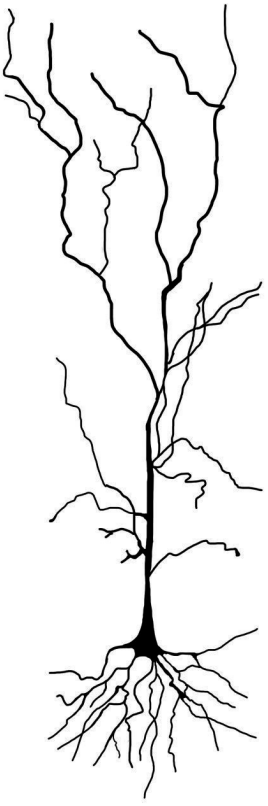
One Model to Rule them All

Bjerva, Johannes

PART III

Multilingual Learning



CHAPTER 6

*Multilingual Semantic Textual Similarity

Abstract | Up until now, we have seen parameter sharing between tasks, in the context of multitask learning. Another possibility is to share parameters between languages, in a sense casting model multilinguality as a type of multitask learning. As a first example of multilingual learning, we look at cross-lingual semantic textual similarity. This task can be approached by leveraging multilingual distributional word representations, in which similar words in different languages are close to each other in semantic space. The availability of parallel data allows us to train such representations for a large number of languages. Such representations have the added advantage of allowing for leveraging semantic similarity data for languages for which no such data exists. In this chapter, the focus is on to what extent such an approach allows for enabling zero-shot learning for the task at hand. We also investigate whether language relatedness has an effect on how successful this is. We train and evaluate on six language pairs for semantic textual similarity, including English, Spanish, Arabic, and Turkish.

*Chapter adapted from: **Bjerva, J.** and Östling, R. (2017a). Cross-lingual Learning of Semantic Textual Similarity with Multilingual Word Representations. In Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden, number 131, pages 211–215. Linköping University Electronic Press, Linköpings universitet.

6.1 Introduction

In order to determine how similar a pair of sentences in two different languages are to one another, it is necessary to have a grasp of multilingual semantics.[2] Continuous space word representations, which lend their strength from distributional semantics, are a clear candidate for this problem, as distances in such a space can be directly interpreted as semantic similarities (see Chapter 3). Given the constantly increasing amount of available parallel data (e.g., Koehn (2005), Tiedemann (2012), Ziemski et al. (2016)), it is possible to learn multilingual word representations for many languages. In this chapter, we approach tasks on cross-lingual semantic textual similarity from SemEval-2016 (Agirre et al., 2016) and SemEval-2017 (Agirre et al., 2017) using such word representations. This approach has the advantage that it allows for zero-shot learning while training on multiple languages, i.e., exploiting data from several source languages for an unseen target language. We aim to answer the following specified research question, in order to answer **RQ 3**:

RQ 3a To what extent can multilingual word representations be used in a simple STS system so as to enable zero-shot learning for unseen languages?

RQ 3b To what extent is the success of zero-shot learning dependent of language relatedness in this setting?

Related work

Semantic Textual Similarity (STS) is the task of assessing the degree to which two sentences are similar in their meanings. In the long-running SemEval STS shared task series, this is measured on a scale ranging from 0, indicating no semantic similarity, to 5, indicating

[2] The exception to this is methods based on machine translation, which are outlined in the following section.

complete semantic similarity (see Agirre et al., 2012; 2013; 2014; 2015; 2016; 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 target translation can aid both system evaluation and development. STS is also a fundamental problem for natural language understanding, as being able to estimate the similarity between two sentences in their meaning content can be seen as a prerequisite for understanding. The task is already a challenging one in a monolingual setting, such as when estimating the similarity between two English sentences. In this chapter, we tackle the more difficult case of cross-lingual STS, e.g., estimating the similarity between an English and an Arabic sentence, in the context of shared tasks on cross-lingual STS at SemEval-2016 (Agirre et al., 2016) and SemEval-2017 (Agirre et al., 2017).[3,4]

Previous approaches to the problem of cross-lingual STS have focussed on two main approaches. The primary, and most successful approach, is to apply a MT system to non-English sentences, and translating these to English (e.g., Tian et al., 2017, and Wu et al., 2017). The advantage of this approach is that the problem essentially boils down to estimating the similarity of two sentences in English. There are at least two advantages to this. First of all, the amount of resources available for English eclipse what is available for most, if not all, other languages. Additionally, comparing two sentences in the same language allows for straight-forward application of features based on the surface forms of the sentences, such as word overlap, common substrings, and so on. MT approaches tend to outperform purely multilingual approaches, with the winner of SemEval-2017, and many of the top systems of SemEval-2016 relying on this approach (Agirre et al., 2016, 2017). There are at least two draw-

[3]SemEval-2016 Task 1: Semantic Textual Similarity: A Unified Framework for Semantic Processing and Evaluation – Cross-lingual STS Subtask.

[4]SemEval-2017 Task 1: Semantic Textual Similarity.

backs to this method, however. Primarily, involving a fully-fledged MT system severely increases the complexity of a system. Furthermore, such methods can be seen as bypassing the problem of cross-lingual STS, rather than tackling it directly, as no actual multilingual similarity assessments are necessarily carried out.

The amount of true multilingual approaches in the literature are somewhat more limited, with notable examples from SemEval-2016 such as Lo et al. (2016), who make use of bilingual embedding space phrase similarities, in combination with cross-lingual machine translation metrics. Another approach is represented by Aldarmaki and Diab (2016), who apply bilingual word representations in a matrix factorisation method, so as to assess STS without translation. The method that bears the most resemblance to the approach taken in this chapter is Ataman et al. (2016), who combine bilingual embeddings with machine translation quality estimation features (Specia et al., 2013). We expand upon this method by using *multilingual* word embeddings as input, rather than bilingual ones. One advantage of our approach, is that it allows for zero-shot learning while training on multiple languages (see Chapter 3), as it does not depend on annotated STS training data for the target language, and only places requirements on the availability of parallel data. This denotes the approach we take to **RQ 3a**, as well as **RQ 3b**. Additionally, our method differs from Ataman et al. (2016) in that we choose a simpler architecture, using only such word representations as input.

6.2 Cross-lingual Semantic Textual Similarity

We will now look at the task of (cross-lingual) STS in more detail. Given two sentences, s_1 and s_2, the task in STS is to assess how semantically similar these are to each other. This is commonly measured using a scale ranging from 0–5, with 0 indicating no semantic overlap, and 5 indicating nearly identical content. In the SemEval

STS shared tasks, the following descriptions are used:

0. The two sentences are completely dissimilar.
1. The two sentences are not equivalent, but are on the same topic.
2. The two sentences are not equivalent, but share some details.
3. The two sentences are roughly equivalent, but some important information differs/missing.
4. The two sentences are mostly equivalent, but some unimportant details differ.
5. The two sentences are completely equivalent, as they mean the same thing.

As an example of sentence similarities, consider the sentence pairs and their human-annotated similarity scores in Table 6.1. These examples are taken from the SemEval-2014 edition of the shared task on STS and Recognising Textual Entailment (RTE), giving us access to entailment information in addition to similarity scores for the purposes of the example (Marelli et al., 2014).[5] Attempting to assess the semantic content of two sentences with a simple score notably does not take important semantic features such as negation into account, and STS can therefore be seen as complimentary to textual entailment. For instance, in sentence No. 219 in Table 6.1, the sentences have a high similarity score, even though their meanings are the opposite of one another. It is also worth to note that STS is highly related to paraphrasing, as replacing an n-gram with a paraphrase thereof ought to alter the semantic similarity of two sentences to a very low degree.

[5] RTE is the task of assessing whether the meaning of one sentence (the *hypothesis*) can be inferred from the other (the *text*).

Table 6.1: Examples of sentence similarities and corresponding entailment judgements.

No.	Text / Hypothesis	Score	Relation
8678	A skateboarder is jumping off a ramp A skateboarder is making a jump off a ramp	4.8	entailment
2709	There is no person boiling noodles A woman is boiling noodles in water	2.9	contradiction
219	There is no girl in white dancing A girl in white is dancing	4.2	contradiction

Table 6.2: Examples of cross-lingual sentence similarities.

English / Spanish	Score
The NATO mission officially ended Oct. 31. La misión de la OTAN terminó oficialmente oct. 31.	5
Mass Slaughter on a Personal Level El sacrificio masivo en un nivel personal	3
Support Workers' Union Will Sue City Over Layoffs Apoyo a los trabajadores "Unión va a demandar ciudad más despidos	1

Successful monolingual approaches in the past have taken advantage of both the relatedness with this task to paraphrasing, and to RTE. Bjerva et al. (2014) attempt to replace words in s_1 with paraphrases obtained from the Paraphrase Database (PPDB, Ganitkevitch et al., 2013), in order to increase the surface similarity with s_2. Additionally, both Bjerva et al. (2014) and Beltagy et al. (2016) make use of (features from) an RTE system to perform the task of STS. Approaches similar to these can be applied in cross-lingual STS, if the sentence pair is translated to a language for which such resources exist.

As an example of sentence similarities in cross-lingual STS, con-

sider the sentence pairs and their human-annotated similarity scores in Table 6.2. The first example contains a Spanish sentence which is a faithful translation of the English one, and has the highest similarity score (5). Although the second example conveys a similar meaning, the translation expresses 'Mass Slaughter' in Spanish as 'A massive sacrifice', resulting in a lower similarity score (3), indicating a loose translation.[6] In the third example sentence, the Spanish sentence conveys the opposite meaning of the English sentence, and has the lowest similarity score (1).

6.3 Method

As mentioned in Section 6.1, we approach the task of multilingual STS in a similar manner to Ataman et al. (2016), with the addition that we use multilingual input representations, rather than bilingual ones. We will now look at how our system is constructed, starting with the input representations.

6.3.1 Multilingual word representations

There are several methods available for obtaining multilingual word representations, as described in Chapter 3. In this chapter, we use a variant of the multilingual skip-gram method (Guo et al., 2016), as detailed in Chapter 3. This method was chosen as it is both relatively simple, and yields high-quality representations for down-stream tasks, as compared to other approaches (Guo et al., 2016). The original method relies on using cross-lingual contexts, with English as a pivot language. For instance, a Spanish word might be used to predict an English context, or the other way around. Our approach differs in that we augment the learning objective so as to include multilingual

[6] See Bos (2014) for a further discussion of faithful, informative, and loose translations in the context of parallel corpora.

contexts, such that we also use, for instance, a Spanish word to predict a French word (Figure 3.5 in Chapter 3).

We train 100-dimensional multilingual embeddings on the Europarl (Koehn, 2005) and UN corpora (Ziemski et al., 2016), including data from bible translations.[7,8] This data was chosen partially since it allows us to learn such embeddings for a large number of languages, in addition to the availability of these corpora. The dimensionality of the embeddings was chosen by balancing a sufficiently high number of dimensions with the computational resources necessary to compute these embeddings with the extended version of the multilingual skip-gram method. Word alignment, which is required for the training of this type of multilingual embeddings, is performed using a tool based on the Efmaral word-alignment tool (Östling and Tiedemann, 2016).[9] This allows us to extract a large amount of multilingual (word, context) pairs. We then use these pairs in order to learn multilingual embeddings, by applying the *word2vecf* tool (Levy and Goldberg, 2014a). In our experiments, we use the same parameter settings as Guo et al. (2016), training using negative sampling (Mikolov et al., 2013a), and with equal weighting of monolingual and cross-lingual contexts.[10]

6.3.2 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. This means that we need a sentence-level representation, based on the multilingual word representations, of-

[7]Using the New Testament (approximately 140,000 tokens), available at `http://homepages.inf.ed.ac.uk/s0787820/bible/`.

[8]Training multilingual embeddings on this data yields a vocabulary coverage of over 85% on the development sets of the languages at hand.

[9]We use the *eflomal* tool, which uses less memory than *efmaral*. Default parameters are used. Available at `https://github.com/robertostling/eflomal`.

[10]Note that we do not use the same implementation as (Guo et al., 2016).

fering us a choice between methods such as those presented in Chapter 2, Section 2.4.2. Given 100-dimensional word representations for each word in our sentence, we opt for the simplistic approach of averaging the vectors across each sentence, such that

$$\vec{s} = \frac{1}{|s|} \sum_{w \in s} \vec{w}, \tag{6.1}$$

where w is a word in the sentence s, and $\vec{w}$ and $\vec{s}$ are their vectorial representations. This is the same approach that is taken by Ataman et al. (2016). The resulting sentence-level representations are then concatenated and passed through two fully connected layers with ReLU activation functions (200 and 100 units, respectively), 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. The system architecture is depicted in Figure 6.1.

We apply dropout ($p = 0.5$) between each layer (Srivastava et al., 2014). All weights are initialised using the approach from Glorot and Bengio (2010). We use the Adam optimisation algorithm (Kingma and Ba, 2014), monitoring the categorical cross-entropy of the 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 monitoring the loss on the validation set. We report results using the model with the lowest validation loss. Hyperparameters are kept constant in all conditions.

Figure 6.1: System architecture used in the semantic textual similarity task.

6.3.3 Data for Semantic Textual Similarity

As the SemEval STS task series has been running for several years, there is a substantial amount of data available. We use data from previous editions of the tasks on (cross-lingual) STS.[11] For English–English this includes data from SemEval 2012 through 2015 (Agirre et al., 2012, 2013, 2014, 2015). For English–Spanish this includes data from SemEval 2016 and 2017 (Agirre et al., 2016, 2017). For Spanish–Spanish this includes data from SemEval 2014 and 2015. For English–Arabic this includes data from SemEval 2017. Finally, for Arabic–Arabic, this includes data from SemEval 2017. We use the concatenation of the training sets of previous editions for training, and validate and test on the most recent data for each language pair. An overview of the available data is shown in Table 6.3.

Table 6.3: Training data used for (cross-lingual) STS from the SemEval shared task series.

Language pair	N sentence pairs	SemEval edition(s)
English – English	3,000	2012 – 2015
English – Spanish	3,900	2016 – 2017
Spanish – Spanish	1,500	2014 – 2015
English – Arabic	2,000	2017
Arabic – Arabic	900	2017

6.4 Experiments and Results

We investigate whether using a multilingual input representation and shared weights allow us to ignore languages in STS, after mapping words to their multilingual representations. This, in turn, is one approach for enabling zero-shot learning for this task (**RQ 3a**). We

[11] This is what is generally recommended by the shared task organisers, and is followed by most participating systems.

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. In doing so, we investigate the extent to which the availability of parallel data allows us to train STS systems without access to STS training data for a given language.

Secondly, we investigate the effect of bundling training data together, in multi-source training, investigating which language pairings are helpful for each other. Concretely, in single-source training, we only train on one out of the language pairs at a time, and evaluate the resulting single-source model on all language pairs. In multi-source training, however, a model is trained on several language pairs at a time, and the resulting model is evaluated as in the single-source training setting. This is done so as to offer insight into **RQ 3b**.

We measure performance between target 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.

6.4.1 Comparison with Monolingual Representations

As a baseline, we compare multilingual embeddings with the performance obtained using the pre-trained monolingual Polyglot embeddings (Al-Rfou et al., 2013). Training and evaluating on the same language pair yields comparable results regardless of embeddings (Table 6.4). This shows that our multilingual embeddings, at the very least, have comparable quality to purely monolingual embeddings in a monolingual setting.

Table 6.4: Single-source training results (Pearson correlations) with monolingual embeddings (polyglot) as compared to multilingual embeddings (multilingual skipgram) on the SemEval-2017 development set. Rows indicate evaluation language, and rows indicate the embeddings used. Bold numbers indicate best results per row.

	Polyglot	**Multilingual Skipgram**
English	0.68	**0.69**
Spanish	0.65	0.65
Arabic	0.70	**0.71**

Table 6.5: Single-source training results with multilingual embeddings on the SemEval-2017 development set (Pearson correlations). Columns indicate training language pairs, and rows indicate testing language pairs. Bold numbers indicate best results per row.

Test \ **Train**	**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

6.4.2 Single-source training

Results when training on a single source corpus, using multilingual embeddings, are shown in Table 6.5. 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. Observing the results from a zero-shot learning perspective, it seems that certain language combinations

can benefit from this approach. Mainly, it seems to be the case that this approach is suitable when training on a monolingual source pair (such as English–English), and evaluating the model on a monolingual target pair (such as Spanish–Spanish). The gap in performance between such cases, and a system where target and source languages are identical is relatively small. One example of this is the results of evaluating on English–English when training on English–English (0.69) as compared to when training on Spanish–Spanish (0.64). We can also observe that zero-shot learning in this setting is more successful between the Indo-European languages Spanish and English, than when involving the Semitic language Arabic.

6.4.3 Multi-source training

We combine training sets from two language pairs in order to investigate how this affects evaluation performance on target language pairs. 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 6.6).

Table 6.6: Results with one source language in addition to target-language data with multilingual embeddings on the SemEval-2017 development set (Pearson correlations). Columns indicate added source language pairs, and rows indicate target language pairs. Bold numbers indicate best results per row.

Test \ Train	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

We observe that the monolingual language pairings (English–English, Spanish–Spanish, Arabic–Arabic) appear to be beneficial for one another, also in this setting. Interestingly, adding Arabic–Arabic training data seems to improve the performance on both English–English and Spanish–Spanish. Although, this difference is small and likely not significant, it is interesting that the models performance does not worsen as an effect of adding this data. This might have been the case, if model capacity is wasted on solving the task for Arabic–Arabic.

Finally, we run a multilingual ablation experiment, in which we train on two out of three of these language pairs, and evaluate on all three. Notably, excluding all Spanish training data yields comparable performance to including it (Table 6.7). Additionally, we see the largest drop in performance when evaluating on Arabic data without having trained on it. This adds further support to the findings in Table 6.5, indicating that language relatedness is of importance for the success of zero-shot learning as applied here.

Table 6.7: Multilingual ablation results with multilingual embeddings on the SemEval-2017 development set (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. The bold diagonal indicates results when the target language pair is not used as a source language pair.

Test \ Ablated	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 6.8: Results with multilingual embeddings on SemEval-2017 Shared Task Test sets. In the multi-source and ablation conditions, we use the systems with the best validation performance for the target language pair. The columns indicate the evaluation language, and the Primary column indicate the aggregated results over all languages, as used in SemEval-2017 (Agirre et al., 2017). The wmt column denotes the es–en test set drawn from WMT's quality estimation track. Rows indicate our three systems, compared with the ECNU system (Tian et al., 2017), and the LIPN–IIMAS system (Arroyo-Fernández and Ruiz, 2017).

	Primary	**ar–ar**	**ar–en**	**es–es**	**es–en**	**wmt**	**en–en**	**en–tr**
Single-source	0.315	0.289	0.105	0.661	0.239	0.030	0.691	0.188
Multi-source	0.294	0.312	0.129	0.692	0.100	0.016	0.688	0.120
Ablation	0.215	0.003	0.110	0.547	0.226	0.020	0.506	0.090
LIPN–IIMAS	0.107	0.047	0.077	0.153	0.172	0.145	0.074	0.080
ECNU	**0.732**	**0.744**	**0.749**	**0.856**	**0.813**	**0.336**	**0.852**	**0.771**

6.4.4 Results on SemEval-2017

In order to compare our system's performance in itself with state-of-the-art systems, we participated in the official shared task results of SemEval-2017 (Bjerva and Östling, 2017b). The results from the official evaluation are shown in Table 6.8. 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, and worse than the best performing ECNU system (Tian et al., 2017). The fact that the system was low in the ranking list for the shared task can be explained by several factors. On the one hand, our approach was very simplistic, whereas other systems took more involved approaches. For instance, the ECNU submission first translates all sentences to English, and then use an ensemble of four deep neural network models and three feature engineered models. The features used included word alignments, summarisa-

tion and MT evaluation metrics, kernel similarities of bags of words, bags of dependencies, n-gram overlap, edit distances, length of common prefixes, suffixes, and substrings, tree kernels, and pooled word embeddings (Tian et al., 2017). In contrast, our system only uses the latter of these features. On the other hand, our system did outperform other systems in individual language tracks. Additionally, in all tracks we outperform the LIPN–IIMAS system, which approaches the task using an attentional LSTM (Arroyo-Fernández and Ruiz, 2017).

Underfitting might be an explanation for the low results obtained as compared to development, indicating that the approach we have taken is simply not sufficient to solve the task of (cross-lingual) STS well. Comparing our approach with other systems, such as the ECNU system, does indeed reveal a staggering difference in complexity.

6.4.5 Results on SemEval-2016

In order to further evaluate our system, we compare with an approach which is relatively similar to ours, namely the FBK HLT-MT submission described by Ataman et al. (2016). We replicate the training, development, and test setting used in Shared task SemEval-2016, with the exception that we only evaluate on one of the domains (Agirre et al., 2016). The results from the *news* domain (English-Spanish) are shown in Table 6.9. On this dataset, the difference between our system and that of Ataman et al. (2016) is relatively small.

Table 6.9: SemEval 2016 system comparison, comparing our single-source system with the FBK HLT-MT system (Ataman et al., 2016). The runs from FBK HLT-MT differ in the features used.

System	**Score**
FBK HLT-MT – Run 1	0.243
FBK HLT-MT – Run 2	0.244
FBK HLT-MT – Run 3	**0.255**
Our system	0.241

6.5 Conclusions

Although our system faired relatively poorly in the official results for SemEval-2017, the multilingual experiments presented in this chapter offer insights into research question (**RQ 3**). Multilingual word representations allow us to leverage a large amount of data from parallel corpora, opening up for multilingual learning of semantic textual similarity. This allows for zero-shot learning, which yielded relatively good performance on unseen target languages on the development sets under consideration. This indicates that multilingual word representations are indeed suitable for enabling zero-shot learning for STS (**RQ 3a**). As for language relatedness, we found that applying zero-shot learning, and sharing parameters, between the Indo-European languages Spanish and English was more beneficial in general than when involving the Semitic language Arabic (**RQ 3b**). Having seen that language similarities to some extent are indicative of performance in zero-shot STS, this raises the question of whether this generalises to other tasks, to what extent language similarities are important for this, and how such language similarities can be quantified. We approach this in the following chapter, where we will look at **RQ 4**.