Towards a multi-purpose gold standard annotation of a multi-parallel corpus

Bettina Schrader, Jonas Kuhn
Institut für Linguistik,
Universität Potsdam
bschrade@uni-potsdam.de
kuhn@ling.uni-potsdam.de

1 Introduction

Parallel corpora have proven useful for a variety of purposes, including but not limited to statistical machine translation (SMT) (cf. Brown et al. 1993) and ``annotation projection" across languages as training data for NLP-tools of various kinds (cf. Yarowsky et al. 2001, Hwa et al. 2002, Mitkov and Barbu 2004, Pado and Lapata 2005). Moreover, if the corpora's annotations include manual word alignment information, they can be used e.g. for evaluating word alignment approaches (cf. Mihalcea and Pedersen 2003, 2005).

Relatively little work had been spent on systematic hand annotation of word alignment information in parallel corpora -- i.e., on explicitly marking correspondences at the word level. Gold standard word alignments have been produced for evaluation purposes (Och and Ney 2003), but until recently with a fairly *ad hoc* annotation, and these gold standard alignments have often been rather small and restricted to few language pairs. Furthermore, these gold standard alignments are rarely used outside the relatively narrow task of evaluating components of SMT systems. Parallel and multi-parallel corpora that offer larger sets of word alignments, or that include additional kinds of information like syntactic trees, are still harder to come by.

A notable exception to the observed lack of systematic approaches to alignment annotation has been the work by the Stockholm group (e.g., Volk and Samuelsson 2004) with their research agenda for building a multi-parallel treebank, including manually annotated links at the word and phrasal level. We may call the Stockholm approach syntax-annotation-driven, since monolingual syntactic treebanking (following guidelines like the Penn Treebank guidelines for English and the NEGRA/TIGER guidelines for German) precedes the annotation of cross-language links. With the possibility of marking higher-level (i.e., phrasal) correspondences, the role of the word-level alignment becomes somewhat less central in the final annotated corpus, such that not-so-clear word alignment links are left out in this approach.

Given the broad applicability of aligned parallel corpora not only in technologically-oriented NLP work, but also in linguistically motivated research, we believe that linguistically grounded alignment annotation of parallel corpora deserves more attention -- and here we include strictly word-based alignment annotation as one important variant. Most of the current non-MT applications of parallel corpora mentioned initially rely on noisy but easy-to-obtain unsupervised word alignments, often leading to surprisingly good performance of the final system trained under the annotation projection idea. However, due to the lack of systematically hand-aligned resources, it is hard to do a more careful error analysis and decide where it is most effective to invest additional development work.

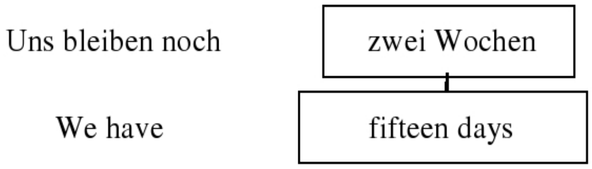
Hence, medium-sized parallel corpora with carefully controlled gold standard annotation of various kinds will be of high value for a number of tasks, and can thus be reused in a highly flexible way, both for evaluations of existing tools and for cross-linguistic research, especially if a multi-parallel corpus is used such that various language pairs are covered. Our longer-term goal is to build such a multi-parallel corpus with multi-level information as the basis for experiments with weakly supervised learning techniques on parallel corpora. For such a corpus, hand-annotated word alignment is an important starting point. Higher-level (phrasal) alignment has the advantage of allowing for a better control of the translational correspondence, but since it depends on a specific syntactic annotation, it is less flexible and cannot be used as a gold standard oracle for statistical word alignment which will continue to play a central role in (semi-)automatic work on parallel corpora.

2 Gold Standard Annotation

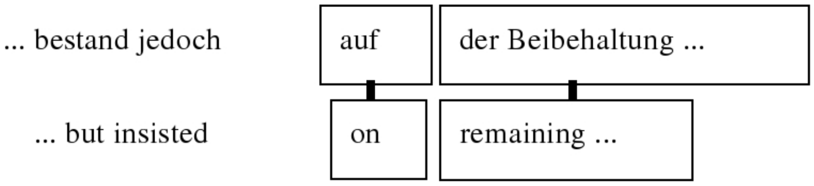
Presently, we have i) formulated annotation guidelines which are based on linguistic principles and are thus largely independent of the language pair, and have ii) annotated a sample of English-German sentence pairs taken from the *Europarl* corpus (Köhn 2005) .

2.1 Annotation Guidelines

The annotation guideline has been inspired both by existing guidelines such as those developed by Melamed (1998), and by the experience of the annotators during a guideline-free test alignment run on 100 sentence pairs. The guiding principle of our guideline is that words or more generally, expressions, should be aligned context-dependently, i.e. if two expressions are used to convey the same meaning *in the given sentence pair*. This principle directly leads to the alignment of multiword units like the following:



Furthermore, the guideline consists of 16 specific rules that specify how to align nominals, prepositions, verbs, verb clusters, etc. Structural divergences are aligned based on context-dependent translational equivalence, as in the following example:



Annotation Process

During gold standard annotation, we corrected the automatic sentence alignment of a 100,000 token sample of the German and English *Europarl* texts, and subsequently annotated 242 sentence pairs at the word level, randomly chosen among the sentence pairs of the Europarl subset. Two annotators annotated each sentence pair independently of each other, and annotation differences were subsequently resolved by discussion. The inter-annotator agreement of the word alignment gold standard – 0.644, corresponding to 0.57 precision and 0.72 recall – was measured using the Dice-coefficient, following a suggestion by Melamed (1998b) , the kappa-statistics not being applicable here.

3 Further Work

With only 242 randomly chosen sentence links, consisting of 4788 German and 5336 English tokens, the word-aligned gold standard is relatively small. However, it is already large enough to assess the usefulness of a given word alignment technique for a specific purpose (such as special-purpose linguistic search), and especially its role in a combined task involving a parallel corpus.

In the future, we will use part of the small gold standard to semi-automatically bootstrap a larger annotated corpus for the same language pair, and we also plan to add word alignment information and other kinds of linguistic annotation for more language pairs, e.g. for the pair German—Dutch. In combination with other annotation levels, the gold standard word alignment will play an important role in our longer-term goal of developing weakly supervised learning techniques for interactive linguistic corpus exploration.

References

Brown, P., Della Pietra, S., Della Pietra, V. and R. Mercer (1993): The mathematics of machine translation: Parameter Estimation. Computational Linguistics 19(2), pp. 263-311

Hwa, R., P. Resnik and A. Weinberg (2002): Breaking the Resource Bottleneck for Multilingual Parsing. In: Proceedings of LREC

Köhn, P. (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of the 10th Machine Translation Summit, Phuket, Thailand, pp. 79-86

Melamed, D (1998): Annotation Style Guide for the BLINKER Project. Tech Report No. 98-06, Institute for Research in Cognitive Science, University of Pennsylvania

Melamed, D (1998b): Manual Annotation of Translational Equivalence: The BLINKER Project, Tech Report No. 98-07, Institute for Research in Cognitive Science, University of Pennsylvania

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. Edmonton, Canada, pp. 1-10

Mihalcea, R. and T. Pedersen (2005): Word Alignment for Languages with Scarce Resources. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts. Ann Arbor, Michigan, pp. 65--74

Mitkov, R. and C. Barbu (2004): Using bilingual corpora to improve pronoun resolution. Languages in Contrast 4(2), pp. 201-211

Och, F. and H. Ney (2003): A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), pp. 19-51

Pado, S. and M. Lapata (2005): Cross-linguistic projection of role-semantic information. In: Proceedings of HLT/EMNLP, Vancouver, Canada

Yarowsky, D., G. Ngai and R. Wicentowski (2001): Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In: Proceedings of the first International Conference on Human Language Technology Research (HLT), San Diego, USA, pp. 200-207

Volk, M. and Y. Samuelsson (2004) : Bootstrapping Parallel Treebanks. In: Proceedings of the Workshop on Linguistically Interpreted Corpora (LINC) at COLING, Geneva