How does morphological complexity translate? A cross-linguistic case study for word alignment

Bettina Schrader
Institute of Cognitive Science – University of Osnabrück, Germany
bschrade@uos.de

Hapax legomena and other so-called rare events present an interesting problem for corpus-based applications: due to their low frequency, they fail to provide enough statistical data for applications like word alignment or statistical machine translation. Simultaneously, hapax legomena are often newly-coined words, i.e. they will probably not be listed in existing bilingual dictionaries. Hence, bilingual dictionaries cannot be used in these applications as a fallback option.

Effort is spent on improving word alignment techniques (cf. Mihalcea and Pedersen 2003, Cherry and Lin 2003, Toutanova et al. 2002). However, detailed linguistic knowledge on what kinds of words hapax legomena are, and how they are translated is not employed.

Here, we are presenting a new word alignment strategy specifically tailored to align German noun compounds with their corresponding English multiword expressions. Its implementation follows from the results of a case study on the morphological features of 512 German hapax legomena and their translations.

For our case-study, we have extracted 512 German hapax legomena from the Europarl corpus (Tiedemann and Nygaard, 2004), and analyzed these hapax legomena in terms of word category membership, morphological complexity, and word length. Additionally, we aligned them manually to their corresponding units in the English part of the corpus. Finally, we added information on the morphological or syntactic properties of the English terms, and characterized which kinds of alignment problems are to be expected if the corpus is automatically aligned.

The analysis yielded that roughly two-thirds of the hapaxes (353 out of 512 words) are nouns or noun compounds, and that their translations tend to be chunk-like multiword expressions in English. These English multiword expressions often consist of a sequence of nouns, like *subsidy mill* (German: *Subventionsmühle*), sometimes of one or more nouns followed by a prepositional phrase, as e.g. *damage to property* (German: *Eigentumsbeschädigungen*).

Additionally, we observed symmetries concerning the numbers of elements in a hapax and its translation: if a hapax is not a compound, then its translation is a single-word unit in 90% of the cases. If the hapax is a compound, however, the number of constituents within the compound will roughly equal those of its translation. In more than 50% of all cases, a hapax consisting of two free morphemes is translated by an expression consisting of two tokens. Most astonishingly, the length of a German nominal, counted in number of characters, closely corresponds to the length of its translation.

Hence we hypothesize that the complexity of an expression, here the morphological complexity of nominal compounds, tends to be retained during translation. Accordingly, if we know the morphological structure of a German compound, we can predict the structural properties of its translation.

We have used these results of our case study to implement a word alignment strategy: on a sentence-aligned, POS-tagged corpus, it recognizes German noun compounds based on their POS-tags and their lengths, counted in characters, and it determines the English translation candidates based on POS-patterns. These POS-patterns cover the most frequent structures in the English translations, namely nouns either preceded by one or more adjectives, or nouns followed by a prepositional phrase. As similarity measure for the alignment, we used the length ratio between the German compounds and their English translation candidates. All aligned translation pairs are used to generate a bilingual dictionary. No attempt is made at computing a complete text alignment, or to filter out incorrect translation pairs.

A first evaluation showed promising results: 67% of the 353 nominal compounds received an entry in the automatically generated dicitonary, with 115 lexicon entries containing the correct translations. After improving the recognition of compounds and multiword expressions, e.g. with respect to hyphenated words, results were even better: 70% of the compounds were now listed in the lexicon, with 175 entries containing the correct translation. Missing entries resulted from failures in compound recognition for very short compounds. Error sources in determining a correct translation were the similarity measure itself, or the fact that the POS-pattern of the correct translation was not accounted for. In very few cases, the translation consisted of a paraphrase of the compound, and hence it was impossible to determine.

As we did not impose any frequency restriction on which compounds or multiword expressions to align, the automatically generated dictionary also contains lexical entries for words with frequencies higher than one. Of these, we have randomly chosen more than 700 additional lexicon entries for evaluation, consisting of both rare events and frequent nouns. Since our implementation recognized compounds based on their POS-tag and word length only, without a proper morphological analysis, not all of these 700 nouns are compounds. We evaluated their lexicon entries nevertheless to see how well our hypothesis holds for them, as well.

Intermediate results show that our strategy is useful for aligning hapax nominals as well as non-hapax nominals, irrespective of their morphological complexity: 602 correct translations were found in 751 lexicon entries, including nominals with multiple translations (e.g. German *Berufsausbildung* was translated both as *vocational training* and *professional training*) . These results so far confirm our hypothesis that morphological complexity is kept during translation.

We are currently carrying out the error analysis on this data. Although this evaluation is not yet complete, we have observed that the frequency of a noun apparently influences the number of alternative translations. These alternative translations involve synonymous nominal constructions, changes of word category, and paraphrases.

Additionally, we are investigating more thoroughly for the nominals in our data set which types of compounds occurred in our data, and how their internal structures relate to the structural properties of the corresponding English expressions, apart from the symmetries already observed.

References

Cherry, C. and D. Lin (2003). A probability model to improve word alignment. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan.

Mihalcea, R. and T. Pedersen (2003). An evaluation exercise for word alignment. In NHLT-NAACL 2003 Workshop: Building and Using parallel Texts. Data Driven Machine Translation and Beyond, pp. 1–10. Edmonton.

Tiedemann, J. and L. Nygaard (2004). The opus corpus – parallel and free. In Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC), pp. 1183–1186. Lisbon, Portugal. Http://logos.uio.no/opus/.

Toutanova, K., H. T. Ilhan, and C. D. Manning (2002). Extensions to hmm-based statistical word alignment models. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 87–94.