Evaluation Datasets for Cross-lingual Semantic Textual Similarity

Tomáš Hercig[‡]
[†] Department of Computer
Science and Engineering,
University of West Bohemia,
Plzeň, Czech Republic
tigi@ntis.zcu.cz

Pavel Král[†‡]
[‡] NTIS – New Technologies
for the Information Society,
University of West Bohemia,
Plzeň, Czech Republic
pkral@kiv.zcu.cz

Abstract

Semantic textual similarity (STS) systems estimate the degree of the meaning similarity between two sentences. Cross-lingual STS systems estimate the degree of the meaning similarity between two sentences, each in a different language.

State-of-the-art algorithms usually employ a strongly supervised, resource-rich approach difficult to use for poorly-resourced languages. However, any approach needs to have evaluation data to confirm the results. In order to simplify the evaluation process for poorly-resourced languages (in terms of STS evaluation datasets), we present new datasets for cross-lingual and monolingual STS for languages without this evaluation data. We also present the results of several state-of-the-art methods on these data which can be used as a baseline for further research.

We believe that this article will not only extend the current STS research to other languages, but will also encourage competition on this new evaluation data.

1 Introduction

Recently, research in natural language understanding is moving beyond monolingual solutions (Wada et al., 2019; Conneau and Lample, 2019; Lin et al., 2019). However, any solution needs some kind of evaluation data. The common source of evaluation data for semantic meaning comparison are the STS tasks from SemEval workshop.

The STS task has a long history at SemEval workshops. Since 2012 till 2017 the STS task has been held annually creating a considerable amount of evaluation datasets. However, most of these datasets are in English and Spanish.

This has led us to the creation of new cross-lingual evaluation data for STS because all algorithms need verification e.g. on evaluation data. The evaluation data consist of pairs of sentences and the degree of their semantic similarity.

This paper presents new STS datasets and thorough experiments using linear transformations for cross-lingual STS. We see four main contributions of our work:

- We show an overview of the existing cross-lingual STS datasets.
- We present newly created datasets for both cross-lingual and monolingual STS.
- We extend experiments for previously published methods replicating the results and validating the conclusions.
- We present initial experiments and provide baseline results on the new datasets.

2 Related Work

This section presents related datasets for STS with main focus on cross-lingual datasets.

Even though most of the datasets is in English and Spanish, there are also available monolingual datasets in Arabic and Czech.

The cross-lingual datasets are also quite focused on English and usually one side of the sentence pair is in English. In summary, the cross-lingual datasets are available in English paired with Arabic, Croatian, Czech, Italian, Spanish, and Turkish. More detailed overview of these datasets can be found further in this section.

In Section 3, we present datasets for both monolingual and cross-lingual STS in Czech, English, French, and German.

- **SemEval 2012:** Agirre et al. (2012) introduced the shared task competition in English. They provided five datasets (paraphrase sentences *MSRpar*,video descriptions *MSRvid*, automatically translated sentences *MTnews* and *MTeuroparl*, and gloss pairs *OnWN*) consisting of 2234 training sentence pairs, 3108 testing sentence pairs, and 6 trial pairs. Glavaš et al. (2017) translated one side (750 sentences each) of two English monolingual datasets (*MSRvid* and *OnWN*) from SemEval 2012 task 6 to Spanish, Italian, and Croatian.

- **SemEval 2013:** Agirre et al. (2013) continued with English STS core task. The whole dataset contains 2250 test and 20 trial new sentence pairs from four datasets (news *Headlines*, mapping of lexical resources *OnWN* and *FNWN*, machine translation evaluations *SMT*). They also introduced typed similarity task for predefined types (e.g. author, location, subject and description).

- **SemEval 2014:** Agirre et al. (2014) divided the task into two subtasks one for English (3750 sentence pairs) and the other for Spanish. The English data contains image descriptions (*Images*), news headlines (*Headlines*), gloss pairs (*OnWN*), news title and tweet comments (*Tweet-news*), discussion forum and news (*Deft-forum* and *Deft-news*). The newly introduced Spanish dataset contained 480 sentence pairs from Wikipedia, 324 sentence pairs from Google News and 65 trial sentence pairs.

- **SemEval 2015:** Agirre et al. (2015) continued with both English (3000 test pairs, 70 trial pairs) and Spanish (751 pairs) subtasks and newly introduced an interpretable STS subtask. The interpretation is evaluated on the alignments of sentence pairs. The English dataset contains image descriptions (*image*), news headlines (*headlines*), student and reference answers (*answers-students*), answers from exchange forums (*answers-forum*), and discussion forum comments (*belief*)

- **SemEval 2016:** Agirre et al. (2016) proposed two Spanish-English datasets in SemEval 2016 task 1. One consists of news headlines *News* (301 sentence pairs) and the other contains sentences from multiple sources including news headlines, question-answering, plagiarism detection, etc. *Multi-source* (294 sentence pairs). Glavaš et al. (2017) translated Spanish sentences from both datasets to Italian and Croatian.

- **SemEval 2017:** Cer et al. (2017) introduced four cross-lingual datasets and three monolingual datasets. Each dataset contains 250 sentence pairs and the data source is Flickr30k image captions (Young et al., 2014), except *Track4b* which is based on data from WMT 2014 quality estimation task (Bojar et al., 2014). The cross-lingual dataset contains the following language pairs: Spanish-English (*Track4a* and *Track4b*), Arabic-English (*Track2*), and Turkish-English (*Track6*). The monolingual dataset consists of English (*Track5*), Spanish (*Track3*), and Arabic (*Track1*).

- **Czech STS:** Svoboda and Brychcín (2018) translated the English *Images* and *Headlines* parts of the dataset from SemEval 2013 - 2015 resulting in a dataset of 575 pairs of news headlines and 850 pairs of image descriptions. However, the links to the English dataset were not preserved leaving us only with a Czech monolingual dataset.

3 Dataset

	First	**Second**
Cs	1612	1576
De	2002	2010
En	2179	2186
Fr	2382	2245

Table 1: Number of tokens for each dataset.

We used the English monolingual dataset (*Track5*) from SemEval 2017 task 1 (Cer et al., 2017) to create new evaluation data for Czech, French and German.

We translated (using Google Translate) and then manually checked both sides of 250 *pairs* from the English STS dataset from SemEval 2017 (*Track5*) into Czech, French, and German.

The translations were manually checked by two upper intermediate (B2) level speakers of the given language in case of French and German and native speakers of Czech. We also asked them to preserve the meaning of the translation as much as possible in relation to the semantic similarity score.

We assume that the translated pairs preserve the same semantic similarity score. The resulting dataset thus contains monolingual and cross-lingual datasets for all language pairs.

Table 1 shows number of tokens in a dataset per side (part) of the evaluation pair.

The dataset consisting of 2000 sentences is available for research purposes at https://gitlab.com/tigi.cz/cross-lingual-sts.

Each file contains one sentence per line. The STS evaluation pair consists of the first and second part (sentence) each stored in a separate file. The semantic similarity score is located in the gold score file. The lines in all files correspond with each other. To get the evaluation data for e.g. EN-CS load the file STS.2017.input.track5.EN.first.txt and STS.2017.input.track5.CS.second.txt and the gold standard in file STS.2017.gs.track5.first-second.txt.

This new resource consists of four monolingual datasets and twelve cross-lingual datasets.

4 Experiments

We follow Brychcín (2020) who used bilingual dictionaries and a new transformation for word level semantic representations which reduces hubness in semantic spaces. He also evaluated his methods on several STS datasets including cross-lingual.

The transformations of semantic spaces and combinations of word representations are too complex and beyond the scope of this short paper, for a thorough description of these methods please refer to the original publication.

In the replicated experiments on new datasets we use monolingual semantic spaces transformed into a unified space using bilingual dictionaries. STS performance is measured by the Pearson correlation between automatically estimated scores and human judgments.

Our experiments start with building monolingual semantic spaces for each of tested languages, namely, Czech (CS), German (DE), English (EN), and French (FR). For all languages we use character-n-gram-based skip-gram model (Bojanowski et al., 2017) pre-trained on Wikipedia[1].

For each language, we construct the vocabulary from 300k most frequent words. We estimate IDF weights on the Wikipedia corpus for every language. Each Wikipedia article represents a document.

The bilingual dictionaries between each pair of languages are created from the 20k most frequent words in the corpus using Google translate.

The global post-processing techniques for semantic spaces used by Brychcín (2020) consist of two steps column-wise mean centering and word vector normalization to unit vectors. This guarantees that all word pairs in the dictionary contribute equally to the optimization criteria of the linear transformation. We always apply this post-processing for both semantic spaces before the linear mapping.

4.1 Linear Transformations

We experiment with the following five techniques for linear mapping to transform the semantic spaces. For detailed description of these methods please see (Brychcín, 2020).

- *Least Squares Transformation* (**LS**)
- *Orthogonal Transformation* (**OT**)
- *Canonical Correlation Analysis* (**CCA**)
- *Ranking Transformation* (**RT**)
- *Orthogonal Ranking Transformation* (**ORT**)

4.2 Word Combinations

Semantic textual similarity is estimated by combining word representations by *Linear Combination* (**LC**), *Principal Angles* (**PA**)[2], and *Optimal Matching* (**OM**)[3]. We evaluate both uniform weighting (for mutual comparison with original methods) and IDF weighting in all three STS approaches.

RT and ORT require special settings to work properly we use the same settings as Brychcín (2020).

4.3 Results

Table 3 shows the mean Pearson correlations for each linear transformation combined with different STS techniques on the created cross-lingual STS datasets. We can state our results support the claims by Brychcín (2020). ORT outperformed other transformations independently of STS technique. IDF weighting boosts the correlations in all cases and together with OM yields the best performance.

In Tables 2, and 4 we show correlations achieved by the best settings, i.e., OM with IDF weighting. In table 2 we compare our results with the top performing system ECNU (Tian et al., 2017) and

[1] Available at https://fasttext.cc.

[2] Using $r = 4$ as recommended by Mu et al. (2017)

[3] For detailed description see(Sultan et al., 2015; Glavaš et al., 2017; Brychcín, 2020)

	Cs-Cs	**De-De**	**En-En**	**Fr-Fr**	**Mean**
Results	0.736	0.706	0.786	0.702	0.732
Tian et al. (2017)	-	-	0.852	-	-
Cer et al. (2017)	-	-	0.728	-	-

Table 2: Individual Pearson correlations for monolingual datasets using OM with IDF weighting.

		LC	**LC IDF**	**PA**	**PA IDF**	**OM**	**OM IDF**
	Monolingual	0.544	0.659	0.695	0.715	0.691	**0.732**
Cross-lingual	**LS**	0.032	0.253	0.382	0.486	0.478	**0.554**
	CCA	0.088	0.319	0.373	0.503	0.498	**0.569**
	OT	0.140	0.361	0.416	0.524	0.511	**0.580**
	RT	0.186	0.385	0.460	0.531	0.519	**0.581**
	ORT	0.320	0.464	0.519	0.560	0.556	**0.608**

Table 3: The mean Pearson correlations over monolingual and cross-lingual datasets. The highest correlations are in bold.

	Cs-De	**Cs-En**	**Cs-Fr**	**De-En**	**De-Fr**	**En-Fr**	**Mean**
LS	0.544	0.588	0.523	0.583	0.515	0.571	0.554
CCA	0.568	0.596	0.539	0.598	0.533	0.580	0.569
OT	0.581	0.613	0.556	0.600	0.544	0.582	0.580
RT	0.564	0.605	0.565	0.607	0.551	0.596	0.581
ORT	0.591	0.630	0.586	0.629	0.583	0.631	0.608

Table 4: The mean Pearson correlations over language pairs of cross-lingual datasets using OM with IDF weighting. The result for Cs-De is the mean value of Cs-De and De-Cs.

with SemEval baseline (Cer et al., 2017). Note that the En-En is equal to the one achieved by Brychcín (2020) and thus validates our implementation.

In Table 4 we can see the mean Pearson correlations over language pairs. The worst results were achieved on De-Fr and Cs-Fr which is not surprising as they are distant language families (Slavic-Romance and Germanic-Romance). In general, French appears to be the most difficult to understand the meaning compared to other language combinations in this dataset.

The linear mapping techniques are sorted by their performance e.g. OT outperforms LS and CCA. The best performing setting is *Orthogonal Ranking Transformation* and *Optimal Matching* with IDF weighting.

5 Conclusion and Future Work

In this paper we presented new STS datasets for both cross-lingual and monolingual STS and provided them to the research community. We extended experiments of previous work on STS using linear transformations to create cross-lingual semantic spaces, by conducting initial experiments on the newly created datasets. We confirmed the findings of Brychcín (2020) by replicating three (previously published) approaches to combine information from word representations.

The used STS system does not require sentence similarity supervision and the only cross-lingual information is a bilingual dictionary. In the future, we intend to investigate the use of unsupervised methods to create the bilingual dictionary.

Acknowledgements

This work has been partly supported from ERDF ”Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)” (no.: CZ.02.1.01/0.0/0.0/17_048/0007267).

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei

Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 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, Denver, Colorado. 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, Dublin, Ireland. Association for Computational Linguistics.

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)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

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. 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. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tomáš Brychcín. 2020. Linear transformations for cross-lingual semantic textual similarity. *Knowledge-Based Systems*, 187:104819.

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)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Goran Glavaš, Marc Franco-Salvador, Simone P. Ponzetto, and Paolo Rosso. 2017. A resource-light method for cross-lingual semantic textual similarity. *Knowledge-Based Systems*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. Representing sentences as low-rank subspaces. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 629–634, Vancouver, Canada. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Lukáš Svoboda and Tomáš Brychcín. 2018. Czech dataset for semantic textual similarity. In *Text, Speech, and Dialogue*, pages 213–221, Cham. Springer International Publishing.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. Ecnu at semeval-2017 task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 191–197, Vancouver, Canada. Association for Computational Linguistics.

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3113–3124, Florence, Italy. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.