Problems Of Reusing An Existing MT System*

Ondřej Bojar, Petr Homola, Vladislav Kuboň
Institute of Formal and Applied Linguistics
ÚFAL MFF UK, Malostranské náměstí 25, Praha 1, CZ-11800
Czech Republic
{bojar,homola,vk}@ufal.mff.cuni.cz

Abstract

This paper describes an attempt to recycle parts of the Czech-to-Russian machine translation system (MT) in the new Czech-to-English MT system. The paper describes the overall architecture of the new system and the details of the modules which have been added. A special attention is paid to the problem of named entity recognition and to the method of automatic acquisition of lexico-syntactic information for the bilingual dictionary of the system. The paper concentrates on the problems encountered in the process of reusing existing modules and their solution.

1 Introduction

The last decade has witnessed several attempts to increase the quality of MT systems by introducing new methods. The strong stress on stochastic methods in the NLP in general and in the MT in particular, the attempts to develop hybrid systems, a wide acceptance of translation-memory based systems among the translation professionals, the aim at limited domain speech-to-speech translation systems, all these (and many other) trends have demonstrated encouraging results in recent years.

Developing and using new methods definitely moves the whole MT field forward, but one should not forget about all the effort invested into the old systems. Reusing at least some parts of those systems may help to decrease the costs of new systems, especially when one of the languages is not a ”big” language and therefore there is not such a wide range of tools, grammars, dictionaries available as for example for English, German, Japanese or Spanish. In this paper we would like to describe one such attempt to reuse the existing system for a new language pair.

2 The original system

One of the systems which was silently abandoned in early nineties was the system for the translation from Czech to Russian called RUSLAN (Oliva, 1989). It was being developed in the second half of eighties with the aim to translate texts from a relatively closed thematic domain, the domain of operating systems of mainframes.

The system used transfer-based architecture. The implementation of the system was almost completely done in Q-systems, a formalism created by Alain Colmerauer (Colmerauer, 1969) for the TAUM-METEO project. The Czech-to Russian system also relied upon a set of dictionaries containing all data exploited by individual modules of the system. Each lexical item in the main (bilingual) dictionary contained not only lexico-syntactic data (valency frames etc.), but also a set of semantic features.

The work on the system RUSLAN has been terminated in 1990, in the final phase of system testing and debugging. The reason was quite simple - after the political changes in 1989 there was no more any commercial demand for Czech to

*The work described in this paper has been supported by the grant of the Grant Agency of the Czech Republic GACR No.405/03/0914 and partially also by the grant of the Grant Agency of the Charles University GAUK No. 351/2005

Russian MT system.

The demand for Czech-English translation has grown dramatically during the years following the abandonment of the system RUSLAN. On the other hand, also the range of methods, tools and resources for MT has grown substantially. Several corpora were created for Czech, the most prominent ones being the morphologically annotated Czech National Corpus and syntactically annotated Prague Dependency Treebank. In 2002 we have started the work on the parallel bilingual Prague Czech English Dependency Treebank (PCEDT) (Cuřín et al., 2004), which contains about a half of the texts from PennTreebank 3 translated into Czech by native speakers. A large morphological dictionary of Czech has been developed (Hajič, 2001), allowing for a good quality morphological analysis of Czech, which has been tested in numerous commercial applications and scientific projects since then.

3 The background of the project

The main motivation for our Czech-English MT experiment was to test several hypotheses. The most prominent of these hypotheses concerns the level, at which it is reasonable to perform the transfer. Due to the differences between both languages it is not sufficient to perform the transfer immediately after the morphological analysis or shallow parsing, as it has been done in the MT system eslko aiming at the translation between closely related (and similar) languages [cf (Hajič et al., 2003)]. On the other hand, it is a question whether the typological differences between Czech and English justify the transfer being performed at the tectogrammatical (deep syntactic) level.

Last but not least, one of our aims was to develop a rule-based MT system with minimal possible costs, either reusing the existing modules or trying to use (semi)automatic methods whenever possible, concentrating on areas where using the human labor would be extremely expensive (for example building a large coverage bilingual dictionary, cf. the following paragraphs.)

4 Czech-English MT system

The main goal of our project is to develop an experimental MT system for the translation of texts from the PCEDT from Czech to English. The system investigates the possibility of reusing the existing resources (grammar, dictionary) in order to decrease the development time. It also exploits the parallel bilingual corpus of syntactically annotated texts, although not as a direct learning material, more like an additional source of linguistic data especially for the dictionary development and for the testing of the system.

The task is complicated by the fact that this translation direction is according to our opinion more complicated than the reverse one. There are several reasons for this claim; the most prominent one is the free word-order nature of the source language. It generally means that it is very often necessary to make substantial changes of the word order if we want to get a grammatical English sentence, while when translating from English to Czech the results are more or less grammatically correct and comprehensible even if we don't change the word order at all.

Another problem of the Czech-English translation is the insertion of articles. Czech doesn't use any articles and it is of course much easier to remove them from the text (when translating from English) than to insert a proper article on a proper place (when translating from Czech).

Let us now look at the individual modules of the new system.

4.1 Morphological analysis

Due to the limited size of the original morphosyntactic dictionary of the system it was necessary to replace the original module by a new one. The new module of morphological analysis of Czech (Hajič, 2001) has been already exploited in numerous applications. It covers almost the entire Czech language, with very few exceptions (it is estimated that it contains about 800 000 lemmas). It is very reliable, due to a really large coverage there are almost no unknown words in the whole PCEDT. The only problem was the incorporation of the new module into the system - the original module of syntactic analysis of Czech from the system RUSLAN was very closely bound to a dictionary lookup and to the morphological module. The new module also uses a different tagset.

4.2 Bilingual dictionary

The bilingual dictionary of the system RUSLAN contained approximately 8000 lexical items with a rich lexico-syntactic information. We have originally assumed that the information contained in the dictionary might be transformed and reused in the new system, but this assumption turned to be false. Although the information contained in the original bilingual dictionary is extremely valuable for the module of syntactic analysis of Czech, we have decided to sacrifice it. The mere 8000 lexical items constitute too small part of the new bilingual dictionary and we have decided to prefer handling the dictionary in a uniform way.

At the moment there are no Czech-English dictionaries exploitable in an MT system. The available machine-readable dictionaries built mainly for a human user (such as WinGED[1] or Svoboda (2001)) suffer from important limitations:

- Sometimes, several variants of translation are combined in one entry[2].
- No clear annotation of meta-language is present, although the entries contain valuable morphological or syntactic information to some extent. (E.g. valency frames are encoded by means of rather inconsistent abbreviations in plain text: *accession to = vstoupení do* or *adjudge sb. to be guilty = uznat vinným koho*.)
- Usually, no morphological information is given along the entries, although the morphological information can be vital for correctly recognizing an occurrence of the entry in a text. For example, an expression *kniha účetní* can be translated as either *an accounting book* or *a book of an accountant* depending whether the Czech word *účetní* is an adjective or a noun.
- No syntactic information is available and no consistent rules have been adopted by the lexicographers to annotate syntactic properties in plain text (such as putting the head of the clause as the first word).

From the point of view of structural machine translation, the lack of syntactic information in the translation dictionary is crucial. In the course of translation, the input sentence is syntactically analyzed before searching for foreign language equivalents. In order to check for presence of multi-word expressions in the input, the dictionary must encode the structural shape of such entries, otherwise the system does not know how to traverse the relevant part of the tree. Similarly, some expressions require some constraints to be met (such as an agreement in case or number) in the input text. If these constraints are not fulfilled, the proposed foreign language equivalent is not applicable.

The importance of valency (subcategorization) frames and their equivalents should be stressed, too. In the described system, already the syntactic analyzer requires verb and adjective valency frames in order to allow for specific syntactic constructions. In general, knowledge of translation equivalents of valencies is important to preserve the meaning (*přijít na nějaký nápad = come at an idea*, literal translation: *come on an idea*; *chodit na housle = attend violin lessons*, lit. *walk on violin*) or to handle auxiliary words properly (*čekat na něhoko = wait for somebody*, lit. *wait on sb.*; *říci něco = tell something* but *přejet něco = run over something*).

4.2.1 Dictionary cleanup

In order to handle the problems mentioned above, we performed an extensive cleanup of the data from available machine-readable dictionaries. The core steps of the cleanup are as follows:

Identifying meta-information.

We manually processed all the entries and searched for frequent words that typically encode some meta-information, such as *sth.*, *st.*, *oneself*. We also checked all entries ending with a word that is potentially a preposition. Based on the expression in the other language, we were able to recognize the meaning and identify, whether the suspicious word expresses a “slot” in the expression or whether it is a fixed part of the expression. (E.g. *mít o sobě vysoké mínění = think something*

[1] `http://www.rewin.cz/`

[2] Throughout the text, we use the term ENTRY as a synonym to translation pair, i.e. a pair of Czech and English expressions.

of oneself, only the word *oneself* encodes a slot, the word *something* is a fixed part of the expression.)

During this phase, entries encoding several translation variants at once were disassembled into separate translation pairs, too.

Part-of-speech disambiguation.

We processed the Czech part of each entry with a morphological analyzer (Hajič, 2001) and we performed manual part-of-speech disambiguation of expressions with ambiguity. It should be noted that automatic tagging would not provide us with satisfactory results due to the lack of sentential context around the expressions.

Adding morphological constraints.

Morphological constraints on word entries describe which values of morphological features are valid for each word of the entry or have to be shared among some words of the entry. Once identified, morphological constraints can be used to check whether a word group in the input text represents an entry or not. With respect to our final task (translation from Czech to English), we aim at Czech constraints only.

We decided to induce morphological constraints automatically, based on corpus examples of the entries. For each entry, we look up sentences that contain all the lemmas of the entry in a close neighborhood (but irrespective to the word order and possible presence of inserted extra words). We weight the instances to promote those with no intervening words and those with connected dependency graph. The list of weighted instances is scanned for both unary (such as "case is accusative", "number is singular") and binary ("the case of the first and second words match") pre-defined constraints selecting those that are satisfied by at least 75% of total weight.

Most of the expressions with at least 10 corpus instances obtain a valid set of constraints. Only expressions containing very common words (so that the words do appear quite often close together without actually forming the expression) obtain too weak constraints. For instance, no case and gender agreement constraints are selected for the expression *bohatý člověk (wealthy man)*.

Adding syntactic information.

Syntactic information (dependency relations among words in the expression) is needed mainly during the analysis of input sentences, therefore we focused on adding the information to the Czech part of entries first. For most of the entries, it was possible to add the dependency structure manually, based on the part-of-speech pattern of the entry. For instance all the entries containing an adjective followed by a noun get the same structure: the noun governs the preceeding adjective. For the remaining entries (with very varied POS patterns), we employ a corpus-based search similar to the automatic procedure of identifying morphological constraints.

4.3 Named entity recognition module

Named entities (NE) are atomic units such as proper names, temporal expressions (e.g., dates) and quantities (e.g., monetary expressions). They occur quite often in various texts and carry important information. Hence, proper analysis of NEs and their translation has an enormous impact on MT quality (Babych and Hartley, 2004). In our system they are extremely important due to the nature of input texts. The Wall Street Journal section of PennTreebank shows much higher density of named entities than ordinary texts. Their correct recognition therefore has a tremendous impact on the performance of the whole system, especially if the evaluation of the translation quality is based on golden standard translations.

NE translation involves both semantic translation and phonetic transliteration. Each type of NE is handled in a different way. For instance, person names do not undergo semantic translation (only transliteration is required), while certain titles and part of names do (e.g., *první dáma Laura Bushová* → *first lady Laura Bush*). In case of organizations, application of regular transfer rules for NPs seems to be sufficient (e.g., *Ústav formální a aplikované lingvistiky* → *Institute of formal and applied linguistics*), although an idiomatic translation may be probably preferable sometimes. With respect to geographical places we apply bilingual glossaries and a set of regular transfer rules as well.

For NE-recognition, we have developed a grammar based on regular expressions that processes typed feature structures. The grammar framework, similarly as the formally a bit weaker platform SProUT (Bering et al., 2003), uses finite-state techniques and unification, i.e., a grammar consists of pattern/action rules, where the left-hand side is a regular expression over typed feature structures (TFS) with variables, representing the recognition pattern, and the right-hand side is a TFS specification of the output structure.

The NE grammar is based on the experiment described in (Piskorski et al., 2004). An example of a simple rule is:

```
#subst[LEMMA: ministerstvo]$s1
+ #top[CASE: gen, PHRASE: $phr]$s2            (1)
== $s1#ministry[ATTR: $s2,
PHRASE: &('ministerstvo ' + phr)]
```

The first TFS matches any morphological variant of the word *ministerstvo* (ministry), followed by a genitive NP. The variables $s1, $s2 and $phr create dynamic value assignments and allow to transport these values to the slots in the output structure of type *ministry*. The output structure contains a new attribute called PHRASE with the lemmatized value of the whole phrase.

If the input phrase is

informace ministerstva zahraničí
o cestování do ohrožených oblastí (2)

then the phrase ”ministerstva zahraničí” will be recognized as a NE and handled as an atomic unit in the whole MT process:

$$\begin{bmatrix} \textit{ministry} & \\ \text{LEMMA} & \text{ministerstvo} \\ \text{FORM} & \text{ministerstva} \\ \text{PHRASE} & \text{ministerstvo zahraničí} \\ \text{ATTR} & \begin{bmatrix} \textit{subst} & \\ \text{LEMMA} & \text{zahraničí} \\ \text{PHRASE} & \text{zahraničí} \\ \text{FORM} & \text{zahraničí} \\ \text{CASE} & \text{gen} \\ \text{NUMBER} & \text{sg} \\ \text{GENDER} & \text{n} \end{bmatrix} \\ \text{CASE} & \text{gen} \\ \text{NUMBER} & \text{sg} \\ \text{GENDER} & \text{n} \end{bmatrix} \tag{3}$$

Lemmatization of NEs is crucial in the context of MT. However, it might pose a serious problem in case of languages with rich inflection due to structural ambiguities, e.g., internal bracketing of complex noun phrases might be difficult to analyze. The core of the framework is based on grammars that have been developed for the MT system Česílko (Hajič et al., 2003).

4.4 Syntactic analysis of Czech

Although we have originally assumed that the module of syntactic analysis of Czech will require only small modifications and its reuse in the new system was one of the goals of our system, it turned out that this module is one of the main sources of problems.

In the course of testing and debugging of the system we had to create a number of new grammar rules covering the phenomena which were not properly accounted for in the original system due to the different nature of the original domain. The texts from PCEDT show for example much higher number of numerals and numeric expressions, some of which require either special grammatical or transfer rules than operating systems manuals from the system RUSLAN. The complexity of input sentences with regard to the number of clauses and their mutual relationship is also much higher. This, of course, decreases the number of sentences which are completely syntactically analyzed and thus degrades the translation quality.

One of the biggest problems of the grammar are the properties of Q-systems. It was quite clear since the start of the project that it is impossible to extract only the knowledge encoded into the grammar, the grammar rules written in Q-systems are so complicated that rewriting them into a different (even chart-parser based) formalism would actually mean to write a completely new grammar. Although we have at our disposal a new, modernized and reimplemented version of a Q-systems compiler and interpreter which overcomes the technical problems of the original version, the nature of the formalism is of course preserved.

4.5 Transfer

The main task of this module is to transform the syntactic structure (syntactic tree) of the input Czech sentence into the syntactic structure (tree) of the corresponding English sentence. The transfer module does not handle the translation of regularly translated lexical units, it is handled by the bilingual dictionary in the earlier phases of the system. The transfer concentrates on three main tasks:

- The transformation of the Czech syntactic tree into the English one reflecting the differences in the word order between both languages.
- The identification and translation of those constructions in Czech, which require specific (irregular) translation into English.
- The insertion of articles (which do not exist in Czech) into the target language sentences.

The development of this module still continues, the initial tests confirmed that a substantial improvement can be achieved in the future.

4.6 Syntactic synthesis of English

The syntactic synthesis of Russian in RUSLAN is very closely bound to transfer, therefore we have tried to use as big portion of the grammar as possible, but of course, substantial modifications of the grammar were necessary. As well as the work on the transfer module, also the work on this module still continues.

4.7 Morphological synthesis of English

Due to the simplicity of English morphology this module has a very limited role in our system. It handles plurals, 3rd persons and irregular words.

5 Conclusion

The problems mentioned in this paper do not allow to formulate an answer to the crucial question - does it really pay off to recycle the old system or not? The integration of existing parts into a new system is so complicated that we are still not able to perform evaluation of results on texts of a reasonable size. One way out of this situation would be the combination of the new modules mentioned in this paper with one of the existing stochastic parsers of Czech instead of the rule-based grammar.

Another possible direction for the future research might be the exploitation of two new modules. The first one will contain partial, but error-free disambiguation of the results of morphological analysis of Czech, which will substantially decrease the morphological ambiguity of individual Czech word forms. This ambiguity (the average number of morphological tags per word form exceeds four in Czech) also negatively influences the performance of the syntactic analysis.

The second way how to decrease the ambiguity is the exploitation of a special module resolving the lexical ambiguity in those cases when the bilingual dictionary provides more than one lexical equivalent. This stochastic module would exploit the context and would suggest the best translation.

References

B. Babych and A. Hartley. 2004. Selecting translation strategies in MT using automatic named entity recognition. In *Proceedings of the Ninth EAMT Workshop, Valetta, Malta*.

C. Bering, W. Drożdżyński, G. Erbach, C. Guasch, P. Homola, S. Lehmann, H. Li, H.-U. Krieger, J. Piskorski, U. Schaefer, A. Shimada, M. Siegel, F. Xu, and D. Ziegler-Eisele. 2003. Corpora and evaluation tools for multilingual named entity grammar development.

Alain Colmerauer. 1969. Les Systemes Q ou un formalisme pour analyser et synthetiser des phrases sur ordinateur.

Jan Cuřín, Martin Čmejrek, Jiří Havelka, and Vladislav Kuboň. 2004. Building a Parallel Bilingual Syntactically Annotated Corpus. In *Proceedings of the 1st International Joint Conference on NLP*.

Jan Hajič. 2001. *Disambiguation of Rich Inflection - Computational Morphology of Czech*, volume I. Prague Karolinum, Charles University Press. 334 pp.

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

Karel Oliva. 1989. A Parser for Czech Implemented in Systems Q. *Explizite Beschreibung der Sprache und automatische Textbearbeitung*.

J. Piskorski, P. Homola, M. Marciniak, A. Mykowiecka, A. Przepiórkowski, and M. Woliński. 2004. Information extraction for Polish using the SProUT platform. In *Proceedings of the International IIS:IIP WM'04 Conference, Zakopane, Poland*.

Milan Svoboda. 2001. GNU/FDL English-Czech Dictionary. `http://slovnik.zcu.cz/`.