Citius at SemEval-2017 Task 2: Cross-Lingual Similarity from Comparable Corpora and Dependency-Based Contexts

Pablo Gamallo
Centro Singular de Investigación en
Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela, Galiza
pablo.gamallo@usc.es

Abstract

This article describes the distributional strategy submitted by the Citius team to the SemEval 2017 Task 2. Even though the team participated in two subtasks, namely monolingual and cross-lingual word similarity, the article is mainly focused on the cross-lingual subtask. Our method uses comparable corpora and syntactic dependencies to extract count-based and transparent bilingual distributional contexts. The evaluation of the results show that our method is competitive with other cross-lingual strategies, even those using aligned and parallel texts.

1 Introduction

A comparable corpus consists of documents in two or more languages or varieties which are not translation of each other and deal with similar topics. Comparable corpora are by definition multilingual and cross-lingual text collections. The use of comparable corpora for word similarity is a well-known task (Fung and McKeown, 1997; Rapp, 1999; Saralegi et al., 2008; Gamallo, 2007; Gamallo and Pichel, 2008; Ansari et al., 2014; Hazem and Morin, 2014). The main advantage of comparable corpora is that the Web can be used as a huge resource of multilingual texts. In contrast, their main drawback is the low performance of the extraction systems based on them. According to (Nakagawa, 2001), word similarity extraction from comparable corpora is a too difficult and ambitious objective, and much more complex than extraction from parallel and aligned corpora. However, the reasonable results our comparable-corpus method achieved in the cross-lingual subtask of SemEval 2017 Task 2 (Camacho-Collados et al., 2017) show that the gap between parallel and comparable corpora for word similarity is shortening. In this article, we describe our comparable-corpus method for cross-lingual similarity in the next section (2).Then Section 3 describes the experiments and the evaluation and, finally, a discusion is addressed in Section 4.

2 The Cross-Lingual Strategy

The best known strategy to extract bilingual correspondences from comparable corpora works as follows: a word w_2 in the target language is semantically related to w_1 in the source language if the context expressions with which w_2 co-occurs tend to be translations of the context expressions with which w_1 co-occurs. The basis of the method is to find the target words that have the most similar distributions with a given source word. The starting point of this strategy is a seed list of bilingual expressions that are used to build the context vectors defining all words in both languages. This seed list is usually provided by an external bilingual dictionary. In our approach, the seed expressions used as cross-language pivot contexts are not bilingual pairs of words as in related work, but bilingual pairs of lexico-syntactic contexts.

The process of building a list of seed bilingual lexico-syntactic contexts consists of two steps: first, we generate a large list of bilingual correlations between lexico-syntactic patterns using an external bilingual dictionary, syntactic parsing and syntactic-based transfer rules. Second, this list is reduced by filtering out those pairs of patterns that do not occur in the comparable corpus. We also remove those that are sparse or unbalanced in the corpus. It results in a list of *seed bilingual contexts*.

To take an example, let us suppose that an English-Spanish dictionary translates the noun *import* into the Spanish counterpart *importación*. To

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 226–229,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

English	Spanish
(*import*, of|to|in|for|by|with, N)	(*importación*, de|a|en|para|por|con, N)
(N, of|to|in|for|by|with, *import*)	(N, de|a|en|para|por|con, *importación*)
(V, obj, *import*)	(V, obj, *importación*)
(V, subj, *import*)	(V, subj, *importación*)
(V, of|to|in|for|by|with, *import*)	(V, de|a|en|para|por|con, *importación*)
(*import*, mod, A)	(*importación*, mod, A)

Table 1: Bilingual correlations between lexico-syntactic patterns generated from the translation pair: import-importación. A patterns is a dependency triple (head, relation, dependent). The head and dependent can be lexical units (e.g. *import*) or Part-of-Speech tags (e.g. N, V, A)

generate bilingual pairs of lexico-syntactic patterns from these two nouns, we follow basic transfer rules such as: (1) if *import* is the subject of a verb, then its Spanish equivalent, *importación*, is also the subject; (2) if *import* is modified by an adjective at the left position, then its Spanish equivalent is modified by an adjective at the right position; (3) if *import* is restricted by a prepositional complement headed by the preposition *in*, then its Spanish counterpart is restricted by a prepositional complement headed by the preposition *en*. The third rule needs a closed list of English prepositions and their more usual Spanish translations. For each entry (noun, verb, or adjective), we only generated a subset of all possible patterns. Notice that prepositions are encoded not as lexical units, but as syntactic dependencies. Table 1 depicts the bilingual pairs of patterns generated from the bilingual word pair *import-importación* and a restricted set of rules.

Finally, the comparable corpus allows us to filter out missing and sparse patterns, for instance: $(import, with, N/importación, con, N)$. The resulting bilingual lexico-syntactic patterns are used as distributional contexts to build the vector space.

The distributional vector space we have adopted is a transparent count-based model with explicit and sparse dimensions. Sparseness reduction is performed by selecting the most relevant contexts per word using a filtering strategy (Bordag, 2008; Gamallo and Bordag, 2011; Gamallo, 2016). The filtering strategy to select the most relevant contexts consists in selecting, for each word, the R (relevant) contexts with highest lexical association scores and computed with loglikelihood measure (Dunning, 1993). The top R contexts are considered to be the most *relevant* and informative for each word. R is a global, arbitrarily defined constant whose usual values range from 10 to 1000 (Biemann et al., 2013; Padró et al., 2014). In short, we keep at most the R most relevant contexts for each target word. This is an explicit and transparent representation giving rise to a non-zero matrix. Methods based on dimensionality reduction and embeddings, by contrast, make the vector space more compact with dimensions that are not transparent in linguistic terms (Gamallo, 2016).

3 Experiments

3.1 Data and Tools

We have participated at the cross-lingual word similarity subtask of SemEval 2017 Task 2 (Camacho-Collados et al., 2017), where each word pair is composed by ten cross-lingual word similarity datasets (Camacho-Collados et al., 2015). More precisely, we sent two different runs to be evaluated against the English-Spanish dataset. In this subtask, we used as comparable corpora the English and Spanish tokenized Wikipedia dumps in text format, which are available at https://sites.google.com/site/rmyeid/projects/polyglot. The difference between the two runs (Citius_run1 and Citius_run2) we have submitted is in the training corpus. While Citius_run1 is only trained with the two above mentioned Wikipedias, Citius_run2 uses additional text created with BootCat (Baroni et al., 2006) and seed words that do not occur in the two Wikipedias.

To process the corpus, we used the multilingual PoS tagger of LinguaKit[1] (Garcia and Gamallo, 2015) and DepPattern, a rule-based and multilingual dependency parser (Gamallo and González, 2011; Gamallo, 2015). Named entities were identified with the NER module provided by LinguaKit while multi-words were extracted by means of an ad-hoc procedure that just selects those appearing in the test dataset.

[1] https://github.com/citiususc/Linguakit

TeamName	Pear.	Spear.	Final
Luminoso_run2	0.75	0.772	0.761
Luminoso_run1	0.748	0.772	0.76
NASARI(baseline)	0.636	0.63	0.633
OoO_run1	0.579	0.59	0.584
Citius_run1&	0.565	0.589	0.577
Citius_run2	0.556	0.576	0.566
SEW_run1	0.495	0.514	0.505
RUFINO_run1&	0.339	0.341	0.34
RUFINO_run2&	0.342	0.333	0.337
UniBuc-Sem_run1*	0.084	0.096	0.09
HCCL_run1*	0.101	0.077	0.087
hjpwhu_run2	0.043	0.041	0.042
hjpwhu_run1	0.043	0.041	0.042
HCCL_run1*	0.04	0.04	0.04

Table 2: Results for the cross-lingual English-Spanish task.

To build the distributional models, target words appearing less than 100 times were filtered out. Similarly, bilingual contexts with frequency less than 50 were removed. The English-Spanish dictionary used to select the seed contexts required by the acquisition algorithm contains 10,828 entries, which is the lexical resource integrated in Apertium, an open source machine translation system[2]. Then, for each word, we selected the 500 most relevant contexts. The final model resulted in a bilingual non-zero matrix of about 440k target words and over 208k different dependency-based contexts. In total we built a non-zero matrix with about 100 billion word-context pairs, which is a relatively easy-to-handle matrix, and even smaller in size than an equivalent dense matrix with 440k words and 500 dimensions. This is the the co-occurrence matrix used by Citius_run1. A version of the system is publicly available at http://gramatica.usc.es/~gamallo/prototypes.htm. A second matrix (used by Citius_run2) was built by searching for new occurrences with BootCat for those test words that were filtered out from the previous co-occurrence matrix.

3.2 Results

Table 2 shows the results for the English-Spanish dataset. Citius_run1 is the 5th best system (out of 14). However, if we only consider the runs using a comparable-based strategy with the Wikipedia dumps (marked with "&" in the table), Citius_run1 is the first out of three, leading by 23 points the second one. It is also noticeable that our comparable-corpus strategy is in a competitive position with other methods based on aligned and parallel corpora, which are most of systems participating at the subtask.

We also participated at the monolingual word similarity task for English and Spanish by making use of the same distributional vector space we have adopted for the cross-lingual task and reported in Gamallo (2016). The results we obtained are reasonable for the two languages, in particular if we only consider the Spearman correlation. According to this measure, Citius_run2 is the 4th best run in English (out of 23), and is also the 4th best system in Spanish (out of 11).

4 Discussion

We have reached interesting results by making use of a traditional and transparent distributional model instead of dense and embedding representations. Besides, in the cross-lingual task, we have built the models with non-parallel corpora instead of using aligned and parallel texts. However, our method is language dependent since it requires syntactic information and specific language processing. Finally, we must also point out that the test dataset is not well suited to the characteristics of our syntax-based strategy. The test dataset includes semantically related word pairs that are not functionally equivalents, such as for instance *globalism / visa* or *nepotism / king* in the English pairs. Even if *globalism* is semantically related to *visa*, they occur in different syntactic positions with different syntactic functions. Models without syntactic contexts (i.e. *bag-of-words* models) tend to perform well in tasks oriented to identify semantic relatedness and analogies (Levy and Goldberg, 2014; Gamallo, 2016). By contrast, syntax-based methods, as the one we have proposed, tend to outperform bag-of-words techniques when the objective is to compute semantic similarity between functional (or paradigmatic) equivalent words, such as detection of synonym, co-hyponym or hypernym word relations (Padó and Lapata, 2007; Peirsman et al., 2007; Gamallo, 2009).

[2] https://sourceforge.net/projects/apertium/

Acknowledgments

This work has received financial support from a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, TelePares (MINECO, ref:FFI2014-51978-C2-1-R), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

References

Ebrahim Ansari, M. H. Sadreddini, Alireza Tabebordbar, and Mehdi Sheikhalishahi. 2014. Combining different seed dictionaries to extract lexicon from comparable corpus. *Indian Journal of Science and Technology* 7(9):1279–1288.

Marco Baroni, Adam Kilgarriff, Jan Pomikálek, and Pavel Rychlý. 2006. Webbootcat: a web tool for instant corpora. In Cristina Onesti Elisa Corino, Carla Marello, editor, *EURALEX International Congress*. Edizioni dell'Orso, Torino, Italy, pages 123–131.

Biemann, C., and Riedl M. 2013. Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling* 1(1):55–95.

Stefan Bordag. 2008. A Comparison of Co-occurrence and Similarity Measures as Simulations of Context. In *9th CICLing*. pages 52–63.

José Camacho-Collados, Mohammad Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *SemEval*. Vancouver, Canada.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *ACL, Beijing, China*. pages 1–7.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translation from non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*. Hong Kong, pages 192–202.

Pablo Gamallo. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Machine Translation SUMMIT XI*. Copenhagen, Denmark.

Pablo Gamallo. 2009. Comparing different properties involved in word similarity extraction. In *14th Portuguese Conference on Artificial Intelligence (EPIA'09), LNCS, Vol. 5816*. Springer-Verlag, Aveiro, Portugal, pages 634–645.

Pablo Gamallo. 2015. Dependency parsing with compression rules. In *International Workshop on Parsing Technology (IWPT 2015)*. Bilbao, Spain.

Pablo Gamallo. 2016. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation* First online: 13 May 2016.

Pablo Gamallo and Stefan Bordag. 2011. Is singular value decomposition useful for word simalirity extraction. *Language Resources and Evaluation* 45(2):95–119.

Pablo Gamallo and Isaac González. 2011. A grammatical formalism based on patterns of part-of-speech tags. *International Journal of Corpus Linguistics* 16(1):45–71.

Pablo Gamallo and José Ramom Pichel. 2008. Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary. *LNCS* 4919:413–423.

Marcos Garcia and Pablo Gamallo. 2015. Yet another suite of multilingual nlp tools. In *Symposium on Languages, Applications and Technologies (SLATE 2015)*. Madrid, Spain, pages 81–90.

Amir Hazem and Emmanuel Morin. 2014. Improving bilingual lexicon extraction from comparable corpora using window-based and syntax-based models. *Lecture Notes in Computer Science* 8404:310–323.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA*. pages 302–308.

Hiroshi Nakagawa. 2001. Disambiguation of single noun translations extracted from bilingual comparable corpora. *Terminology* 7(1):63–83.

Sebastian Padó and Mirella Lapata. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics* 33(2):161–199.

Muntsa Padró, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. 2014. Nothing like good old frequency: Studying context filters for distributional thesauri. In *EMNLP*. pages 419–424.

Yves Peirsman, Kris Heylen, and Dirk Speelman. 2007. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *CoSMO Workshop*. Roskilde, Denmark, pages 9–16.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *ACL'99*. pages 519–526.

X. Saralegi, I. San Vicente, and A. Gurrutxaga. 2008. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. In *LREC 2008 Workshop on Building and Using Comparable Corpora*.