Identifying Word Translations in Non-Parallel Texts

Reinhard Rapp
ISSCO, Université de Genève
54 route des Acacias
Genève, Switzerland
rapp@divsun.unige.ch

Abstract

Common algorithms for sentence and word-alignment allow the automatic identification of word translations from parallel texts. This study suggests that the identification of word translations should also be possible with non-parallel and even unrelated texts. The method proposed is based on the assumption that there is a correlation between the patterns of word co-occurrences in texts of different languages.

1 Introduction

In a number of recent studies it has been shown that word translations can be automatically derived from the statistical distribution of words in bilingual parallel texts (e. g. Catizone, Russell & Warwick, 1989; Brown et al., 1990; Dagan, Church & Gale, 1993; Kay & Röscheisen, 1993). Most of the proposed algorithms first conduct an alignment of sentences, i. e. those pairs of sentences are located that are translations of each other. In a second step a word alignment is performed by analyzing the correspondences of words in each pair of sentences.

The results achieved with these algorithms have been found useful for the compilation of dictionaries, for checking the consistency of terminological usage in translations, and for assisting the terminological work of translators and interpreters.

However, despite serious efforts in the compilation of corpora (Church & Mercer, 1993; Armstrong & Thompson, 1995) the availability of a large enough parallel corpus in a specific field and for a given pair of languages will always be the exception, not the rule. Since the acquisition of non-parallel texts is usually much easier, it would be desirable to have a program that can determine the translations of words from comparable or even unrelated texts.

2 Approach

It is assumed that there is a correlation between the co-occurrences of words which are translations of each other. If – for example – in a text of one language two words A and B co-occur more often than expected from chance, then in a text of another language those words which are translations of A and B should also co-occur more frequently than expected. This assumption is reasonable for parallel texts. However, in this paper it is further assumed that the co-occurrence patterns in original texts are not fundamentally different from those in translated texts.

Starting from an English vocabulary of six words and the corresponding German translations, table 1a and b show an English and a German co-occurrence matrix. In these matrices the entries belonging to those pairs of words that in texts co-occur more frequently than expected have been marked with a dot. In general, word order in the lines and columns of a co-occurrence matrix is independent of each other, but for the purpose of this paper can always be assumed to be equal without loss of generality.

If now the word order of the English matrix is permuted until the resulting pattern of dots is most similar to that of the German matrix (see table 1c), then this increases the likelihood that the English and German words are in corresponding order. Word n in the English matrix is then the translation of word n in the German matrix.

3 Simulation

A simulation experiment was conducted in order to see whether the above assumptions concerning the similarity of co-occurrence patterns actually hold. In this experiment, for an equivalent English and German vocabulary two co-occurrence matrices were computed and then compared. As the English vocabulary a list of 100 words was used, which had been suggested by Kent & Rosanoff (1910) for association experiments. The German vocabulary consisted of one by one translations of these words as chosen by Russell (1970).

The word co-occurrences were computed on the basis of an English corpus of 33 and a German corpus of 46 million words. The English corpus consists of

Table 1: When the word orders of the English and the German matrix correspond, the dot patterns of the two matrices are identical.

(a)

		1	2	3	4	5	6
blue	1		•			•	
green	2	•		•			
plant	3		•				
school	4						•
sky	5	•					
teacher	6				•		

(b)

		1	2	3	4	5	6
blau	1		•	•			
grün	2	•				•	
Himmel	3	•					
Lehrer	4						•
Pflanze	5		•				
Schule	6				•		

(c)

		1	2	5	6	3	4
blue	1		•	•			
green	2	•				•	
sky	5	•					
teacher	6						•
plant	3		•				
school	4				•		

the *Brown Corpus*, texts from the *Wall Street Journal*, *Grolier's Electronic Encyclopedia* and scientific abstracts from different fields. The German corpus is a compilation of mainly newspaper texts from *Frankfurter Rundschau*, *Die Zeit* and *Mannheimer Morgen*. To the knowledge of the author, the English and German corpora contain no parallel passages.

For each pair of words in the English vocabulary its frequency of common occurrence in the English corpus was counted. The common occurrence of two words was defined as both words being separated by at most 11 other words. The co-occurrence frequencies obtained in this way were used to build up the English matrix. Equivalently, the German co-occurrence matrix was created by counting the co-occurrences of German word pairs in the German corpus. As a starting point, word order in the two matrices was chosen such that word n in the German matrix was the translation of word n in the English matrix.

Co-occurrence studies like that conducted by Wettler & Rapp (1993) have shown that for many purposes it is desirable to reduce the influence of word frequency on the co-occurrence counts. For the prediction of word associations they achieved best results when modifying each entry in the co-occurrence matrix using the following formula:

$$A_{i,j} = \frac{(f(i\&j))^2}{f(i) \cdot f(j)} \quad (1)$$

Hereby $f(i\&j)$ is the frequency of common occurrence of the two words i and j, and $f(i)$ is the corpus frequency of word i. However, for comparison, the simulations described below were also conducted using the original co-occurrence matrices (formula 2) and a measure similar to mutual information (formula 3).[1]

$$A_{i,j} = f(i\&j) \quad (2)$$

$$A_{i,j} = \frac{f(i\&j)}{f(i) \cdot f(j)} \quad (3)$$

Regardless of the formula applied, the English and the German matrix where both normalized.[2] Starting from the normalized English and German matrices, the aim was to determine how far the similarity of the two matrices depends on the correspondence of word order. As a measure for matrix similarity the sum of the absolute differences of the values at corresponding matrix positions was used.

$$s = \sum_{i=1}^{N} \sum_{j=1}^{N} |E_{i,j} - G_{i,j}| \quad (4)$$

This similarity measure leads to a value of zero for identical matrices, and to a value of 20 000 in the case that a non-zero entry in one of the 100 * 100 matrices always corresponds to a zero-value in the other.

4 Results

The simulation was conducted by randomly permuting the word order of the German matrix and then computing the similarity s to the English matrix. For each permutation it was determined how many words c had been shifted to positions different from those in the original German matrix. The simulation was continued until for each value of c a set of 1000 similarity values was available.[3] Figure 1 shows for the three formulas how the average similarity $\bar{s}$ between the English and the German matrix depends on the number of non-corresponding word positions c. Each of the curves increases monotonically, with formula 1 having the steepest, i. e. best discriminating characteristic. The dotted curves in figure 1 are the minimum and maximum values in each set of 1000 similarity values for formula 1.

[1] The logarithm has been removed from the mutual information measure since it is not defined for zero co-occurrences.

[2] Normalization was conducted in such a way that the sum of all matrix entries adds up to the number of fields in the matrix.

[3] $c = 1$ is not possible and was not taken into account.

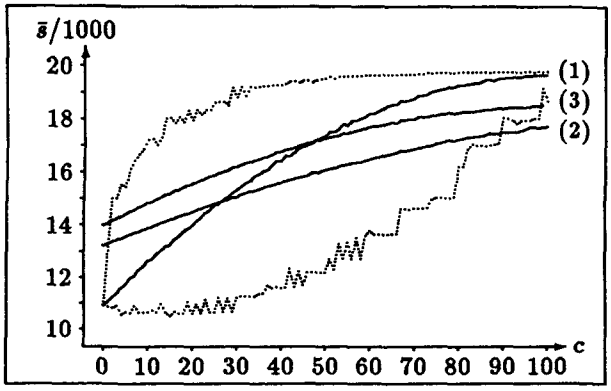

Figure 1: Dependency between the mean similarity $\bar{s}$ of the English and the German matrix and the number of non-corresponding word positions c for 3 formulas. The dotted lines are the minimum and maximum values of each sample of 1000 for formula 1.

5 Discussion and prospects

It could be shown that even for unrelated English and German texts the patterns of word co-occurrences strongly correlate. The monotonically increasing character of the curves in figure 1 indicates that in principle it should be possible to find word correspondences in two matrices of different languages by randomly permuting one of the matrices until the similarity function s reaches a minimum and thus indicates maximum similarity. However, the minimum-curve in figure 1 suggests that there are some deep minima of the similarity function even in cases when many word correspondences are incorrect. An algorithm currently under construction therefore searches for many local minima, and tries to find out what word correspondences are the most reliable ones. In order to limit the search space, translations that are known beforehand can be used as anchor points.

Future work will deal with the following as yet unresolved problems:

- Computational limitations require the vocabularies to be limited to subsets of all word types in large corpora. With criteria like the corpus frequency of a word, its specificity for a given domain, and the salience of its co-occurrence patterns, it should be possible to make a selection of corresponding vocabularies in the two languages. If morphological tools and disambiguators are available, preliminary lemmatization of the corpora would be desirable.
- Ambiguities in word translations can be taken into account by working with continuous probabilities to judge whether a word translation is correct instead of making a binary decision. Thereby, different sizes of the two matrices could be allowed for.

It can be expected that with such a method the quality of the results depends on the thematic comparability of the corpora, but not on their degree of parallelism. As a further step, even with non parallel corpora it should be possible to locate comparable passages of text.

Acknowledgements

I thank Susan Armstrong and Manfred Wettler for their support of this project. Thanks also to Graham Russell and three anonymous reviewers for valuable comments on the manuscript.

References

Armstrong, Susan; Thompson, Henry (1995). A presentation of MLCC: Multilingual Corpora for Cooperation. *Linguistic Database Workshop,* Groningen.

Brown, Peter; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Fredrick; Lafferty, John D.; Mercer, Robert L.; Rossin, Paul S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2), 79–85.

Catizone, Roberta; Russell, Graham; Warwick, Susan (1989). Deriving translation data from bilingual texts. In: U. Zernik (ed.): *Proceedings of the First International Lexical Acquisition Workshop*, Detroit.

Church, Kenneth W.; Mercer, Robert L. (1993). Introduction to the special issue on Computational Linguistics using large corpora. *Computational Linguistics*, 19(1), 1–24.

Dagan, Ido; Church, Kenneth W.; Gale, William A. (1993). Robust bilingual word alignment for machine aided translation. *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives.* Columbus, Ohio, 1–8.

Kay, Martin; Röscheisen, Martin (1993). Text–Translation Alignment. *Computational Linguistics,* 19(1), 121–142.

Kent, G.H.; Rosanoff, A.J. (1910). A study of association in insanity. *American Journal of Insanity,* 67, 37–96, 317–390.

Russell, Wallace A. (1970). The complete German language norms for responses to 100 words from the Kent-Rosanoff word association test. In: L. Postman, G. Keppel (eds.): *Norms of Word Association.* New York: Academic Press, 53–94.

Wettler, Manfred; Rapp, Reinhard (1993). Computation of word associations based on the co-occurrences of words in large corpora. In: *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, 84–93.