Extraction of Translation Equivalents from Parallel Corpora

Jörg Tiedemann
Department of Linguistics
Uppsala University
joerg@stp.ling.uu.se
January 25, 1998

1 Introduction

In the past much effort was devoted to the compilation of multi-lingual parallel corpora for the purpose of linguistic information retrieval. This paper aims to introduce and evaluate three simple strategies for the extraction of translation equivalents from structured parallel texts. The goal is to support the production of bilingual dictionaries for domain-specific applications. The approaches described in the paper assume sentence alignment, strict translations, and historical relations between considered language pairs. They take advantage of corpus characteristics like short aligned units and structural & orthographic similarities in order to obtain results with a high level of precision. Furthermore, it will be shown that automatic filtering can be used to improve the precision of the extracted material. Simple techniques are used to detect translation candidates that are most likely wrong.

2 Pre-processing the Corpus

Important pre-processing steps include paragraph and sentence boundary detection, tokenization, compilation of collocations, and multilingual sentence alignment. The precision of extracted material depends highly on the quality of these pre-processing steps. All three extraction methods introduced in the paper assume tokenised and aligned corpora.

Tokenization is one of the fundamental tasks in preparing the corpus for linguistic information retrieval. The separation of lexical tokens from punctuation symbols is important. Here, special pattern can be defined to segment the corpus. Difficulties arise due to the ambiguity of the used symbols. For further discussions see [GT94].

Automatic sentence alignment can be done by statistical methods like those proposed in [GC93], [FC94], [FM94], [Chu93], [SFI92], and [Mel96]. These approaches yield high precision especially for closely related language pairs.

3 Extraction by Iterative Size Reduction

In this section an iterative extraction method is introduced that takes advantage of aligned parallel corpora with a large number of short aligned text structures. This phenomenon often appears in highly structured texts like technical documentation. The following investigations are based on the parallel Scania corpus ([Sca93]) which contains many short structures due to brief descriptions, list elements, tables, picture captions, and headers. These elements can be easily recognised, aligned, and analysed. Among the Swedish/English alignments of the Scania corpus with a total number of 35,818 alignments there are 2,628 aligned header structures, 8,558 aligned

table cell structures, and 8,423 aligned list item alignments. About every third alignment represents a 1:1 token alignment, although many of them are copies of each other.

First, we extracted 1:1, 1:x, and x:1 token alignments from the corpus to compile a basic set of translation equivalents. This basic dictionary was used to analyse the remaining alignments in an iterative process by removing known translations from the total set of corpus alignments. As the result, the size of the alignments (in tokens) decreases. Then, we extracted newly obtained 1:1 token alignments and added them to the set of known translations. This extended set of translation pairs was used in the next step to analyse the remaining alignments from the former step. This procedure may be repeated until no new 1:1 token alignments appear.

This simple algorithm produces a large number of new translation equivalents in case of many short aligned structures. When applied to the Scania corpus the method provides results of high precision as shown in table 1.

Step	*Swedish/English*	Swedish/German
1	184 (96.2 % correct)	223 (96.7 % correct)
2	90 (93.3 % correct)	231 (98.7 % correct)
3	26 (100 % correct)	165 (98.8 % correct)
4	7 (85.7 % correct)	85 (95.3 % correct)
5	2 (100 % correct)	31 (93.5 % correct)
recall (incl. basic dictionary)[1]	12.2. %	14. 6 %

Table 1: The number of translation equivalents (1:1 token pairs) extracted with the iterative size reduction method.

The quality of the results in each step highly depends on the set of translation equivalents obtained in former steps. In general, wrong pairs will produce wrong alignments in the next reduction step. Table 1 shows that the precision of Swedish/German results decreases slightly in steps 4 and 5. However, the numbers of newly discovered Swedish/English pairs are too small in the final steps to observe a similar behaviour.

The algorithm so far is limited to the extraction of 1:1 pairs. However, in language pairs with different compounding rules (like Swedish/English) many 1:x correspondences appear. Unfortunately, remaining 1:x alignments cannot be considered as correct translation equivalents automatically. Some tokens may belong to previously removed tokens. The remaining tokens are usually grammatical function words. Fortunately, most of the 1:x alignments can be transformed to correct translation equivalents by simply removing these words using language specific function word lists. Adequate lists may be compiled by extracting the most frequent words from the corpus. The efficiency of this simple approach can be seen in table 2.

[1]The recall value is estimated by comparing the number of extracted pairs with the total number of source tokens used in the corpus.

	before	after
Swedish/English 1:x	46.9 %	84.5 %
Swedish/English x:1	1.5 %	79.1 %
Swedish/German 1:x	4.4 %	84.6 %
Swedish/German x:1	44 %	69.2 %

Table 2: Precision estimates for extracted Swedish/English and Swedish/German 1:x respectively x:1 alignments before and after the removal of function words.

The removal of function words improves the precision values greatly. Applying a more sophisticated list of function words would produce even better results and would admit the usage of 1:x pairs for the iterative extraction process. This would represent a major improvement especially for language pairs like Swedish/English.

Finally, it is worth mentioning that any bilingual dictionary may be used by the algorithm, provided that it suits the domain of the corpus under consideration.

4 Considerations to String Similarities

Evaluations to string similarities can be used for different tasks in computational linguistics. One application is the identification of morphological deviations. This may be used in the size reduction method, which is introduced in the previous section, to identify slightly modified translation pairs (see further [Tie97]). Another application is the identification of cognates from bilingual texts in case of similar character sets and historical relations between the languages under consideration. In both cases simple string matching algorithms can be used to compare word pairs. Applications to technical texts are especially profitable because of internationalisation and similarities in the origin of technical terminology.

One simple algorithm based on character comparison is the 'longest common sub-sequence ratio' (LCSR) which is defined as follows: The LCSR score is calculated by the length of the longest common, not necessarily contiguous, sub-sequence of characters divided by the character length of the longer string ([Mel95]).

Mostly, even more simple algorithms represent sufficient measures for string similarity. We used algorithms to search initial and final character sequences, and algorithms to compare fixed character sequences. More complex algorithms may raise the recall but not the precision of results when applied with similar threshold values. See further [Bor98] for a comparison of different approaches.

The selection of word pairs, which are considered for the string comparison, is important for the quality of the extraction of cognates. A major improvement represents sentence alignment. The chance of getting 'false friends' remains very low if only word pairs from aligned text structures are considered. When applied to the Swedish/English and Swedish/German alignments of the

Scania corpus with a threshold of 70% the method provides results with high precision as shown in table 3.

	Swedish/English	Swedish/German
pairs	360	843
precision	94.7 %	97.9 %
recall[2]	2.3 %	5.5 %

Table 3: The number of cognates that were identified with string similarity measures.

Sentence alignment also allows further evaluation of similarity measures. Scores from token pairs can be combined for further analyses (see also [Tie97]). Usually, words that are very similar to only one word of the corresponding alignment are more likely translation equivalents than words that show similarities to several words from the corresponding alignment. This assumption can be expressed by the following similarity score combinations:

$$diff_s(x,y) = sim(x,y) - \sum_{z \neq y} sim(x,z)$$

$$diff_t(x,y) = sim(x,y) - \sum_{z \neq x} sim(z,y)$$

Here, the term sim(x,y) represents the similarity score for the token pair (x,y). In this way new scores for each token pair are calculated which may be evaluated by specific threshold filters. Including evaluations of similarity score combinations the similarity method yielded a recall of 3.2% with an estimated precision of 92.3% for Swedish/English alignments, and a recall of 7.7% with an estimated precision of 96.2% for Swedish/German alignments.

5 Extraction based on Statistical Measures

Many statistical measures are based on co-occurrence measures. The basic input data is the absolute frequencies of the occurrence of single words or word groups and the co-occurrence frequencies of word pairs and pairs of word groups in corresponding subparts of the text (for instance aligned sentences). Using these values, statistical measures can be estimated for monolingual as well as for multi-lingual input.

Similar to word distribution measurements based on Mutual Information like in [FC94] another statistical ratio, the Dice coefficient can be used to measure the co-occurrence of words or word groups. This ratio is used, for instance, in the collocation compiler XTract ([Sma93]) and in the lexicon extraction system Champollion ([SMH96]). It is defined as follows:

[2] The recall value is hard to estimate. The value mentioned in table 3 is measured by comparing the number of resulting pairs with the set of source words used in the corpus. According to expectations, the recall is very small for both language pairs because only a small subset of word pairs actually represent cognates.

$$Dice(x,y) = \frac{2P(x,y)}{P(x)+P(y)}$$

P(x,y) represents the probability for the occurrence of x and y in corresponding parts; P(x) and P(y) are the single probabilities that x and y occur. All these values can be estimated by corresponding frequency values. Notice that x and y may be single words as well as word groups.

The major advantage of statistical measures is language independence. The major problem, however, is the proper selection of monolingual text units for the considerations. The chosen text units have to be comparable in their semantic complexity, otherwise statistical measures produce incorrect and incomplete results. Consider for instance the different usage of compounds in Swedish and English. In Swedish, compounds are mostly used in the form of compositions while in English corresponding expressions appear as sequences of separated words. Therefore, a sophisticated compilation of monolingual collocations should be performed for the identification of non-compositional compounds before performing statistical extraction.

A second problem is the different variety of morphological forms in different languages. To solve this problem all word forms should be reduced to their stem forms before counting occurrence frequencies.

A last problem arises with infrequent words. Because of its definition the introduced ratio may produce high scores for pairs containing infrequent text units although they do not correspond. This fact is due to the high probability that infrequent words appear in corresponding text parts by chance. Therefore, infrequent text units should be excluded from this statistical consideration.

However, infrequent text units can be analysed in a different way. Like in the 'size reduction' method (see section 3) we presume short aligned text units. By using monolingual frequency counts, all terms with a higher frequency than a specific threshold value t_1 are removed from the complete set of alignments. Then, the remaining alignments are examined. With the assumption that infrequent terms are translated into infrequent terms in the other language, all remaining infrequent one-to-one term pairs (a term may be a word or a word phrase) are extracted which occur less frequently than another threshold value t_2. These pairs may be considered as translation equivalent candidates. Again, the biggest difficulty is the selection of proper terms for frequency counts.

Due to time constraints the algorithms were applied without previous compilation of collocations and without stemming. Sentences were simply segmented into space separated tokens and frequency counts were based on these text elements. Filtering the Dice coefficient with a threshold of 0.7 provided results with a recall of 7.8% and an estimated precision of 86% for Swedish/German alignments and a recall of 4.1% with an estimated precision of 70% for Swedish/English alignments. The extraction of low frequent word pairs with t_1=100 and t_2=10 provided results with a recall of 6.7% and a good precision of 93.5% for Swedish/German alignments, and a recall of 2.8% and an unsatisfactory precision of 46% for Swedish/English alignments.

6 Automatic Evaluation Filter

The results from the three extraction methods introduced above are certainly not perfect and free of mistakes, although precision yielded promising values. Fortunately, automatic filters can be used to remove pairs that are most likely wrong. Statistical and empirical evaluations can be applied to analyse the set of translation candidates. Special considerations should be attached to terms, for which multiple translation candidates were extracted. Most errors are to be found here. The translation candidates can be compared with each other and the most unlikely candidates can be removed automatically. Several approaches are applicable:

Length based filter: A length difference ratio can be calculated by dividing the length of the shorter string by the length of the longer string. In case of multiple translation candidates the pair with the highest score identifies the most likely translation. Now, all scores for alternative candidates are compared with this score and all pairs, which do not pass a chosen threshold, are removed. The length difference can also be used to compare alternative translations with the most likely translation. Furthermore, it may be also used to exclude translation candidates with unreasonable length differences to the source language term.

Similarity filter: Word comparison algorithms can be applied to extracted pairs. The highest resulting score identifies the most likely translation among multiple translation candidates. Then, all similarity scores of alternative candidates are compared with the similarity score of the most likely translation. Another possibility is to calculate similarities between the most likely translation and alternative candidates. This presumes multiple translations to appear because of slightly morphological modifications rather than because of synonym translations.

Frequency based filter: Absolute and co-occurrence frequencies are calculated for extracted pairs. Using these values, the Dice coefficient is estimated and used for the removal of translation candidates with unreasonable low scores. In case of multiple translations the Dice score can also be used to identify the most likely translation and the score for alternative candidates is compared with the score of this translation.

Combined filter: The filter described above may be combined. Generally, one approach may be used for the identification of the most likely translation and another approach may be used for the comparison of alternative translations with this candidate. One possible combination is for instance the usage of the Dice coefficient for the identification of the most likely translation and a similarity filter to compare this candidate with alternative ones.

Subset filter: Translation candidates may be incomplete. Therefore, if one translation candidate is completely included in another candidate it should be removed. Translation candidates can be considered as sets of words. Translations are removed if the complete set of words is included in an alternative translation candidate.

A non-trivial task is to adjust these filters to obtain the best result. The improvement level depends on the quality of the previously applied extraction methods.

7 Discussion and Conclusion

In this paper three methods are described for the retrieval of translation equivalents from aligned bilingual corpora. They represent independent approaches that use different assumptions. Table 4 shows corpus characteristics and preparations that are necessary for each single approach.

method	highly structured text	similar character set	tokeni-zation	sentence alignment	compilation of collocations	removal of function words
size reduction	X		X	X		X
similarity measures		X	X			
score combinations		X	X	X		
Dice statistics			X	X	X	
low frequent terms	X		X	X	X	

Table 4: Necessary characteristics and preparations in order to apply the different extraction methods.

The size reduction method takes advantage of corpora with many short aligned text structures. Quantity and quality will decrease if less structured texts are used even if a large set of translation equivalents is used in the initial step. For highly structured texts this method provides fast and precise results. The method is not necessarily limited to 1:1 word pairs, although the extraction of phrase alignments needs additionally the removal of function words. An advantage is that any otherwise compiled dictionary may be used by the algorithm as long as it suits the domain of the corpus. In this way, the size reduction method was applied again to the Scania corpus using an initial dictionary that was compiled by different extraction methods. The recall[3] of the final dictionary amounts to 28.3% with an estimated precision of 96.5% for Swedish/English and 49.4% with an estimated precision 96.7% for Swedish/German.

The string similarity approach aims to extract closely related word pairs. It is limited to 1:1 token pairs. The method is applicable to historically related language pairs only. Precision and recall can easily be adjusted by modifying the threshold value. The application described in the paper provides results with precision and recall similar to the size reduction method. However, the extracted pairs represent different sets for both approaches. In contrast to the other approaches the cognate extraction method depends highly on the language pair and assumes similar character sets.

Statistical extractions depend on the quality of the text segmentation and the frequency counts. The simple approach described in the paper provides results with recall values similar to the other two extraction methods. However, the precision is much lower for both language pairs. Especially for Swedish/English alignments many incomplete pairs appear. A more sophisticated segmentation and previously applied stemming would improve the precision greatly.

[3]The number of extracted pairs is compared with the total number of source tokens that are used in the corpus to estimate the recall.

When applied to the Swedish/German alignments all three approaches provided results with much higher precision and recall than when applied to Swedish/English alignments, although the size reduction method and statistical evaluations should be mostly language independent. However, essential pre-processing steps like collocation compilation and stemming were not carried out before applying these algorithms and therefore, remarkable differences in quality and quantity of the achieved results arose.

References

[Bor98] Lars Borin. Linguistics isn't always the answer: Word comparison in computational linguistics. In Proceedings of the 11th Nordic Conference on Computational Linguistics NODALIDA '98, Copenhagen, 1998.

[Chu93] Kenneth W. Church. Char align: A Program for Aligning Parallel Texts at the Character Level. In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, ACL. Association for Computational Linguistics, 1993.

[FC94] Pascale Fung and Kenneth W. Church. K-vec: A New Approach for Aligning Parallel Texts. In Proceedings of the 15th International Conference on Computational Linguistics, Kyoto/Japan, 1994.

[FM94] Pascale Fung and Kathleen R. McKeown. Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping. In Proceedings of the 1st Conference of the AMTA, Columbia/Maryland, 1994. Association for Machine Translation in the Americas.

[GC93] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1), 1993.

[GT94] Gregory Grefenstette and Pasi Tapainen. What is a word, What is a sentence? Problems of Tokenization. In Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX '94), Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 1994. Rank Xerox Research Centre, Grenoble Laboratory.

[Mel95] I. Dan Melamed. Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons. In Proceedings of the 3rd Workshop on Very Large Corpora, Boston/Massachusetts, 1995.

[Mel96] I. Dan Melamed. A Geometric Approach to Mapping Bitext Correspondence. In Conference on Empirical Methods in Natural Language Processing, Philadelphia/USA, 1996.

[Sca93] Scania project - Homepage. http://stp.ling.uu.se/~corpora/scania/ Uppsala University, Linguistics Department, 1997.

[SFI92] Michael Simard, George F. Foster, and Pierre Isabelle. Using Cognates to Align Sentences in Bilingual Corpora. In Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal/Canada, 1992.

[Sma93] Frank Smadja. Retrieving Collocations from Text: XTRACT. Computational Linguistics, 1993.

[SMH96] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translation Collocations for Bilingual Lexicons: A Statistical Approach. In Association for Computational Linguistics. Association for Computational Linguistics, 1996.

[Tie97] Jörg Tiedemann. Automatical Lexicon Extraction from Aligned Bilingual Corpora. Diploma thesis, Magdeburg University, Department of Computer Science, 1997.