Non-probabilistic alignment of rare German and English nominal expressions

Bettina Schrader

Institute of Cognitive Science
University of Osnabrück
Germany
bschrade@uos.de

Abstract

We present an alignment strategy that specifically deals with the correct alignment of rare German nominal compounds to their English multiword translations. It recognizes compounds and multiwords based on their character lengths and on their most frequent POS-patterns, and aligns them based on their length ratios. Our approach is designed on the basis of a data analysis on roughly 500 German hapax legomena, and as it does not use any frequency or co-occurrence information, it is well-suited to align rare compounds, but also achieves good results for more frequent expressions. Experiment results show that the strategy is able to correctly identify correct translations for 70% of the compound hapaxes in our data set. Additionally, we checked on 700 randomly chosen entries in the dictionary that was automatically generated by our alignment tool. Results of this experiment also indicate that our strategy works for non-hapaxes as well, including finding multiple correct translations for the same head compound.

1. Introduction

Word alignment is a very useful technique for preprocessing parallel corpora for a range of applications, including but not limited to machine translation (Brown et al., 1993), cross-language information retrieval (Hiemstra, 1996), dictionary creation (Smadja et al., 1996; Melamed, 2001) and induction of NLP-tools (Kuhn, 2004).

Effort is spent on improving word alignment techniques (Mihalcea and Pedersen, 2003; Cherry and Lin, 2003; Toutanova et al., 2002), which is of major importance in areas where it is very difficult to establish correct correspondences between source and target language words. One of these areas is the correct alignment of multiword sequences as these require n:m alignment beads, where n, m or both are higher than 1. Another problem is the alignment of so-called rare events, i.e. types that occur too rarely in order to align them based on statistical information only.

In this paper, we suggest a comparatively simple technique that tackles both problems with word alignment, at least with respect to nominal expressions: it aligns nominals based on word length and part of speech patterns in an English-German parallel corpus, without taking frequency counts or any other kind of statistical information into account. Hence, it is useful for aligning nominals with a frequency of one, so-called hapax legomena, as well as nominals with higher frequencies.

Our alignment strategy has been incorporated into a new text alignment system, which, due to its modular and flexible architecture, is well suited for testing new strategies and evaluating their strengths and failures irrespective of overall text alignment quality.

In the following sections, we first give an overview over strategies for dealing with noun compounds and rare events. Secondly, we describe the corpus from which we extracted roughly 500 hapax legomena, and how we analysed them. Fourthly, we explain how we used analysis results for implementing an alignment strategy in our word alignment tool. Finally, we report on the results of our alignment experiments.

2. Problems for Statistical Word Alignment

2.1. Multiword Units

In order to handle multiwords, Brown et al. (1993) had to introduce so-called n:m-alignments into their models, thus resulting in much more complex statistical computation.

Kupiec (1993) aligned noun chunks in a French-English parallel corpus using the EM-algorithm[1]: his algorithm uses POS-patterns for recognizing nominals including post-modifying prepositional phrases (PPs). Secondly, the EM-algorithm is used to statistically learn the correct alignment of the chunks. As Kupiec (1993)'s strategy relies on statistics, however, it has difficulties dealing with rare chunks.

For the purposes of training statistical machine translation models on alignment data, (Nießen and Ney, 2001) try to work around n:m alignments by splitting German compounds into their components. However, this approach has theoretical disadvantages: This approach implicitly assumes that a compound's meaning is made up compositionally, and that (the same degree of) compositionality also holds for its translation. However, this is not necessarily true, as examples like

(1) Personen|stand (marital status)
personal status

readily show. Hence it would seem more advisable to *not* split compounds but finding means to align them correctly to the entire equivalent expression in the other language.

Tschorn and Lüdeling (2003), on the other hand, argue that many unknown words are due to productive word formation processes, i.e. that compositionality holds, and that translations for compound components can be found in existing dictionaries. In their approach, unknown words are morphologically analyzed, their components are looked up in a bilingual dictionary, and, if possible, aligned with their translations. This yields, in the best case, complete matches between a German compound and its English translation.

[1] see (Manning and Schütze, 1999), chapter 14.2.2 for a general introduction

Even partial matches between the compound and a part of its translation helps to improve sentence alignment quality, which is the task the authors have in mind. Unfortunately, partial matches are insufficient for word alignment, and the strategy of Tschorn and Lüdeling (2003) necessarily fails if compositionality does not hold.

2.2. Hapax Legomena and Rare Events

Another problem for statistical approaches to word alignment, namely how to align rare words, seems to have been neglected in the literature. Dejean et al. (2003) report that lemmatizing those types that are rare improves results, although they do not give an explanation for the phenomenon. Others exclude words below a certain frequency from evaluations of their alignment tools (Merkel et al., 2002), thereby admitting that rare words are a problem for their approaches.

3. Corpus description

For our experiments, we have used the *Europarl* corpus which has been incorporated into the *OPUS* parallel corpus collection (Tiedemann and Nygaard, 2004). The *Europarl* corpus consists of verbatim protocols of European Parliament sessions, and it is aligned at the sentence level. For the purpose of our experiments, we have tagged the English and German parts of the corpus using the publicly available *tree-tagger* (Schmid, 1994). Additionally, we corrected the sentence alignments of five randomly chosen protocol files manually.

Table 1 shows the size of our corpus, as well as the proportions of hapax legomena, other rare events, i.e. types with a frequency between 2 and 10, and of types occurring more than 10 times:

The five protocol files for which we corrected the sentence alignments comprise 103,091 tokens of German and 109,732 tokens of English text.

4. Data analysis

In order to find a strategy that is able to align hapax legomena successfully, we extracted all German hapax legomena from the Europarl corpus, and analyzed those hapax legomena that occurred within the corpus subset for which we had corrected the sentence alignment.

We analyzed these 512 hapaxes with respect to word category membership, morphological complexity and word length. We also aligned them manually to their English correspondences and analyzed which categories they belonged to, how they were structured and how long they were. We also hypothesized which kinds of alignment problems are to be expected if the corpus is automatically aligned.

The analysis yielded that 353 of the 512 German hapax legomena, or 68.95%, are noun compounds, and that their translations are chunk-like multiword expressions in 68% of all cases. These expressions are most often a sequence of nouns, either preceded by an adjective or followed by a PP. Paraphrases or nouns followed by clauses are rare.

Additionally, we found a strong correlation between nouns and their translations both in morphological complexity and in their respective lengths, counted in characters[2]: If a German noun contains n elements, then its translation most often also contains n elements.

# elements	1	2	3	4	other
1	59	2	1	0	3
2	30	119	56	15	10
3	2	20	15	8	10
4	0	1	0	0	2

Table 2: Expression complexity. Rows show the number of components in English multiword units, columns give the equivalent numbers for German compounds.

If it contains one more element, this is mainly due to the structure of the English translation. In these cases, it often contains a noun plus PP as in

(2) Kongreß|vorlage ↔ submission to Congress

With respect to the lengths of the German compounds and the nominals into which they were translated, counted in characters, we found that the median of the length ratios is 1, and the average of the ratio

$$\text{length ratio} = \frac{\text{German compound length}}{\text{english multiword length}}$$

equals 1.139.

5. Implementation

5.1. The compound alignment strategy

We used the results of our data analysis to implement a *compound alignment strategy*. The input is a sentence-aligned, POS-tagged corpus in English and German. For each sentence bead in the corpus, the algorithm recognizes German noun compounds and English multiword nominals, and sets them into correspondence using their length ratios. A German token is considered a compound if it is tagged as a noun and if it is at least 12 characters long. This threshold corresponds to the first quartile of the average hapax noun length in our data set. For each English noun, identified by its POS-tag, the algorithm seeks to construct several candidates: i) the noun or noun sequence itself (the nominal), ii) the nominal preceded by an adjective, iii) the nominal followed by a prepositional phrase, and iv) the nominal preceded by an adjective *and* followed by a PP.

In a final step, the length ratios between each German compound and each English candidate translation of a sentence bead are computed. If the similarity

$$\text{sim (compound,multiword)} = 1 - |\text{length ratio}|$$

is greater than zero, the translation pair is added to a bilingual dictionary. No further filtering, e.g. with respect to frequency of a compound, is employed, i.e. both hapax and non-hapax nominals are aligned.

[2] including blanks for multiword nouns

Language	Tokens	Types	Hapax Legomena	Other Rare Events	Frequent Types
English	29.077,024	101,967	39,200 (38.44%)	35,608 (34.92%)	27,159 (26.64%)
German	27.643,792	286,330	140,826 (49.18%)	98,126 (34.27%)	47,378 (16.55%)

Table 1: Corpus characteristics of the Europarl corpus

5.2. The ATLAS text alignment system

The compound alignment strategy has been integrated into ATLAS, a modular and flexible text alignment system that allows for the easy adding and testing of alignment modules.

The input to ATLAS is a bilingual, parallel corpus that may be annotated on various linguistic levels, including but not limited to sentence alignment information, POS-tags, lemmas, or chunks. All levels of annotation are accessible to the alignment modules, as well as central data bases with dictionary or alignment information. During an alignment process, a system-internal dictionary is created, and a process-final filtering step discards translation pairs that have been computed, but are not used for the text alignment. The output of the aligner is a bilingual dictionary, along with corpus alignment information.

6. Test and Results

After the implementation of the compound alignment strategy, we have conducted a test on the 100,000 token *Europarl* subset on which whe had carried out the data analysis: only this subset of the corpus was submitted to the text aligner, including POS-annotation and sentence alignment information. Within this subset, our strategy was used to find and align noun compounds and their translations, and all results were used to construct a bilingual German-English dictionary. No other alignment strategy was used, nor did the program compute a full text alignment, or filter the results in any way.

Afterwards, we semi-automatically evaluated whether the dictionary contained lexical entries for the 353 hapax nouns in our analysis, whether they contained correct translations, partial translations, and why correct translations were missing in the lexicon entries. We also tested why lexicon entries were missing in the automatically generated dictionary. We did not evaluate the translation direction English→German.

Results are that the dictionary contains lexical entries for 236 of the 353 compounds in our data set (66.86%), and more than 1600 additional entries with German headwords. This is not surprising given that our implementation of the strategy does not include any frequency restriction, i.e. it aligns nouns irrespective of how often they occur in the input data. With respect to the missing entries, we found out that in most cases, the German compounds did not pass the length threshold, and hence no aligment was computed for them. Other error sources like tokenization problems, compound recognition errors due to paraphrases or hyphenation occurred but rarely. Fortunately, there were only 11 cases where we could not attribute errors to tokenization or POS-patterns, and hence had to attribute them to the length-based similarity measure.

With respect to the hapaxes that received an entry in the dictionary, we found out that the entries contained the correct translation in 47% of all cases. Additionally, we found 306 translation suggestions where the correct translation was either partially present, or a substring of a suggestion. 121 entries, however, did not contain any correct translation. Our error analysis showed that these fully incorrect lexicon entries were partially due to the similarity measure itself, and partially due to nominal recognition not working optimally. In detail, the nominal recognition failed because hyphenated words in both languages had been split into their components during tokenization, because the POS-tagger treated words that occurred sentence-internally *and* in upper case as names, but did not tag them as common nouns, because our algorithm did not account for all POS-patterns, and because of paraphrasing, deletion or category changes during the translation process. An informal evaluation of the additional 1600 lexicon entries confirmed the overall impression: in principle, the method works well, but the recognition of the nominals can be improved.

Accordingly, we revised those parts of the implementation that dealt with the recognition of German and English nominals. In detail, we repaired the over-eager tokenization with respect to German and English hyphenated nominals.

After these modifications, the nominal recognition components is able to recognize, and hence to align, hyphenated words like

(3) Geldwäsche-Bekämpfungsrichtlinie
(English: anti-money laundering directive)

or

(4) anti-riot act (German: Antiterrorgesetz)

We also allowed English "names", i.e. upper case nouns, to be aligned by the algorithm. Afterwards, we re-ran the testing.

Overall, performance increased: Now, 248 of 353 compounds (70.25%) were headwords in the automatically generated dictionary, with 175 entries containing the correct translations. With respect to the missing entries, error numbers decreased with respect to unaccounted-for POS-patterns for the English expressions, and with respect to the similarity measure, although we did not change it at all. This could be an effect of the similarity measure actually discarding errors made during the nominal recognition in the first test run.

A final analysis was carried out on 760 compounds that appeard not in the original data set, but had been aligned nevertheless: We found 561 correct translations, including multiple translations for some head words, as in

(5) Berufsausbildung ↔ vocational training, profes-

sional training

These 760 compounds come from all frequency ranges within the 100,000 token subset that we used for this evaluation, i.e. the set contains hapaxes as well as other rare events, but also frequent compounds like *Geschäftsordnung* (Rules of Procedure), which occurred 32 times in this subset. For these frequent nouns, we noticed that their lexicon entries contained many more translation candidates than for rare words. Additionally, we noticed that frequent compounds tended to have multiple translations in the corpus, and they had often been included in the automatically generated dictionary.

7. Summary and Further Work

Summed up, our analysis on 512 German hapaxes has shown that most of them are nouns with English multiword translations. Consequently, we have implemented an alignment strategy that uses POS-patterns and word lengths to recognize German compounds and English multiword nominals, and to subsequently align them.

Test results are good – nominals are recognized correctly and are assigned correct translations in 70% of all cases, with errors due to either unusual length ratios between expressions or to POS-patterns that are unaccounted for in our algorithm.

With respect to integrating our compound alignment strategy into standard statistical aligners, two steps have to be taken: i) multiword expressions must be recognized in a corpus, and ii) a similarity measure must be added to the statistical alignment model. The multiword recognition can be done in a preprocessing step on a POS-tagged corpus, but disambiguation between different possible multiword structures, i.e. whether some multiword is a noun preceded by an adjective, or whether it is a noun followed by a PP, may be difficult. Regarding the incorporation of word length into a statistical model, it should first be tested whether there is generally a correlation between all kinds of words and their translations, and not just between nominals and their translations.

Currently, we are carrying out experiments on how to use morphological analyses for the recognition of German compounds, and we have started to carry out similar experiments for the language pairs Swedish-German, German-French, and English-French. We also want to refine our similarity measure.

8. References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Colin Cherry and Dekang Lin. 2003. A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 88–95, Sapporo, Japan.

Herve Dejean, Eric Gaussier, Cyril Goutte, and Kenji Yamada. 2003. Reducing parameter space for word alignment. In Rada Mihalcea and Ted Pedersen, editors, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 23–26, Edmonton, Canada, May 31.

D. Hiemstra. 1996. Using statistical methods to create a bilingual dictionary. Master's thesis, Universiteit Twente.

Jonas Kuhn. 2004. Exploiting parallel corpora for monolingual grammar induction – a pilot study. In *Workshop proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 54–57, Lisbon, Portugal. LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, Massachusetts, London.

I. Dan Melamed. 2001. *Empirical Methods for exploiting parallel texts*. MIT Press, Cambridge, MA.

Magnus Merkel, Mikael Andersson, and Lars Ahrenberg. 2002. The PLUG link annotator – interactive construction of data from parallel corpora. In Lars Borin, editor, *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999*, pages 151–168. Rodopi, Amsterdam/New York, NY.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *NHLT-NAACL 2003 Workshop: Building and Using parallel Texts. Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton, Canada.

Sonja Nießen and Hermann Ney. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of the Machine Translation Summit VIII*, pages 247–252, Santiago de Compostela, Spain.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, England.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus – parallel and free. In *Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1183–1186, Lisbon, Portugal. http://logos.uio.no/opus/.

Kristina Toutanova, H. Tolga Ilhan, and Christopher D. Manning. 2002. Extensions to HMM-based statistical word alignment models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 87–94, Philadelphia, USA.

Patrick Tschorn and Anke Lüdeling. 2003. Morphological knowledge and alignment of english-german parallel corpora. In *Proceedings of the 2003 Corpus Linguistics Conference*, pages 818–827.