Measuring Semantic Relatedness Across Languages

Alistair Kennedy
Department of Computer Science
University of Toronto
Toronto, Canada
akennedy@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, Canada
gh@cs.toronto.edu

Abstract

Measures of Semantic Relatedness are well established in Natural Language Processing. Their purpose is to determine the degree of relatedness between two words without specifying the nature of their relationship. Most of these measures work only between pairs of words in a single language. We propose a novel method of measuring semantic relatedness between pairs of words in two different languages. This method does not use a parallel corpus but is rather seeded with a set of known translations. For evaluation we construct a data set of cross-language word pairs with similarity scores from French and English versions of Rubenstein & Goodenough's data set. We found that our new cross-language measure correlates more closely with averaged human scores than our unilingual baselines.

1 Introduction

There are two dominant methods of measuring semantic relatedness between pairs of words within a single language. One method is to use existing thesauri and determine relatedness of words by how close together they appear within the thesaurus. The second method is to use the distributional similarity of the words. If two words frequently appear in the same contexts then they are likely to be similar in meaning [1]. Often the context of a word is taken to mean the set of surrounding words. We enhance these distributional Measures of Semantic Relatedness (MSRs) to work across languages.

There are two main contributions in this paper. One is a novel Cross-Language Measure of Semantic Relatedness (CL-MSR) that works between two languages without the need for a parallel corpus. The second is a new cross-lingual evaluation data set in the style of Rubenstein & Goodenough [2] for words in French and English. Ultimately, work of this kind will have many applications. It could be used as part of a system for parallel corpus alignment for machine translation, or for cross-language information retrieval, to name just a couple.

1.1 Previous Work

There has been much research on learning translations of words between languages. Our goal is slightly different, as we aim to measure degrees of relatedness; nonetheless, our system should also be usable for finding translations.

Although some related measures have been built from parallel corpora [3, 4], we would like to avoid using such a resource as they are expensive to construct and do not exist for all language pairs. Graph-based approaches have had some success in learning translations [5, 6, 7, 8]. Graphs are built in two languages where the nodes are words and the edges indicate a high degree of relatedness between two words. A set of known translations is used to map together nodes between these two graphs. New translations can be inferred from the graph structure.

The method closest to what we are proposing can be found in the works of Haghighi et al. [9] and Daumé & Jagarlamudi [10]. A set of known translations is used to seed a method of mapping contexts between languages and also mapping between substrings from each word using Canonical Correlation Analysis (CCA). This method maps contexts and also maps substrings of characters to learn translations.

Another method is to use a bilingual lexicon to map words acting as contexts from one language to another [11, 12]. Similarly Explicit Semantic Analysis (ESA) [13] has been enhanced by using the cross-language links of Wikipedia to map words from different languages into the same vector space made up of Wikipedia articles [14, 15, 16]. Mohammad & Hirst [17] use a German-English bilingual lexicon to map contexts of words in German into concepts in the Macquarie Thesaurus. None of these methods actually learn from a known set of translations, as the only contexts with explicitly labeled translations can be used. Contexts made up of short phrases might be very difficult to map between languages. Our method is more general in that any features should be usable.

2 Methodology

The theory behind distributional MSRs is that if two words tend to appear in the same contexts, then they are more likely to be semantically related. One obvious problem with applying this to a cross-lingual domain is that contexts differ by being in different languages (translations will rarely share the same contexts). We create a mapping between contexts of two different languages with the help of a set of known translations. If two contexts in two different languages tend to contain pairs of words that are known to be translations of each other then we can infer that these contexts are related too and should be mapped together. This way contexts from one language can be mapped to multiple contexts in another language.

2.1 Building Term-Context Matrices

For corpora we selected the French and English editions of Wikipedia (downloaded in early July 2012). We tagged these corpora with the French and English Stanford POS taggers [18, 19]. In these experiments we measured only semantic relatedness between nouns, leaving verbs and adjectives for future work. As contexts we selected all nouns, verbs and adjectives that appear within a window of 5 words of each noun. We decided to include only words and contexts that appeared at least 100 times in our matrix. This produced an English term-context matrix with 79,221 nouns and 143,645 contexts, while the French matrix had 33,646 nouns and 71,704 contexts.

Raw co-occurrence counts are generally noisy and do not produce a very good term-context matrix for building a MSR, so we used Pointwise Mutual Information (PMI) to re-weight every term-context pair. To construct and re-weight our matrix, we used a modified version of the Generalized Term Semantics (GenTS) system [20]. To measure the distance between two word vectors we used cosine similarity.

2.2 Mapping Matrices between Languages

Next we acquired a set of known translations. We extracted translations using the French Wordnet Libre du Francais (WOLF) v0.1.5 [21] and Princeton WordNet v2.0 [22]. The synsets of each resource are aligned, allowing for one to easily extract translations. In total this produced 29,826 noun translations, 10,400 of which contained one word in each of the French and English matrices. These translations will be referred to as the training data.

Using this translation set we found an association between the English and French contexts. Associations were measured only between two contexts of the same part-of-speech. We created a confusion matrix of observed co-occurrences to measure the association between two contexts c_e and c_f in English and French. We use $w \in c$ to denote a word found in a context, see equation 1.

$$\begin{array}{c} \\ w_f \in c_f \\ w_f \notin c_f \end{array} \begin{array}{c} \begin{array}{cc} w_e \in c_e & w_e \notin c_e \end{array} \\ \begin{bmatrix} O_{0,0} & O_{0,1} \\ O_{1,0} & O_{1,1} \end{bmatrix} \end{array} \quad (1) \qquad E_{i,j} = \frac{\sum_y O_{i,y} \ \sum_x O_{x,j}}{\sum_{x,y} O_{x,y}} \quad (2) \qquad PMI = \log \frac{O_{0,0}}{E_{0,0}} \quad (3)$$

English			French			Bilingual		
gem	jewel	3.94	joyau	bijou	3.22	gem	bijou	3.58
midday	noon	3.94	midi	dîner	2.17	*midday*	*dîner*	*3.05*
cemetery	mound	1.69	cimetière	monticule	0.22	*cemetery*	*monticule*	*0.96*
car	journey	1.55	auto	voyage	0.33	*car*	*voyage*	*0.94*
noon	string	0.04	midi	ficelle	0.00	noon	ficelle	0.02
cord	smile	0.02	corde	sourire	0.00	cord	sourire	0.01

Table 1: Examples of similarity measurements on a scale of 0 to 4 from the English, French, and bilingual versions of Rubenstein and Goodenough's data set. Italics indicates pairs that were discarded from the evaluation.

The counts of English-French translations $\langle w_e, f_e \rangle$ were as follows;

- $O_{0,0}$ number of translations $\langle w_e, f_e \rangle$ where $w_e \in c_e$ and $w_f \in c_f$;
- $O_{0,1}$ number of translations $\langle w_e, f_e \rangle$ where $w_e \in c_e$ but $w_f \notin c_f$;
- $O_{1,0}$ number of translations $\langle w_e, f_e \rangle$ where $w_f \in c_f$ but $w_e \notin c_e$;
- $O_{1,1}$ number of translations $\langle w_e, f_e \rangle$ where $w_e \notin c_e$ and $w_f \notin c_f$.

When counting translations, we took the weight of each word into account. The weight of a translation $\langle w_e, w_f \rangle$ was found by taking the product of the PMI score of w_e in c_e and the PMI score of w_f in c_f from the term-context matrix. In the case of $O_{0,1}, O_{1,0}, O_{1,1}$ we took the sum of the products of each translation with every other context in the entire matrix.

From this matrix of observed counts we calculated expected values (equation 2) corresponding to every observed value in the matrix. Then we calculated the PMI (equation 3) to weight the dependency between the two contexts and generated a large matrix of weighted mappings between the French and English contexts (the notation here is taken from [23]). This gave us a translation matrix recording the PMI weight of the association between the context pairs in each language.

Next we mapped the French matrix into the context-space of the English matrix. This process was done separately for each vector representing each French word. For each context c_f that a French word appears in, the weight of c_f was distributed across all English contexts that it is mapped to. The PMI weights in the translation matrix were normalized, so that the sum of the weights of all English contexts that a French context was mapped into is equal to the original French context.

One problem is that this will create an extremely dense matrix. If all mappings are used, the new French matrix can become so large that it would require over a hundred gigabytes of RAM to load. There are two parameters that can be adjusted to reduce the size of this matrix. One parameter is to keep only mappings if their score is beyond some threshold; we chose 0.05. The second is to set a threshold for the minimum PMI score between the English and French contexts. We experimented with 5 different PMI thresholds, 1.0, 2.0, 3.0, 4.0, and 5.0.

The next step is simply to merge the two matrices. This is fairly straightforward; however, we took care to label each word as being either French or English.

3 Evaluating Cross-Language Measures of Semantic Relatedness

3.1 Evaluation Data

Often data sets in the style of Rubenstein and Goodenough are used to evaluate MSRs within a single language. Their data set contains a list of 65 of English word pairs with human-assigned similarity scores, ranging from 0 to 4, averaged between a number of human annotators. Some translations of this data set have been created for other languages, including German [24] and French [25] versions. Examples of such pairs of words and their scores in English and French can be seen in Table 1.

A cross-language version would contain pairs where the first word comes from the French set and the second word comes from the English set and vice versa. New values for each word pair will have to be manually validated; however, one might expect they will be fairly close to those scores provided with the unilingual data sets. With this intuition in mind, we created a new cross-lingual

Measure	Pearson	Spearman	Kendall
English-PMI	0.192	0.143	0.104
French-PMI	0.117	0.013	0.011
CL-MSR-1.0	0.295	0.320	0.224
CL-MSR-2.0	**0.301**	**0.333**	**0.225**
CL-MSR-3.0	0.294	0.332	0.224
CL-MSR-4.0	0.258	0.312	0.213
CL-MSR-5.0	0.185	0.299	0.206

Table 2: Correlation scores on bilingual Rubenstein and Goodenough style data set.

data set. All pairs from the French and English versions, where their scores are within ± 1 of each other, were used. From this, we created a triple of $\langle word_{en}, word_{fr}, score_{avg} \rangle$ where the two words are from corresponding French and English pairs and $score_{avg}$ is the average of the scores from the French and English versions. Every corresponding pair in French and English can produce two cross-lingual pairings giving us a set of 100 cross-language word pairs with their averaged similarity. Some examples are shown in Table 1 where the instances in italics are pairs that had to be thrown out as the unilingual scores were not within ± 1. We removed every pair of words used in our evaluation data from the training set before mapping English and French contexts together.

3.2 Results

Next we evaluate the MSRs on our cross-language data set. For this we created five versions of the CL-MSR, where PMI thresholds of $1.0, 2.0, \ldots, 5.0$ were used to determine which contexts should be mapped together. We also tested two baselines; MSRs using the PMI-weighted English and French matrices. The baseline correlations for these experiments should not actually be zero as French and English have a large number of cognates – words spelled similarly or identically with similar meaning. Therefore, even a unilingual MSR should be able to perform on part of this data set. We evaluate the correlation with Pearson's product-moment correlation, Spearman's rank correlation ρ and Kendall's rank correlation τ. Since the scores for each word pair are averages, we believe that the rank-based correlations of Spearman and Kendall are more informative than Pearson's.

The results in Table 2 show that even the baseline systems have some correlation. The English baseline outperforms the French one, perhaps because the English term-context matrix contains many more words than the French version. The CL-MSRs were much more successful than the baselines on this data set. The best correlations came when using a threshold of 2.0 for mapping between contexts; however, thresholds of 1.0 and 3.0 were not noticeably different. As the threshold increased to 5.0, the correlations decreased more noticeably.

4 Conclusion

We have presented a novel method of mapping contexts from one language to another using a set of known translations. Although we used a sliding window of POS-tagged unigrams, any context could be used in this situation. Our CL-MSR is general enough that it could be ported to other language pairs, provided there is a suitably sized training set. As future work we would like to examine how many translation pairs are actually needed and how corpus imbalance affects the CL-MSR.

For our evaluation we produced a new cross-language word-similarity data set. Our best CL-MSR shows a noticeable improvement over the unilingual baselines on three different correlation measures. There are a number of other evaluations that could be tried, such as selecting the correct translation of a word from a set of candidates. It could also be useful to build a larger cross-language word-similarity data set with new manually assigned scores. Likewise our CL-MSR could be applied to larger problems such as cross-language information retrieval or machine translation

Acknowledgments

This work was financially supported by the Defense Advanced Research Projects Agency, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto.

References

[1] John Rupert Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32, 1957.

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

[3] António Ribeiro, José Gabriel Pereira Lopes, and João Mexia. Extracting equivalents from aligned parallel texts: Comparison of measures of similarity. In *Proceedings of the International Joint Conference, 7th Ibero-American Conference on AI: Advances in Artificial Intelligence*, IBERAMIA-SBIA '00, pages 339–349, London, UK, 2000. Springer-Verlag.

[4] Hans Hjelm. Identifying cross language term equivalents using statistical machine translation and distributional association measures. In Joakim Nivre, Kadri Muischnek Heiki-Jaan Kaalep, and Mare Koit, editors, *Proceedings of NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics*, pages 97–104, Tartu, Estonia, May 2007.

[5] Oren Etzioni, Kobi Reiter, Stephen Soderl, and Marcus Sammer. Lexical translation with application to image search on the web. In *Proceedings of the 11th Machine Translation Summit*, pages 175–182, 2006.

[6] Lukas Michelbacher, Florian Laws, Beate Dorow, Ulrich Heid, and Hinrich Schütze. Building a cross-lingual relatedness thesaurus using a graph similarity measure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[7] Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, and Jeff Bilmes. Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174(9-10):619–637, June 2010.

[8] Tiziano Flati and Roberto Navigli. The CQC algorithm: Cycling in graphs to semantically enrich and enhance a bilingual dictionary. *Journal of Artificial Intelligence Research*, 43:135–171, 2012.

[9] Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of The Association of Computational Linguistics: Human Language Technologies*, pages 771–779, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[10] Hal Daumé, III and Jagadeesh Jagarlamudi. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 407–412, Portland, OR, 2011. Association for Computational Linguistics.

[11] Reinhard Rapp. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526, College Park, Maryland, 1999. Association for Computational Linguistics.

[12] Nikesh Garera, Chris Callison-Burch, and David Yarowsky. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 129–137, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[13] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[14] Samer Hassan and Rada Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, (EMNLP) 2009*, pages 1192–1201. ACL, 2009.

[15] Philipp Sorg and Philipp Cimiano. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In Helmut Horacek, Elisabeth Métais, Rafael Muñoz, and Magdalena Wolska, editors, *Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB)*, pages 36–48. Springer, June 2009.

[16] Philipp Sorg and Philipp Cimiano. Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data Knowledge Engineering*, 74:26–45, April 2012.

[17] Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 571–580, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[18] Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics – Volume 13*, EMNLP '00, pages 63–70, Hong Kong, 2000. Association for Computational Linguistics.

[19] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Edmonton, Canada, 2003. Association for Computational Linguistics.

[20] Alistair Kennedy and Stan Szpakowicz. Supervised distributional semantic relatedness. In Petr Sojka, Aleš Horák, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue, 15th International Conference, TSD 2012*, pages 207–214, Brno, Czech Republic, September 2012. Springer.

[21] Benoît Sagot and Darja Fišer. Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Marocco, 2008.

[22] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA; London, May 1998.

[23] Stefan Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2004.

[24] Iryna Gurevych. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, IJCNLP'05, pages 767–778, Berlin, Heidelberg, 2005. Springer-Verlag.

[25] Colette Joubarne and Diana Inkpen. Comparison of semantic similarity for different languages using the Google N-gram corpus and second-order co-occurrence measures. In *Canadian Conference on Artificial Intelligence*, pages 216–221, 2011.