language pairs jointly. In order to reduce the clutter of the different data distributions, we devised six binary features that mark the languages of each pair. $lang1$, $lang2$ and $lang3$ are set to 1 if s is written in either ar, en, or es, respectively. The other three features, $lang4$, $lang5$, and $lang6$, provide the same information for t. For instance, the value for the six features for a pair en–ar would be 0 1 0 1 0 0.

2.2 Lengths (3 feats.)

Intuitively, if s and t have a similar length, being semantically similar is more plausible. Hence, we consider two integer features tok_s and tok_t: the number of tokens in s and t. We also use a length model (Pouliquen et al., 2003) len, defined as

$$\varrho(s,t) = e^{-0.5\left(\frac{\frac{|t|}{|s|}-\mu}{\sigma}\right)^2}, \quad (1)$$

[1]http://xgboost.readthedocs.io

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 144–149,
Vancouver, Canada, August 3 - 4, 2017.

where μ and σ are the mean and standard deviation of the character lengths ratios between translations of documents from L into L'; $|\cdot|$ represents the length of $\cdot$ in characters. If the ratio of lengths of s and t is far from the mean for related snippets, $\varrho(s,t)$ is rather low. This has shown useful in similar cross-language tasks (Barrón-Cedeño et al., 2010; Potthast et al., 2011). The parameters for the different language pairs are $\mu_{en-ar} = 1.23 \pm 0.60$, $\mu_{en-es} = 1.13 \pm 0.41$, $\mu_{en-tr} = 1.04 \pm 0.56$, and $\mu_{x-x} = 1.00 \pm 0.32$ for monolingual pairs.

2.3 Lexical Similarities (5 feats.)

We compute cosine similarities between character n-gram representations of s and t, with $n = [2, 5]$ ($2grm$,. . .,$5grm$). The pre-processing in this case is casefolding and diacritics removal. The fifth feature *cog* is the cosine similarity computed over "pseudo-cognate" representations. From an NLP point of view, cognates are "words that are similar across languages" (Manning and Schütze, 1999). We relax this concept and consider as pseudo-cognates any words in two languages that share prefixes. To do so, we discard tokens shorter than four characters, unless they contain non-alphabetical characters, and cut off the resulting tokens to four characters (Simard et al., 1992).

This kind of representations is used on European languages with similar alphabets (McNamee and Mayfield, 2004; Simard et al., 1992). We apply Buckwalter transliteration to texts in `ar` and remove vowels from the snippets written in latin alphabets. For the pseudo-cognates computations, we use three characters instead of four.

2.4 Explicit Semantic Analysis (1 feat.)

We compute the similarity between s and t by means of explicit semantic analysis (ESA) (Gabrilovich and Markovitch, 2007). ESA is a distributional-semantics model in which texts are represented by means of their similarity against a large reference collection. CL-ESA —its cross-language extension (Potthast et al., 2008)— relies on a comparable collection. We compute a standard cosine similarity of the resulting vectorial representations of s and t. Our reference collection consists of $12k$ comparable Wikipedia articles from the `ar`, `en`, and `es` 2015 editions. We did not compile a reference collection for `tr`.

2.5 Context Vectors in a Neural Machine Translation Engine (2 feats.)

Hidden units in neural networks learn to interpret the input and generate a new representation of it. We take advantage of this characteristic and train a multilingual neural machine translation (NMT) system to obtain a representation in a common space for sentences in all the languages. We build the NMT system in the same philosophy of Johnson et al. (2016) using and adapting the Nematus engine (Sennrich et al., 2016). The multilingual system is able to translate between any combination of languages `ar`, `en`, and `es`. It was trained on 60 k parallel sentences (20 k per language pair) using 512-dimensional word embeddings, 1024 hidden units, a minibatch of 200 samples, and applying Adadelta optimisation. The parallel corpus includes data from United Nations (Rafalovitch and Dale, 2009), Common Crawl[2], News Commentary[3] and IWSLT.[4]

We are not interested in the translations but in the context vectors output of the hidden layers of the encoder, as these are supposed to have learnt an interlingua representation of the input. We compute the cosine similarity between 2048-dimensional context vectors from the internal representation when the encoder is fed with s and t. Two independent systems, one trained with words and another one trained with lemmas[5] provide our two features $lNMT$ and $wNMT$.

2.6 Embeddings for Babel Synsets (2 feats.)

BabelNet is a multilingual semantic network connecting concepts via *Babel synsets* (Navigli and Ponzetto, 2012). Each concept, or word, is identified by its ID irrespective of its language, making these IDs interlingua. For this feature, we gather corpora in the three languages, convert them into sequences of BabelNet IDs, and estimate 300-dimensional word embeddings using the CBOW algorithm, as implemented in the `Word2Vec`

[2] `http://commoncrawl.org/`

[3] `http://www.casmacat.eu/corpus/news-commentary.html`

[4] `https://sites.google.com/site/iwsltevaluation2016/mt-track/`

[5] We built a version of the lemma translator with an extra language: Babel synsets (cf. Section 2.6), representing sentences with BabelNet IDs instead of words. The purpose was to extract also this feature for the `tr` surprise language, since it could be used for every language once the input sentences are converted into BabelNet IDs. However, the training was not advanced enough before the deadline and we could not include the results.

2017 Track	L–L'	Instances	Pctge.
1	ar–ar	1,081	5.11
2	ar–en	2,162	10.21
3	es–es	1,555	7.34
4	es–en	1,595	7.53
5	en–en	14,778	69.80
6	tr–en	0*	0.00
total		21,171	100

Table 1: Instances provided in the history of STS. (*No training data exists for this pair.)

package (Mikolov et al., 2013), with a 5-token window. We use the same corpora described before for training the NMT system with the addition of parts of Wikipedia and Gigaword to estimate the embeddings. For these experiments we annotated $1.7\,G$ tokens for ar, $1.1\,G$ for en, and $0.9\,G$ for es. As we are not interested in all the words of a sentence to represent its semantics, we restrict the extraction of Babel synsets to adjectives, adverbs, nouns, and verbs. Negations are considered tagging them with a special label. The global embeddings are then estimated from $1.9\,G$ synsets ($0.9\,G$ ar, $0.5\,G$ en, and $0.4\,G$ es).

Our two features consist of the cosine similarity between the embeddings of the two snippets. The full embedding of a snippet is obtained as the sum of the embeddings if its tokens. The difference between the two features lies in the corpus from which we estimate the embeddings. For $BNall$ we used the full collection of corpora in the three languages. For $BNsub$ we only used the subcollection of data coming from the languages involved in the pair. Significant differences in the performance of these two features will allow us to discern weather the interlinguality of these embeddings is a fair assumption or not (even if synsets are interlingua, its embeddings do not need to be).

2.7 Additional Features

We produced variations of the described features. We used other similarity measures than cosine: modified versions of the weighted Jaccard similarity, and the Kullback–Leibler and Jensen–Shannon divergences). We replicated the features described in Sections 2.3 to 2.6 with their monolingual counterpart. We obtained the counterpart translating ar and es snippets into en for Tracks 1-4 and 6, and en snippets into es for Track 5 with the multilingual NMT system (cf. Section 2.5). We used Google Translate for tr.

3 Experiments

For training, we used all the annotated datasets released both in the current and in previous editions.[6] Table 1 shows the size of the different language collections. Note the important imbalance: there are more than ten times more instances available in en only than in the rest of languages. We used the test set from the 2016 edition (only in English) (Agirre et al., 2016) as our internal test set.

Using the features in Sections 2.1 to 2.6, we train two regressors by:

Sys1 learning one SVM per each language pair

Sys2 learning one single SVM for all the language pairs together.

We experiment with a third system using all the extensions of Section 2.7 on XGBoost. The purpose of this system is to analyse and compare different assumptions made for Sys1 and Sys2:

Sys3 learning one single XGB for all the language pairs with an extended set of features.

Table 2 shows the results of the three settings; including the average Pearson correlation for mono- and cross-language tracks. Comparing Sys1 and Sys2, we see that in the case of en–en the best performance is obtained when training on en only. Adding instances in other languages slightly confuses our regressor, but differences are small; the number of examples added is only a 30%. Nevertheless, considering together different language pairs does help when dealing with less-represented pairs. This is the case of ar–ar, es–es, and es–en where the inclusion of more than ten times more instances in other languages boosts the performance. We did not observe this behaviour in the rest of language pairs. The worst case is that of the surprise pair tr–en. The reason could be that we could not compute all the features for these instances and instead, we used equivalents for en. Regarding the performance of our models on mono- and cross-language pairs, considering one single classifier versus one per language pair makes no difference when dealing with monolingual instances. This reflects the nature of the data: 82% of the training set is monolingual. The story is different when dealing with cross-language instances. Further experiments are necessary using one classifier with cross-language instances only.

[6] In order to combine all the datasets we had to do some cleaning and adaptation. For instance, the similarity values in some of the subsets ranged $[0, 4]$ rather than $[0, 5]$.

Track	L–L'	Sys1	Sys2	Sys3
Primary	all	0.4725	0.4438	0.4704
1	ar–ar	0.6052	0.6287	0.5508
2	ar–en	0.1829	0.1805	0.1357
3	es–es	0.7574	0.7380	0.7676
4a	es–en	0.4327	0.4447	0.4825
4b	es–en	0.0116	0.0151	0.1112
5	en–en	0.7376	0.7347	0.7269
6	tr–en	0.5800	0.3652	0.5179
avg_{mono}		0.7001	0.7005	0.6818
avg_{cross}		0.3359	0.2899	0.3435

Table 2: Official Pearson correlation performance for our three submissions. Average correlations for *mono*- and *cross*-language tracks at the bottom.

Regarding Sys3, we observe a lost in performance with respect to Sys1 and Sys2, except for the tracks involving es. The system introduces three variations with respect to Sys2: the learning model, the addition of several similarity measures for each representation, and the addition of new representations obtained after translating the input into en (es). A deeper analysis shows that the performance drop is due to the learning algorithm. XGBoost is performing better than SVM in our cross-validation. However, the loss function we use is a mean squared error and the evaluation is done via Pearson correlation. We attribute the discrepancy to this fact. Still, except for en–en, the inclusion of the two families of features improves the results of the basic features set.

Gradient boosting methods allow to estimate the importance of each feature in a very natural way: the more a feature is used to take the decisions in the construction of the boosted trees, the more important it is (Hastie, 2013). The complete analysis is out of the scope of this paper, but some comments and remarks can be made in the light of their relative importance. Figure 1 shows the relative importance of the features given by three XGBoost regressors: one trained only with en monolingual data, one for en–es cross-language data, and one for all the languages trained together. The concrete distribution of features depends on the specific language pair, but the set $\{len, 2grm, (CL)ESA, lNMT, wNMT, BNsub, BNall\}$ stands out among the full set. Notice that language identifiers are not relevant at all for the joint model and the regressor practically neglects them.

In general, the internal representation of the neural network is more important for cross-language pairs and Babel embeddings are more relevant for monolingual pairs. In the latter, we observe almost no difference between the relative importance of $BNsub$ and $BNall$, confirming the assumption of the interlinguality of the embeddings. (CL-)ESA is always among the most informative features. Finally, the high contribution of two simple scores is worth noting: len and $2grm$. This comes at no surprise for len (Barrón-Cedeño et al., 2014). Regarding the n-grams similarity, in general $\{3, 4\}$-grams perform better in similar tasks (e.g., comparable corpora parallelisation (Barrón-Cedeño et al., 2015)), but no important difference exist with respect to using 2-grams.

4 Conclusions and Future Work

Our approach to the SemEval 2017 task on semantic textual similarity focused on designing text representations which could be equivalent across languages. For example, instead of using standard monolingual or bilingual word embeddings, we build embeddings for the interlingua Babel synsets or let an autoencoder learn representations in the multilingual space. In internal experiments, monolingual word embeddings performed better than BabelNet embeddings for the monolingual tracks, but the advantage of the latter is that the same embeddings can be used for the seven tracks. This is useful for less-resourced languages and for easy porting of the system to new languages. That was true for the tr–en track but, at the moment, the huge difference between the performance of our systems across tracks does not allow us to go further with this conclusion.

In the future we want to take advantage of the amount of information that BabelNet has and we aim at including synsets for multiword expressions and exploiting translations to be able to extract the same sense in all the languages. We are also studying the behaviour of the internal representation of NMT systems in order to determine the appropriate configuration of the translation system to be used for this purpose. To the best of our knowledge, the internal representation and the importance of its dimensionality has not been studied as an interlingual space.

Acknowledgements

Part of this work has been funded by the Leibniz Gemeinschaft via the SAW-2016-ZPID-2 project. The research work of the second author is carried out in the framework of the Interactive sYstems for Answer Search project (IYAS), at the Qatar Computing Research Institute, HBKU.

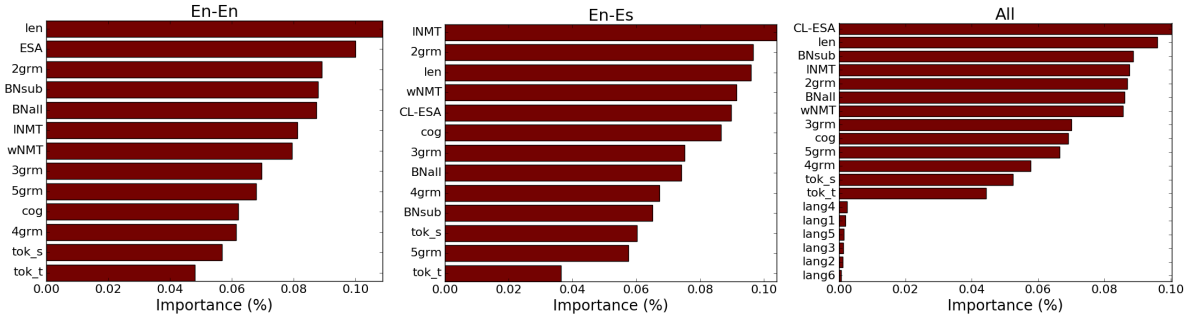

Figure 1: Relative importance of the features in the XGBoost regressors for the monolingual `en–en` Track 5, the cross-language `en–es` Track 4, and the all joint training.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, CA, pages 497–511.

Alberto Barrón-Cedeño, Cristina Españ-Bonet, Josu Boldoba, and Lluís Màrquez. 2015. A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*. Beijing, China.

Alberto Barrón-Cedeño, Monica Paramita Lestari, Paul Clough, and Paolo Rosso. 2014. A Comparison of Approaches for Measuring Cross-Lingual Similarity of Wikipedia Articles. In Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, Springer International Publishing, volume 8416 of *Lecture Notes in Computer Science*, pages 424–429.

Alberto Barrón-Cedeño, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism Detection across Distant Language Pairs. In Chu-Ren Huang and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. COLING 2010 Organizing Committee, Beijing, China, pages 37–45.

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.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 1606–1611.

Trevor Hastie. 2013. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York, NY, 2nd ed. corr. 7th printing 2013 edition.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR* abs/1611.04558.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Paul McNamee and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7(1-2):73–97.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*. http://code.google.com/p/word2vec.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Cross-language plagiarism detection. *Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis* 45(1):1–18.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A Wikipedia-Based Multilingual Retrieval Model. *Advances in Information Retrieval, 30th European Conference on IR Research* LNCS (4956):522–530. Springer-Verlag.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. 2003. Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*. Borovets, Bulgaria, pages 401–408.

Alexandre Rafalovitch and Robert Dale. 2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In *Proceedings of the Machine Translation Summit XII*. International Association of Machine Translation, pages 292–299.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT16)*.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.