ČESÍLKO – an MT system for closely related languages

Jan HAJIČ
ÚFAL MFF UK
Malostranské nám.25
Praha 1, Czech Republic, 11800
hajic@ufal.mff.cuni.cz

Jan HRIC
KTI MFF UK
Malostranské nám.25
Praha 1, Czech Republic, 11800
hric@barbora.mff.cuni.cz

Vladislav KUBOŇ
ÚFAL MFF UK
Malostranské nám.25
Praha 1, Czech Republic, 11800
vk@ufal.mff.cuni.cz

Abstract

The demonstration of our system addresses one very important part of the translation business – the localization of texts and programs from one source language into a group of mutually related target languages. It shows step by step a simple method of machine translation between related languages and its incorporation into an existing commercial translation aid using the concept of translation memory. It is quite clear that the localization of the same source into several typologically similar target languages, one language pair after another, is a waste of money and effort. In the translation process it is necessary to solve very similar problems for each source-target language pair. The use of one language from the target group as a pivot and to perform the translation and localization through this language seems to be quite natural solution of these problems. It is of course much easier to translate texts from Czech to Polish or from Russian to Bulgarian than from English or German to any of these languages.

Introduction

As part of our "pivot" language solution, we are using a combination of an MT system with a commercial machine aided translation (MAT) system. We are using the TRADOS™ system, although any such system will do. The system uses the concept of translation memory, which contains pairs of previously translated sentences from a source to a target language. When a human translator starts translating a new sentence, the system tries to match (with a degree of similarity set by a user) the source with sentences already stored in the translation memory. If found, the human translator decides whether to use it, to modify it or to reject it.

1 Translation Memory Integration

The segmentation of the translation memory (the texts are stored as relevant pairs of source/target language sentences) is the key feature of our method. The translation memory may be exported into a text file and thus allows for an easy manipulation with its content. Let us suppose that we have at our disposal two translation memories – one human made for the source/pivot language pair and the other created by an MT system for the pivot/target language pair. The substitution of segments of a pivot language by the segments of a target language is then only a routine procedure. The human translator translating from the source to the target language then gets a translation memory for the required pair (source/target); there is no trace of the pivot language left. The system of penalties applied in TRADOS Translator's Workbench guarantees that a previous human-made translation present in the memory gets higher priority than the automatic translation. This method has at least three advantages:

- The use of machine-made translation memory only as a resource supporting the direct human translation from the source to the target language has no negative effect on the quality of translation and from the user's point of view.
- There is no difference (except for the small difference in the quality of translation memories) when our method is used compared to the original process of working with the support of solely human-made translation memories.
- The third advantage is the fact that given a sufficient quality of the MT from the pivot to the target language, our method may substantially increase the speed and reduce the costs of the translation from the source to the target languages.

2 The System ČESÍLKO

The system ČESÍLKO (cf. Hajič, Hric, Kuboň (2000)) was designed for the method of translation and localization described above with Czech as a pivot language for other Slavic languages. Up to now we have fully implemented Czech-to-Slovak translation (this is the pair of two most closely related Slavic languages) by means of word-for-word translation using stochastic disambiguation of Czech word forms. The tagging system is based on statistics, it uses an exponential model of probability distribution – see Hajič, Hladká (1998). The morphological analysis of Czech is based on the morphological dictionary developed by Jan Hajič and Hana Skoumalová in 1988-99 (for latest description, see Hajič (1998)). The dictionary covers over 700 000 items and it is able to recognize more than 15 mil. word-forms. We have already started work on Czech-to-Polish translation and we also plan to incorporate into our system the already existing Czech-to-Russian MT system RUSLAN (cf. Oliva (1989)).

2.1 Solving Input Ambiguity

The greatest problem of the word-for-word translation approach is the problem of ambiguity of individual word forms. The type of ambiguity is slightly different in languages with a rich inflection (majority of Slavic languages) and in languages which do not have such a wide variety of forms derived from a single lemma. The main problem is that even though several Slavic languages have the same property as Czech, the ambiguity is not matching, it is distributed in a different manner and the "form-for-form" translation is not applicable.

Although we believe that a more profound linguistic analysis is needed to get even the subtle differences right, for now we are using an alternative way to the solution of this problem, namely, the application of a stochastically based morphological disambiguator for Czech whose success rate is close to 92%. Our system therefore consists of the following modules:

1. Import of the input from empty translation memory
2. Morphological analysis of Czech
3. Morphological disambiguation
4. Domain-related bilingual glossaries
5. General bilingual dictionary
6. Morphological synthesis of Slovak
7. Export of the output to the original translation memory

2.2 Evaluation

For the evaluation of our system we have exploited the close connection between our system and the TRADOS Translator's Workbench. The method is simple – the human translator receives the translation memory created by our system and translates the text using this memory. The translator is free to make any changes to the text proposed by the translation memory. The target text created by a human translator is then compared with the text created by the mechanical application of translation memory to the source text. TRADOS then evaluates the percentage of matching in the same manner as it normally evaluates the percentage of matching of source text with sentences in translation memory. In the first testing on relatively large texts (tens of thousands words) the translation created by our system achieved about 90% match (as defined by the TRADOS match module) with the results of human translation

Acknowledgements

This project was supported by the grant GAČR 405/96/K214 and partially by the grant GAČR 201/99/0236 and project of the Ministry of Education No. VS96151.

References

Hajič, Jan (1998). *Building and Using a Syntactically Annotated Corpus: The Prague Dependency Treebank.* In: Festschrift for Jarmila Panevová, Karolinum Press, Charles University, Prague. pp. 106—132.

Hajič, Jan and Barbora Hladká (1998). *Tagging Inflective Languages. Prediction of Morphological Categories for a Rich, Structured Tagset.* ACL-Coling'98, Montreal, Canada, August 1998, pp. 483-490.

Hajič, Jan, Hric, Jan and Kuboň, Vladislav (2000). *Machine Translation of Very Close Languages.* In: Proceedings of the ANLP 2000, Seattle, U.S.A., April 2000, pp. 7-12.

Oliva, Karel (1989). *A Parser for Czech Implemented in Systems Q*; Explizite Beschreibung der Sprache und automatische Textbearbeitung XVI, MFF UK Prague