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.

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.

4	Very similar	The two words are synonyms (e.g., *midday-noon* or *motherboard-mainboard*).
3	Similar	The two words share many of the important ideas of their meaning but include slightly different details. They refer to similar but not identical concepts (e.g., *lion-zebra* or *firefighter-policeman*).
2	Slightly similar	The two words do not have a very similar meaning, but share a common topic/domain/function and ideas or concepts that are related (e.g., *house-window* or *airplane-pilot*).
1	Dissimilar	The two words describe clearly dissimilar concepts, but may share some small details, a far relationship or a domain in common and might be likely to be found together in a longer document on the same topic (e.g., *software-keyboard* or *driver-suspension*).
0	Totally dissimilar and unrelated	The two words do not mean the same thing and are not on the same topic (e.g., *pencil-frog* or *PlayStation-monarchy*).

Table 2: The five-point Likert scale used to rate the similarity of item pairs. See Table 4 for examples.

Similarity (Jurgens et al., 2014) which was designed to systematically order a broad range of semantic relations: synonymy, similarity, relatedness, topical association, and unrelatedness. Table 2 describes the five points in the similarity scale along with example word pairs.

Pairing word selection. Having the initial 500-word seed set at hand, we selected a pair for each word. The selection was carried out in such a way as to ensure a uniform distribution of pairs across the similarity scale. In order to do this, we first assigned a random intended similarity to each pair. The annotator then had to pick the second word so as to match the intended score. In order to allow the annotator to have a broader range of candidate words, the intended score was considered as a similarity interval, one of [0-1], [1-2], [2-3] and [3,4]. For instance, if the first word was *helicopter* and the presumed similarity was [3-4], the annotator had to pick a pairing word which was "semantically similar" (see Table 2) to *helicopter*, e.g., *plane*. Of the 500 pairing words, 45 (9%) and 71 (14%) were named entities and multiwords, respectively. This resulted in an English dataset comprising 500 word pairs, 105 (21%) and 112 (22%) of which have at least one named entity and multiword, respectively.

2.1.2 Dataset translation

The remaining four multilingual datasets (i.e., Farsi, German, Italian, and Spanish) were constructed by translating words in the English dataset to the target language. We had two goals in mind while selecting translation as the construction strategy of these datasets (as opposed to independent word samplings per language): (1) to have comparable datasets across languages in terms of domain coverage, multiword and named entity distribution[4] and (2) to enable an automatic construction of cross-lingual datasets (see Section 2.2).

Each English word pair was translated by two independent annotators. In the case of disagreement, a third annotator was asked to pick the preferred translation. While translating, the annotators were shown the word pair along with their initial similarity score, which was provided to help them in selecting the correct translation for the intended meanings of the words.

2.1.3 Scoring

The annotators were instructed to follow the guidelines, with special emphasis on distinguishing between similarity and relatedness. Furthermore, although the similarity scale was originally designed as a Likert scale, annotators were given flexibility to assign values between the defined points in the scale (with a step size of 0.25), indicating a blend of two relations. As a result of this procedure, we obtained 500 word pairs for each of the five languages. The pairs in each language were shuffled and their initial scores were discarded. Three annotators were then asked to assign a similarity score to each pair according to our similarity scale (see Section 2.1.1).

Table 3 (first row) reports the average pairwise Pearson correlation among the three annotators for each of the five languages. Given the fact that our word pairs spanned a wide range of domains, and that there was a possibility for annotators to misunderstand some words, we devised a procedure to check the quality of the annotations and to improve the reliability of the similarity scores. To this end, for each dataset and for each annotator

[4] Apart from the German dataset in which the proportion of multiwords significantly reduces (from 22% of English to around 11%) due to the compounding nature of the German language, other datasets maintain similar proportions of multiwords to those of the English dataset.

	English	Farsi	German	Italian	Spanish
Initial scores	0.836	0.839	0.864	0.798	0.829
Revised scores	0.893	0.906	0.916	0.900	0.890

Table 3: Average pairwise Pearson correlation among annotators for the five monolingual datasets.

MONOLINGUAL			
DE	Tuberkulose	LED	0.25
ES	zumo	batido	3.00
EN	Multiple Sclerosis	MS	4.00
IT	Nazioni Unite	Ban Ki-moon	2.25
FA	شام آخر	لئوناردو دا وینچی	2.08
CROSS-LINGUAL			
DE-ES	Sessel	taburete	3.08
DE-FA	Lawine	برف	2.25
DE-IT	Taifun	ciclone	3.46
EN-DE	pancreatic cancer	Chemotherapie	1.75
EN-ES	Jupiter	Mercurio	3.25
EN-FA	film	پوچ‌گرایی	0.25
EN-IT	island	penisola	3.08
ES-FA	duna	بیابان	2.25
ES-IT	estrella	pianeta	2.83
IT-FA	avvocato	نمایشگر	0.08

Table 4: Example pairs and their ratings (EN: English, DE: German, ES: Spanish, IT: Italian, FA: Farsi).

we picked the subset of pairs for which the difference between the assigned similarity score and the average of the other two annotations was more than 1.0, according to our similarity scale. The annotator was then asked to revise this subset performing a more careful investigation of the possible meanings of the word pairs contained therein, and change the score if necessary. This procedure resulted in considerable improvements in the consistency of the scores. The second row in Table 3 ("Revised scores") shows the average pairwise Pearson correlation among the three revised sets of scores for each of the five languages. The inter-annotator agreement for all the datasets is consistently in the 0.9 ballpark, which demonstrates the high quality of our multilingual datasets thanks to careful annotation of word pairs by experts.

2.2 Cross-lingual datasets

The cross-lingual datasets were automatically created on the basis of the translations obtained with the method described in Section 2.1.2 and using the approach of Camacho-Collados et al. (2015).[5] By intersecting two aligned translated pairs across two languages (e.g., *mind-brain* in English and *mente-cerebro* in Spanish), the approach creates two cross-lingual pairs between the two languages (*mind-cerebro* and *brain-mente* in the example). The similarity scores for the constructed cross-lingual pairs are computed as the average of the corresponding language-specific scores in the monolingual datasets. In order to avoid semantic shifts between languages interfering in the process, these pairs are only created if the difference between the corresponding language-specific scores is lower than 1.0. The full details of the algorithm can be found in Camacho-Collados et al. (2015). The approach has been validated by human judges and shown to achieve agreements of around 0.90 with human judges, which is similar to inter-annotator agreements reported in Section 2.1.3. See Table 4 for some sample pairs in all monolingual and cross-lingual datasets. Table 5 shows the final number of pairs for each language pair.

[5] http://lcl.uniroma1.it/similarity-datasets/

	EN	DE	ES	IT	FA
EN	500	914	978	970	952
DE	-	500	956	912	888
ES	-	-	500	967	967
IT	-	-	-	500	916
FA	-	-	-	-	500

Table 5: Number of word pairs in each dataset. The cells in the main diagonal of the table (e.g., EN-EN) correspond the monolingual datasets of subtask 1.

3 Evaluation

We carried out the evaluation on the datasets described in the previous section. The experimental setting is described in Section 3.1 and the results are presented in Section 3.2.

3.1 Experimental setting

3.1.1 Evaluation measures and official scores

Participating systems were evaluated according to standard Pearson and Spearman correlation mea-

sures on all word similarity datasets, with the final official score being calculated as the harmonic mean of Pearson and Spearman correlations (Jurgens et al., 2014). Systems were allowed to participate in either multilingual word similarity, cross-lingual word similarity, or both. Each participating system was allowed to submit a maximum of two runs.

For the multilingual word similarity subtask, some systems were multilingual (applicable to different languages), whereas others were monolingual (only applicable to a single language). While monolingual approaches were evaluated in their respective languages, multilingual and language-independent approaches were additionally given a global ranking provided that they tested their systems on at least four languages. The final score of a system was calculated as the average harmonic mean of Pearson and Spearman correlations of the four languages on which it performed best.

Likewise, the participating systems of the cross-lingual semantic similarity subtask were allowed to provide a score for a single cross-lingual dataset, but must have provided results for at least six cross-lingual word similarity datasets in order to be considered for the final ranking. For each system, the global score was computed as the average harmonic mean of Pearson and Spearman correlation on the six cross-lingual datasets on which it provided the best performance.

3.1.2 Shared training corpus

We encouraged the participants to use a shared text corpus for the training of their systems. The use of the shared corpus was intended to mitigate the influence that the underlying training corpus might have upon the quality of obtained representations, laying a common ground for a fair comparison of the systems.

- **Subtask 1.** The common corpus for subtask 1 was the Wikipedia corpus of the target language. Specifically, systems made use of the Wikipedia dumps released by Al-Rfou et al. (2013).[6]

- **Subtask 2.** The common corpus for subtask 2 was the Europarl parallel corpus[7]. This corpus is available for all languages except Farsi. For pairs involving Farsi, participants were allowed to use the OpenSubtitles2016 parallel corpora[8]. Additionally, we proposed a second type of multilingual corpus to allow the use of different techniques exploiting comparable corpora. To this end, some participants made use of Wikipedia.

3.1.3 Participating systems

This task was targeted at evaluating multilingual and cross-lingual word similarity measurement techniques. However, it was not only limited to this area of research, as other fields such as semantic representation consider word similarity as one of their most direct benchmarks for evaluation. All kinds of semantic representation techniques and semantic similarity systems were encouraged to participate.

In the end we received a wide variety of participants: proposing distributional semantic models learnt directly from raw corpora, using syntactic features, exploiting knowledge from lexical resources, and hybrid approaches combining corpus-based and knowledge-based clues. Due to lack of space we cannot describe all the systems in detail, but we recommend the reader to refer to the system description papers for more information about the individual systems: OoO (Xu et al., 2017), HCCL (He et al., 2017), Citius (Gamallo, 2017), jmp8 (Melka and Bernard, 2017), l2f (Fialho et al., 2017), QLUT (Meng et al., 2017), RUFINO (Jimenez et al., 2017), MERALI (Mensa et al., 2017), Luminoso (Speer and Lowry-Duda, 2017), hhu (Qasemizadeh and Kallmeyer, 2017), Mahtab (Ranjbar et al., 2017), SEW (Bovi and Raganato, 2017) and Wild_Devs (Rotari et al., 2017).

3.1.4 Baseline

As the baseline system we included the results of the concept and entity embeddings of NASARI (Camacho-Collados et al., 2016). These embeddings were obtained by exploiting knowledge from Wikipedia and WordNet coupled with general domain corpus-based Word2Vec embeddings (Mikolov et al., 2013). We performed the evaluation with the 300-dimensional English embedded vectors (version 3.0)[9] and used them for all languages. For the comparison within and

[6] https://sites.google.com/site/rmyeid/projects/polyglot

[7] http://opus.lingfil.uu.se/Europarl.php

[8] http://opus.lingfil.uu.se/OpenSubtitles2016.php

[9] http://lcl.uniroma1.it/nasari/

System	English			Farsi			German			Italian			Spanish		
	r	ρ	Final	r	ρ	Final	r	ρ	Final	r	ρ	Final	r	ρ	Final
Luminoso_run2	**0.78**	**0.80**	**0.79**	0.51	0.50	0.50	**0.70**	**0.70**	**0.70**	**0.73**	**0.75**	**0.74**	**0.73**	**0.75**	**0.74**
Luminoso_run1	**0.78**	0.79	**0.79**	0.51	0.50	0.50	0.69	0.69	0.69	**0.73**	**0.75**	**0.74**	**0.73**	**0.75**	**0.74**
QLUT_run1*	**0.78**	0.78	0.78	-	-	-	-	-	-	-	-	-	-	-	-
hhu_run1*	0.71	0.70	0.70	0.54	0.59	0.56	-	-	-	-	-	-	-	-	-
HCCL_run1*	0.68	0.70	0.69	0.42	0.45	0.44	0.58	0.61	0.59	0.63	0.67	0.65	0.69	0.72	0.70
NASARI (baseline)	0.68	0.68	0.68	0.41	0.40	0.41	0.51	0.51	0.51	0.60	0.59	0.60	0.60	0.60	0.60
hhu_run2*	0.66	0.70	0.68	0.61	0.60	0.60	-	-	-	-	-	-	-	-	-
QLUT_run2*	0.67	0.67	0.67	-	-	-	-	-	-		-	-	-	-	-
RUFINO_run1*	0.65	0.66	0.66	0.38	0.34	0.36	0.54	0.54	0.54	0.48	0.47	0.48	0.53	0.57	0.55
Citius_run2	0.60	0.71	0.65	-	-	-	-	-	-	-	-	-	0.44	0.64	0.52
l2f_run2 (a.d.)	0.64	0.65	0.65	-	-	-	-	-	-	-	-	-	-	-	-
l2f_run1 (a.d.)	0.64	0.65	0.64	-	-	-	-	-	-	-	-	-	-	-	-
Citius_run1*	0.57	0.65	0.61	-	-	-	-	-	-	-	-	-	0.44	0.63	0.51
MERALI_run1*	0.59	0.60	0.59	-	-	-	-	-	-	-	-	-	-	-	-
Amateur_run1*	0.58	0.59	0.59	-	-	-	-	-	-	-	-	-	-	-	-
Amateur_run2*	0.58	0.59	0.59	-	-	-	-	-	-	-	-	-	-	-	-
MERALI_run2*	0.57	0.58	0.58	-	-	-	-	-	-	-	-	-	-	-	-
SEW_run2 (a.d.)	0.56	0.58	0.57	0.38	0.40	0.39	0.45	0.45	0.45	0.57	0.57	0.57	0.61	0.62	0.62
jmp8_run1*	0.47	0.69	0.56	-	-	-	0.26	0.51	0.35	0.41	0.64	0.50	-	-	-
Wild_Devs_run1	0.46	0.48	0.47	-	-	-	-	-	-	-	-	-	-	-	-
RUFINO_run2*	0.39	0.40	0.39	0.25	0.26	0.26	0.38	0.36	0.37	0.30	0.31	0.31	0.40	0.41	0.41
SEW_run1	0.37	0.41	0.39	0.38	0.40	0.39	0.45	0.45	0.45	0.57	0.57	0.57	0.61	0.62	0.62
hjpwhuer_run1	-0.04	-0.03	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.05	0.05	0.05	-0.06	-0.06	0.00
Mahtab_run2*	-	-	-	**0.72**	**0.71**	**0.71**	-	-	-	-	-	-	-	-	-
Mahtab_run1*	-	-	-	**0.72**	**0.71**	**0.71**	-	-	-	-	-	-	-	-	-

Table 6: Pearson (r), Spearman (ρ) and official (Final) results of participating systems on the five monolingual word similarity datasets (subtask 1).

across languages NASARI relies on the lexicalizations provided by BabelNet (Navigli and Ponzetto, 2012) for the concepts and entities in each language. Then, the final score was computed through the conventional closest senses strategy (Resnik, 1995; Budanitsky and Hirst, 2006), using cosine similarity as the comparison measure.

3.2 Results

We present the results of subtask 1 in Section 3.2.1 and subtask 2 in Section 3.2.2.

3.2.1 Subtask 1

Table 6 lists the results on all monolingual datasets.[10] The systems which made use of the shared Wikipedia corpus are marked with * in Table 6. Luminoso achieved the best results in all languages except Farsi. Luminoso couples word embeddings with knowledge from ConceptNet (Speer et al., 2017) using an extension of Retrofitting (Faruqui et al., 2015), which proved highly effective. This system additionally proposed two fallback strategies to handle out-of-vocabulary (OOV) instances based on loanwords and cognates. These two fallback strategies proved essential given the amount of rare words or domain-specific words which were present in the datasets. In fact, most systems fail to provide scores for all pairs in the datasets, with OOV rates close to 10% in some cases.

The combination of corpus-based and knowledge-based features was not unique to

[10] Systems followed by (a.d.) submitted their results after the official deadline.

System	Score	Official Rank
Luminoso_run2	0.743	1
Luminoso_run1	0.740	2
HCCL_run1*	0.658	3
NASARI (baseline)	0.598	-
RUFINO_run1*	0.555	4
SEW_run2 (a.d.)	0.552	-
SEW_run1	0.506	5
RUFINO_run2*	0.369	6
hjpwhuer_run1	0.018	7

Table 7: Global results of participating systems on subtask 1 (multilingual word similarity).

System	German-Spanish			German-Farsi			German-Italian			English-German			English-Spanish		
	r	ρ	Final	r	ρ	Final	r	ρ	Final	r	ρ	Final	r	ρ	Final
Luminoso_run2	**0.72**	**0.74**	**0.73**	**0.59**	**0.59**	**0.59**	**0.74**	**0.75**	**0.74**	**0.76**	**0.77**	**0.76**	**0.75**	**0.77**	**0.76**
Luminoso_run1	**0.72**	0.73	0.72	**0.59**	**0.59**	**0.59**	0.73	0.74	0.73	0.75	**0.77**	**0.76**	**0.75**	**0.77**	**0.76**
NASARI (baseline)	0.55	0.55	0.55	0.46	0.45	0.46	0.56	0.56	0.56	0.60	0.59	0.60	0.64	0.63	0.63
OoO_run1	0.54	0.56	0.55	-	-	-	0.54	0.55	0.55	0.56	0.58	0.57	0.58	0.59	0.58
SEW_run2 (a.d.)	0.52	0.54	0.53	0.42	0.44	0.43	0.52	0.52	0.52	0.50	0.53	0.51	0.59	0.60	0.59
SEW_run1	0.52	0.54	0.53	0.42	0.44	0.43	0.52	0.52	0.52	0.46	0.47	0.46	0.50	0.51	0.50
HCCL_run2* (a.d.)	0.42	0.39	0.41	0.33	0.28	0.30	0.38	0.34	0.36	0.49	0.48	0.48	0.55	0.56	0.55
RUFINO_run1†	0.31	0.32	0.32	0.23	0.25	0.24	0.32	0.33	0.33	0.33	0.34	0.33	0.34	0.34	0.34
RUFINO_run2†	0.30	0.30	0.30	0.26	0.27	0.27	0.22	0.24	0.23	0.30	0.30	0.30	0.34	0.33	0.34
hjpwhu_run2	0.05	0.05	0.05	0.01	0.01	0.01	0.06	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04
hjpwhu_run1	0.05	0.05	0.05	0.01	0.01	0.01	0.06	0.05	0.05	-0.01	-0.01	0.00	0.04	0.04	0.04
HCCL_run1*	0.03	0.02	0.02	0.03	0.02	0.02	0.03	-0.01	0.00	0.34	0.28	0.31	0.10	0.08	0.09
UniBuc-Sem_run1*	–	–	–	-	-	-	-	-	-	0.05	0.06	0.06	0.08	0.10	0.09
Citius_run1†	–	–	–	-	-	-	-	-	-	-	-	-	0.57	0.59	0.58
Citius_run2†	–	–	–	-	-	-	-	-	-	-	-	-	0.56	0.58	0.57

System	English-Farsi			English-Italian			Spanish-Farsi			Spanish-Italian			Italian-Farsi		
	r	ρ	Final	r	ρ	Final	r	ρ	Final	r	ρ	Final	r	ρ	Final
Luminoso_run2	**0.60**	**0.59**	**0.60**	**0.77**	**0.79**	**0.78**	**0.62**	**0.63**	**0.63**	**0.74**	**0.77**	**0.75**	**0.60**	**0.61**	**0.60**
Luminoso_run1	**0.60**	**0.59**	**0.60**	0.76	0.78	0.77	**0.62**	**0.63**	**0.63**	**0.74**	0.76	**0.75**	**0.60**	0.60	**0.60**
hhu_run1	0.49	0.54	0.51	-	-	-	-	-	-	-	-	-	-	-	-
NASARI (baseline)	0.52	0.49	0.51	0.65	0.65	0.65	0.49	0.47	0.48	0.60	0.59	0.60	0.50	0.48	0.49
hhu_run2	0.43	0.58	0.49	-	-	-	-	-	-	-	-	-	-	-	-
SEW_run2 (a.d.)	0.46	0.49	0.48	0.58	0.60	0.59	0.50	0.53	0.52	0.59	0.60	0.60	0.48	0.50	0.49
HCCL_run2* (a.d.)	0.44	0.42	0.43	0.50	0.49	0.49	0.37	0.33	0.35	0.43	0.41	0.42	0.33	0.28	0.30
SEW_run1	0.41	0.43	0.42	0.52	0.53	0.53	0.50	0.53	0.52	0.59	0.60	0.60	0.48	0.50	0.49
RUFINO_run2†	0.37	0.37	0.37	0.24	0.23	0.24	0.30	0.30	0.30	0.28	0.29	0.29	0.21	0.21	0.21
RUFINO_run1†	0.26	0.25	0.25	0.34	0.34	0.34	0.25	0.26	0.26	0.35	0.36	0.36	0.25	0.25	0.25
HCCL_run1*	0.02	0.01	0.01	0.12	0.07	0.09	0.05	0.05	0.05	0.08	0.06	0.06	0.02	0.00	0.00
hjpwhu_run1	0.00	-0.01	0.00	-0.05	-0.05	0.00	0.01	0.00	0.01	0.03	0.03	0.03	0.02	0.02	0.02
hjpwhu_run2	0.00	-0.01	0.00	-0.05	-0.05	0.00	0.01	0.00	0.01	0.03	0.03	0.03	0.02	0.02	0.02
OoO_run1	-	-	-	0.58	0.59	0.58	-	-	-	0.57	0.57	0.57	-	-	-
UniBuc-Sem_run1*	-	-	-	0.08	0.10	0.09	-	-	-	-	-	-	-	-	-

Table 8: Pearson (r), Spearman (ρ) and the official (Final) results of participating systems on the ten cross-lingual word similarity datasets (subtask 2).

Luminoso. In fact, most top performing systems combined these two sources of information. For Farsi, the best performing system was Mahtab, which couples information from Word2Vec word embeddings (Mikolov et al., 2013) and knowledge resources, in this case FarsNet (Shamsfard et al., 2010) and BabelNet. For English, the only system that came close to Luminoso was QLUT, which was the best-performing system that made use of the shared Wikipedia corpus for training. The best configuration of this system exploits the Skip-Gram model of Word2Vec with an additive compositional function for computing the similarity of multiwords. However, Mahtab and QLUT only performed their experiments in a single language (Farsi and English, respectively).

For the systems that performed experiments in at least four of the five languages we computed a global score (see Section 3.1.1). Global rankings and results are displayed in Table 7. Luminoso clearly achieves the best overall results. The second-best performing system was HCCL, which also managed to outperform the baseline. HCCL exploited the Skip-Gram model of Word2Vec and performed hyperparameter tuning on existing word similarity datasets. This system did not make use of external resources apart from the shared Wikipedia corpus for training. RUFINO, which also made use of the Wikipedia corpus only, attained the third overall position. The system exploits PMI and an association measure to capture second-order relations between words based on the Jaccard distance (Jimenez et al., 2016).

3.2.2 Subtask 2

The results for all ten cross-lingual datasets are shown in Table 8. Systems that made use of the shared Europarl parallel corpus are marked with * in the table, while systems making use of

System	Score	Official Rank
Luminoso_run2	0.754	1
Luminoso_run1	0.750	2
NASARI (baseline)	0.598	-
OoO_run1*	0.567	3
SEW_run2 (a.d.)	0.558	-
SEW_run1	0.532	4
HCCL_run2* (a.d.)	0.464	-
RUFINO_run1†	0.336	5
RUFINO_run2†	0.317	6
HCCL_run1*	0.103	7
hjpwhu_run2	0.039	8
hjpwhu_run1	0.034	9

Table 9: Global results of participating systems in subtask 2 (cross-lingual word similarity).

Wikipedia are marked with †. Luminoso, the best-performing system in Subtask 1, also achieved the best overall results on the ten cross-lingual datasets. This shows that the combination of knowledge from word embeddings and the ConceptNet graph is equally effective in the cross-lingual setting.

The global ranking for this subtask was computed by averaging the results of the six datasets on which each system performed best. The global rankings are displayed in Table 9. Luminoso was the only system outperforming the baseline, achieving the best overall results. OoO achieved the second best overall performance using an extension of the Bilingual Bag-of-Words without Alignments (BilBOWA) approach of Gouws et al. (2015) on the shared Europarl corpus. The third overall system was SEW, which leveraged Wikipedia-based concept vectors (Raganato et al., 2016) and pre-trained word embeddings for learning language-independent concept embeddings.

4 Conclusion

In this paper we have presented the SemEval 2017 task on *Multilingual and Cross-lingual Semantic Word Similarity*. We provided a reliable framework to measure the similarity between nominal instances within and across five different languages (English, Farsi, German, Italian, and Spanish). We hope this framework will contribute to the development of distributional semantics in general and for languages other than English in particular, with a special emphasis on multilingual and cross-lingual approaches. All evaluation datasets are available for download at http://alt.qcri.org/semeval2017/task2/.

The best overall system in both tasks was Luminoso, which is a hybrid system that effectively integrates word embeddings and information from knowledge resources. In general, this combination proved effective in this task, as most other top systems somehow combined knowledge from text corpora and lexical resources.

Acknowledgments

The authors gratefully acknowledge the support of the MRC grant No. MR/M025160/1 for PheneBank and ERC Starting Grant MultiJEDI No. 259234. Jose Camacho-Collados is supported by a Google Doctoral Fellowship in Natural Language Processing.

We would also like to thank Ángela Collados Ais, Claudio Delli Bovi, Afsaneh Hojjat, Ignacio Iacobacci, Tommaso Pasini, Valentina Pyatkin, Alessandro Raganato, Zahra Pilehvar, Milan Gritta and Sabine Ullrich for their help in the construction of the datasets. Finally, we also thank Jim McManus for his suggestions on the manuscript and the anonymous reviewers for their helpful comments.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL*. pages 19–27.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria, pages 183–192.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925* .

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of EMNLP*. pages 278–289.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*. Berlin, Germany, pages 7–12.

Claudio Delli Bovi and Alessandro Raganato. 2017. Sew-Embed at SemEval-2017 Task 2: Language-Independent Concept Representations from a Semantically Enriched Wikipedia. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)* 49(1-47).

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of Lexical Semantic Relatedness. *Computational Linguistics* 32(1):13–47.

José Camacho-Collados and Roberto Navigli. 2016. Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*. Berlin, Germany, pages 43–50.

Jose Camacho-Collados and Roberto Navigli. 2017. BabelDomains: Large-Scale Domain Labeling of Lexical Resources. In *Proceedings of EACL (2)*. Valencia, Spain, pages 223–228.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets. In *Proceedings of ACL (2)*. Beijing, China, pages 1–7.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence* 240:36–64.

Gerard de Melo. 2015. Wiktionary-based word embeddings. *Proceedings of MT Summit XV* pages 346–359.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*. pages 1606–1615.

Pedro Fialho, Hugo Patinho Rodrigues, Lusa Coheur, and Paulo Quaresma. 2017. L2F/INESC-ID at SemEval-2017 Tasks 1 and 2: Lexical and semantic features in word and textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Lev Finkelstein, Gabrilovich Evgenly, Matias Yossi, Rivlin Ehud, Solan Zach, Wolfman Gadi, and Ruppin Eytan. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems* 20(1):116–131.

Marc Franco-Salvador, Paolo Rosso, and Manuel Montes-y Gómez. 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management* 52(4):550–570.

Pablo Gamallo. 2017. Citius at SemEval-2017 Task 2: Cross-Lingual Similarity from Comparable Corpora and Dependency-Based Contexts. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*. Austin, USA.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. pages 748–756.

Roger Granada, Cassia Trojahn, and Renata Vieira. 2014. Comparing semantic relatedness between word pairs in Portuguese using Wikipedia. In *Computational Processing of the Portuguese Language*, Springer, pages 170–175.

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Natural Language Processing–IJCNLP 2005*, Springer, pages 767–778.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of EMNLP*. pages 1192–1201.

Junqing He, Long Wu, Xuemin Zhao, Yonghong Yan, and Yonghong Yan. 2017. HCCL at SemEval-2017 Task 2: Combining Multilingual Word Embeddings and Transliteration Model for Semantic Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* .

Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides GM Petrakis, and Evangelos Milios. 2006. Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems* 2(3):55–73.

Sergio Jimenez, George Dueñas, Lorena Gaitan, and Jorge Segura. 2017. RUFINO at SemEval-2017 Task 2: Cross-lingual lexical similarity by extending PMI and word embeddings systems with a Swadesh's-like list. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Sergio Jimenez, Fabio A. Gonzalez, and Alexander Gelbukh. 2016. Mathematical properties of soft cardinality: Enhancing jaccard, dice and cosine similarity measures with element-wise distance. *Information Sciences* 367:373–389.

Colette Joubarne and Diana Inkpen. 2011. Comparison of semantic similarity for different languages using the Google n-gram corpus and second-order co-occurrence measures. In *Advances in Artificial Intelligence*, Springer, pages 216–221.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. *SemEval 2014* pages 17–26.

Alon Lavie and Michael J. Denkowski. 2009. The Meteor metric for automatic evaluation of Machine Translation. *Machine Translation* 23(2-3):105–115.

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR, abs/1508.00106* .

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation* 43(2):139–159.

Josu Melka and Gilles Bernard. 2017. Jmp8 at SemEval-2017 Task 2: A simple and general distributional approach to estimate word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Fanqing Meng, Wenpeng Lu, Yuteng Zhang, Ping Jian, Shumin Shi, and Heyan Huang. 2017. QLUT at SemEval-2017 Task 2: Word Similarity Based on Word Embedding and Knowledge Base. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. 2017. MERALI at SemEval-2017 Task 2 Subtask 1: a Cognitively Inspired approach. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an online lexical database. *International Journal of Lexicography* 3(4):235–244.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1):1–28.

Tristan Miller, Chris Biemann, Torsten Zesch, and Iryna Gurevych. 2012. Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of COLING*. pages 1781–1796.

Saif Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *CoRR* abs/1203.1858. http://arxiv.org/abs/1203.1858.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Portland, Oregon, HLT'11, pages 752–762.

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.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proc. of ACL*. pages 454–459.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*. pages 1532–1543.

Nghia The Pham, Angeliki Lazaridou, and Marco Baroni. 2015. A multitask objective to inject lexical contrast into distributional semantics. In *Proceedings of ACL*. pages 21–26.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of ACL*. pages 468–478.

Behrang Qasemizadeh and Laura Kallmeyer. 2017. HHU at SemEval-2017 Task 2: Fast Hash-Based Embeddings for Semantic Word Similarity Assessment. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2016. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In *Proceedings of IJCAI*. New York City, USA, pages 2894–2900.

Niloofar Ranjbar, Fatemeh Mashhadirajab, Mehrnoush Shamsfard, Rayeheh Hosseini pour, and Aryan Vahid pour. 2017. Mahtab at SemEval-2017 Task 2: Combination of corpus based and knowledge-based methods to measure semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*. pages 448–453.

Răzvan-Gabriel Rotari, Ionuţ Hulub, Ştefan Oprea, Mihaela Plămadă-Onofrei, Alina Beatrice Lorenç, Raluca Preisler, Adrian Iftene, and Diana Trandabăç. 2017. Wild Devs at SemEval-2017 Task 2: Using neuronal networks to discover word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10):627–633.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. *CoNLL 2015* pages 258–267.

Mehrnoush Shamsfard, Akbar Hesabi, Hakimeh Fadaei, Niloofar Mansoory, Ali Famian, Somayeh Bagherbeigi, Elham Fekri, Maliheh Monshizadeh, and S Mostafa Assi. 2010. Semi automatic development of farsnet; the persian wordnet. In *Proceedings of 5th Global WordNet Conference, Mumbai, India*. volume 29.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*. San Francisco, USA.

Robert Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37:141–188.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of ACL*. Berlin, Germany, pages 1661–1670. http://www.aclweb.org/anthology/P16-1157.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research* 55:953–994.

Min Xiao and Yuhong Guo. 2014. Semi-supervised matrix completion for cross-lingual text classification. In *Proceedings of AAAI*. pages 1607–1614.

Zhirong Xu, Da Xu, and Song Zhang. 2017. BoT at SemEval-2017 Task 2: Cross-lingual Word Similarity Based on Translation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Vancouver, Canada.

Dongqiang Yang and David MW Powers. 2006. Verb similarity on the taxonomy of wordnet. In *Proceedings of the Third International WordNet Conference*. Jeju Island, Korea, pages 121–128.

Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*. pages 1393–1398.