Using Significant Word Co-occurences for the Lexical Access Problem

Rico Feist and **Daniel Gerighausen** and **Manuel Konrad** and **Georg Richter** and
Department of Computer Science,
University of Leipzig,
Germany
rf@ricofeist.de, daniel@bioinf.uni-leipzig.de
manuel.konrad, georg.richter @studserv.uni-leipzig.de

Thomas Eckart and **Dirk Goldhahn** and **Uwe Quasthoff**
Natural Language Processing Group,
University of Leipzig,
Germany
teckart, dgoldhahn, quasthoff
@informatik.uni-leipzig.de

Abstract

One way to analyse word relations is to examine their co-occurrence in the same context. This allows for the identification of potential semantic or lexical relationships between words. As previous studies showed word co-occurrences often reflect human stimuli-response pairs. In this paper significant sentence co-occurrences on word level were used to identify potential responses for word stimuli based on three automatically generated text corpora of the Leipzig Corpora Collection.

1 Introduction

Conventional dictionaries have very limited possibilities for retrieving information. By contrast electronic dictionaries offer a much wider and more dynamic range of access strategies. One important task in dictionary lookup is to retrieve a word starting just with the corresponding meaning. For this purpose the flexibility of electronic dictionaries should be advantageous. In the following the related task of retrieving a word based just on a stimulus of five related input words is addressed. Based on the assumption that word co-occurrences in the same context can be used to analyse word relations and to identify potential semantic or lexical relationships between words an automatic system is built based on an electronic dictionary extracted from Web corpora. As previous studies showed word co-occurrences often reflect human stimuli-response pairs (Spence, 1990; Schulte im Walde, 2008). In this paper significant sentence co-occurrences on word level were used to identify potential responses for word stimuli based on three automatically generated text corpora of the Leipzig Corpora Collection (LCC).

2 Used Methods and Resources

2.1 Used Corpora

The text corpora of the Leipzig Corpora Collection (Biemann, 2007; Goldhahn, 2012) were used as data basis. As the origin of the stimuli data was unknown corpora based on different text material were exploited:

- eng_wikipedia_2010: a corpus based on the English Wikipedia generated in 2010 containing 23 million sentences
- eng_news_2008: 49 million sentences from English newspaper articles collected in 2008
- eng_web_2002: 57 million sentences of English Web text crawled in 2002

Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon, pages 46–49,
Dublin, Ireland, August 23, 2014.

All of these corpora were generated by the standard preprocessing toolchain of the LCC. This toolchain contains different procedures to ensure corpus quality like language identification and pattern based removal of invalid text material (Goldhahn, 2012). Furthermore all corpora were annotated with statistical information about word co-occurrences based on co-occurrence in the same sentence or direct neighbourhood. These word relations were generated by using the log-likelihood ratio (Buechler, 2006) as measure of significance. Complete sentences were used as co-occurrence window.

2.2 Raw Results Generation

For each of the five stimulus words and every corpus all co-occurrent words were extracted. For extracted terms that significantly co-occurred with more than one of the stimulus words the significance of co-occurrence were combined. Based on the sum of the significance values a ranking of the most relevant terms for every stimulus was created for every corpus. The most significant 15 words were considered as raw result for every corpus and stimulus 5-tuple.

2.3 Postprocessing

The raw results were combined by replacing the result ranks in the three intermediate result lists l_i ($1 <= i <= 3$) with a weight ($weight_i(w) = 16 - rank_i(w)$). These weights were merged by generating the combined weight for all three corpora $c\_weight(w) = \sum_{i=1}^{3} weight_i(w)$. The word with the highest combined weight was chosen as response for a stimulus tupel.

The same procedure was used in two variations:

- Rankings were generated based on the combination of all inflected terms of the same word stem (by using the Porter stemmer(Porter, 1980)).
- Stop words were removed from the result lists to reduce the influence of high frequent function words[1].

For some stimuli only stop words were extracted as response. Here not only the 15 most significant terms were extracted from every corpus but the 45 most significant terms. This lead to more useful results in most cases.

2.4 Results

All three variants were evaluated on the training data set. The evaluation lead to the conclusion that a stop word filter is a useful preprocessing step, whereas stemming lead to unsatisfactory results (cp. table 1). As a consequence only the stop word filter (without stemming) was used for the test data set where 281 (14.05%) of the responses were correctly predicted.

Used Data	Correctly Predicted	Percentage
Original corpus data (incl. stop words, unstemmed)	61	3.05%
Removed stop words	262	13.1%
Stemmed	43	2.15%

Table 1: Evaluation of the different approaches based on the training data set

3 Conclusion

It is noteworthy that corpora where solely stop words were removed yielded better results than corpora where additional stemming took place. One reason for this observation is most likely that by using the Porter algorithm an overstemming occurred. Some word pairs were reduced to identical word stems

[1]For this purpose the stop word list of the database management system MySQL was used (https://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html).

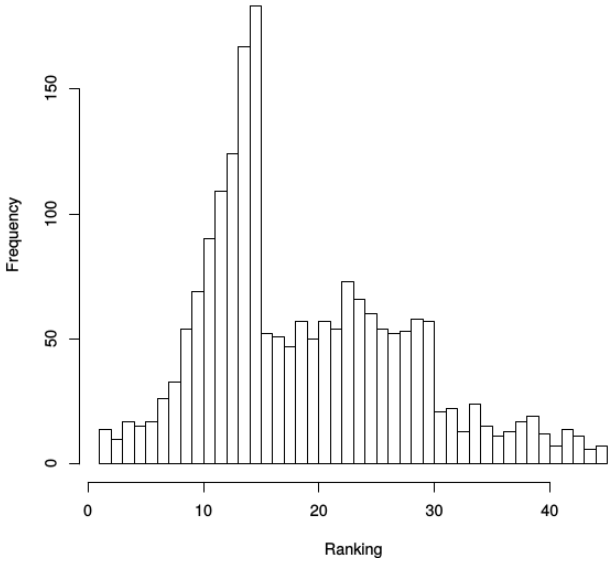

Figure 1: Histogramm of the combined weights for the training data set

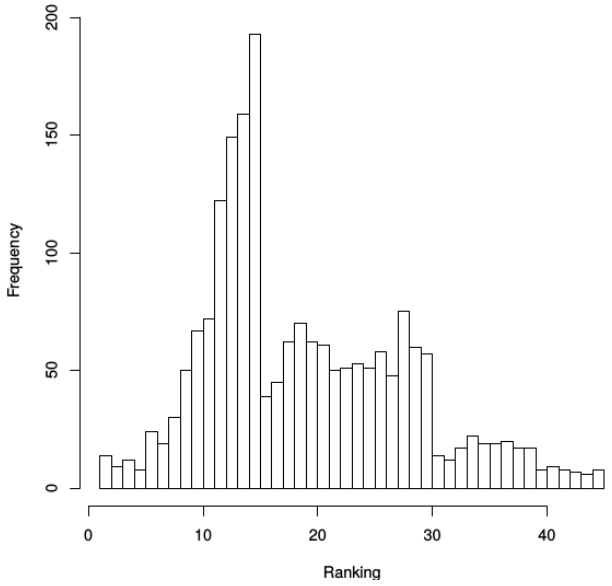

Figure 2: Histogramm of the combined weights for the test data set

although having no semantic relationship.
The final results also contained a disproportionately high number of specific terms. As an example the word "god" was chosen 38 times as response. An analysis of the corpora showed that the word "god" was especially frequent in the Web corpus (330,276 of 56,523,369 sentences (0.59%)) and the Wikipedia-based corpus (58,605 of 22,675,331 sentences (0.26%)). In contrast, the newspaper corpus had only 29 occurrences of the term "god" (in 48,903,372 sentences (0.00006%)).

The evaluation for both the training (figure 1) and the test data set (figure 2) shows that there is a peak for the combined weight of 15. This behaviour originates in terms that have the maximum rank in one of the three corpora but being no significant co-occurrent term in the other two.

4 Further Improvements

The evaluation showed that the used corpora generated results of different quality. This was especially demonstrated at the example of the term "god". As a consequence a stricter selection of the corpus material combined with a weighting of the specific results from each corpus could improve the predictions.

The used corpora reflect a specific selection of input material (in this case written text material from different sources of the years 2002 to 2010). A corpus that reflects more of the details of the test data (most notably being significantly older) would very likely enhance the results. This is especially the case as words that became prominent over the last decades (like technical terms or words strongly related to more recent political developments) would not have occurred in the generated results. A deeper analysis of the input material and the availability of a comparable corpus would have been the prerequisites.

An examination of the results also showed that in many cases a synonym of the correct response was identified. Hence the usage of a synonym database could also lead to further improvements. Furthermore using part of speech information could be beneficial for the weighting of intermediate results. The basic idea is that part of speech of stimulus and response are very likely to be the same. This would have eliminated parts of the generated result sets. Furthermore a deeper analysis of the ranking procedure may have reduced the effect which manifests in many terms having a weight of 15 in the results.

References

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig Corpora Collection - Monolingual corpora of standard size. *Proceedings of Corpus Linguistic 2007*, Birmingham, UK.

Marco Buechler. 2006. Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten. *Diploma Thesis*, University of Leipzig.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Martin F. Porter. 1980. An algorithm for suffix stripping. *electronic library and information systems*, 14.3 (1980): 130-137.

Donald P. Spence, and Kimberley C. Owens. 1990. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, Volume 19(5):317–330.

Sabine Schulte im Walde, and Alissa Melinger. 2008. An in-depth look into the co-occurrence distribution of semantic associates. *Italian Journal of Linguistics, Special Issue on From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*.