A Machine Translation System into a Minority Language

Petr Homola and Vladislav Kuboň
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
Prague 1, Czech Republic
{homola,vk}@ufal.mff.cuni.cz

Abstract

The paper presents a machine translation system from Czech to Lower Sorbian, a minority language spoken in a region around Cottbus in Germany. This West Slavonic language, which is spoken by less than 20,000 people, is very archaic, it, has supine, dual and some other grammatical forms, which disappeared in most Slavonic languages. The paper describes the architecture of the system and focuses on morphological disambiguation, partial syntactic parser and lexical and structural transfer. First evaluation results on a small set of sentences are also presented.

Keywords: machine translation, minority languages, shallow NLP

1 Introduction

There are several major problems of minority languages in the modern society. In the age of globalization, there is a strong pressure to use a majority language everywhere, and although the democratic governments usually pay a great deal of attention to the needs of minorities, minority languages always are in danger of dissolving. One of the possible ways how to help to preserve a minority language might be using an MT system for producing relatively cheap translations from other languages, thus making available the texts which would not normally be translated.

This paper suggests a solution how to exploit the proximity of related languages for such a relatively simple MT system. The system is not new — it already exists for several language pairs (Czech-Slovak, Czech-Polish. Czech-Lithuanian), cf. (Hajič *et al.* 03). the extension concerns Lower Sorbian, a minority language spoken in Germany in the area around Cottbus.

2 Česílko — a multilingual MT system for related languages

The system Česílko has been developed as an answer to a growing need of translation and localization from one source language to many target languages. It is quite clear that the independent translation or localization of the same document into several typologically similar target languages is a waste of effort and money. Our solution proposes to use one language from the target group as a pivot and to perform the translation through this language. It is quite true that applying the pivot language approach has a serious drawback — the translation quality, which needs to be very high, may deteriorate in this two-step process. A negligible shift of the meaning during the translation into a pivot language may be amplified by a subsequent translation from the pivot language to the actual target language.

In order to overcome these problems we have suggested an approach combining the human-made translation from the source language into a pivot language with a machine translation between a pivot and a (closely related) target language. The reviewer of the target language text may then review the translation against the original source language text and he thus can eliminate any problem caused by the translation from the source into the pivot language.

The system consists of the following steps:

1. Morphological analysis of Czech
2. Morphological disambiguation of Czech by means of a stochastic tagger
3. Search in the domain-related bilingual glossaries
4. Search in the general bilingual dictionary
5. Morphological synthesis of the target language

The necessity to account for phenomena which cannot be handled by this very simple architecture led us to the inclusion of a shallow parsing module for Czech for some of the language pairs. This module directly follows the morphological disambiguation of Czech.

2.1 Czech-to-Slovak

The architecture described above has been in fact inspired by the almost absolute syntactic similarity between Czech and Slovak. It was therefore quite natural to apply it for the first time for the translation between Czech as a source language and Slovak as a target language. The basic premise of the system was to use as simple method of analysis and transfer as possible. The system Česílko therefore uses the method of direct word-for-word translation, the use of which is justified by the similarity (even though not identity) of syntactic constructions in both languages.

The system has been tested on texts from the domain of documentation of corporate information systems. It is, however, not limited to any specific domain; it has also undergone thorough testing on rather difficult texts of a Czech general encyclopedia, and in an cross-lingual treebank annotation transfer project. Its primary task is, however, to provide support for translation and localization of various technical texts.

Since Czech and Slovak have almost the same syntax, the greatest problem of the word-for-word translation approach is the problem of ambiguity of word forms. 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(imper.)]), however, the ambiguity of gender, number and case is very high (for example, the form of the adjective *jarní* [spring] is 27-way ambiguous). Even though several Slavic languages have the same property as Czech, the ambiguity is not preserved at all or it is preserved only partially, it is distributed in a different manner and the "form-for-form" translation is not applicable.

2.2 Czech-to-Polish

After the initial success with Slovak, the best candidate for a new target language was Polish. It is close enough to Czech but it contains several phenomena that are different and provide thus the natural "next step".

The Polish morphological data was kindly provided to us by Morphologic, Inc. (Budapest, Hungary). We converted the data for use with our morphological generator. In general, according to our expectations, with the decreasing similarity level also the quality of results has decreased.

The main problems concerned word-order. agreement and different verbal valency frames.

2.3 Czech-to-Lithuanian

The tests of the Czech-to-Polish module confirmed our assumption that with decreasing similarity of both languages the quality of results will also decrease. It was also confirmed by an analysis of the planned Czech-to-Russian module described in (Homola 02). The paper suggested that one possible way of improving the quality of the translation would be an exploitation of a partial transfer.

The interesting question was whether it is possible to cross a borderline between different language groups. Due to the fact that Slavic and Baltic languages are relatively typologically similar (rich morphology, relatively free word order), it was decided to test, the limits of the method by developing a Czech-to-Lithuanian module.

The initial comparative study showed that for Czech-to-Lithuanian translation it is necessary to enrich the scheme of the system by creating a shallow parser working with the results of the tagger and preceding the dictionary lookup phase.

The module of a shallow syntactic analysis of Czech is based on the LFG formalism, even though it does not use the complete LFG framework, as described in (Bresnan 01). We leave out e.g. the completeness and coherence conditions and anaphoric binding. The main goal of the module is to analyze only the simpler parts (constituents) of the sentence, such as nominal and prepositional phrases. The result of this module is an underspecified dependency tree.[1]

3 Basic facts about Lower Sorbian

Sorbian is a West Slavonic minority language spoken in Lusatia in Germany. It splits into many dialects which differ significantly from each other. Two written standards are used in the present, Upper Sorbian in Saxony and Lower Sorbian in Brandenburg. We have chosen Lower Sorbian for our experiments, mainly because there exists a morphological tool capable of generating inflected forms from many lemmas obtained as a result of the translation process.

Both morphology and syntax of Lower Sorbian are very similar to Czech, nevertheless the grammar of Lower Sorbian is more complicated than

[1] This language pair has also been extended by a named entity recognition component (Homola & Piskorski 04).

the Czech one since the Lower Sorbian language is much more archaic. In the following text we describe some aspects of Lower Sorbian which are important with respect to MT from Czech.

- Lower Sorbian has *dual,* a special number used instead of plural for the amount 2, e.g. *dub (1), duba* (2), *duby* (more than 2). We ignore this number because the number of persons or objects can only be decided with a proper understanding of the context. This may result in an translation error although the sentence as such is grammatical, but such a strategy is unavoidable if we want to keep the whole system as simple as possible.

- The *supine* is another grammatical form which is not present in Czech. It is an infinite verb form used to express a goal or decisions, usually together with a verb of movement, e.g., *źi spat* "go to sleep" (cf. the infinitive form *spaś).*

- The system of tenses is richer in Lower Sorbian. Whereas Czech only uses one periphrastic past form, Lower Sorbian also has synthetic past forms, *aorist* and *imperfect.* Nevertheless these forms are rarely used in contemporary texts, i.e.. one can use the periphrastic form to translate past tense.

- Lower Sorbian does not drop the auxiliary verb *byś* in the third person of the past form (cf. Czech *převzala* "took over" vs. Lower Sorbian *jo pśiwzeła).* We ignore this difference in the current version of the system, since the participle forms are the same for all persons, therefore the shallow parser does not deliver the information about the person at all.

One of the important things which really may substantially decrease the quality of output provided by our system is the word order. Due to the typological similarity of both languages and the fact that both Czech and Lower Sorbian use the word order to express topic-focus distribution, we can preserve the word order of the source (Czech) text. Word order would be an issue if one would like, for example, to insert syntactic elements (e.g.. auxiliary verbs in periphrastic tenses, see above) which are dropped in Czech, but we use no transfer rules for this phenomenon in this initial version of the system.

4 Implementation

In order to cope with some syntactic differences between Czech and some of the target languages, we have implemented an environment for interpreting context free grammars on feature structures (similar to LFG). The input of the grammar is supposed to be morphologically disambiguated. The completeness and coherence conditions (as defined in the LFG) are not applied, as most f-structures will be incomplete. Moreover, we use no valence lexicon. Partial f-structures (a chain of f-structures) are accepted as a result, but they must cover the whole sentence continuously.

The grammar consists of a set of phrase structure rules. Constraints (equations) are assigned to every element of the right-hand side of the rules. The application of phrase structure rules produces c-structures (which are not used in the further process), whereas constraints define the associated f-structures.

Rules consist of the left-hand and right-hand side. The left-hand side contains one non-terminal symbol of the grammar, the right-hand side can consist of several symbols.

Since the input of the grammar is supposed to be morphologically disambiguated, we use a stochastical tagger (Hajič 01). This solution has the disadvantage that the output of the tagger contains errors, but there is no possibility to get better results at the moment.[2]

The syntactic analysis of Czech only uses few rules. Our main goal is to analyze simpler parts (constituents) of the sentence, such as noun and prepositional phrases. Thus, every sentence is syntactically represented by a chain of f-structures in our system.

The f-structures are then processed by the transfer component, whose main task is to convert all lexical entries. The conversion involves translation of the lemmas (basic word forms) and changing morphological tags, if necessary, especially the gender of nouns. It is obvious that changing the gender can break an agreement within a constituent (typical for the scheme *adjective + noun),* if the noun governs attributes that have to agree with the governor (correct transla-

[2] (Žáčková 02) has proven that it, is not possible to disambiguate Czech texts by means of shallow syntactic: parsing.

tion is given in brackets):

(1)

srbský$_{masc}$	jazyk$_{masc}$
Sorbian	*language*
*serbski$_{masc}$	*rěc$_{fem}$
(serbska	rěc)

Thus, another task of the transfer component is to adapt morphological categories of dependents of the translated item to preserve the agreement. The same also concerns the agreement between prepositions and their objects.

Converted f-structures are linearized. The word order of the source sentence is preserved. Finally, linearized sentences are processed by the morphological synthesis, which gives the final output.

5 Evaluation

As it has been explained in previous sections, our translation method is very simple, the translation is not expected to be perfect, and post-editing of the result is necessary. Nevertheless, the result of translation can be understood without problems and can serve, for example, as raw translation.

In order to get results comparable to other language pairs, we have translated a small set of sentences from Czech to Lower Sorbian using our system and proof-read the result so that it was grammatical. The proof-read version has then been used to compute the accuracy of the translation using the Trados Translator's Workbench. This evaluation method has been described in a more detail in (Hajič *et al.* 03). We have used two parameters in the MT process:

- *Manual disambguation* Usually, the MT process is fully automatic. But since the stochastical tagger causes many errors which make it impossible for the parser to recognize constituents, we have disambiguated the source text manually to see how big the accuracy drop caused by the tagger errors is.
- *Shallow parser* We have made some experiments without the shallow parser to see how this component increases the accuracy. The accuracy without the parser is comparable to the results for Slovak (Hajič *et al.* 00) and Polish (Dębowski *et al.* 02), whereas the accuracy with the parser is comparable to the result for Lithuanian (Hajič *et al.* 03).

The accuracy for all four combinations of the parameters is given in Table 1.

	tagger	manual disamb.
no parser	92%	93%"
shallow parser	93%	95%

Table 1: Evaluation of the pair Czech-Lower Sorbian

\Ve see that the result without parsing is similar to the accuracy achieved for Slovak. However, the improvement if the parser is used is quite low. The reason are tagger errors which break the agreement in noun phrases. The consequence is wrong gender of adjectives in the translated text. Without the tagger errors, the parsers improved the accuracy up to 95%. This serious problem could be solved by a 'deeper' parser which would use non-disambiguated input. The development or integration of such a component will be included in our future work, one possibility could be the component described in (Zeman 01). The most common translation errors are dropped auxiliary verbs in periphrastic tense construction.

Table 2 summarizes the results from Czech to four different target languages using Česílko. The results achieved with the shallow parser are emphasized, the baseline for English (a commercial MT system has been used), which allows for a comparison of translation results among the related and non-related languages, is presented in italics and it is taken from (Hajič *et al.* 03).

target language	accuracy
English	30%
Slovak	90%.
Polish	71.4%
Lithuanian	**87.6%**
Lower Sorbian	92%/**93%**

Table 2: Evaluation of implemented target languages

6 Conclusions

This paper documents a fact that with a relatively simple method it is possible to achieve surprisingly good quality of machine translation even for a minority language spoken only by tens of thousands people. It also shows that the initial presupposition that it will be relatively easy to extend the original Czech-to-Slovak MT system to other related or syntactically similar languages was correct. The new language pair reuses

modules originally developed for other language pairs (the module of shallow syntactic analysis of Czech was for the first time included for Czech-to-Lithuanian translation). Most of the efforts devoted to extending a system goes towards building a bilingual dictionary and towards morphological synthesis of the new target language. We hope that it will be possible to extend the system even further by adding a new target language in the same way as we did for Lower Sorbian.

Acknowledgements

We are very indebted to Gerat Nagora and Georg Müller for their morphological tool *Sorborto* for Lower Sorbian and for the support with evaluation.

The work described in this paper has been supported by the grant of the Ministry of education of the Czech Republic No. MSM 0021620838.

References

(Bresnan 01) Joan Bresnan. *Lexical-functional syntax.* Blackwell Publishers, Oxford, 2001.

(Dębowski *et al*. 02) L. Dębowski, J. Hajič, and V. Kuboň. Testing the limits — Adding a new language to an MT system. In *Prague Bulletin of Mathematical Linguistics, pp. 95-102, Prague,* 2002.

(Hajič 01) J. Hajič. *Disambiguation of rich inflection (computational morphology of Czech).* Karolinum, Charles University Press, Prague. 2001.

(Hajič *et al*. 00) J. Hajič, J. Hric, and V. Kuboň. Machine translation of very close languages. In *Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, Washington, USA, April 2000, pp. 7-12,* 2000.

(Hajič *et al*. 03) J. Hajič, P. Homola, and V. Kuboň. A simple multilingual machine translation system, 2003.

(Homola & Piskorski 04) P. Homola and J. Piskorski. How can shallow NLP help a machine translation system. In *Proceedings of the Conference Human Language Technologies - The Baltic Perspective, Riga, Latvia,* 2004.

(Homola 02) P. Homola. Machine translation among Slavic languages. 2002.

(Žáčková 02) E. Žáčková. *Parciální syntaktická analýza (češtiny).* PhD thesis. Fakulta informatiky Masarykovy univerzity, Brno, 2002.

(Zeman 01) D. Zeman. How much will a RE-based preprocessor help a statistical parser? In *Proceedings of the Seventh International Workshop on Parsing Technologies, Tsinghua University Press, ISBN 7-302-04925-4, Beijing Daxue, Beijing,* 2001.