

Machine Translation of Very Close Languages

Jan HAJIČ
Computer Science Dept.
Johns Hopkins University
3400 N. Charles St., Baltimore,
MD 21218, USA
hajic@cs.jhu.edu

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

Using examples of the transfer-based MT system between Czech and Russian RUSLAN and the word-for-word MT system with morphological disambiguation between Czech and Slovak ČESILKO we argue that for really close languages it is possible to obtain better translation quality by means of simpler methods. The problem of translation to a group of typologically similar languages using a pivot language is also discussed here.

Introduction

Although the field of machine translation has a very long history, the number of really successful systems is not very impressive. Most of the funds invested into the development of various MT systems have been wasted and have not stimulated a development of techniques which would allow to translate at least technical texts from a certain limited domain. There were, of course, exceptions, which demonstrated that under certain conditions it is possible to develop a system which will save money and efforts invested into human translation. The main reason why the field of MT has not met the expectations of sci-fi literature, but also the expectations of scientific community, is the complexity of the task itself. A successful automatic translation system requires an application of techniques from several areas of computational linguistics (morphology, syntax, semantics, discourse analysis etc.) as a necessary, but not a sufficient condition. The general opinion is that it is easier to create an MT system for a pair of related languages. In our contribution we would like to

demonstrate that this assumption holds only for really very closely related languages.

1. Czech-to-Russian MT system RUSLAN

1.1 History

The first attempt to verify the hypothesis that related languages are easier to translate started in mid 80s at Charles University in Prague. The project was called RUSLAN and aimed at the translation of documentation in the domain of operating systems for mainframe computers. It was developed in cooperation with the Research Institute of Mathematical Machines in Prague. At that time in former COMECON countries it was obligatory to translate any kind of documentation to such systems into Russian. The work on the Czech-to-Russian MT system RUSLAN (cf. Oliva (1989)) started in 1985. It was terminated in 1990 (with COMECON gone) for the lack of funding.

1.2 System description

The system was rule-based, implemented in Colmerauer's Q-systems. It contained a full-fledged morphological and syntactic analysis of Czech, a transfer and a syntactic and morphological generation of Russian. There was almost no transfer at the beginning of the project due to the assumption that both languages are similar to the extent that does not require any transfer phase at all. This assumption turned to be wrong and several phenomena were covered by the transfer in the later stage of the project (for example the translation of the Czech verb "být" [to be] into one of the three possible Russian equivalents: empty form, the form "byť" in future

tense and the verb “javljat6sja”; or the translation of verbal negation).

At the time when the work was terminated in 1990, the system had a main translation dictionary of about 8000 words, accompanied by so called transducing dictionary covering another 2000 words. The transducing dictionary was based on the original idea described in Kirschner (1987). It aimed at the exploitation of the fact that technical terms are based (in a majority of European languages) on Greek or Latin stems, adopted according to the particular derivational rules of the given languages. This fact allows for the “translation” of technical terms by means of a direct transcription of productive endings and a slight (regular) adjustment of the spelling of the stem. For example, the English words *localization* and *discrimination* can be transcribed into Czech as “lokalizace” and “diskriminace” with a productive ending -ation being transcribed to -ace. It was generally assumed that for the pair Czech/Russian the transducing dictionary would be able to profit from a substantially greater number of productive rules. This hypothesis proved to be wrong, too (see Bémová, Kuboň (1990)). The set of productive endings for both pairs (English/Czech, as developed for an earlier MT system from English to Czech, and Czech/Russian) was very similar.

The evaluation of results of RUSLAN showed that roughly 40% of input sentences were translated correctly, about 40% with minor errors correctable by a human post-editor and about 20% of the input required substantial editing or re-translation. There were two main factors that caused a deterioration of the translation. The first factor was the incompleteness of the main dictionary of the system. Even though the system contained a set of so-called fail-soft rules, whose task was to handle such situations, an unknown word typically caused a failure of the module of syntactic analysis, because the dictionary entries contained - besides the translation equivalents and morphological information - very important syntactic information.

The second factor was the module of syntactic analysis of Czech. There were several reasons of parsing failures. Apart from the common inability of most rule-based formal grammars to cover a

particular natural language to the finest detail of its syntax there were other problems. One of them was the existence of non-projective constructions, which are quite common in Czech even in relatively short sentences. Even though they account only for 1.7% of syntactic dependencies, every third Czech sentence contains at least one, and in a news corpus, we discovered as much as 15 non-projective dependencies; see also Hajič et al. (1998). An example of a non-projective construction is “Soubor se nepodařilo otevřít.” [lit.: File *Refl.* was_not_possible to_open. – It was not possible to open the file]. The formalism used for the implementation (Q-systems) was not meant to handle non-projective constructions. Another source of trouble was the use of so-called semantic features. These features were based on lexical semantics of individual words. Their main task was to support a semantically plausible analysis and to block the implausible ones. It turned out that the question of implausible combinations of semantic features is also more complex than it was supposed to be. The practical outcome of the use of semantic features was a higher ratio of parsing failures – semantic features often blocked a plausible analysis. For example, human lexicographers assigned the verb ‘to run’ a semantic feature stating that only a noun with semantic features of a human or other living being may be assigned the role of subject of this verb. The input text was however full of sentences with ‘programs’ or ‘systems’ running etc. It was of course very easy to correct the semantic feature in the dictionary, but the problem was that there were far too many corrections required.

On the other hand, the fact that both languages allow a high degree of word-order freedom accounted for a certain simplification of the translation process. The grammar relied on the fact that there are only minor word-order differences between Czech and Russian.

1.3 Lessons learned from RUSLAN

We have learned several lessons regarding the MT of closely related languages:

- The transfer-based approach provides a similar quality of translation both for closely related and typologically different languages
- Two main bottlenecks of full-fledged transfer-based systems are:

- complexity of the syntactic dictionary
- relative unreliability of the syntactic analysis of the source language
- Even a relatively simple component (transducing dictionary) was equally complex for English-to-Czech and Czech-to-Russian translation
- Limited text domains do not exist in real life, it is necessary to work with a high coverage dictionary at least for the source language.

2. Translation and localization

2.1 A pivot language

Localization of products and their documentation is a great problem for any company, which wants to strengthen its position on foreign language market, especially for companies producing various kinds of software. The amounts of texts being localized are huge and the localization costs are huge as well.

It is quite clear that the localization from one source language to several target languages, which are typologically similar, but different from the source language, is a waste of money and effort. 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. There are several reasons, why localization and translation is not being performed through some pivot language, representing a certain group of closely related languages. Apart from political reasons the translation through a pivot language has several drawbacks. The most important one is the problem of the loss of translation quality. Each translation may to a certain extent shift the meaning of the translated text and thus each subsequent translation provides results more and more different from the original. The second most important reason is the lack of translators from the pivot to the target language, while this is usually no problem for the translation from the source directly to the target language.

2.2 Translation memory is the key

The main goal of this paper is to suggest how to overcome these obstacles by means of a combination of an MT system with commercial

MAHT (Machine-aided human translation) systems. We have chosen the TRADOS Translator's Workbench as a representative system of a class of these products, which can be characterized as an example-based translation tools. IBM's Translation Manager and other products also belong to this class. Such systems uses so-called 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 the source with sentences already stored in the translation memory. If it is successful, it suggests the translation and the human translator decides whether to use it, to modify it or to reject it.

The segmentation of a translation memory is a key feature for our system. The translation memory may be exported into a text file and thus allows 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 language to the target language then gets a translation memory for the required pair (source/target). The system of penalties applied in TRADOS Translator's Workbench (or a similar system) guarantees that if there is already a human-made translation present, then it gets higher priority than the translation obtained as a result of the automatic MT. This system solves both problems mentioned above – the human translators from the pivot to the target language are not needed at all and the machine-made translation memory serves only as a resource supporting the direct human translation from the source to the target language.

3. Machine translation of (very) closely related Slavic languages

In the group of Slavic languages, there are more closely related languages than Czech and Russian. Apart from the pair of Serbian and Croatian languages, which are almost identical and were

considered one language just a few years ago, the most closely related languages in this group are Czech and Slovak.

This fact has led us to an experiment with automatic translation between Czech and Slovak. It was clear that application of a similar method to that one used in the system RUSLAN would lead to similar results. Due to the closeness of both languages we have decided to apply a simpler method. Our new system, ČESÍLKO, aims at a maximal exploitation of the similarity of both languages. The system uses the method of direct word-for-word translation, justified by the similarity of syntactic constructions of both languages.

Although the system is currently being tested on texts from the domain of documentation to corporate information systems, it is not limited to any specific domain. Its primary task is, however, to provide support for translation and localization of various technical texts.

3.1 System ČESÍLKO

The greatest problem of the word-for-word translation approach (for languages with very similar syntax and word order, but different morphological system) is the problem of morphological 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. For example, in Czech there are only rare cases of part-of-speech ambiguities (stát [to stay/the state], žena [woman/chasing] or tři [three/rub(imperative)]), much more frequent is the ambiguity of gender, number and case (for example, the form of the adjective *jarní* [spring] is 27-times ambiguous). The main problem is that even though several Slavic languages have the same property as Czech, the ambiguity is not preserved. It is distributed in a different manner and the “form-for-form” translation is not applicable.

Without the analysis of at least nominal groups it is often very difficult to solve this problem, because for example the actual morphemic categories of adjectives are in Czech distinguishable only on the basis of gender, number and case agreement between an adjective

and its governing noun. An alternative way to the solution of this problem was the application of a stochastically based morphological disambiguator (morphological tagger) for Czech whose success rate is close to 92%. Our system therefore consists of the following modules:

1. Import of the input from so-called ‘empty’ translation memory
2. Morphological analysis of Czech
3. Morphological disambiguation
4. Domain-related bilingual glossaries (incl. single- and multiword terminology)
5. General bilingual dictionary
6. Morphological synthesis of Slovak
7. Export of the output to the original translation memory

Let us now look in a more detail at the individual modules of the system:

ad 1. The input text is extracted out of a translation memory previously exported into an ASCII file. The exported translation memory (of TRADOS) has a SGML-like notation with a relatively simple structure (cf. the following example):

Example 1. – A sample of the exported translation memory

```
<RTF Preamble>...</RTF Preamble>
<TrU>
<CrD>23051999
<CrU>VK
<Seg L=CS_01>Pomocí výkazu ad-hoc můžete rychle a jednoduše vytvářet řešerše.
<Seg L=SK_01>n/a
</TrU>
```

Our system uses only the segments marked by <Seg L=CS_01>, which contain one source language sentence each, and <Seg L=SK_01>, which is empty and which will later contain the same sentence translated into the target language by ČESÍLKO.

ad 2. 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 contains over 700 000 dictionary entries and its typical coverage varies between

99% (novels) to 95% (technical texts). The morphological analysis uses the system of positional tags with 15 positions (each morphological category, such as Part-of-speech, Number, Gender, Case, etc. has a fixed, single-symbol place in the tag).

Example 2 – tags assigned to the word-form “pomoci” (help/by means of)

pomoci:

NFP2-----A---- | NFS7-----A---- | R--2-----

where :

N – noun; R – preposition

F – feminine gender

S – singular, P – plural

7, 2 – case (7 - instrumental, 2 - genitive)

A – affirmative (non negative)

ad 3. The module of morphological disambiguation is a key to the success of the translation. It gets an average number of 3.58 tags per token (word form in text) as an input. The tagging system is purely statistical, and it uses a log-linear model of probability distribution – see Hajič, Hladká (1998). The learning is based on a manually tagged corpus of Czech texts (mostly from the general newspaper domain). The system learns contextual rules (features) automatically and also automatically determines feature weights. The average accuracy of tagging is between 91 and 93% and remains the same even for technical texts (if we disregard the unknown names and foreign-language terms that are not ambiguous anyway).

The lemmatization immediately follows tagging; it chooses the first lemma with a possible tag corresponding to the tag selected. Despite this simple lemmatization method, and also thanks to the fact that Czech words are rarely ambiguous in their Part-of-speech, it works with an accuracy exceeding 98%.

ad 4. The domain-related bilingual glossaries contain pairs of individual words and pairs of multiple-word terms. The glossaries are organized into a hierarchy specified by the user; typically, the glossaries for the most specific domain are applied first. There is one general matching rule for all levels of glossaries – the longest match wins.

The multiple-word terms are sequences of lemmas (not word forms). This structure has several advantages, among others it allows to minimize the size of the dictionary and also, due to the simplicity of the structure, it allows modifications of the glossaries by the linguistically naive user. The necessary morphological information is introduced into the domain-related glossary in an off-line preprocessing stage, which does not require user intervention. This makes a big difference when compared to the RUSLAN Czech-to-Russian MT system, when each multiword dictionary entry cost about 30 minutes of linguistic expert’s time on average.

ad 5. The main bilingual dictionary contains data necessary for the translation of both lemmas and tags. The translation of tags (from the Czech into the Slovak morphological system) is necessary, because due to the morphological differences both systems use close, but slightly different tagsets. Currently the system handles the 1:1 translation of tags (and 2:2, 3:3, etc.). Different ratio of translation is very rare between Czech and Slovak, but nevertheless an advanced system of dictionary items is under construction (for the translation 1:2, 2:1 etc.). It is quite interesting that the lexically homonymous words often preserve their homonymy even after the translation, so no special treatment of homonyms is deemed necessary.

ad 6. The morphological synthesis of Slovak is based on a monolingual dictionary of Slovak, developed by J.Hric (1991-99), covering more than 100,000 dictionary entries. The coverage of the dictionary is not as high as of the Czech one, but it is still growing. It aims at a similar coverage of Slovak as we enjoy for Czech.

ad 7. The export of the output of the system ČESÍLKO into the translation memory (of TRADOS Translator’s Workbench) amounts mainly to cleaning of all irrelevant SGML markers. The whole resulting Slovak sentence is inserted into the appropriate location in the original translation memory file. The following example also shows that the marker <CrU> contains an information that the target language sentence was created by an MT system.

Example 3. – A sample of the translation memory containing the results of MT

```
<RTF Preamble>...</RTF Preamble>
<TrU>
<CrD>23051999
<CrU>MT!
<Seg L=CS_01>Pomocí výkazu ad-hoc můžete rychle a jednoduše vytvářet řešerše.
<Seg L=SK_01>Pomocí výkazov ad-hoc můžete rýchlo a jednoducho vytvárať rešerše.
</TrU>
```

3.2 Evaluation of results

The problem how to evaluate results of automatic translation is very difficult. 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. Our system achieved about 90% match (as defined by the TRADOS match module) with the results of human translation, based on a relatively large (more than 10,000 words) test sample.

4. Conclusions

The accuracy of the translation achieved by our system justifies the hypothesis that word-for-word translation might be a solution for MT of really closely related languages. The remaining problems to be solved are problems with the one-to-many or many-to-many translation, where the lack of information in glossaries and dictionaries sometimes causes an unnecessary translation error.

The success of the system ČESÍLKO has encouraged the investigation of the possibility to use the same method for other pairs of Slavic

languages, namely for Czech-to-Polish translation. Although these languages are not so similar as Czech and Slovak, we hope that an addition of a simple partial noun phrase parsing might provide results with the quality comparable to the full-fledged syntactic analysis based system RUSLAN (this is of course true also for the Czech-to-Slovak translation). The first results of Czech-to-Polish translation are quite encouraging in this respect, even though we could not perform as rigorous testing as we did for Slovak.

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

- Bémová, Alevtina and Kuboň, Vladislav (1990). *Czech-to-Russian Transducing Dictionary*; In: Proceedings of the XIIIth COLING conference, Helsinki 1990
- Hajič, Jan (1998). *Building and Using a Syntactically Annotated Coprus: 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; Brill, Eric; Collins, Michael; Hladká Barbora; Jones, Douglas; Kuo, Cynthia; Ramshaw, Lance; Schwartz, Oren; Tillman, Christoph; and Zeman, Daniel: *Core Natural Language Processing Technology Applicable to Multiple Languages*. The Workshop'98 Final Report. CLSP JHU. Also at: <http://www.clsp.jhu.edu/ws98/projects/nlp/report>.
- Kirschner, Zdeněk (1987). *APAC3-2: An English-to-Czech Machine Translation System*; Explizite Beschreibung der Sprache und automatische Textbearbeitung XIII, MFF UK Prague
- Oliva, Karel (1989). *A Parser for Czech Implemented in Systems Q*; Explizite Beschreibung der Sprache und automatische Textbearbeitung XVI, MFF UK Prague