Towards a Rule-Based Machine Translation System Between Czech and Russian

1. Introduction

In this paper we describe the on-going work on developing the Machine Translation (MT) system between Czech and Russian. This system is to be implemented within the project Česílko – an MT system between the closely-related languages, in which the relatedness of Slavic languages is exploited (Hajič et al. 2003). Česílko already includes Czech-to-Slovak, Czech-to-Polish and Czech-to-Lithuanian MT systems. The first one was based merely on the direct word-to-word translation, and as the languages are very closely-related and share most syntactic features, the results were more than satisfactory. Czech and Russian are more distant languages, so the additional step of a syntactic transfer may be needed. Here we will present the development of the dictionary and the initial steps in writing transfer rules.

2. Czech-Russian Dictionary

2.1 Looking for the dictionary

It was not so easy to find a Czech-Russian dictionary in a plain text format, because almost all the dictionaries were within the on-line search. We could not convert the commercial dictionaries into a machine-readable format either. The only way was to induce it with the help of the available resources. First we intended to make use of Ruslan Czech-Russian dictionary (Oliva 1989), but it contains only 6000 words and is adjusted to the special format of that MT system, that was never finished.
The way that we chose was to extract the dictionary from the freely available parallel Czech-Russian corpus UMC 0.1 (Klyueva and Bojar 2008). We used sentences, that were aligned 1-to-1 as those most reliable.

2.2 Processing the corpus

Over 88,000 sentences parallel in Czech and Russian were processed by taggers. For Russian we used TreeTagger and for Czech the Hajic's tagger in order to get lemmas. In our study we did not use tags. Word forms were therefore substituted by lemmas. The output of the parallelly lemmatized text looks like the following:
(1) *v zoufalý snaha udržet se u moc hodit Parvíz Mušaraf pákistánský ústavní rámec za hlava a vyhlásit v země výjimečný stav . || в отчаянный стремление удерживать власть , Первез Мушарраф отвергнуть конституционный система пакистан и объявить о введение чрезвычайный положение .*

Therefore we ran GIZA++ word alignment (Och and Ney 2003) on the lemmatized corpus and got the table of word correspondences, which are exactly the machine-readable dictionary we need. Example 2 shows the first word entries from the dictionary where translation pairs are sorted by the frequency.
(2)
a|и
být|быть
země|страна|gender=fem
tento|этот

který|который
...
The nouns on the russian side were enriched with the morphological feature of gender and animatedness.

2.3 Cleaning the dictionary

The number of entries that we got was over than 400,000 words, and we faced the question how much we should trust the word alignment. During manually processing we found lots of incorrespondences so we decided to delete the double entries on the Czech side even at a price of non-resolved ambiguity.

3. Syntactic Transfer Module

As we mentioned earlier, the syntax of Czech and Russian differs in a number of constructions, though not so radically as for example Czech and English. So our next step will be to think over the transfer rules, that can capture the syntactic differences between the two languages. Here we will name only some of them.

(3) Construction with "to be"
Я студент or *Я - студент* (ru) vs. *Jsem student.* (cz)
'*I am a student*.'

(4) Past tence of the verb "to be"
Я был дома (ru) vs. *Byl jsem doma* (cz)

(5) The verb "to have"
У меня есть кошка (ru) vs. *Mám kočku* (cz)

In the future we plan to make a list of transfer rules, that is going to capture the most frequent and evident incorrespondences.

4. Conclusion
So now we have a Czech-Russian dictionary with over 40,000 entries which is enough to run the first experiments on a rule-based MT system. We are currently working on the set of rules on syntactic transfer between Czech and Russian.

References

Klyueva Natalia and Ondřej Bojar: UMC 0.1: Czech-Russian-English Multilingual Corpus. Proceedings of the Conference "Corpora 2008". St.Petersburg, 2008

Hajič, J., Homola, P. and Kuboň, V. : A simple multilingual machine translation system. In Proceedings of the MT Summit IX, New Orleans, 2003.

Och F.J. and Ney H.. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, March 2003.

Oliva, K.: A Parser for Czech Implemented in Systems Q, in Explizite Beschreibung der Sprache und automatische Textbearbeitung, MFF UK Prague, 1989.