Back to the Future: The Case for Direct English-Swedish MT

Lars Ahrenberg
Maria Holmqvist

Department of Computer and Information Science,
Linköpings universitet
SE – 58183 Linköping

In this presentation we will report work on a data-driven direct machine translation system that has been under development for some years in the Linköping part of the VINNOVA-sponsored project “Corpus-based Machine Translation”. An important aim for the system is that it should support quick development of domain-restricted machine translation. Thus, the system relies on a development environment with alignment and analysis tools that can produce the data required for a given translation task from a parallel corpus.

We currently call the system T4F, which is an abbreviation for tokenization, tagging, transfer, transposition and filtering, all of which are main phases/modules of the system. Other modules such as ranking of alternative translations are used in the system, but have not (yet) been deemed to be important enough to enter into the acronym.

The presentation will have three parts that are shortly summarized below.

1. The case for direct English-Swedish translation

For a long time direct systems have been out of vogue. As Yorick Wilks writes in his introduction to Part II: Theoretical and Metodological Issues of the Readings in Machine Translation (p. 204):

"... no one seems to defend direct coupling methods since SYSTRAN declared itself to be a transfer system some years ago."

Yet, SMT, Statistical Machine Translation, which enjoys much interest and a high status at the moment, is the most direct approach of all, since it claims that you can build MT systems from nothing else but data and alignment tools. Statistical MT is usually presented as an approach of its own, but it has much in common with the classical direct approaches, in particular the reliance on word correspondences.

Of the three classical models for machine translation, the interlingua approach clearly stands out as an extreme and easily defined model. The difference between a direct approach and a syntactic transfer approach is more difficult to tell, at least in practice if not in principle. There are however some traits that tend to re-appear when direct systems are characterized. They are:

1. they are designed for a specific language-pair and direction of translation;
2. they exploit word correspondences and similarities between the two languages as far as possible, and use syntactic and semantic analyses only to the extent that it is necessary for the translation quality;
3. the central data source of a direct system is the dictionary, where correspondences are stored; rules for solving ambiguities are closely linked to entries in the dictionary;
4. they tend to work in a stepwise fashion, performing sequential substitutions of the input, often on a word-for-word basis, and, at least in the early days, with no clear separation of analysis and generation;
5. in matching an input sentence to the dictionary, the longest possible match is chosen;
6. differences in word order are handled by transposition rules;

In contrast, syntactic transfer approaches are based on an analysis phase that aims at capturing a complete formal syntactic representation for the sentence as a whole. Once this information has been acquired it is also used as the object of transfer. This difference in the object of transfer: a sentence analysis vs. a set of word analyses may actually be regarded as the main difference between syntactic transfer systems and direct systems. But it must be stressed that direct systems do not preclude syntactic or semantic analyses. There is a pragmatic constraint on the analysis, though, that it is subordinated to the translation task.

Another difference concerns generation. A pure transfer system relies on a grammar for the target language to derive target sentences, while a direct system uses the word order of the source sentence as the point of departure for deriving a proper word order for the translation.

The choice between direct models and transfer models for a given language pair must of course also consider empirical issues. If the two languages are structurally similar, in particular as regards lexical correspondences, morphology and word order, the case for abstract syntactic analysis seems less convincing. To introduce an abstract representation of a concrete object for the sole purpose of creating another concrete object which is more or less isomorphic to the first is just an unnecessary roundabout. The question is where to draw the line. In the case of translation from English to Swedish, the following seems to hold:

- Translation units correspond well in source and target. Of course there are cases of deletions, additions and shifts of various kinds but they are relatively rare and a large majority of them can be tied to lexical entries.
- Grammatical morphemes correspond fairly well in numbers and use. This means that to a morphological variant of an English word corresponds only a small number of morphological variants for any Swedish translation.
- There are seldom no more than two word order shifts from source to target sentence. Moreover, most of them are obligatory and regular.

2. The T4F model

T4F has three distinct phases: analysis, transfer and selection. The analysis phase includes tokenization, i.e., the identification and normalization of the appropriate translation units from the source sentence, and tagging, i.e., the assignment of relevant linguistic information to the translation units. We make a distinction between inherent (or language-internal) information and contextual information. A parser for inherent information can be re-used for other language pairs, while the contextual information is motivated by the target language. As an example we categorize English nouns as definite or indefinite on the basis of their context so as to make the choice of a corresponding Swedish noun easier.

In the current system we use the FDG parser of Connexor with some post-processing to obtain the inherent information and a specialized set of rules for contextual tagging. The latter makes use of
the dependency information provided by the parser, and word order, to determine values for contextual tags.

In the transfer phase the English source tokens are considered one by one. For any English token e all Swedish tokens that (i) are defined as possible translations of e, and (ii) have properties that match the (inherent and contextual) properties of e, are retrieved. Null words as well as multi-word expressions are possible Swedish tokens. If no Swedish token can be found, the English token is used. In this way we get a set of possible translations for each source token. The number of elements in this set may be quite high, e.g., in the case of prepositions. To reduce the size of the sets a filtering module is applied. In addition, there are rules for expanding multi-word expressions and rules for changing word order. Filtering rules may be applied also to re-ordered sets.

When the sets cannot be reduced further, target sentences are derived by combining all remaining tokens. As alternative translations and bigrams can be assigned probabilities on the basis of their frequencies in the corpus, this process may produce ranked lists of translations.

### 3. Evaluation and discussion

The current system has been evaluated on two corpora. The first one is the ATIS corpus used by Jonsson (2001) and Ahrenberg & Jonsson (2001), where the purpose was to investigate whether the additional information provided by dependency relations would actually be of help. The second corpus is an English-Swedish parallel corpus of on-line help texts for MS Access XP. In this case, comparisons were made with a baseline system using only the token dictionary of T4F and the procedure for deriving ranked translations. The BLEU measure has been used as a numerical measure of translation quality. T4F generally performs better than the baseline, but the difference varies with the quality of the lexicon, which in turn depends on the output from the word alignment systems. If automatic alignment is used and the alignment file is not revised, the filtering and reordering rules are less applicable and the burden of selecting a translation falls on the probabilistic ranking procedure.

More detailed results from these evaluations will be discussed in the presentation. We will also discuss some cases that are problematic for the current model, such as negations, word order differences between main and subordinate clauses, prepositions and genitives. The latter introduce dependencies that are hard to sort out without recourse to a constituent level.

### References

