A Hybrid Machine Translation System for Typologically Related Languages

Petr Homola and **Vladislav Kuboň**
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
110 00 Praha 1, Czech Republic
{homola,vk}@ufal.mff.cuni.cz

Abstract

This paper describes a shallow parsing formalism aiming at machine translation between closely related languages. The formalism allows to write grammar rules helping to (partially) disambiguate chunks in input sentences. The chunks are then translatred into the target language without any deep syntactic or semantic processing. A stochastic ranker then selects the best translation according to the target language model. The results obtained for Czech and Slovak are presented.

Introduction

Although the automatic translation of closely related languages is a subject considered by many linguists as slightly inferior compared to the full-fledged machine translation (MT) of unrelated language pairs, it is at the same time a stimulating field providing a number of interesting research topics. It has been investigated recently for numerous language groups — for Slavic langauges in (Homola & Kuboň 2004), for Turkic languages in (Altintas & Cicekli 2002) and for Scandinavian languages in (Dyvik 1995).

All MT systems for closely related languages mentioned above are based on 'shallow' methods. The close relatedness of languages from the same language group guarantees that there are usually only minor syntactic differences which can be handled by shallow parsing of a source language or by 'shallow transfer'.

Data structures

In our framework, the basic data structure for the representation of linguistic data is a typed feature structure. It is used for the representation of individual words. The interaction between words is modelled by a chain graph. It is used as an auxiliary data structure for both parsing (analysis) and synthesis. It represents all hypotheses that are valid up to a certain point in the parsing process. At the end of it, the remaining valid hypotheses build up the result set (the parsing process is described in detail in (Colmerauer 1969)).

Rules

The structure of the rules

The grammar for analysis and synthesis consists of declarative rules that prescribe how to combine phrases into complex structures or how to decompose them. In our system, the rules are context-free.

A rule can be applied if the right-hand side matches the categories of a subchain in the chain graph and all conditions associated with the rule are met. In such a case, a new edge or a subchain of new edges is added to the chain graph which spans over the edges that are covered by the right-hand side of the rule. The feature structure the new edge is labeled with is based on one of the feature structures of the covered edges and possibly extended by unification according to the conditions associated with the rule.

There are several kinds of syntactic rules which are described in more detail in the following subsections.

Shallow analysis

'Shallow' rules are used to identify the so-called chunks and their internal syntactic structure. Chunks may be, for instance, nominal or prepositional phrases. A simple but very frequent example may be the combination of an adjective and a noun that are adjacent and agree in gender, case and number. Shallow rules do not usually reflect the relationship (mainly dependencies) between the main verb and its complements.

Other analytical rules

Other analytical rules concern mainly verb phrases. Constituents of the sentence are attached to verbs and the edges are labeled with grammatical functions (such as subject, object etc.). In general, the result of the parsing process up to here is a set of feature structures that cover continuous segments of the input sentence. Ideally, the result is one feature structure that covers the whole sentence. But more often, the result is a forest of dependency trees.

Deep syntactic rules and verbal valency

Although the presented framework is focused on shallow parsing, it also allows for coping with deep syntax in some extent in case it is needed.

Deep syntactic rules are used to represent the linguistic meaning of a sentence. These rules mainly map grammatical functions to thematic roles such as agent, patient etc. Each verb in the sentence has to be analyzed deeply so that it becomes a dependent of another verb (this also holds for participles).

Structural transfer and synthesis

Structural transfer rules are used to adapt the structure of subtrees so that it is grammatical in the target language. In simple cases, such a rule only linearizes a subtree. Here is an example of a generative rule that linearizes the subject from a verb phrase:

```
(t_vp -> t_vp / diff np (subj) (
  (^ (subj) ! .)
  (. (t_case) = nom)
))
```

Statistical postprocessing

An essential part of the whole MT system is the statistical postprocessor. The main problem with the formalism described in the previous section is that both the analyzer and the transfer are non-deterministic in general, i.e., they generate multiple results. It would be very complicated to resolve this kind of ambiguity by hand-written rules. That is why we have implemented a stochastic post-processor which aims at the selection one particular sentence that suits best the given context.

We use a simple language model based on trigrams (trained on word forms without any morphological annotation) which is intended to sort out "wrong" target sentences (these include grammatically ill-formed sentences as well as inappropriate lexical mapping). The current model has been trained on a corpus of 18.8 million words which have been randomly chosen from the Slovak Wikipedia.

Let us present an example of how this component of the system works. In the source text we had the following Czech segment:

(1) *Společnost* *ve* *zprávě*
company-FEM,SG,NOM in report-FEM,SG,LOC
uvedla
inform-LPART,FEM,SG
"The company informed in the report, . . . "

The rule-based part of the system has generated two target segments:

1) *Spoločnosť vo zpráve uviedli*,
2) *Spoločnosť vo zpráve uviedla*.

The Czech word *uvedla* is ambiguous (fem.sg and neu.pl). According to the language model, the ranker has (correctly) chosen the second sentence as the most probable result.

There are also many homonymic word forms that result in different lemmas in the target languages. For example, the word *pak* means both "then" and "fool-pl.gen", the word *tři* means "three" and the imperative of "to scrub", *ženu* means "wife-sg.acc" and "(I'm) hurrying out" etc. The ranker is supposed to sort out the contextually wrong meaning in all these cases.

We have evaluated the system on approx. 300 text segments from technical and news domain. The metrics we are using is the Levenshtein edit distance between the automatic translation and a reference translation. Given accuracies for all sentences we use the arithmetic mean as the translation accuracy of the whole text. In our test data, 35% of segments have been translated 'perfectly'. For 12% of segments, the system has generated a perfect translation but the ranker has chosen a different one. In general, the accuracy of the translation is 95.33%.

Conclusions

The formalism presented in this paper provides sufficient expressive power for shallow parsing of natural languages, with a specific attention to the issues present in richly inflected languages. The formalism allows to parse syntactic constructions which pose a problem for very simple systems of MT between closely related languages. These systems are based mainly on morphological tagging and adding a shallow parser module may help to increase the quality of the output.

The quality of the output of our experimental system has been significantly improved by a stochastical postprocessor. The combination of a comparatively simple context-free grammar with hand-written rules (based on unification) and a statistical language model seems to ensure reasonable translation quality for pairs of relatively close languages.

There are some topics left for the future research. The current language model is very simple, and although it improves the translation significantly, more experiments with various (possibly morphologically enriched) data have to be performed so that the optimal strategy of combining rule-base and stochastical methods can be found. Moreover an extensive evaluation based on a standard metrics would help to compare the performance of the system with other MT frameworks. This work is currently in progress.

Acknowledgments

The research presented in this paper has been supported by the grant No. 1ET100300517 of the GAAV ČR.

References

Altintas, K., and Cicekli, I. 2002. A Machine Translation System between a Pair of Closely Related Languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*, 192–196.

Colmerauer, A. 1969. Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. Technical report, Mimeo, Montréal.

Dyvik, H. 1995. Exploiting Structural Similarities in Machine Translation. *Computers and Humanities* 28:225–245.

Homola, P., and Kuboň, V. 2004. A translation model for languages of acceding countries. In *Proceedings of the IX EAMT Workshop*. La Valetta: University of Malta.