RUFINO at SemEval-2017 Task 2: Cross-lingual lexical similarity by extending PMI and word embeddings systems with a Swadesh's-like list

Sergio Jimenez
Instituto Caro y Cuervo
Bogotá, Colombia
sergio.jimenez@caroycuervo.gov.co

George Dueñas
Instituto Caro y Cuervo
Bogotá, Colombia
george.duenas@caroycuervo.gov.co

Lorena Gaitan
Universidad Cooperativa de Colombia
Bogotá, Colombia
lorena.gaitanb@campusucc.edu.co

Jorge Segura
Universidad Nacional de Colombia
Bogotá, Colombia
jaseguran@unal.edu.co

Abstract

The RUFINO team proposed a non-supervised, conceptually-simple and low-cost approach for addressing the Multilingual and Cross-lingual Semantic Word Similarity challenge at SemEval 2017. The proposed systems were cross-lingual extensions of popular monolingual lexical similarity approaches such as PMI and word2vec. The extensions were possible by means of a small parallel list of concepts similar to the Swadesh's list, which we obtained in a semi-automatic way. In spite of its simplicity, our approach showed to be effective obtaining statistically-significant and consistent results in all datasets proposed for the task. Besides, we provide some research directions for improving this novel and affordable approach.

1 Introduction

Pairwise semantic lexical similarity is a core component in NLP systems that tackle fundamental NLP tasks such as word sense disambiguation (Camacho-Collados et al., 2015), semantic textual similarity (Agirre et al., 2017) and many others. Since more than two decades, the problem has been addressed mainly for the English language, but only recently, other languages have been considered. The task 2 in SemEval 2017 (Camacho-Collados et al., 2017) proposes a public challenge for this task in 5 languages (English, Spanish, Italian, German and Farsi) and an additional cross-lingual challenge in their 10 possible combinations. This paper describes the participating systems of the RUFINO team in these challenges.

Lexical-similarity systems receive two words as input and return a numerical score that reflects the similarity or relatedness between them. Cross-lingual systems extends the idea to words in different languages. The evaluation of such systems consist in measuring the correlation of the scores obtained by several word pairs against the consensus of human judgments (gold standard).

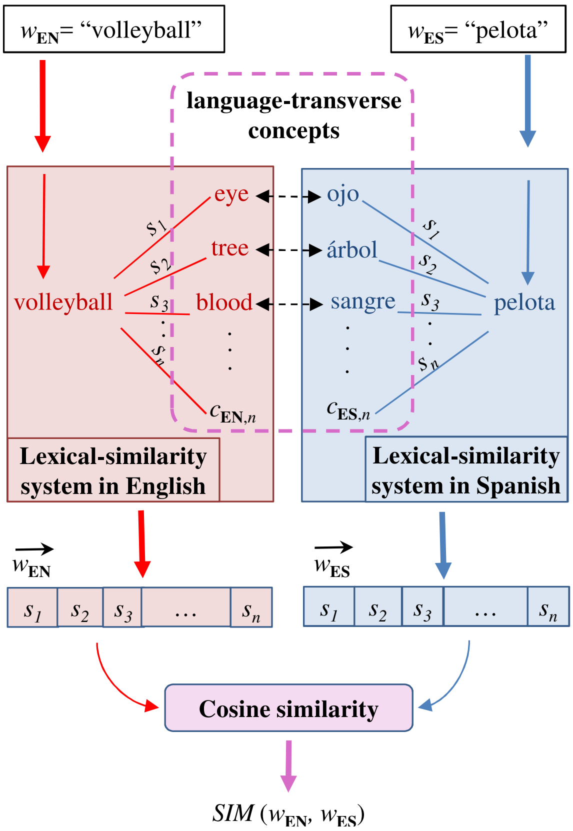

Figure 1: Architecture of the cross-lingual system

The main fundamental resources used by lexical similarity systems are monolingual corpora, parallel corpora and knowledge-based resources such as WordNet (Miller, 1995) and Babelnet (Navigli and Ponzetto, 2012). Among them, monolingual corpora are the cheapest and most available resource in the majority of the languages. Aiming to propose a lexical similarity system with easy

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 239–244,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

replicability, beyond the 5 languages of the challenge, the RUFINO team proposed a system based mainly on monolingual corpora.

Like monoligual lexical similarity, the cross-lingual variant of this task aims to establish quantitatively the degree of similarity between two words, but with the added complexity of being in different languages. This task contributes to solve other higher level task such as cross-lingual text similarity and entailment (Jimenez et al., 2012, 2013). However, to the best of our knowledge, it is not possible to build a cross-lingual system between two non-similar languages based solely on monolingual corpora. For that, we proposed a resource inspired by the well-known Swadesh's list. The Swadesh's list (Swadesh, 1950, 1952) is comprised of approximately 200 concepts aimed to be universal, culturally independent and transverse to almost any language for the purposes of comparative linguistics. We used Wikipedia and Google Translate to build a list for the 5 languages of the competition containing 66 concepts with similar properties to the ones proposed by Swadesh. Since the alignment of concepts grouped into synsets among WordNets in different languages is not always available, we decided to use Google Translate. In case of considering a not included language among the supported ones by Google Translate (they are more than 100), we estimated it could be comparatively more feasible and economic to build an automatic translator from a parallel corpus than the manual construction of a WordNet for that language. Nevertheless, a resource like BabelNet, for instance, could also provide accurate translations of transverse concepts. Our goal, is to build cross-lingual systems starting from monolingual systems connected across languages by the proposed list of concepts. Figure 1 provides a general overview of the general architecture of the proposed system.

The organizers of the challenge proposed a benchmark corpus for the sake of comparison of the participating systems. The systems proposed by our team used for training the Wikipedias in the 5 languages, which is the benchmark corpora for the monolingual sub-task. The benchmark corpus for the cross-lingual systems is the Europarl parallel corpus[1].Alternatively, our cross-lingual systems used the proposed list of language-traverse concepts, which is considerably smaller, simpler and cheaper than the Europarl corpus. Although, the results obtained by our systems were in the middle range of the general ranking of official results, all of them were statistically significant and consistent across all datasets. Moreover, in some cases our results were comparable to other systems relying in considerably larger, more complex and more expensive resources.

The rest of the paper contains the following sections. In section 2, we present the motivation for our approach. Section 3 contains the detailed description of our participating systems. In section 4 the obtained results are presented and discussed. Finally, in section 5 we provide some concluding remarks.

2 Motivation

A concept list of basic vocabulary items showing the universality of certain parts of the lexicon of human languages was initially proposed by Morris Swadesh (Swadesh, 1950, 1952). Swadesh claimed that certain morphemes and everyday words such as *mother*, *son*, *hand*, *head*, *sun*, *warm*, *water*, *tree*, etc. connected with concepts and experiences common to all human groups are relatively stable over time. Since then, many concepts lists following the same characteristics have been compiled for several purposes in descriptive linguistics.

Considering that concepts are not only transverse to languages, but they also share some proximity when they are semantically close. For instance, *mother* and *son* are more semantically close than *mother* and *sun* independently of the language. Our approach is based on the idea that a set of transverse concepts to languages serve as a support to index a vectorial representation of the words of a given language. In order to obtain such representation for just one language, it is required the lexicalization of that set of transverse concepts and a lexical-similarity (or distance) system of that language. This semantic representation is cross-lingual since it only depends on the relative similarities (or distances) of each one of the words to be represented to the set of transverse concepts. Therefore, the representation of a particular word w is a vector where each dimension corresponds to the similarity score between w and each word from the set of transversal concepts.

Intuitively, three conditions that a set of transverse concepts for a set of languages should follow

[1] http://www.statmt.org/europarl/

were considered. First, these concepts should be relatively frequent in all the given languages due to the fact that infrequent words tend to produce low-quality measurements in the required lexical similarity systems based either on knowledge or corpus. Second, it is preferable that the transverse concepts are lexicalized in each one of the languages with just one word. This condition could anticipate problems with the rules of the usage of multi-words in each language. Third, the monolingual lexical similarity systems should be similar in their construction and used resources. The latter improves the conditions so that the distances and similarities among concepts could be proportional through the different languages.

As a result, the list of transverse concepts, a relatively simple resource to obtain, can be useful to turn a set of monolingual systems into a cross-lingual system.

3 Methods

We build two groups of monolingual lexical similarity systems and other two groups of cross-lingual systems. For both, monolingual and cross-lingual sub-tasks, the systems labeled as *run1* rely mainly on Pointwise Mutual Information (PMI) (Church and Hanks, 1990), and those labeled as *run2* were based on Polyglot's word embeddings (Al-Rfou et al., 2013). The following subsections describe such systems.

3.1 Monolingual systems

3.1.1 *run1*: PMI and common contexts

PMI is a simple corpus-based information-theoretical method for finding associations between pairs of words using the distributional hypothesis, which states that associations between words depend on the coocurrences of the words in a large corpus. The PMI score between two words a and b can be computed with this formula:

$$PMI(a,b) = -\log\left(\frac{P(a\wedge b)}{P(a)P(b)}\right).$$

Probabilities can be estimated by the following expressions:

$$P(a)=\frac{o_a}{N};\ P(b)=\frac{o_b}{N};\ P(a\wedge b)=\frac{o_{a\wedge b}}{N-1}$$

Where o_a and o_b are the number or occurrences of words a and b in the corpus, $o_{a\wedge b}$ is the number of coocurrences, and N the total number of words in the corpus (all occurrences). We used the benchmark corpora proposed for the task, that is, Wikipedia's dumps for the 5 languages downloaded in October 2016. The preprocessing comprised lower-casing and stopwords[2] removal. For obtaining $o_{a\wedge b}$, each coocurrence of a followed by b or vice-versa was counted. N was the total number of non-stopwords on each corpus.

The PMI scores computed using coocurrences is a low-cost and effective tool for finding word associations. However, associations between synonyms or words in the same category cannot be detected with such method because they do not tend occur consecutively in text. For capturing these second-order relationships, we proposed an association measure based in the proportion of common contexts between pairs of words. For that, we defined the context of a word as a duple of its left and right neighbor words (after removing stopwords). During the process of context definiction, we also tried other context settings such as two neighbor words before and after, two before/one after, one before/ two after, just one before, just one after, just two before and just two after (we even attempted not to remove the stopwords). However, we observed that when using the trial data, the setting with the best performance was a neighbor word before and after. Thus, we collected for each word a the set of its contexts C_a. The Jaccard coefficient (Jaccard, 1901; Jimenez et al., 2016) was used for comparing pairs of words represented as their sets of contexts:

$$JCC(a,b)=\frac{|C_a\cap C_b|}{|C_a\cup C_b|}$$

The final similarity score for a pair of words was the average of previously scaled PMI and JCC scores.

$$SIM(a,b)=0.5\left(\frac{PMI(a,b)}{\max PMI}+\frac{JCC(a,b)}{\max JCC}\right)$$

Here, $\max PMI$ and $\max JCC$ are the maximum scores of the corresponding measures within the entire dataset of word pairs being compared. In our implementation, if PMI produced a mathematical error such as division by zero or logarithm of a negative number, then the PMI score was replaced by the average of the scores obtained by the same measure for the other non-erroneous word pairs in the dataset.

[2] urls of stopwords

English	Spanish	Italian	German	Farsi	English	Spanish	Italian	German	Farsi
angel	ángel	angelo	Engel	فرشته	background	fondo	sfondo	Hintergrund	زمینه
animal	animal	animale	Tier	حیوان	ministry	ministerio	ministero	Ministerium	وزارت
artist	artista	artista	Künstler	هنرمند	birth	nacimiento	nascita	Geburt	تولد
metal	metal	metallo	Metall	فلز	musician	músico	musicista	Musiker	نوازنده
poet	poeta	poeta	Dichter	شاعر	newspaper	periódico	giornale	Zeitung	روزنامه
minute	minuto	minuto	Minute	دقیقه	novel	novela	romanzo	Roman	رمان
blood	sangre	sangue	Blut	خون	christians	cristianos	cristiani	Christen	مسیحیان
ring	anillo	anello	Ring	حلقه	piano	piano	pianoforte	Klavier	پیانو
painter	pintor	pittore	Maler	نقاش	beautiful	hermoso	bello	schön	زیبا
comedy	comedia	commedia	Komödie	کمدی	composer	compositor	compositore	Komponist	آهنگساز
price	precio	prezzo	Preis	قیمت	quality	calidad	qualitŕ	Qualität	کیفیت
concert	concierto	concerto	Konzert	کنسرت	contract	contrato	contratto	Vertrag	قرارداد
sale	venta	vendita	Verkauf	فروش	religion	religión	religione	Religion	دین
read	leer	leggere	lesen	خواندن	candidate	candidato	candidato	Kandidat	نامزد
crisis	crisis	crisi	Krise	بحران	congress	congreso	congresso	Kongress	کنگره
train	tren	treno	Zug	قطار	scene	escena	scena	Szene	صحنه
tree	árbol	albero	Baum	درخت	shipyard	astillero	cantiere navale	Werft	کشتیسازی کارخانه
texts	textos	testi	Texte	متون	sister	hermana	sorella	Schwester	خواهر
domain	dominio	dominio	Domain	دامنه	soldier	soldado	soldato	Soldat	سرباز
doubt	duda	dubbio	Zweifel	شک	speed	velocidad	velocità	Geschwindigkeit	سرعت
drama	drama	dramma	Drama	درام	engines	motores	motori	Motoren	موتورهای
statue	estatua	statua	Statue	مجسمه	structure	estructura	struttura	Struktur	ساختار
error	error	errore	Fehler	خطا	discovery	descubrimiento	scoperta	Entdeckung	کشف
eye	ojo	occhio	Auge	چشم	depth	profundidad	profonditŕ	Tiefe	عمق
factory	fábrica	fabbrica	Fabrik	کارخانه	translation	traducción	traduzione	Übersetzung	ترجمه
weapon	arma	arma	Waffe	سلاح	device	dispositivo	dispositivo	Gerät	دستگاه
friend	amigo	amico	Freund	دوست	identity	identidad	identitŕ	Identität	هویت
guitar	guitarra	chitarra	Gitarre	گیتار	violence	violencia	violenza	Gewalt	خشونت
hand	mano	mano	Hand	دست	founder	fundador	fondatore	Gründer	موسس
value	valor	valore	Wert	ارزش	weight	peso	peso	Gewicht	وزن
wind	viento	vento	Wind	باد	important	importante	importante	wichtig	مهم
window	ventana	finestra	Fenster	پنجره	marriage	matrimonio	matrimonio	Ehe	ازدواج
word	palabra	parola	Wort	کلمه	message	mensaje	messaggio	Nachricht	پیام

Table 1: List of 66 language-transverse concepts in the 5 target languages

3.1.2 *run2*: Polyglot's embeddings

Our second monolingual system used the pre-trained Polyglot's word embeddings (Al-Rfou et al., 2013), which were obtained using the *word2vec* algorithm (Mikolov et al., 2013) applied to Wikipedia as corpus for a large number of languages. For each pair of target words a and b, their 64-dimensional vector representations (64 is the number of dimmensions in Polyglot's vectors) were obtained from Polyglot's files and then compared using cosine similarity. If a target word started with a capital letter and it was not found in the database of embeddings, then the word is lowercased and searched again. Similarity, if multi-words targets are not found we used the vectorial summation of the representations of the composing words. After that, if some target is still not found, as before, we used the average score of non-erroneous word pairs in the dataset.

3.2 Obtaining a Swadesh-like list

For obtaining a list of concepts with similar properties to the Swadesh's list, first we collected the lists of the top-5000 more frequent terms from the Wikipedia for each one of the 5 target languages. Next, each word on each list was translated to the other 4 languages and the translations were translated back to the original language. All translations were obtained using the GOOGLE-TRANSLATE() function in the spreadsheet editor of Google Drive. On each list, we preserved only the rows whose all 4 back translations coincided with the original word. Finally, the 5 list were merged and aligned for identify terms that occurred in the 5 languages. Only the terms occurring exactly in the 5 languages were preserved.

From the previous selection, we obtained a list containing 172 concepts with their lexicalizations in the 5 target languages. This initial list was purged manually by removing proper names, cardinals, stopwords and other unwanted forms. The final result is an aligned list of 66 concepts of fre-

Method→	PMI-JCC	Polyglot	NASARI
language	***run1***	***run2***	**baseline**
English	0.656	0.394	0.682
Spanish	0.549	0.406	0.600
Italian	0.476	0.306	0.596
German	0.539	0.369	0.514
Farsi	0.360	0.256	0.405
Average	0.481	0.334	0.529

Table 2: Results for the monolingual sub-task (values are the harmonic mean between Pearson's and Spearman's correlation coefficients).

quent words in 5 languages. Besides, all possible combination pair from the 5 words on each concept are common translations of the others. The obtained list is shown in Table 1. That is the proposed list of language-traverse concepts used for enhancing the previously described monolingual lexical-similarity systems to support cross-linguality.

3.3 Cross-lingual systems

The proposed lexical cross-lingual systems were built by combining the monolingual systems described in subsections 3.1.1 and 3.1.2, with the list of 66 language-traverse concepts proposed in the previous subsection. The method for that is straightforward and depicted in Figure 1. Basically, for obtaining a vectorial representation of a word in a particular language, such word is compared using a monolingual lexical-similarity system for that language, against the 66 lexicalizations of the transverse concepts in that language. The result is a 66-dimensional vector, which is a language-independent representation the word. For comparing a pair of words in two different languages, their language-independent vectorial representations are obtained using their respective monolingual systems and the aligned list of concepts. Then the final similarity score is obtained computing the cosine similarity between the two vectors. We built two cross-lingual systems labeled as *run1*, using the monolingual systems described in subsection 3.1.1, and *run2*, with the systems described in subsection 3.1.2.

4 Results and discussion

Results obtained by our monolingual systems (*run1* and run2) are shown in Table 2. *Run1* averaged relatively close to the baseline, which in fact, is a very strong baseline based in knowledge from Babelnet (Camacho-Collados et al., 2016). The system that outperformed the baseline was the $PMI\text{-}JCC$ monolingual system (*run1*) for German. *Run2*, based on Polyglot's embeddings, was consistently worse than *run1*. Although, both systems use the same corpora, the difference in performance is significant. As regards our runs, we suggest that $PMI\text{-}JCC$ is a method that takes better advantage of small corpora in comparison with the *word2vec* algorithm used in the construction of Polyglot's embeddings.

Method→	PMI-JCC	Polyglot	NASARI
languages	***run1***	***run2***	**baseline**
it-fa	0.249	0.210	0.486
es-it	0.356	0.288	0.595
es-fa	0.257	0.300	0.479
en-it	0.342	0.238	0.648
en-fa	0.253	0.373	0.505
en-es	0.340	0.337	0.633
en-de	0.330	0.303	0.598
de-it	0.327	0.232	0.561
de-fa	0.240	0.267	0.458
de-es	0.318	0.302	0.549
Average	0.301	0.285	0.551

Table 3: Results for the cross-lingual sub-task (values are the harmonic mean between Pearson's and Spearman's correlation coefficients).

Unlike the results of monolingual systems, the results for *run1* and *run2* in the cross-lingual task had a similar performance and were considerably less than the baseline (see Table 3). An interesting question we asked was to what extent the results of monolingual systems predict the performance of bilingual systems. In order to answer this question, we measured Pearson's correlation (r) between the result of the bilingual system and the minimum between the results of the two monolingual systems for the 10 language combinations. The result was $r_{run1} = 0.883$, $r_{run2} = 0.263$, and $r_{baseline} = 0.950$. Clearly, the results of monolingual systems based on $PMI\text{-}JCC$ and NASARI are good predictors of the results of bilingual systems.

5 Conclusions and future directions

From our participation in the task 2 in SemEval 2017 we can gather several conclusions. First, the proposed lexical-monolingual systems based respectively on PMI-JCC and Polyglot's embeddings (i.e. *word2vec*) obtained considerably different results, in spite of being constructed using the same corpus (i.e. Wikipedia). This result suggest that, for inferring lexical relationships, relatively small corpora can be better exploited by simpler methods such as PMI, which is convenient for under-resourced languages. Second, the proposed approach of using a parallel list of language-transverse concepts for building lexical cross-lingual systems from monolingual resources showed to be effective with a good cost-benefit ratio. Third, there is an important performance gap between the proposed approach and the knowledge-based baseline approach.

However, the monolingual versions of both our approach (*run1*) and that baseline share the property of being good predictors of the performance of the cross-lingual versions. Therefore, we conclude that a straightforward way to improve the proposed system is to use better monolingual systems. Additionally, the method for selecting the set of language-traverse concepts can be improved by considering the transversality of the relationships and by the use of size-balanced multilingual corpora.

Acknowledgments

We are grateful to Nestor Ruiz of the Instituto Caro y Cuervo for supporting and promoting the participation in this competition by the students of the graduate course in computational analysis of the language. We would also like to thank María Catalina Suárez, Karen González and María Camila Escobar for their helpful discussions, valuable comments and suggestions.

References

Eneko Agirre, Daniel Cer, Mona Diab, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval*.

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*. ACL, pages 183–192.

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.

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 the Association for Computational Linguistics*. 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.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1):22–29.

Paul Jaccard. 1901. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft cardinality+ ml: learning adaptive similarity functions for cross-lingual textual entailment. In *Proceedings of the Sixth International Workshop on Semantic Evaluation, SemEval*. ACL, pages 684–688.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2013. Softcardinality: Learning to identify directional cross-lingual entailment from cardinalities and smt. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*. volume 2, pages 34–38.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.

George Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

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.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics* 16(4):157–167.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos. *Proceedings of the American philosophical society* 96(4):452–463.