CNRC at SemEval-2016 Task 1: Experiments in Crosslingual Semantic Textual Similarity

Chi-kiu Lo **Cyril Goutte** **Michel Simard**

National Research Council Canada
Multilingual Text Processing
1200 Montreal Road, Ottawa, Ontario K1A 0R6, Canada
FirstName.LastName@nrc.ca

Abstract

We describe the systems entered by the National Research Council Canada in the SemEval-2016 Task1: Crosslingual Semantic Textual Similarity. We tried two approaches: One computes a true crosslingual similarity based on features extracted from lexical semantics and shallow semantic structures of the source and target fragments, combined using a linear model. The other approach relies on Statistical Machine Translation, followed by a monolingual semantic similarity, relying again on syntactic and semantic features. We report our experiments using trial data, as well as official final results on the evaluation data.

1 Introduction

The SemEval-2016 Semantic Textual Similarity (STS) evaluation (task1) introduced a crosslingual track. Given a Spanish-English bilingual fragment pair, the goal is to compute the degree of equivalence between them. This offers additional challenges compared to the "STS Core" track, where both fragments are in the same language (English or Spanish in 2014-15). The crosslingual track requires potentially to detect which fragment is in which language, perform further language processing accordingly, and estimate lexical and semantic similarities across languages.

In our work, we investigated two approaches. In the first approach, we try to build a true crosslingual similarity based on a number of features computed from both fragments. One of these features projects Spanish words into an English embedding space in order to compute similarities in that space. Other features compute various kinds of syntactic and semantic overlap between the fragments. These features are combined in a linear model estimated on the trial data, and combined with an isotonic regression (de Leeuw et al., 2009) layer in order to account for non-linearity in the scores.

The second approach uses a Statistical Machine Translation system to map Spanish fragments to English, then relies on a monolingual semantic similarity between the translated fragment and the English fragment. Various monolingual similarity features, using embeddings, syntactic and semantic information, are computed and combined again using a linear model followed by an isotonic regression layer.

In the next section, we describe the two approaches and their components: SMT system, crosslingual and monolingual feature extraction, and the output layer fitting the features to the semantic similarity scores. We then describe the corpora used to fit the features to the output similarity score, and make various modeling choices (Section 3). We present our experimental results on the trial and test data in Section 4.

2 System description

We describe our two approaches: the direct crosslingual similarity using embedding mapping (`EMAP` run) and the use of Machine Translation followed by a monolingual semantic similarity (`MT1` and `MT2`).

2.1 Crosslingual Embedding Mapping

Feature Extraction: We evaluate the semantic similarity of the given text based on two levels: lex-

Proceedings of SemEval-2016, pages 668–673,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

ical semantics and shallow semantic structure.

One of the trivial ways to evaluate the crosslingual lexical similarity is using the alignment probability of an alignment model trained with a large-scale parallel corpus.[1] However, the alignment model does not evaluate the meaning similarity of words as well as the vector space model which is explicitly trained to evaluate semantic similarity. We therefore propose to combine the two models: given a Spanish word, we first look up in the alignment model for a list of the most probable (5-best) aligned English word (the *mapping* step); we then evaluate the lexical similarity of each entry in the 5-best list against the target English word using a word embeddings model. In our experiments, we used pretrained `word2vec` (Mikolov et al., 2013) embeddings.[2] The resulting crosslingual lexical similarity of the targeted pair of Spanish and English words is the highest similarity between the 5 mapped words and the target English word. We then reconstruct the semantic phrasal similarity by averaging the English-idf-weighted crosslingual embeddings mapped lexical similarity according to the 1-1 maximal matching alignment of the lexicons in the two phrases.

In addition to the flat lexical semantic feature, we use XMEANT (Lo et al., 2014), the crosslingual semantic frame based machine translation evaluation metric, for generating shallow structural semantic features. We use MATE (Björkelund et al., 2009) for Spanish shallow semantic parsing and SENNA (Collobert et al., 2011) for English shallow semantic parsing. In evaluating machine translation quality, the confusion of semantic roles is a major source of errors due to reordering. However, in evaluating STS, confusion of semantic roles is less frequent while missing information in one of the test fragments is more frequent. This motivates a further simplification of the 12 semantic role types (Lo et al., 2014) into 5 semantic role types: action, agent, patient, beneficiary and others. The same phrasal semantic similarity function mentioned above is used for evaluating the role fillers similarity, instead of the ITG-constrained crosslingual phrasal similarity function (Lo et al., 2014).

As a result, for each pair of the test sentences, we extracted 12 features (Table 1). The first feature is the simple phrasal similarity by considering the whole string of the testing sentences as one phrase. The second feature is the XMEANT score. The remaining 10 features are the precision and recall of the 5 semantic role types used in XMEANT.

Feature	Description
1	Embedding-based phrasal similarity
2	XMEANT score
3,4	p,r for semantic role: action
5,6	p,r for semantic role: agent
7,8	p,r for semantic role: patient
9,10	p,r for semantic role: beneficiary
11,12	p,r for semantic role: others

Table 1: Features used by the cross-lingual and monolingual semantic similarity. p,r stands for precision and recall.

Fitted Output Layer: Most of the extracted features are correlated with the gold standard semantic similarity score, by capturing various aspects of the similarity. In order to combine these features, we estimate a linear combination by fitting a least mean squares linear regression on the trial data gold standard similarity score.[3] Although it may be desirable to combine features in a non linear way, the amount of available annotated data severely limits our capacity to estimate non-linear models. We improve the modeling slightly by fitting a non-linear transformation of the estimated score produced by the linear combination, with the constraint that the transformation preserves the ordering of scores. This is done through isotonic regression, using the efficient implementation available in `R` (de Leeuw et al., 2009). In order to avoid overfitting to the limited number of training example, we use a 10-fold cross-validation estimator on the trial data to select the appropriate features and check the performance of the isotonic regression layer.

2.2 MT + Monolingual Similarity

Statistical Machine Translation. All Spanish text was translated to English using an SMT system based on *Portage*, the NRC's phrase-based SMT technology (Larkin et al., 2010). The system was

[1] Part of the MT system described in Section 2.2.

[2] `https://code.google.com/archive/p/word2vec/`

[3] We use the `glm` function in `R`.

trained using standard resources – Europarl, Common Crawl (CC) and News & Commentary (NC) – totaling approximately 110M words in each language. Phrase extraction was done by aligning the corpora at the word level using HMM, IBM2 and IBM4 models, using the union of phrases extracted from these separate alignments for the phrase table, with a maximum phrase length of 7 tokens. Phrase pairs were filtered so that the top 30 translations for each source phrase were retained. The following feature functions are used in the log-linear model: three 5-gram language models with Kneser-Ney smoothing (Kneser and Ney, 1995), i.e. one for each of Europarl, CC and NC data, combined linearly (Foster and Kuhn, 2007) to best fit NC data; lexical estimates of the forward and backward translation probabilities obtained either by relative frequencies or using the method of (Zens and Ney, 2004); lexicalized distortion (Tillmann, 2004; Koehn et al., 2005); and word count. The parameters of the log-linear model were tuned by optimizing BLEU on the development set using the batch variant of MIRA (Cherry and Foster, 2012). Decoding uses the cube-pruning algorithm of (Huang and Chiang, 2007) with a 7-word distortion limit.

For any given input in Spanish, the SMT system produces the translation that is most likely with regard to its own training data; that translation may be arbitrarily distant from the English sentence to which it will be compared in the STS task. These arbitrary surface differences may complicate the task of measuring semantic similarity. To alleviate this problem, we bias the MT system to produce a translation that is as close as possible on the surface to the English sentence. This is done by means of log-linear model features that aim at maximizing n-gram precision between the MT output and the English sentence. The relative weights of these features are set to maximize BLEU on the trial data. This optimization is performed separately from that of the other features, using a simple grid-search approach. Systems `MT1` and `MT2` are the top-ranking systems in this regard: `MT1` uses unigram/bigram/trigram weights 16/4/1, while `MT2` uses 16/2/2.

Feature Extraction. The features extracted in this run are essentially the same as those in the crosslingual approach, except that the lexical semantic similarity is now directly evaluated using the monolingual word embeddings model. Similarly, the structural semantic similarity is now evaluated using MEANT (Lo et al., 2015) instead of XMEANT, and the semantic role similarity features are obtained by evaluating the semantic parses in the same languages.

Fitted Output Layer. This output layer is essentially the same as the one in the crosslingual approach, but estimated on the monolingual features. One key advantage here is that we rely on a monolingual semantic similarity. We can then fit the monolingual models on English trial and test data from previous STS tasks. We combine the 2012 to 2014 development and test sets totaling 10,662 examples, instead of the 103 trial bilingual pairs available for the crosslingual task.

3 Textual Similarity Data

Monolingual. In order to estimate parameters of the monolingual similarity, we use 10,662 English pairs from the 2012 to 2014 development and test sets. We test this similarity on the 2015 test data, comprising 3000 examples in 5 different test sets.

Crosslingual. We use the 103 Spanish-English pairs provided as trial data for the crosslingual task for two purposes. We estimate the crosslingual similarity output layer, and we compute performance estimates for all our runs. In order to tune the output layer (in particular to select the relevant features), we compute an unbiased estimator of the prediction performance using cross-validation on the data used for fitting (103 examples in crosslingual, 10,662 in monolingual).

4 Experimental results

4.1 Results on trial data

For the crosslingual embedding mapping (`EMAP` run), the only gold standard data available is the 103 trial set pairs. In order to get an unbiased estimate of the performance, we compute a 10-fold cross-validation estimator. We use it to select the best subset of features. Figure 1 shows that the choice of features has a large impact, with estimated performance ranging from 0.64 (for all features) to 0.71 with the four features with highest correlation with the gold

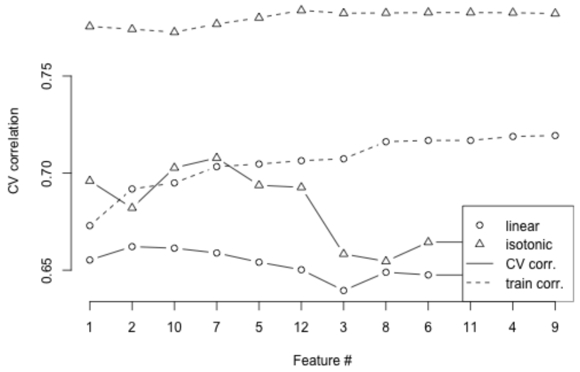

Figure 1: Cross-validated correlation when adding new features, ranked by individual correlation with the gold standard. The optimal CV estimate (solid+triangle curve) is obtained with four features (1, 2, 7 and 10).

	Top1	Linear	Isotonic
EMAP	.674	.659*	.708*
MT1	.714	.723	.731
MT2	.713	.720	.727

Table 2: Estimated results on the trial data (* estimates from 10-fold cross-validation).

standard. These four features are: the crosslingual similarity score, the XMEANT score and two semantic role features. The cross-validated correlation for the linear model and isotonic regression are 0.659 and 0.708, respectively (Table 2). For our submitted run EMAP, we re-estimated the linear mode and isotonic regression on all 103 trial examples and used those models to estimate scores on the two test sets (301 and 2973 examples, respectively).

For the MT1 and MT2 runs, the monolingual similarity is trained on the available monolingual STS data from 2012 to 2014. In order to test the resulting similarity, we apply it to the 2015 test data, and obtain an average correlation of .713 on the five test sets, significantly below the best performing system at the 2015 task (.801 average). We also apply the monolingual STS to the trial data after forced decoding and feature extraction and obtain an estimated trial correlation of .727–.731 (Table 2).

Run	News (301)	MultiSource (294)	Mean
MT1	0.876	0.646	0.762
MT2	0.878	0.631	0.756
EMAP	0.719	0.411	0.567

Table 3: Official evaluation results for our three runs.

4.2 Test results

Test results computed by the organizers are shown in Table 3. Average test results and results on the News part are significantly higher than estimated on the trial data. A large difference is not unexpected considering the test data was clearly very different from the trial data. The average fragment length is 3-4 times larger on News than on trial, for example. We can conjecture that the higher performance on the News test set may actually be due to the longer fragments providing more information to estimate the similarity. On the Multisource test set, on the other hand, all our runs perform poorly. We also note that the performance of the EMAP run is much worse, relative to the MT runs, than could be anticipated from the trial data performance (Table 2). We analyse these differences below.

4.3 Analysis

MT vs EMAP: As noted above, performance of EMAP was disappointing on the test set, and the gap with MT runs much larger than expected from the trial data. Two main reasons can explain this. First, MT uses monolingual word embeddings directly, and second, we could use 10,662 monolingual pairs with reference STS scores to fit and tune the feature combination model. By contrast, for lack of cross-lingual embedding vectors, the EMAP run had to rely on word alignment to map Spanish words to English embeddings, and the feature combination model could only use 103 cross-lingual pairs with reference STS scores.

MultiSource performance is lower than performance on the News part of the test set for almost all systems involved in the evaluation (FBK being the notable exception). The difference is particularly pronounced for our runs: our MT runs are only .035 below the top run on News, which is likely not significant on 301 examples;[4] our performance on MultiSource, on the other hand, is .17 to .40 below the top runs. Given that this gap is especially pronounced for our systems, we can not rule out formatting errors of inconsistencies on our side. Further investigation will be needed to clarify this.

Running Time: One attractive feature of both our approaches is that they rely on shallow semantic features which are easy and fast to obtain using semantic role labeling. The `MT` runs rely on a SMT system which is expensive to train, but this can be done offline. Once trained, producing translations is of the order of ten sentences per second. The linear feature combination model is also fast to train and apply, requiring a single dot product between the 12 features and a similarly-sized parameter vector.[5]

5 Conclusion

We described the systems used for the submissions of the National Research Council Canada to the crosslingual semantic textual similarity task. We experimented with two approaches. The first estimates a true crosslingual similarity combining lexical semantics and shallow semantic structure. The second uses Machine Translation in combination with a monolingual semantic similarity. We found that the latter outperforms the former. We conjecture that this may be due to the very limited amount of crosslingual data available. By contrast, there are very large corpora available for training a reasonably efficient MT system, and we can rely on a lot of data from previous STS tasks to build a monolingual semantic similarity. Test results indicate that our approach performs relatively well on the `News` test set, but suffers on the `Multisource` test set. In addition, the crosslingual approach performs much worse on average on the test data than estimated on the limited trial data. This suggests that it is hard to build a competitive, truly crosslingual approach from little reference data, when it is possible to rely on thousands to millions of examples to build a SMT+monolingual similarity pipeline.

[4] Assessing statistical significance would require access to the predictions of the best run.

[5] Prediction for the 301+2973 test examples takes 24ms.

References

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398.

Jan de Leeuw, Kurt Hornik, and Patrick Mair. 2009. Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(1):1–24.

George Foster and Roland Kuhn. 2007. Mixture-Model Adaptation for SMT. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135, Prague, Czech Republic, June.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic, June. Association for Computational Linguistics.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of IWSLT-2005*, Pittsburgh, PA.

S. Larkin, B. Chen, G. Foster, U. Germann, E. Joanis, H. Johnson, and R. Kuhn. 2010. Lessons from NRC's Portage system at WMT 2010. In *the Joint Fifth Workshop on Statistical Machine Translation and Metrics-MATR*, pages 127–132.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. XMEANT: Better semantic MT evaluation without reference translations. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 765–771.

Chi-kiu Lo, Philipp C Dowling, and Dekai Wu. 2015. Improving evaluation and optimization of MT systems against MEANT. In *10th Workshop on Statistical Machine Translation (WMT 2015)*, page 434.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Christoph Tillmann. 2004. A Unigram Orientation Model for Statistical Machine Translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Richard Zens and Hermann Ney. 2004. Improvements in Phrase-Based Statistical Machine Translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 257–264, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.