A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets

José Camacho-Collados, Mohammad Taher Pilehvar and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
{collados,pilehvar,navigli}@di.uniroma1.it

Abstract

Despite being one of the most popular tasks in lexical semantics, word similarity has often been limited to the English language. Other languages, even those that are widely spoken such as Spanish, do not have a reliable word similarity evaluation framework. We put forward robust methodologies for the extension of existing English datasets to other languages, both at monolingual and cross-lingual levels. We propose an automatic standardization for the construction of cross-lingual similarity datasets, and provide an evaluation, demonstrating its reliability and robustness. Based on our procedure and taking the RG-65 word similarity dataset as a reference, we release two high-quality Spanish and Farsi (Persian) monolingual datasets, and fifteen cross-lingual datasets for six languages: English, Spanish, French, German, Portuguese, and Farsi.

1 Introduction

Semantic similarity is a field of Natural Language Processing which measures the extent to which two linguistic items are similar. In particular, word similarity is one of the most popular benchmarks for the evaluation of word or sense representations. Applications of word similarity range from Word Sense Disambiguation (Patwardhan et al., 2003) to Machine Translation (Lavie and Denkowski, 2009), Information Retrieval (Hliaoutakis et al., 2006), Question Answering (Mohler et al., 2011), Text Summarization (Mohammad and Hirst, 2012), Ontology Alignment (Pilehvar and Navigli, 2014), and Lexical Substitution (McCarthy and Navigli, 2009).

However, due to the lack of standard multilingual benchmarks, word similarity systems had in the main been limited to the English language (Mihalcea and Moldovan, 1999; Agirre and Lopez, 2003; Agirre and de Lacalle, 2004; Strube and Ponzetto, 2006; Gabrilovich and Markovitch, 2007; Mihalcea, 2007; Pilehvar et al., 2013; Baroni et al., 2014), up until the recent creation of datasets built by translating the English RG-65 dataset (Rubenstein and Goodenough, 1965) into French (Joubarne and Inkpen, 2011), German (Gurevych, 2005), and Portuguese (Granada et al., 2014). And what is more, cross-lingual applications have grown in importance over the last few years (Hassan and Mihalcea, 2009; Navigli and Ponzetto, 2012; Franco-Salvador et al., 2014; Camacho-Collados et al., 2015b). Unfortunately, very few reliable datasets exist for evaluating cross-lingual systems.

This paper provides two contributions: Firstly, we construct Spanish and Farsi versions of the standard RG-65 dataset scored by twelve annotators with high inter-annotator agreements of 0.83 and 0.88, respectively, in terms of Pearson correlation, and secondly, we create fifteen cross-lingual word similarity datasets based on RG-65, covering six languages, by proposing an improved version of the approach of Kennedy and Hirst (2012) for the automatic construction of cross-lingual datasets from aligned monolingual datasets.

The paper is structured as follows. We first briefly review some of the major monolingual and cross-lingual word similarity datasets in Section 2. We then discuss the details of our procedure for the construction of the Spanish and Farsi word similarity datasets in Section 3. Section 4 provides the details of our algorithm for the automatic construction of the cross-lingual datasets. We report the results of the evaluation performed on the generated datasets in Section 5. Finally, we specify the released resources in Section 6, followed by concluding remarks in Section 7.

2 Related Work

Multiple word similarity datasets have been constructed for the English language: MC-30 (Miller and Charles, 1991), WordSim-353 (Finkelstein et al., 2002), MEN (Bruni et al., 2014), and Simlex-999 (Hill et al., 2014). The RG-65 dataset (Rubenstein and Goodenough, 1965) is one of the oldest and most popular word similarity datasets, and has been used as a standard benchmark for measuring the reliability of word and sense representations (Agirre and de Lacalle, 2004; Gabrilovich and Markovitch, 2007; Hassan and Mihalcea, 2011; Pilehvar et al., 2013; Baroni et al., 2014; Camacho-Collados et al., 2015a). The original RG-65 dataset was constructed with the aim of evaluating the degree to which contextual information is correlated with semantic similarity for the English language. Rubenstein and Goodenough (1965) reported an inter-annotator agreement of 0.85 for a subset of fifteen judges (no final inter-annotator agreement for the total fifty-one judges was calculated). The original English RG-65 has also been used as a base for different languages: French (Joubarne and Inkpen, 2011), German (Gurevych, 2005), and Portuguese (Granada et al., 2014). No inter-annotator agreement was calculated for the French version, while the German and Portuguese were reported to have the respective inter-annotator agreements of 0.81 and 0.71 in terms of average pairwise Pearson correlation. Our Spanish version of the RG-65 dataset reports a high inter-annotator agreement of 0.83, while the Farsi version achieves 0.88.

A few works have also focused on the construction of cross-lingual resources. Hassan and Mihalcea (2009) built two sets of cross-lingual datasets by translating the English MC-30 (Miller and Charles, 1991) and the WordSim-353 (Finkelstein et al., 2002) datasets into three languages. However, these datasets have several issues due to their construction procedure. The main problem arises from keeping the original scores from the English dataset in the translated datasets. For instance, the Spanish dataset contains the identical pair *mediodia-mediodia* with a similarity score of 3.42 (in the 0-4 scale). Furthermore, the datasets contain orthographic errors such as *despliege* and the previously mentioned *mediodia* (instead of *despliegue* and *mediodía*), and nouns translated into words with a different part of speech (e.g., *implement* from the English noun dataset MC-30 translated to the Spanish verb *implementar*). Additionally, the selection of the datasets was not ideal: MC-30 is a small subset of RG-65 and WordSim-353 has been criticized for its annotation scheme, which conflates similarity and relatedness (Hill et al., 2014).

Kennedy and Hirst (2012) proposed an automatic procedure for the construction of a French-English version of RG-65. We refine their approach by also dealing with some issues that may arise in the automatic process. Additionally, we provide an evaluation of the automatic procedure on different languages.

3 Building Monolingual Word Similarity Datasets

In this section we explain our methodology for the construction of the Spanish and Farsi versions of the English RG-65 dataset (Rubenstein and Goodenough, 1965). The methodology is divided into two main steps: First, the original English dataset is translated into the target language (Section 3.1) and then, the newly translated pairs are scored by human annotators (Section 3.2).

3.1 Translating from English to Spanish/Farsi

The translation of RG-65 from English to Spanish and Farsi was performed by, respectively, three English-Spanish and three English-Farsi annotators who were fluent English speakers and native speakers of the target language. The translation procedure was as follows. First, two annotators translated each English pair in the dataset into the target language. Then a third annotator checked for disagreements between the first two translators and picked the more appropriate translation among the two options.

Finally, all three translators met and performed a final check, with specific focus on the following two cases: (1) duplicate pairs in the dataset, and (2) pairs with repeated words. Our goal was to reduce these two cases as much as possible. A final adjudication was performed accordingly. We note that there remain three pairs with identical words in both Spanish and Farsi datasets, as no suitable translation could be found to distinguish the words in the English pair. For instance, the two words in the pair *midday-noon* translate to the same Spanish word *mediodía*.

English			Spanish			Farsi		
noon	string	0.04	mediodía	cuerda	0.00	ظهر	نخ	0.00
cemetery	woodland	0.79	cementerio	bosque	1.18	قبرستان	جنگل	0.50
mound	shore	0.97	loma	orilla	1.21	ماهور	ساحل	1.17
food	rooster	1.09	comida	gallo	1.54	غذا	خروس	1.00
bird	woodland	1.24	pájaro	bosque	1.67	پرنده	بیشه زار	1.79
glass	jewel	1.78	cristal	joya	1.96	شیشه	جواهر	1.29
bird	crane	2.63	pájaro	grulla	2.92	پرنده	درنا	2.83
autograph	signature	3.59	autógrafo	firma	3.46	امضا	امضا	4.00
automobile	car	3.92	automóvil	coche	3.92	خودرو	ماشین	3.88

Table 1: Sample word pairs from the English and the newly created Spanish and Farsi RG-65 datasets.

3.2 Scoring the dataset

Twelve native Spanish speakers were asked to evaluate the similarity for the Spanish translations. In order to obtain a more global distribution of judges, we included judges both both Spain and Latin America. As far as the Farsi dataset was concerned, twelve Farsi native speakers scored the newly translated pairs. The guidelines provided to the annotators were based on the recent SemEval task on Cross-Level Semantic Similarity (Jurgens et al., 2014), which provides clear indications in order to distinguish similarity and relatedness. The annotators were allowed to give scores from 0 to 4, with a step size of 0.5.

Table 1 shows example pairs with their corresponding scores from the English and the newly created Spanish and Farsi versions of the RG-65 dataset. As we can see from the table, the scores across languages are not necessarily identical, with small, in a few cases significant, differences between the corresponding scores. This is due to the fact that associated senses with words do not hold one-to-one correspondence across different languages. This renders the approach of Hassan and Mihalcea (2009) insufficiently accurate for handling these differences.

4 Automatic Creation of Cross-lingual Similarity Datasets

In this section we present our automatic method for building cross-lingual datasets. Although being targeted at building semantic similarity datasets, the algorithm is task-independent, so it may also be used for any task which measures any kind of relation between two linguistic items in a numerical way.

Kennedy and Hirst (2012) proposed a method which exploits two aligned monolingual word similarity datasets for the construction of a French-English cross-lingual dataset. We followed their initial idea and proposed a generalization of the approach which would be capable of automatically constructing reliable cross-lingual similarity datasets for any pair of languages.

Algorithm. Algorithm 1 shows our procedure for constructing a cross-lingual dataset starting from two monolingual datasets. Note that the pairs in the two monolingual datasets should be previously aligned. Specifically, we refer to each dataset D as $\{P_D, S_D\}$, where P_D is the set of pairs and S_D is a function mapping each pair in P_D to a value on a similarity scale (0-4 for RG-65). For each two aligned pairs *a-b* and *a'-b'* across the two datasets, if the difference in the corresponding scores is greater than a quarter of the similarity scale size (1.0 in RG-65), the pairs are not considered (line 7) and therefore discarded. Otherwise, two new pairs *a-b'* and *a'-b* are created with a score equal to the average of the two original pairs' scores (lines 8-11 and 15-18). In the case of repeated pairs, we merge them into a single pair with a similarity equal to their average score (lines 12-14 and lines 19-21).

By following this procedure we created fifteen cross-lingual datasets based on the RG-65 word similarity datasets for English, French, German, Spanish, Portuguese, and Farsi. Table 2 shows

Algorithm 1 Automatic construction of cross-lingual similarity datasets

Input: two aligned datasets $D = \{P_D, S_D\}$ and $D' = \{P_{D'}, S_{D'}\}$, where P_X is the set of pairs in dataset X and S_X is the mapping of these pairs to their corresponding scores.

Output: a cross-lingual semantic similarity dataset $C = \{P_C, S_C\}$

$P_C \leftarrow \emptyset$
Define *Cnt*, which counts how many times an output cross-lingual pair is repeated
for each aligned pairs $(a, b) \in P_D, (a', b') \in P_{D'}$
  $score = S_D(a, b)$
  $score' = S_{D'}(a', b')$
  $avg\_score = (score + score')/2$
  if $|score - score'| \leq size(sim\_scale)/4$ **then**
    if $(a, b') \notin P_C$ **then**
      $P_C \leftarrow P_C \cup \{(a, b')\}$
      $S_C(a, b') = avg\_score$
      $Cnt(a, b') = 1$
    else
      $S_C(a, b') = \frac{(S_C(a,b') \times Cnt(a,b')) + avg\_score}{Cnt(a,b')+1}$
      $Cnt(a, b') + +$
    if $(a', b) \notin P_C$ **then**
      $P_C \leftarrow P_C \cup \{(a', b)\}$
      $S_C(a', b) = avg\_score$
      $Cnt(a', b) = 1$
    else
      $S_C(a', b) = \frac{(S_C(a',b) \times Cnt(a',b)) + avg\_score}{Cnt(a',b)+1}$
      $Cnt(a', b) + +$
return $\{P_C, S_C\}$

the number of word pairs for each cross-lingual dataset. Note that there is not a single pair of languages whose total count reaches the maximum number of possible word pairs, i.e., 130. This is due, on the one hand, to language peculiarities resulting in some pairs having significant score difference across languages (higher than 1 on the 0-4 scale), and, on the other hand, to the repetition of some pairs occurring as a result of the automatic creation process, a problem which is handled by our algorithm.

Table 3 shows sample pairs with their corresponding similarity scores from four of the cross-lingual datasets: Spanish-English, Spanish-French, Spanish-German, and English-Farsi. These cross-lingual datasets are constructed on the basis of our newly-generated Spanish and Farsi monolingual datasets (see Section 3). The quality of these four datasets is evaluated in Section 5.2.

	FR	**DE**	**ES**	**PT**	**FA**
EN	100	125	126	120	120
FR	-	96	103	92	100
DE	-	-	125	118	122
ES	-	-	-	113	122
PT	-	-	-	-	122

Table 2: Number of word pairs for each cross-lingual dataset (EN: English, FR: French, DE: German, ES: Spanish, PT: Portuguese, FA: Farsi).

5 Evaluation

5.1 Spanish and Farsi Monolingual Datasets

The inter-annotator agreements according to the average pairwise Pearson correlation among the judges for the newly created Spanish and Farsi datasets are, respectively, 0.83 and 0.88, which may be used as upper bounds for evaluating automatic systems. Our further analysis revealed that for both datasets no annotator obtained an average Pearson correlation with the rest of the annotators lower than 0.80, which attests to the reliability of our judges and guidelines. The German (Gurevych, 2005) and Portuguese (Granada et al., 2014) versions of the RG-65 dataset reported a lower inter-annotator agreement of 0.81 and 0.71, respectively, whereas the original English RG-65 (Rubenstein and Goodenough, 1965) reported an inter-annotator agreement of 0.85 for a subset of fifteen judges. As also mentioned earlier, the French version (Joubarne and Inkpen, 2011) did not report any inter-annotator agreement.

5.2 Cross-lingual Datasets

Along with the monolingual evaluation, we also performed an evaluation on four of the automatically created cross-lingual datasets. The evaluated language pairs were Spanish-English, Spanish-French, Spanish-German, and English-Farsi. In each case a proficient speaker of both languages was selected to carry out the evaluation. The Pearson correlations of the human judges with the automatically generated scores were 0.89 for Spanish-English, 0.94 for Spanish-French, 0.91 for Spanish-German, and 0.92 for English-Farsi, showing the reliability of our cross-lingual dataset creation process and reinforcing the quality of the newly created monolingual datasets.

ES	EN		ES	FR	
monje	assylum	0.41	cuerda	midi	0.00
bosque	bird	1.46	chico	sage	0.54
viaje	car	1.74	comida	coq	1.08
hermano	monk	2.25	hermano	gars	1.71
pollo	rooster	3.36	grulla	oiseau	2.67
cementerio	graveyard	3.94	chaval	garçon	3.88

ES	DE		EN	FA	
orilla	autogramm	0.02	mound	اجاق	0.07
caldera	werkzeug	1.04	coast	جنگل	1.03
pájaro	wald	1.65	journey	ماشین	1.53
coche	fahrt	2.34	food	میوه	2.56
cojín	kissen	3.21	stove	کوره	3.10
colina	berg	3.61	car	خودرو	3.90

Table 3: Example pairs from the Spanish-English, Spanish-French, Spanish-German, and English-Farsi cross-lingual word similarity datasets (EN: English, FR: French, DE: German, ES: Spanish, FA: Farsi).

6 Release of the Resources

All the resources obtained as a result of this work are freely downloadable and available to the research community at `http://lcl.uniroma1.it/similarity-datasets/`.

Among these resources we include the newly created Spanish and Farsi word similarity datasets, together with the annotation guidelines used during the creation of the datasets. Our algorithm for the automatic creation of cross-lingual datasets (Algorithm 1) is provided as an easy-to-use Python script. Finally, we also release the fifteen cross-lingual datasets built by using this algorithm, including Spanish, English, French, German, Portuguese, and Farsi languages.

7 Conclusion

We developed two versions of the standard RG-65 dataset in Spanish and Farsi. We also proposed and evaluated an automatic method for creating cross-lingual semantic similarity datasets. Thanks to this method, we release fifteen cross-lingual datasets for pairs of languages including English, Spanish, French, German, Portuguese, and Farsi. All these datasets are intended for use as a standard benchmark (as RG-65 already is for the English language) for evaluating word or sense representations and, more specifically, word similarity systems, not only for languages other than English, but also across different languages.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

We would like to thank Leyli Badiei, Núria Bel, Alicia Burga, Marcos Camacho, Ángela Collados, José Cuenca, Luis Espinosa, Darío Garigliotti, Afsaneh Hojjat, Ignacio Iacobacci, Amin Lak, Montserrat Marimon, Javier Martínez, Ali Orang, Ana Osorio, Lluís Padró, Abdolhamid, Razieh, and Zahra Pilehvar, Mohammad Sadegh Rasooli, Molood Sadat Safavi, Mohammad Shojafar, Hossein Soleimani, and Fatemeh Torabi Asr for their help in the construction and evaluation of the word similarity datasets. We would also like to thank Jim McManus for his comments on the manuscript.

References

Eneko Agirre and Oier Lopez de Lacalle. 2004. Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of LREC*, pages 1123–1126, Lisbon, Portugal.

Eneko Agirre and Oier Lopez. 2003. Clustering WordNet word senses. In *Proceedings of Recent Advances in Natural Language Processing*, pages 121–130, Borovets, Bulgaria.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247, Baltimore, Maryland.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015a. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577, Denver, USA.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015b. A unified multilingual semantic representation of concepts. In *Proceedings of ACL*, Beijing, China.

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 Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the* 14^{th} *Conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 414–423, Gothenburg, Sweden.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606–1611, Hyderabad, India.

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*, pages 170–175. São Carlos/SP, Brazil.

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of IJCNLP*, pages 767–778. Jeju Island, Korea.

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

Samer Hassan and Rada Mihalcea. 2011. Semantic Relatedness Using Salient Semantic Analysis. In *Proceedings of AAAI*, pages 884–889, San Francisco, USA.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. arXiv:1408.3456.

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.

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*, pages 216–221. Perth, Australia.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the* 8^{th} *International Workshop on Semantic Evaluation (SemEval 2014), in conjunction with COLING 2014*, pages 17–26, Dublin, Ireland.

Alistair Kennedy and Graeme Hirst. 2012. Measuring semantic relatedness across languages. In *Proceedings of xLiTe: Cross-Lingual Technologies Workshop at the Neural Information Processing Systems Conference*, pages 1–6, Lake Tahoe, USA.

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

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

Rada Mihalcea and Dan Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of AAAI*, pages 461–466, Orlando, Florida, USA.

Rada Mihalcea. 2007. Using Wikipedia for automatic Word Sense Disambiguation. In *Proc. of NAACL-HLT-07*, pages 196–203, Rochester, NY.

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

Saif Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *CoRR*, 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 ACL*, pages 752–762, Portland, Oregon.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelRelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of AAAI*, pages 108–114, Toronto, Canada.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for Word Sense Disambiguation. In *Proceedings of CICLing*, pages 241–257, Mexico City, Mexico.

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

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of ACL*, pages 1341–1351, Sofia, Bulgaria.

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

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of AAAI*, pages 1419–1424, Boston, USA.