SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity

Jose Camacho-Collados*[1], Mohammad Taher Pilehvar*[2], Nigel Collier[2] and Roberto Navigli[1]

[1]Department of Computer Science, Sapienza University of Rome
[2]Department of Theoretical and Applied Linguistics, University of Cambridge
[1]{collados,navigli}@di.uniroma1.it
[2]{mp792,nhc30}@cam.ac.uk

Abstract

This paper introduces a new task on Multilingual and Cross-lingual Semantic Word Similarity which measures the semantic similarity of word pairs within and across five languages: English, Farsi, German, Italian and Spanish. High quality datasets were manually curated for the five languages with high inter-annotator agreements (consistently in the 0.9 ballpark). These were used for semi-automatic construction of ten cross-lingual datasets. 17 teams participated in the task, submitting 24 systems in subtask 1 and 14 systems in subtask 2. Results show that systems that combine statistical knowledge from text corpora, in the form of word embeddings, and external knowledge from lexical resources are best performers in both subtasks. More information can be found on the task website: http://alt.qcri.org/semeval2017/task2/.

1 Introduction

Measuring the extent to which two words are semantically similar is one of the most popular research fields in lexical semantics, with a wide range of Natural Language Processing (NLP) applications. Examples include Word Sense Disambiguation (Miller et al., 2012), Information Retrieval (Hliaoutakis et al., 2006), Machine Translation (Lavie and Denkowski, 2009), Lexical Substitution (McCarthy and Navigli, 2009), Question Answering (Mohler et al., 2011), Text Summarization (Mohammad and Hirst, 2012), and Ontology Alignment (Pilehvar and Navigli, 2014). Moreover, word similarity is generally accepted as the most direct in-vitro evaluation framework for word representation, a research field that has recently received massive research attention mainly as a result of the advancements in the use of neural networks for learning dense low-dimensional semantic representations, often referred to as word embeddings (Mikolov et al., 2013; Pennington et al., 2014). Almost any application in NLP that deals with semantics can benefit from efficient semantic representation of words (Turney and Pantel, 2010).

However, research in semantic representation has in the main focused on the English language only. This is partly due to the limited availability of word similarity benchmarks in languages other than English. Given the central role of similarity datasets in lexical semantics, and given the importance of moving beyond the barriers of the English language and developing language-independent and multilingual techniques, we felt that this was an appropriate time to conduct a task that provides a reliable framework for evaluating multilingual and cross-lingual semantic representation and similarity techniques. The task has two related subtasks: multilingual semantic similarity (Section 1.1), which focuses on representation learning for individual languages, and cross-lingual semantic similarity (Section 1.2), which provides a benchmark for multilingual research that learns unified representations for multiple languages.

1.1 Subtask 1: Multilingual Semantic Similarity

While the English community has been using standard word similarity datasets as a common evaluation benchmark, semantic representation for other languages has generally proved difficult to evaluate. A reliable multilingual word similarity benchmark can be hugely beneficial in evaluating the robustness and reliability of semantic

Authors marked with * contributed equally.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 15–26,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

representation techniques across languages. Despite this, very few word similarity datasets exist for languages other than English: The original English RG-65 (Rubenstein and Goodenough, 1965) and WordSim-353 (Finkelstein et al., 2002) datasets have been translated into other languages, either by experts (Gurevych, 2005; Joubarne and Inkpen, 2011; Granada et al., 2014; Camacho-Collados et al., 2015), or by means of crowdsourcing (Leviant and Reichart, 2015), thereby creating equivalent datasets in languages other than English. However, the existing English word similarity datasets suffer from various issues:

1. The similarity scale used for the annotation of WordSim-353 and MEN (Bruni et al., 2014) does not distinguish between similarity and relatedness, and hence conflates these two. As a result, the datasets contain pairs that are judged to be highly similar even if they are not of similar type or nature. For instance, the WordSim-353 dataset contains the pairs *weather-forecast* or *clothes-closet* with assigned similarity scores of 8.34 and 8.00 (on the [0,10] scale), respectively. Clearly, the words in the two pairs are (highly) related, but they are not similar.

2. The performance of state-of-the-art systems have already surpassed the levels of human inter-annotator agreement (IAA) for many of the old datasets, e.g., for RG-65 and WordSim-353. This makes these datasets unreliable benchmarks for the evaluation of newly-developed systems.

3. Conventional datasets such as RG-65, MC-30 (Miller and Charles, 1991), and WS-Sim (Agirre et al., 2009) (the similarity portion of WordSim-353) are relatively small, containing 65, 30, and 200 word pairs, respectively. Hence, these benchmarks do not allow reliable conclusions to be drawn, since performance improvements have to be large to be statistically significant (Batchkarov et al., 2016).

4. The recent SimLex-999 dataset (Hill et al., 2015) improves both the size and consistency issues of the conventional datasets by providing word similarity scores for 999 word pairs on a consistent scale that focuses on similarity only (and not relatedness). However, the dataset suffers from other issues. First, given that SimLex-999 has been annotated by turkers, and not by human experts, the similarity scores assigned to individual word pairs have a high variance, resulting in relatively low IAA (Camacho-Collados and Navigli, 2016). In fact, the reported IAA for this dataset is 0.67 in terms of average pairwise correlation, which is considerably lower than conventional expert-based datasets whose IAA are generally above 0.80 (Rubenstein and Goodenough, 1965; Camacho-Collados et al., 2015). Second, similarly to many of the above-mentioned datasets, SimLex-999 does not contain named entities (e.g., *Microsoft*), or multiword expressions (e.g., *black hole*). In fact, the dataset includes only words that are defined in WordNet's vocabulary (Miller et al., 1990), and therefore lacks the ability to test the reliability of systems for WordNet out-of-vocabulary words. Third, the dataset contains a large number of antonymy pairs. Indeed, several recent works have shown how significant performance improvements can be obtained on this dataset by simply tweaking usual word embedding approaches to handle antonymy (Schwartz et al., 2015; Pham et al., 2015; Nguyen et al., 2016).

Since most existing multilingual word similarity datasets are constructed on the basis of conventional English datasets, any issues associated with the latter tend simply to be transferred to the former. This is the reason why we proposed this task and constructed new challenging datasets for five different languages (i.e., English, Farsi, German, Italian, and Spanish) addressing all the above-mentioned issues. Given that multiple large and high-quality verb similarity datasets have been created in recent years (Yang and Powers, 2006; Baker et al., 2014; Gerz et al., 2016), we decided to focus on nominal words.

1.2 Subtask 2: Cross-lingual Semantic Similarity

Over the past few years multilingual embeddings that represent lexical items from multiple languages in a unified semantic space have garnered considerable research attention (Zou et al., 2013; de Melo, 2015; Vulić and Moens, 2016; Ammar et al., 2016; Upadhyay et al., 2016), while at the same time cross-lingual applications have also been increasingly studied (Xiao and Guo, 2014; Franco-Salvador et al., 2016). However, there have been very few reliable datasets for evaluating cross-lingual systems. Similarly to the case of multilingual datasets, these cross-lingual datasets have been constructed on the basis of conventional English word similarity datasets: MC-30 and WordSim-353 (Hassan and Mihalcea, 2009), and RG-65 (Camacho-Collados et al., 2015). As a result, they inherit the issues affecting their parent datasets mentioned in the previous subsection: while MC-30 and RG-65 are composed of only 30 and 65 pairs, WordSim-353 conflates similarity and relatedness in different languages. Moreover, the datasets of Hassan and Mihalcea (2009) were not re-scored after having been translated to the other languages, thus ignoring possible semantic shifts across languages and producing unreliable scores for many translated word pairs.

For this subtask we provided ten high quality cross-lingual datasets, constructed according to the procedure of Camacho-Collados et al. (2015), in a semi-automatic manner exploiting the monolingual datasets of subtask 1. These datasets constitute a reliable evaluation framework across five languages.

2 Task Data

Subtask 1, i.e., multilingual semantic similarity, has five datasets for the five languages of the task, i.e., English, Farsi, German, Italian, and Spanish. These datasets were manually created with the help of trained annotators (as opposed to Mechanical Turk) that were native or fluent speakers of the target language. Based on these five datasets, 10 cross-lingual datasets were automatically generated (described in Section 2.2) for subtask 2, i.e., cross-lingual semantic similarity.

In this section we focus on the creation of the evaluation test sets. We additionally created a set of small trial datasets by following a similar process. These datasets were used by some participants during system development.

2.1 Monolingual datasets

As for monolingual datasets, we opted for a size of 500 word pairs in order to provide a large enough set to allow reliable evaluation and comparison of the systems. The following procedure was used for the construction of multilingual datasets: (1) we first collected 500 English word pairs from a wide range of domains (Section 2.1.1), (2) through translation of these pairs, we obtained word pairs for the other four languages (Section 2.1.2) and, (3) all word pairs of each dataset were manually scored by multiple annotators (Section 2.1.3).

Animals	Language and linguistics
Art, architecture and archaeology	Law and crime
Biology	Literature and theatre
Business, economics, and finance	Mathematics
Chemistry and mineralogy	Media
Computing	Meteorology
Culture and society	Music
Education	Numismatics and currencies
Engineering and technology	Philosophy and psychology
Farming	Physics and astronomy
Food and drink	Politics and government
Games and video games	Religion, mysticism and mythology
Geography and places	Royalty and nobility
Geology and geophysics	Sport and recreation
Health and medicine	Textile and clothing
Heraldry, honors, and vexillology	Transport and travel
History	Warfare and defense

Table 1: The set of thirty-four domains.

2.1.1 English dataset creation

Seed set selection. The dataset creation started with the selection of 500 English words. One of the main objectives of the task was to provide an evaluation framework that contains named entities and multiword expressions and covers a wide range of domains. To achieve this, we considered the 34 different domains available in BabelDomains[1] (Camacho-Collados and Navigli, 2017), which in the main correspond to the domains of the *Wikipedia featured articles page*[2]. Table 1 shows the list of all the 34 domains used for the creation of the datasets. From each domain, 12 words were sampled in such a way as to have at least one multiword expression and two named entities. In order to include words that may not belong to any of the pre-defined domains, we added 92 extra words whose domain was not decided beforehand. We also tried to sample these seed words in such a way as to have a balanced set across occurrence frequency.[3] Of the 500 English seed words, 84 (17%) and 83 were, respectively, named entities and multiwords.

Similarity scale. For the annotation of the datasets, we adopted the five-point Likert scale of the SemEval-2014 task on Cross-Level Semantic

[1] http://lcl.uniroma1.it/babeldomains/

[2] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

[3] We used the Wikipedia corpus for word frequency calculation during the dataset construction.