Jmp8 at SemEval-2017 Task 2: A simple and general distributional approach to estimate word similarity

Josué Melka
LIASD - Université Paris 8
jmelka@ai.univ-paris8.fr

Gilles Bernard
LIASD - Université Paris 8
gb@ai.univ-paris8.fr

Abstract

We have built a simple corpus-based system to estimate words similarity in multiple languages with a count-based approach. After training on Wikipedia corpora, our system was evaluated on the multilingual subtask of SemEval-2017 Task 2 and achieved a good level of performance, despite its great simplicity. Our results tend to demonstrate the power of the distributional approach in semantic similarity tasks, even without knowledge of the underlying language. We also show that dimensionality reduction has a considerable impact on the results.

1 Introduction

Despite the crucial importance of semantic similarity in NLP, the vast majority of experiments have been conducted on the English language, which raises the question whether the developed approaches can be generalized.

SemEval-2017 Task 2 provides us with a framework for evaluating semantic representations in multiple languages and compare them. We focus here on the **multilingual** subtask, which consists of five monolingual word similarity datasets.

Our submission is based on the well known statistical approach which uses *bag-of-contexts* representation of words in a vector space model. We run two versions of our system, the first one using a direct sparse representation and the second one with compressed dense representation (detailed below). This second version was evaluated after the official evaluation deadline, and produced superior results as will appear below.

We briefly describe the multilingual subtask in section 2. Next, in section 3, we detail our system and its parameters. The results are presented and analyzed in section 4, and then we conclude in section 5.

2 Task description

Camacho-Collados et al. (2017) describes the task as follows:

> Given a pair of words, the task is to automatically measure their semantic similarity. All pairs in our datasets are scored according to a [0-4] similarity scale, where 4 denotes that the two words are synonymous and 0 indicates that they are completely dissimilar.
>
> **Multilingual word similarity** This subtask provides five monolingual word similarity datasets in English, German, Italian, Spanish and Farsi. The subtask is intended to test not only monolingual approaches but also multilingual and language-independent techniques.

The individual score of the systems is defined by the authors as the harmonic mean of Pearson and Spearman correlations on the corresponding dataset. However, as our analysis lead us to take into account the separate behavior of both measures, we did not focus here on the final score.

3 Our system

Our system is *corpus-based* only, and uses a few well known ideas from the distributional approach in word semantic similarity.

3.1 The training corpus

We have used the *Wikipedia corpus* taken from https://sites.google.com/site/rmyeid/projects/polyglot as recommended by the authors of the task in order to compare fairly with other corpus-based systems.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 230–234,
Vancouver, Canada, August 3 - 4, 2017.

Some properties of these corpora are given in Table 1. It should be noted that no preprocessing was made on the corpora documents.

Table 1: Statistics of the Wikipedia corpora

	size	lines	words	uniques
en	8.7G	70.9M	1 392M	5.3M
de	3.5G	32.2M	482M	5.7M
it	1.8G	11.9M	265M	1.9M
es	2.1G	14.8M	338M	2.3M

3.2 Language model

Our model is *count-based*, and we have used the same parameters for all languages.

First, we counted occurrences of alphabetic words in each corpus (barring words with non alphabetic characters), and kept the 100,000 most frequent for context words and the 300,000 most frequent as vocabulary. These arbitrary limits are justified by physical constraints of memory and time.

Contexts

The context we use for a given word w_i is defined as $w_{i-L}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+L}$. In this work we use a context length $L = 4$.

For each context word w_{i-k} we apply a weight of $\frac{1}{k}$ to give a stronger influence to nearest words in the context.

Then we built a *word-context* matrix by summing the weighted context occurrences for each word in the vocabulary.

PPMI

Pointwise Mutual Information (PMI) introduced by Ward Church and Hanks (1989) is one of the popular ways to measure the semantic association between words and their textual context as defined above, and can be easily estimated from the word-context matrix M, as:

$$\text{PMI}(w_i, c_j) = \log \frac{M_{ij} \sum_k \sum_n M_{kn}}{\sum_k M_{ik} \sum_n M_{nj}}$$

Bullinaria and Levy (2007) argue that the Positive PMI (PPMI) outperforms the other variants of PMI for semantic similarity tasks.

$$\text{PPMI}(w, c) = \max(0, \text{PMI}(w, c))$$

Vector compression

A common approach inspired by Latent Semantic Analysis (Deerwester et al., 1990) is to use truncated singular value decomposition (SVD) to reduce the vector dimensionality. The SVD factorization of the PPMI matrix is $M^{\text{PPMI}} = U \cdot \Sigma \cdot V^\top$, and can be truncated to the first d components.

In our experiments, we have used the symmetric variant proposed by Levy et al. (2015) using only the U_d matrix for representing word vectors, and we chose $d = 500$. Randomized SVD (Halko et al., 2009) from Scikit-learn was used to produce the matrix decomposition.

3.3 Evaluating word pairs similarity

Basically, we have used the cosine similarity to compare the word vectors.

Multi-word expressions

While some special features of the present task (such as *domain-specific* terms and *named entities*) do not necessarily require a special adaptation, *multi-word* expressions cannot be compared directly with single-word vectors. For this reason, we simply sum the vectors of every word in a multi-word expression to give the corresponding vector estimation.

See in section 4.3 a discussion about the results of this method.

Out of vocabulary words

Some words of the test dataset do not appear in our vocabulary, and we choose to give the median value .5 to the similarity of pairs including one or more out of vocabulary (OOV) words. Table 2 shows the numbers of such pairs for each language.

Table 2: pairs with OOV words

	pairs	*%*
en	21	4.2
de	68	13.6
it	24	4.8
es	17	3.4

A closer look shows that some words (such as "Brexit" or "DeepMind") were missed because they appeared too recently to be in our corpus, others because they contain non-alphabetic characters (like apostrophes or dashes), and the main part because they were not frequent enough to have been retained in our vocabulary.

The fact that the German language presents a higher OOV rate is not surprising, due to the morphological richness of this language. This can be improved by using a larger vocabulary and/or

using morphological approaches such as Bojanowski et al. (2016).

4 Results

We report the results obtained with our system (**Jmp8**) on four different languages in Table 3 and Table 4. Note that, due to a bug correction, the data is not exactly the same as in the official evaluation, though the magnitudes are similar. Moreover, the results of our second version have not been submitted for the challenge due to lack of time.

Luminoso is the best performer on this subtask, and **HCCL** is, to our knowledge, the best system which is corpus based and uses the shared training corpora. **NASARI** (Camacho-Collados et al., 2016) is the baseline proposed by the authors of the task.

Table 3: Pearson correlation

	en	de	it	es
Luminoso	0.783	0.7	0.728	0.732
HCCL	0.675	0.576	0.635	0.688
Jmp8-1	0.516	0.286	0.436	0.455
Jmp8-2	0.687	0.578	0.652	0.685
NASARI	0.683	0.513	0.597	0.602

Table 4: Spearman correlation

	en	de	it	es
Luminoso	0.795	0.7	0.754	0.754
HCCL	0.7	0.614	0.668	0.715
Jmp8-1	0.652	0.502	0.635	0.643
Jmp8-2	0.731	0.604	0.695	0.727
NASARI	0.681	0.514	0.594	0.597

4.1 Comparison of both Jmp8 versions

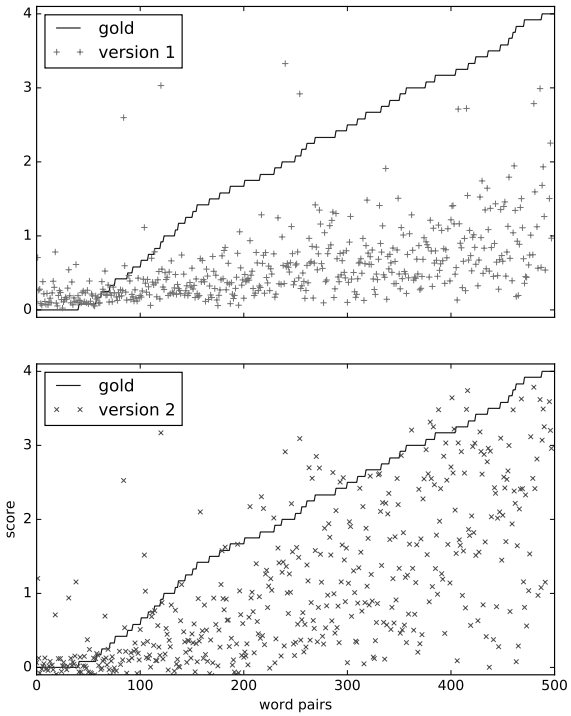
Jmp8-1 simply uses the PPMI matrix to compute similarities with sparse vectors of 100,000 components, while the second version, Jmp8-2, is based on a truncated SVD matrix which represents words as dense vectors of 500 components.

It turns out that Jmp8-1 produces a very important difference between Pearson and Spearman correlations, while Jmp8-2 provides more consistent results, and also better ones. In fact, Jmp8-2 outperforms **NASARI** in all cases, and achieves similar performance to **HCCL**.

Interpretation

The important difference between both Jmp8 versions is explained by the fact that Jmp8-1 presents a non-linear relationship with the gold standard, as depicted in Figure 1.

Figure 1: Comparison of both versions (en)



4.2 Language independence

These results suggest that our method (especially the second version) generalizes well for different languages, even if there are differences.

Our interpretation is that the English language is favored because its corpus is the biggest; Italian and Spanish results indicate that our approach remains interesting even with a much smaller corpus. The results are significantly lower for the German language despite the size of its corpus (this is true for all methods mentioned here), presumably because there are many out of vocabulary words.

This is supported by the fact that we found the correlations to be much higher (comparable to Italian and Spanish values), if, instead of using .5 median value for OOV pairs, we simply deleted these pairs from the dataset.

With SVD approach, these deletions improved correlations by about 20% ($p = 0.65$ and $s = 0.67$) for German and by less than 3% for other languages. Note that these numbers should be ta-

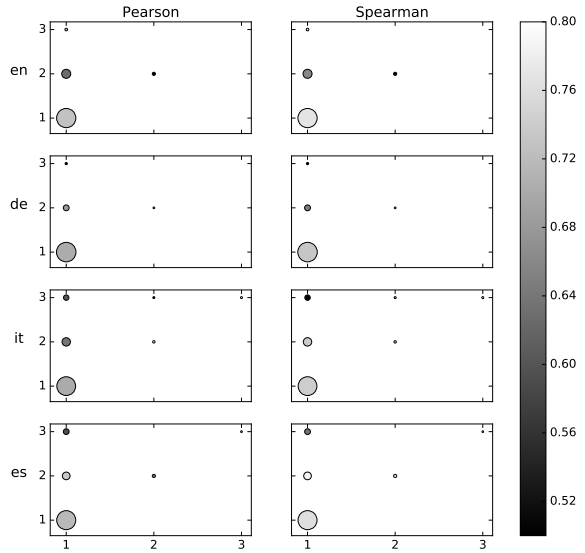
ken with caution because missing data can introduce bias.

4.3 Multi-word influence

To show the effect of the number of words in expressions on global performances, we calculated and plotted for each language the correlation with the gold standard separately for each number of words by expression (Figure 2). The size of the circles indicates the amount of pairs in each group.

For multiple reasons, it is somewhat difficult to analyze the influence of multi-word expressions on the overall performance. However, as one can expect, our simplistic method appears to degrade performance when the number of words in expressions increases. It is rather surprising that our results are still quite good despite this negative influence, but this should be mitigated by the number of pairs involved.

Figure 2: Multi-word influence in a pair (Jmp8-2)



Another approach such as phrasing (Mikolov et al., 2013) can be applied as well to address this issue.

4.4 Comparison with WordSim-353 dataset

Leviant and Reichart (2015) has translated the WordSim-353 dataset into several languages [1], and we have tested our system with the similarity subset (Agirre et al., 2009), which contains 201 pairs of words. It should be noted that WS353 uses single words only, and we have very few OOV words (0 in English, 4 in German and 1 in Italian). Table 5 shows our results.

Table 5: Correlations on WS353-sim dataset

		en	de	it
Jmp8-1	P	0.608	0.461	0.447
	S	0.667	0.547	0.592
Jmp8-2	P	0.722	0.654	0.600
	S	0.737	0.676	0.602

The gap between Pearson and Spearman correlations is still present for Jmp8-1, confirming that sparse vectors do not perform well in semantic similarity tasks.

Another interesting point is that the correlations for the German language are significantly higher than for the present task, which can be explained by the lower OOV rate in this dataset, as discussed above (section 4.2).

Surprisingly, contrary to the results of the present task, Italian results are significantly lower than for the other languages, though less so than were German results in the present task. We have not yet found a good explanation for this, as it is clear that OOV words are out of the picture.

5 Conclusion

We have shown that it is possible to achieve a good level of performance in multilingual word semantic similarity task with a rather simple but generalist approach.

While one should take these results with caution, some important conclusions can be drawn from our work. First, it is confirmed that the raw sparse PPMI representation is less adapted to similarity measure than the compressed dense SVD representation. Second, a specific approach needs to be developed to address multi-word expressions, although the vector addition seems to work moderately well for 2-words. And last, we have seen that OOV pairs can be problematic for a systematic comparison between systems and/or languages.

The ability of our method to handle multiple languages seems good, but needs further investigation in those directions with more extensive test sets in order to yield a refined analysis.

Finally, we are considering the combination of this method with other approaches, both from *word embeddings* methods and from supervised techniques.

[1] http://technion.ac.il/~ira.leviant/Multilingual_SimLex_Wordsim.html

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 19–27. http://aclweb.org/anthology/N09-1003.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR* abs/1607.04606. http://arxiv.org/abs/1607.04606.

John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods* 39(3):510–526.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 15–26. http://www.aclweb.org/anthology/S17-2002.

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.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6):391.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2009. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *CoRR* abs/0909.4061. http://arxiv.org/abs/0909.4061.

Ira Leviant and Roi Reichart. 2015. Separated by an Un-common Language: Towards Judgment Language Informed Vector Space Modeling. *CoRR* abs/1508.00106. http://arxiv.org/abs/1508.00106.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association of Computational Linguistics* 3:211–225. http://aclweb.org/anthology/Q15-1016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

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

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*. http://aclweb.org/anthology/P89-1010.

A Supplemental Material

We made our source code and outputs available at `https://github.com/yoch/jmp8`