Comparison of Semantic Similarity for Different Languages using the Google n-gram Corpus and Second-Order Co-Occurrence Measures

Colette Joubarne and Diana Inkpen

School of Information Technology and Engineering
University of Ottawa, ON, Canada, K1N 6N5
mjoub063@uottawa.ca, diana@site.uottawa.ca

Abstract. Despite the growth in digitization of data, there are still many languages without sufficient corpora to achieve valid measures of semantic similarity. If it could be shown that manually-assigned similarity scores from one language can be transferred to another language, then semantic similarity values could be used for languages with fewer resources. We test an automatic word similarity measure based on second-order co-occurrences in the Google n-gram corpus, for English, German, and French. We show that the scores manually-assigned in the experiments of Rubenstein and Goodenough's for 65 English word pairs can be transferred directly into German and French. We do this by conducting human evaluation experiments for French word pairs (and by using similarly produced scores for German). We show that the correlation between the automatically-assigned semantic similarity scores and the scores assigned by human evaluators is not very different when using the Rubenstein and Goodenough's scores across language, compared to the language-specific scores.

1 Introduction

Semantic similarity refers to the degree to which two words are related. Measures of semantic similarity are useful for techniques such as information retrieval, data-mining, question answering, and text summarization. As indicated by Irene Cramer [2] many studies such as question answering, topic detection, and text summarization,, rely on semantic relatedness measures based on word nets and/or corpus statistics as a resource. However, these approaches require large and various amounts of corpora, which are often not available for languages other than English. If it could be shown that measures of semantic similarity have a high correlation across languages, then values for semantic similarity could be assigned to translated n-grams; thus enabling one set of values to be applied to many languages.

Determining semantic similarity is routinely performed by humans, but it is a complex task for computers. Gabrilovich and Markovitch [3] point out that humans do not judge text relatedness only based on words. Identification of similarity involves reasoning at a much deeper level that manipulates concepts. Measures of similarity for humans are based on the larger context of their background and experience. Language is not merely a different collection of characters, but is founded on a culture that impacts the variety and subtlety of semantically similar words. For example, in French, often described as the “language of love”, the verbs “to like” and “to love” both translate to “aimer”. The word pair “cock, rooster” from Rubenstein and Goodenough [10] translate to “coq, coq” in French, and “Hahn, Hahn” in German.

Rubenstein and Goodenough [10] defined the baseline for the comparison of semantic similarity measures. However, the fact that translation is not a 1:1 relation introduces difficulty in the use of a baseline. Understanding whether it is possible to use translated words to measure semantic similarity using corpora from another language is the goal of this experiment.

2 Related Work

Automatically assigning a value to the degree of semantic similarity between two words has been shown to be quite difficult [5]. Rubenstein and Goodenough [10] presented human subjects with 65 noun pairs and asked them how similar they were on a scale from 0 to 4. Miller and Charles [8] took a subset of this data (30 pairs) and repeated this experiment. Their results were highly correlated (97%) to those of the previous study.

Semantic similarity is a fundamental task in Natural Language Processing, therefore many different approaches to automate measures of semantic similarity of words have been studied. Jarmasz and Szpakowicz, [7] used a computerized version of Roget’s Thesaurus to calculate the semantic distance between the word pairs. They achieved correlation of 0.82 with Rubenstein and Goodenough’s [10] results. Budanitsky and Hirst [1] compared 5 different measures of semantic similarity based on WordNet. They found that when comparing the correlation of each measure with Rubenstein and Goodenough’s [10] human evaluator scores, the difference between the automatic measures was small (within 0.05). Islam and Inkpen [6] introduced Second Order Co-occurrence PMI as a measure of semantic similarity, and achieved results with a 0.71 correlation to Rubenstein and Goodenough [10] when measured using the British National Corpus (BNC)[a].

Hassan and Mihalcea [4] use the interlanguage links found in Wikipedia to produce a measure of relatedness using explicit semantic analysis. They achieved a correlation with Miller and Charles [8] word pairs between 0.32 and 0.50 for Spanish, Arabic and Romanian. Not surprisingly, they found that better results were achieved for languages with a larger Wikipedia. Mohammad et al [9] proposed a new method to determine semantic distance combining text from a language, such as German, which has fewer corpora available, with a knowledge source in a language with large corpora available, such as English. They combined German text with an English thesaurus to create cross-lingual distributional profiles of concepts to achieve a correlation of 0.81 with Rubenstein and Goodenough’s word pairs [10].

Typically, two approaches have been used to solve multilingual problems, rule-based systems and statistical learning from parallel corpora. Rule-based systems usually have low accuracy, and parallel corpora can be difficult to find. Our approach will be to use manual translation and language-specific corpora, in order to measure

[a] http://www.natcorp.ox.ac.uk/

and compare semantic similarity for English, French and German, using second-order co-occurrence.

3 Data

The data used was the Google n-gram corpus, which included n-grams (n=1-5) generated from roughly 100 billion word tokens from the web for each language. Only the unigrams, bigrams and 5-grams were used for this project. Since the purpose is to compare the semantic similarity of nouns only, and to compare results achieved on the same data, it was decided that removal of non-alphabetic characters and stemming of plurals was sufficient for our purposes.[b]

The word pairs were taken from Rubenstein and Goodenough [10] and translated into French using a combination of Larousse French-English dictionary, Le Grand dictionnaire terminologique, maintained by the Office quebecois de la langue francaise, a couple of native speakers and a human translator. In some cases where the semantic similarity of the word pair was high, the direct translation of each word in the word pair resulted in the same word. In these cases the pair was left out completely. The semantic similarity of the translated words was then evaluated by human judges.

The 18 evaluators, who had French as their first language, were asked to judge the similarity of the French word pairs. They were instructed to indicate, for each pair, their opinion of how similar in meaning the two words are on a scale of 0-4, with 4 for words that mean the same thing, and 0 for words that mean completely different things. The results were averaged over the 18 responses (with the exception of three word pairs, where the respondents left their scores blank, so these were only averaged over 17). For 71% of the word pairs there was good agreement amongst the evaluators, with over half of the respondents agreeing on their scores; however in 23% of the cases, there was high disagreement with scores ranging from 0-4. The results can be seen in Appendix A[c], which presents the words pairs for the three languages used in our study together with the similarity scores according to human judges.

The German translation of the word pairs, including human evaluation of similarity, was borrowed from Mohammad et al [9]. Some of the word pairs do not match exact translations. Since the focus of their study was on the comparison between scores from human evaluators and automated results, they addressed the issue of semantically similar words resulting in identical words during translation, by choosing another related word.

A comparison of the frequencies for similarity values amongst all evaluators for each language, presented in Table 1, shows that the English and German scores are similarly distributed, whereas the French scores are more heavily weighted around a score of 0 and 1.

[b] Stopword removal and stemming was performed during further research, but it was found that results were significantly worse for stopword removal and stemming, and relatively unchanged for stopword removal alone. Stopword lists were taken from Multilingual Resources at University of Neuchatel. The Lingua stemming algorithms was used.

[c] Available at http://www.site.uottawa.ca/~mjoub063/wordsims.htm

Table 1: Frequency of similarity scores

Similarity Score	Frequency		
	English	German	French
0	0	4	15
1	25	19	23
2	12	16	5
3	8	4	10
4	20	22	12

4 Methodology

Unigram and bigram counts were taken directly from the 1-gram and 2-gram files, taking into account characters and accents in the French and German alphabets. Second order counts were generated from the 5-gram data.

Two measures of semantic similarity were used, point-wise mutual information and second order co-occurrence point-wise mutual information. These measures were calculated for each set of word pairs, and compared to the baseline measures from the original data set, as well as the new values generated by human evaluators.

Point-wise mutual information (PMI) measure is a corpus-based measure, as opposed to a dictionary-based measure of semantic similarity. PMI measures the more general sense of semantic relatedness where two words are related by their proximity of use without necessarily being similar. The PMI score between 2 words w_1 and w_2 is defined as the probability of the 2 words appearing together divided by the probability of each word occurring separately. PMI was chosen because it scales well to larger corpora, and it has been shown to have the highest correlation amongst corpus-based measures [6].

Second order co-occurrence PMI (SOC-PMI) is also a corpus-based measure that determines a measure of semantic relatedness, based on how many words appear in the neighbourhood of both words. The SOC-PMI score between 2 words w1 and w2 is defined as the probability of word y appearing with w_1 and of y appearing with $w_{2,}$ within a given window in separate contexts. SOC-PMI was chosen because it fits well with the Google n-gram corpora. The frequencies for a window of size 5 are easily obtained from the 5-gram counts. The formula can be found in Islam and Inkpen [6].

5 Results

The PMI and SOC-PMI scores were calculated for each set of word pairs and compared to both the scores collected by Rubenstein and Goodenough [10] and the language specific scores collected from human evaluators (see Table 2).

Table 2: Pearson correlation of calculated PMI and SOC-PMI scores with R&G scores and new human evaluator scores

Language	vs. R&G		vs. Evaluators	
	PMI	SOC-PMI	PMI	SOC-PMI
English	0.41	0.61	n/a	n/a
French	0.34	0.19	0.29	0.17
German	0.40	0.27	0.47	0.31

6 Discussion

Our best correlation of 0.61 for the English SOC-PMI is not as good as that achieved by Islam and Inkpen [6]. However, their correlation of 0.73 was achieved using the BNC. The higher results could possibly be explained by the lack of noise in the BNC (discussion of noise issues found in Google n-gram corpus appears in Section 7), as well as the ability to use a larger window than supported by the Google 5-grams.

The correlation of the SOC-PMI scores and the original scores was slightly lower than for the human scores for the German word pairs, and slightly higher for the French ones. Almost 2/3 of the French and German word pairs had a SOC-PMI of 0. This is reflected in the poor correlation values and is likely due to the fact that the French and German corpora were approximately 1/10 the size of the English corpus.

7 Conclusion and Future Work

Given the lack of data for over 2/3 of the French and German pairs, it is not possible to make any claims with any certainty; however, since the results were not significantly improved by using language specific human evaluation, the results do suggest that it might be possible to transfer semantic similarity across languages. While further work needs to be done to confirm our hypothesis, we have produced a set of human evaluator scores for French which can be used for future work.

Although results were improved from earlier work, given the larger corpora for English, it appears that larger French and German corpora are still required to draw any significant conclusions. The Google n-gram corpora, for both French and German, contain approximately 13 billion tokens each; however, many of these tokens are not words. There are strings of repeating combinations of letters and many instances of multiple words in one token. For example, there are roughly 500-1000 tokens containing “abab” or “cdcd” and every other combination. There are 2000 occurrences of “voyageurdumonde” and 5000 of “filleougarcon”. Future work of this type with the Google n-grams should consider using a dictionary to filter out these kinds of tokens.

Another approach would be to select words that are common in all of the languages of interest, and that result in unique word pairs after translation. A new baseline would have to be created. This would require some study of word frequencies, and effort being spent in having the semantic similarity of the word pairs evaluated by human evaluators.

Budanitsky and Hirst [1] suggest a different approach. In their comparison of 5 different measures of semantic similarity, they suggest that comparing only to human evaluator scores is not a sufficient comparison, and that what we are really interested in is the relationship between the concepts for which the words are merely surrogates; the human judgments that we need are of the relatedness of word senses, not for words. They attempt to define such an experiment, and find that the effectiveness of the 5 measures varies considerably when compared this way. The idea of using the relatedness of word senses, not of words, could possibly overcome some of the issues[d] encountered when translating the word pairs.

Acknowledgements

We address our thanks to the Social Science Research Council (SSHRC) and to the Natural Sciences and Engineering Research Council (NSERC) of Canada for supporting this research work. We thank Aminul Islam for sharing his code for SOC-PMI. We thank Saif Mohammad for sharing the German word pair similarity scores. We also thank Stan Szpakowicz for his comments on the draft of this paper.

References

1. Budanitsky, A. and Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh.
2. Cramer, I (2008). How Well Do Semantic Relatedness Measures Perform? A Meta-Study. In *Proceedings of STEP 2008 Conference*, Vol. 1, pp. 59-70.
3. Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, January 2007.
4. Hassan, S. and Mihalcea, R. (2009), Cross-lingual Relatedness using Encyclopedic Knowledge, to appear in *Proceedings of the Conference on Empirical Methods in Natural Language* Processing (EMNLP 2009), pp. 1192-1201, Singapore, August 2009.
5. Inkpen, D. and Desliets, A. (2005). Semantic Similarity for Detecting Recognition Errors in Automatic Speech Transcripts. *EMNLP 2005*, Vancouver, Canada.
6. Islam, A. and Inkpen, D. (2006). Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words, in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, pp. 1033-1038, May 2006.
7. Jarmasz, M. and Szpakowicz, S. (2003). Roget’s Thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 212–219, Borovets, Bulgaria.
8. Miller, G. A. and W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1), 1–28.
9. Mohammad, S., Gurevych, I., Hirst, G. and Zesch, T. (2007). Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL-2007)*, Prague, Czech Republic.
10. Rubenstein, H. and J. B. Goodenough (1965). Contextual correlates of synonymy. *Communications of the ACM* 8(10), 627–633.

[d] In some cases the translation of word pairs resulted in the same word, and in other cases the result produced a phrase, or a more obscure word. For example – “midday, noon” = “midi, midi”, “woodland” = “region boisée”, and “mound” = “monticle”.