Annotation Guidelines for Czech-English Word Alignment

Ivana Kruijff-Korbayová[*], Klára Chvátalová[†], Oana Postolache[*]

a[*]Saarland University, Saarbrücken, Germany
{korbay, oana@coli.uni-sb.de }
a[†]Charles University, Prague, Czech Republic
klara.chvatalova@centrum.cz

Abstract

Here needs to come an abstract.

1. Introduction

Parallel multilingual corpora aligned at the sentence- or word-level are a valuable resource for developing machine translation systems and, recently, projecting annotations across word alignments. Our goals are in the latter group. In particular, we experiment with the projection of *information structure* on a Czech-English parallel corpus, namely a portion of the Prague Czech-English Dependency Treebank version 1.0 (**?**). We annotate information structure in Czech automatically (**?**). In order to project the annotation, we need an alignment of the tree nodes or at least of the surface words.

We first created automatic word alignment of the PCEDT data by GIZA++ (**?**). However, an informal examination established that the quality is too low for our purposes. Therefore, we decided for manual alignment. Since there existed no guidelines for aligning Czech and English, we took the Annotation Style Guide of the Blinker Project (henceforth BASG) (**?**) as a starting point, because it has been reused in several projects dealing with word alignment.

In this paper we report on our experience with applying BASG to word alignment of Czech and English text. Overall, we found that the general rules in BASG which were originally developed for English and French can be applied for English and Czech as well. We identified a range of systematically occurring differences between the two languages, for which we felt the need to add more specific guidelines. In Section 2. we describe the PCEDT corpus, the alignment annotation process and the tool we used. In Section 3. we overview our extensions of BASG .

2. Manual Word Alignment on the PCEDT

Manual word alignment was performed on the text part of the Prague Czech-English Dependency Treebank 1.0 (PCEDT) (**?**). The English sentences originate from the Wall-Street Journal part of the Penn Treebank corpus. They were translated by native speakers of Czech, who were instructed to translate sentence-by-sentence, and keep the translation both accurate and as close to the English original as possible.

We used a word alignment annotation and visualization tool implemented by Chris Callison-Burch (University of Edinburgh). The tool presents for each pair of sentences as a matrix of clickable squares. Aligned word-pairs (or phrases) are represented by filled squares. The filling has two color degrees (black and grey), representing whether the annotator is *sure* or *unsure* of the alignment link.

We encountered systematically occurring cases for which we wished to be able to distinguish between *strong* and *weak* alignment. Some of them are discussed in Section 3.. In the current version of the tool we used the grey squares for weak alignment, thus overloading their semantics to encode both weak alignment and annotator's uncertainty.

The process leading to the formulation of the present guidelines involved a coordinator supervising the project and one annotator, both native speakers of Czech proficient in English. First the coordinator annotated a trial set of 20 sentences according to BASG and sketched several additional rules for the annotator. The annotator then annotated the same trial set. The annotations were automatically compared and the differences and rules discussed. The annotator wrote the first version of the additional guidelines, and then annotated two more data sets. The annotator discussed additional guidelines with the coordinator regularly, and updated the guidelines.

The aligned data set consists of 285 sentences. This covers all files in the development set, and some of the training set. The Czech files contain 7,706 words, the English files 7,902 (including punctuation marks). To compare the automatic and manual alignment, we computed the Alignment Error Rate (AER) (**?**) for GIZA++ against the annotator: The average AER is 0.348 with a standard deviation of 0.071.

3. Extensions of BASG

In our guidelines we discuss about 20 types of cases for which we extended or elaborated BASG , plus a few miscellaneous instances, and some additional examples; the guidelines contain 97 examples of alignment. They are organized similarly to BASG and we use similar headings when possible. The complete guidelines are available online: `www.coli.uni-sb.de/~korbay/alignment/`.

Included among the cases discussed in the guidelines are the following phenomena:

Articles and Determiners English uses articles whereas Czech does not. There are also differences in the use of possessive pronouns as determiners. We align articles and determiners present only in one language to the head of the corresponding noun phrase.

Case marking English often uses prepositions or possessive markers where Czech inflects the head noun of a phrase. We therefore align the former with the latter.

Zero subject Czech is a subject pro-drop language. We align the subject in English with the corresponding main verb in Czech, which also carries the agreement features.

Different types of attributes Attributes are often expressed differently in the two languages, particularly when English uses nominal premodifiers of nouns. Czech then often uses a construction with a non-congruent attribute, which involves an additional head noun. We align such additional nouns with the corresponding head noun by grey squares (weak alignment).

Additional common nouns Czech often uses a common noun in addition to a proper noun, such as the name of an institution, company, date expression, etc. in order to avoid inflecting them. We align the common noun with the corresponding head noun by grey squares (weak alignment).

Negation involving pronouns Unlike English, Czech employs negative congruence and uses a negative verb form along with a negative pronoun. It is straightforward to align the pronouns and the verb forms. However, this alignment results in aligning a negative verb form in Czech with a positive one in English. In order to explicitly encode the involvement of all parts in negation, we additionally align as a phrase all words reflecting the negation.

In the full paper, we will discuss these and several other phenomena (e.g., reflexive pronouns, numerals, various types of subordinate clauses) in more detail, and present examples.

4. Conclusions and Further Work

We presented our experience from manual word alignment of a Czech-English parallel corpus. We briefly discussed several cases of systematic differences between these languages for which we extended the existing guidelines (**?**). The obvious next step is to evaluate whether these guidelines lead to more consistent annotations and an improvement in inter-annotator agreement.

We have also experienced the need to make a distinction between strong and weak alignment, in order to adequately represent certain systematically occurring cases of cross-lingual correspondence. Typically, such correspondence involves a part which fits the concept of word-to-word semantic equivalence, and another part where the relationship is weaker, e.g., added words. Leaving the weakly equivalent part unaligned means losing some information, but annotating such cases as phrase alignment also means losing the information about the strongly equivalent parts. Therefore, we propose to include a labeling of strong vs. weak alignment besides the already commonly used labeling of sure vs. unsure alignment.

5. References