English–Hungarian NP Alignment in a Linguistically Enriched Translation Memory

Example based machine translation (EBMT, Nagao 1984, Somers 2003) requires sub-sentence level alignment, the process of automatic identification of corresponding sub-sentence level segment pairs in human translations. In the proposed presentation we will describe and analyze the noun phrase (NP) alignment techniques developed for MetaMorpho TM a linguistically enriched translation memory system (Hodász–Pohl 2005).

Our translation memory (TM) differs from the mostly language independent commercially available TM products in the following features. (1) In order to find the stored segment most similar to the searched one, a morpho-syntactic similarity measure is applied. (2) Not only whole sentences are searched in the memory. NPs, and the sentence skeleton (derived from the sentence by substituting NPs with symbolic NP slots) are also searched in the database, and their most probable translations are morphologically altered and combined to form a possible translation sentence. Hence the recall of the translation memory is increased by suggesting translations built up from NP and sentence skeleton translations looked up separately. In such an EBMT system, not only full sentence translations but also NP and sentence skeleton translations have to be stored in the memory. NPs of the stored sentence pair have to be aligned either by the translator or by an automatic means. Leaving the tedious task of NP alignment to the translator would decrease productivity, thus an automatic means of NP alignment was developed for MetaMorpho TM.

Previous and related works include corpus-based statistical phrase alignment methods, and parse tree alignment techniques for EBMT systems. Recent advances in tree alignment (e.g. Groves, 2004) are promising but English and Hungarian parsers are too different for such methods depending on the internal structure of NP trees. Corpus-based statistical means, like the one developed by Kupiec (1993), are more robust, but they require reprocessing of the whole translation memory after a new sentence pair is stored. In a TM product a new sentence pair should be stored in less than one second, so in our NP aligner we substituted the time consuming statistical data collection for dictionary usage.

In our first experiment NPs of the stored sentence pair were extracted by the MetaMorpho English and Hungarian parser (Prószéky–Tihanyi 2002). The Hungarian grammar was in its early development stage, so later on we developed another means of Hungarian NP extraction, which aims to find Hungarian counterparts of English NPs without using a Hungarian parser.

Our heuristic NP aligner algorithm calculates a matching score for all possible English-Hungarian NP pairs, extracts the best-matching pairs, and applies a threshold to filter ambiguous alignment links. The matching score is based on tokenized dictionary matching, cognate matching (Simard et al. 1992) and POS matching. The similarity measure allows limited insertion or deletion of grammar words (prepositions, pronouns, determinants).

First dictionary matching is done then cognates are searched among the unmatched words. Finally POS matching is calculated among the previously unmatched words. If any function word remains unmatched, it is discarded with a small penalty in the matching score.

During the dictionary-based matching, all possible stems of words in an English NP are looked up in the dictionary using a stem index. The dictionary index also contains references to phrases (or multi-word lexemes). Longer matches are preferred so matching is started from the longest dictionary entries. This way if "*hard disk drive*" is in an English NP and in the dictionary too, the shorter matching dictionary entries, "*hard disk*", "*disk drive*", "*hard*", "*disk*", "*drive*" are not matched to words in the NP. A dictionary entry found in an English NP is considered to be found in a Hungarian NP if at least one possible stem of all tokens of the entry can be matched to an unmatched token of the Hungarian NP.

Cognate matching is done similarly to the process described by Simard et al.

By matching the part of speech tags of words not found in the dictionary, the recall of the aligner can be increased. In our experiment we used only basic POS categories like noun, verb, adjective, preposition, determinant, pronoun, etc.

Elimination of unmatched function words is important because there are differences in grammar words of English and Hungarian. In Hungarian different cases are used instead of prepositions of English, and English possessive pronouns usually correspond to a determinant and some case marking on the possessed Hungarian noun. Handling unmatched words without any distinction would result in lower recall.

At the time of our first experiments the Hungarian grammar was only partially implemented and had low recall and precision. Therefore, we developed a simple heuristic means of extracting Hungarian NP candidates by mapping the words of the English NPs to the words of the Hungarian sentence. Each word of an English NP is mapped to all possible word positions in the Hungarian sentence using stemmed dictionary matching and cognate matching. The word mappings producing the smallest number of words between the mapped Hungarian words are selected. After selecting the matched words in the Hungarian sentence, unmatched words of the English NP are matched to the unmatched words between matched Hungarian words if their POS categories are "compatible". Later on the Hungarian NP may be expanded to the left (preferably) or to the right depending on the POS of unmatched words and a basic Hungarian NP grammar. (The grammar contains simple rules, as determinants on the left side of the NP are considered part of the Hungarian NP even if they had no counterpart in the English NP.) After "guessing" the Hungarian NPs, we calculate English–Hungarian NP matching scores the same way we do with parsed Hungarian NPs.

The first experiments with the new NP aligner and Hungarian NP guesser algorithms were carried out on a small part of the SZAK corpus (Kis–Kis, 2003) containing 40 sentence pairs (with the average length of 23 words measured on the English side). We used a small bilingual dictionary containing 116,000 word and phrase mappings. Alignment precision was 85% and could have been increased to 91% if the translator marked the sentences, where more than half of the NPs were translated into VPs. Precision was 65% among the hand-alignable NPs (56% of all English NPs in the small test corpus).

Work that will be finished before the conference: Currently we are building a larger hand-aligned test corpus, so more results can be expected before the conference takes place. Recently the Hungarian parser (available to us) improved a lot, so we are going to compare the results of using the parser instead of our NP guesser algorithm.

References

Groves, D.; Hearne, M.; Way, A. (2004): Robust Sub-Sentential Alignment of Phrase-Structure Trees. COLING '04, Geneva, Switzerland.

Hodász, G.; Pohl, G (2005): MetaMorpho TM: a linguistically enriched translation memory. In: *International Workshop, Modern Approaches in Translation Technologies* (W. Hahn, J. Hutchins and C. Vertan, Eds.) Borovets, 2005. pp. 26-30.

Kis, Á; Kis, B. (2003): A Prescriptive Corpus-based Technical Dictionary. In: *Papers in Computational Lexicography: Proceedings of COMPLEX 2003.* Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest.

Kupiec, J. (1993): An Algorithm for finding Noun Phrase Correspondences in Bilingual Corpora. In: *Proceedings of the 31st annual meeting on Association for Computational Linguistic*s, pp. 17-22

Nagao, M. (1984): A framework of a mechanical translation between Japanese and English by analogy principle. In: A. Elithorn and R. Banerji (eds.) *Artificial and human intelligence*, 173-180. Amsterdam, North-Holland.

Prószéky, G.; Tihanyi, L. (2002): MetaMorpho: A Pattern-Based Machine Translation Project. In: *24th 'Translating and the Computer' Conference* London, 2002. pp. 19-24.

Simard, M.; Foster, G.; Isabelle, P. (1992): Using Cognates to Align Sentences in Bilingual Corpora. In: *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine translation*, (TMI92), Montreal, pp. 67-81.

Somers, H. (2003): An Overview of EBMT. In M. Carl. and A. Way. (eds.) *Recent Advances in Example-based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.3-57.