Testing the Limits - Adding a New Language to an MT System

Lukasz Debowski

Institute of Computer Science PAS
ul. Ordona 21, 01-237 Warszawa,
Poland
ldebowsk@ipipan.waw.pl

Jan Hajič, Vladislav Kuboň

ÚFAL MFF UK
Malostranské nám. 25, 11800 Praha 1
Czech Republic
{hajic,vk}@ufal.mff.cuni.cz

Abstract. This paper deals with a problem of an application of an MT method developed for a pair of very closely related languages to a pair of languages whose degree of relatedness (and thus also the degree of similarity) is lower. The close relatedness of the original language pair (Czech and Slovak) allowed a substantial simplification of the translation method used. This paper provides an overview of problems (and outlines their solution) that arise when adding a less similar language (Polish).

Introduction

One of the most widely used techniques of machine-aided human translation of the last decade or so is without doubts a method of human translation supported by a translation memory. This technique can substantially speed up the translation process especially when it concerns the translation and localization of various kinds of technical documentation.

At the same time, the difficulty of machine translation undoubtedly increases with the "distance" of languages in question. Fortunately, the reverse is also true: the closer the languages, the more chances there are that the translation quality will be reasonable.

In the system Česílko, described in (Hajič, Hric & Kuboň 2000), we have suggested that the translation memory (TM) can be used in a creative way for making the translation process more automatic (in a way which in fact does not depend on the languages used). In the same paper we have also described a method of "triangular" translation for a group of closely related languages through a pivot language using both human and machine translation, and its implementation for Czech and Slovak.

In this paper we would like to concentrate on the problem of adaptation of our method for a new language (Polish).

The use of the translation memory in the system Česílko

Use of a pivot

Localization of the same document into several typologically similar target languages separately is a waste of effort and money, since identical source language problems are being solved several times. The use of one language from the target group as a pivot and to perform the translation through this language seems to be quite a natural solution for 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.

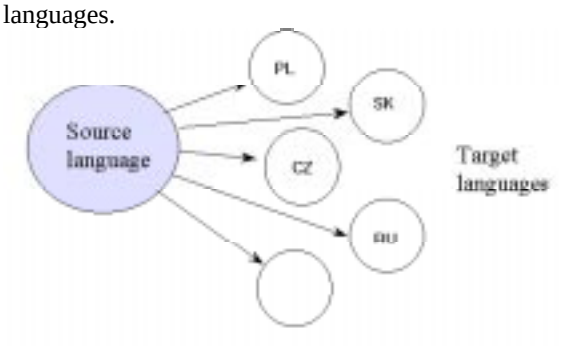

Fig. 1. A traditional model of localization

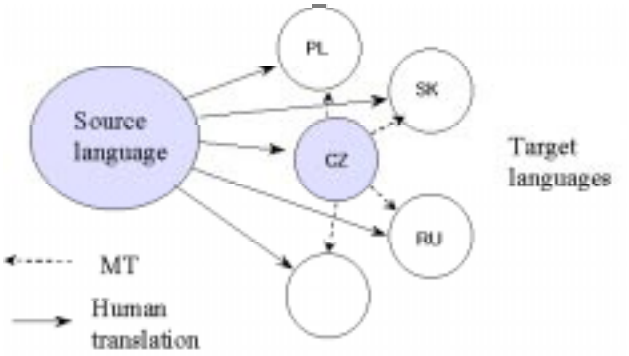

Fig. 2. Our model for translation

Creating a memory

The system Česílko was designed as a tool that allows building automatically translation “memories” for human translators between very closely related languages

(such as Czech and Slovak). Such translation "memory" would then be used as if created by humans, but appropriately marked for the human translators.

If we have at our disposal two translation memories – one human-made for the source/pivot language pair (say, English/Czech) and the other one created by an MT system for the pivot/target language pair (Czech/Slovak or Czech/Polish), the substitution of segments of a pivot language (Czech) by the segments of a target language (Slovak or Polish) is then only a routine procedure. The human translator translating from the source (English) to the target language (Slovak or Polish) then gets a translation memory for the required source/target pair.

The system of penalties is used to give preference to human translation in case when it exists in the TM.

Basic properties of the system

Our basic premise is to as simple method of analysis and transfer as possible. Our experience from an existing MT system RUSLAN (Czech-to-Russian MT system) aimed at the translation of software manuals for operating systems of mainframes – cf. (Oliva, 1989)) made it apparent that a full-fledged syntactic analysis of Czech is both unnecessary and too unreliable and costly. The present system 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; currently it is being tested on rather difficult texts of a Czech general encyclopedia. Its primary task is, however, to provide support for translation and localization of various technical texts.

This original pair of languages (Czech to Slovak)

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.

We have applied a stochastically based morphological disambiguate for Czech whose accuracy seems to be sufficient. Thus the system consists of the following steps:

1. Import of the source (Czech input) sentence (a segment from an "empty" translation memory")
2. Morphological analysis of Czech
3. Morphological disambiguation of Czech

4. Domain-related bilingual glossaries
5. General bilingual dictionary
6. Morphological synthesis of Slovak
7. Export to the original translation memory (Slovak target sentence) with an appropriate marking.

Morphological analysis of Czech

The morphological analysis of Czech is based on the morphological dictionary described in (Hajič, 2001). The dictionary covers over 800,000 lemmas and it is able to recognize about 20 mil. word forms. The morphological analysis uses a system of 15 positional tags: each morphological category, such as Part of speech, Gender, Case, etc., has a fixed one-letter place in the tag.

Morphological disambiguation of Czech

The module of morphological disambiguation (tagging) is a key to the success of the translation. The tagging system is based on an exponential probabilistic model (Hajič and Hladká, 1998), trained on roughly one million words using the level 1 manual annotation of the Prague Dependency Treebank (Hajič, 1998). The average accuracy of tagging is now over 94% (measured on tokens of running text). Lemmatization chooses the first lemma with a possible corresponding tag and works with accuracy close to 98%. This works well for lemma homonymy with a different part of speech, but for true polysemy resolution (word sense disambiguation for words with the same part of speech) we will have to add word sense disambiguation described in (Cikhart & Hajič, 1999).

Domain-related bilingual dictionaries (glossaries)

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.

Currently, the system handles well *n:n* term translation, uses heuristic guessing for asymmetric cases (*m:n*) and a more sophisticated system for handling the tags correctly in an *n:m* translation case is under development.

General bilingual dictionary

The main bilingual dictionary contains data necessary for the translation of both lemmas and tags. The translation of tags is necessary, because both tagsets use similar but slightly different tag sets. Also, the tags do not always correspond exactly, e.g. there are some Slovak nouns that have different gender, or tags with variants that do not exist in the other language. Therefore, a Czech tag is not translated into a single tag, but into a priority-ordered list of tags.

Morphological synthesis of Slovak

The morphological synthesis of Slovak is based on a monolingual dictionary of Slovak, developed by J. Hric (1991-99), covering more than 100,000 lemmas. The

coverage of the dictionary is still growing. It aims at a similar coverage of Slovak as has currently been achieved for Czech.

Evaluation of results

For the evaluation of our system, we have exploited 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 Slovak system achieved about 90% match (as defined by the TRADOS match module) with the human translation.

Testing the limits of the approach: Polish

It is clear that a word-for-word approach to MT as it was described in previous sections is applicable only to languages with high degree of similarity. An open question is where is the real limit of applicability of our method, which pairs of languages are close enough for our method to provide reasonable quality of translation and which are not. It is therefore quite natural to extend our system to other Slavic languages.

Due to the fact that, as far as we know, no other Slavic language has so many resources for stochastic natural language processing, it is quite natural that we are going to stick to Czech as a source language. The 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".

In order to obtain results comparable to the Czech-to-Slovak system we have used the same set of test data and the same evaluation method. The Polish morphological data was kindly provided to us by Morphologic, Inc. (Budapest, Hungary). We converted the data for use with our morphological generator. The comparison of the output from our system with the text post-edited by a Polish native speaker led to following results:

- 25,6% of sentences from the test sample did not require any postediting
- 16,7% of sentences were marked with less than 50% match against the correct post-edited sentences
- 33,3% of sentences achieved a match between 75% and 99%
- 24,4% of translated sentences had a match between 50% and 75%

The weighted average match (the length of a particular sentence was used as a weight) throughout the testing sample reached **71,4%**.

A match lower than 50% does not mean that the sentences were not usable for postediting. For example, one of the sentences with very low match was the following sentence:

Czech original:

Požadavky starší třiceti dnů se mažou.

[The requests older than 30 days are deleted.]

The result of MT:

Żądania starszy trzydziestu dzieni się smarują.

Post-edited Polish sentence:

Żądania starsze niż trzydzieści dni są wymazywane.

The match between the result of MT and the correct Polish sentence was 32% (according to TRADOS Translators Workbench standard computation), even though we need only 21 elementary operations to get the correct sentence (50 characters long) from the automatically translated one.

Word-order problems

The difference in quality of results obtained for Polish and Slovak as target languages mirrors the degree of similarity of both languages and the source language (Czech). While Slovak has almost identical word order as Czech, Polish contains several phenomena causing the necessity of word-order adjustments during the translation. The most obvious difference is the change of the word order in some types of nominal groups. Concerning congruent attributes, Czech prefers in most cases the order <Adj N>, adjective noun, while Polish typically uses the order <N Adj> for adjectives defining a "species" of the nominal head, while the order <Adj N> is reserved for adjectives defining a "feature" of the noun.

This problem is quite frequent, as the word-for-word translation method preserves the original order of words in all cases and thus it is a source of numerous errors. The general solution of this problem is very complicated; full solution would require even semantic analysis of the source text in our system, which is definitely beyond the intended basic design of our system.

A partial solution may be based on the exploitation of domain-related bilingual glossaries. It might be worth considering to include into this dictionary at least the most frequent terms of a particular domain, namely those that have different word order than the original one. In this case we would get a (correct) term-for-term translation instead of the word-for-word one.

Problems of agreement

All kinds of differences in gender or case are another source of relatively frequent errors. Both Czech and Polish are languages with strong requirements of gender, number and case agreement not only between subject and verb (gender and number agreement), but also in several other kinds of constructions. As an example of those constructions we may take e.g. gender, number and case agreement of the nominal group or the gender and number agreement of relative pronouns and their antecedents etc.

An example:

Czech original:
Počet dialogových procesů by měl pokrývat pracující uživatele.
[The number of dialog processes should cover working users.]
MT result:
Ilość dialogowych procesów *miałby* pokrywać pracujących użytkowników.
Polish (correct) sentence: Ilość procesów dialogowych *powinna* pokrywać ilość pracujących użytkowników.

In this case the translation of the masculine Czech form *počet* [number] is translated into feminine Polish form *ilość* causes diasgreement in gender in the target sentence obtained by MT (*miałby – powinna*). The situation is even more complicated by the incorrect translation of the conditional.

Similarly, the agreement within a noun group is broken in such cases, too. This inadequacy of our system could be solved at least partially through the introduction of a module of partial parsing of nominal groups. At the moment it doesn't seem efficient to aim at the solution of more complex agreement problems, like the problems of subject-verb agreement or the problem of assigning correct gender and number to relative pronouns.

Differences in cases

The first problem is the difference of valency frames between source and target words. Unlike Slovak, Polish contains several words that have different valency frame than their Czech counterparts. This of course results in a translation error, because the main bilingual dictionary does not contain any valency information.

Vast majority of word pairs (source – target) in both languages, however, have identical valency frames. Adding the valency frames only to the pairs of words with different valency should improve the quality of results for a reasonable price.

The second problem is the difference in prepositional constructions. For example, the Czech preposition *pro* [for] requires the use of the accusative case, while the corresponding Polish preposition *dla* requires the genitive case. Similarly (or even worse), some Czech cases are expressed by Polish prepositions.

Also in this case it is possible to solve the problem by listing the information about the required case in the main bilingual dictionary only for those prepositions where the cases differ in both languages.

Lexical problems

Quite serious is also the problem of lexical transfer in those cases where more Polish lexical units correspond to a single Czech one. A typical example is the Czech copula *nebo* [or], which may be translated either as *lub*, *bądź* (in more complex coordinations) or *czy* (yes-no questions only). It seems that there is no simple solution to this problem, apart from a simple frequency-based default selection; even a word-sense disambiguation based on the usual local context would fail here.

Addressing the reader

One very interesting problem is the use of the gender-based *Pan/Pani* ([Mr./Mrs.] in the Polish 3rd Pers. Sg.) rather than genderless Czech polite form *vy* [You] (2nd Pers Pl. (auxiliary verb) / Sg.(predicate)). In Polish, one can use gender-unbiased plural-only *Państwo* (both 2./3. Pers Pl.(verb+predicate)) in public lectures, and "friendly" *ty* in non-written advertisments, but none of those is suitable for written manuals or instructions. In Czech, there is even an alternative: *my* [we] can be used instead of *vy*; in Polish, however, it is not possible.

A Polish convention is to use slashed forms *Pan/Pani/Państwo* e.g. in forms that are supposed to be filled, but it looks ugly in regular sentences.

For the translation of technical texts, either a fully "impersonal" style is preferred (PL mediopassive in present/future and *-no/to* (*by*) in past/conditional) or "fuzzy" 2nd pers. pl. style is used, when explicit forms *Wy* [you] or *Państwo* are avoided, but it requires substantial sentence structure change.

Miscellaneous

- In Polish, copula *być* [to be] usually cannot be omitted as it is in Czech, therefore, it must be inserted at appropriate places.
- Polish 1st and 2nd person clitics, however rare in technical writing, are another problem. Czech forms *jsem* [I am], *jsi*, *jste* [you are], *jsme* [we are] are clitized to Polish floating suffixes *-(e)m*, *-(e)s*, *-(e)smy*, *-(e)scie*. These suffixes can attach to almost any word before the main verbal form but usually they go after the verbal form being: past participle, *powinien* and *jest* (present tense of *być* is reduplicated).
- For expressing that "something is something" Polish grammar admits only:
  NP(Nom.)+ *być* (finite form)+NP(Instr.)
  NP(Nom.)+ *być* (finite form)+Adj(Nom.)
  Inf.+ *jest*(finite form)+adverb.
  NP(Nom.)+ *to* (finite form)+NP(Nom.) (here *to* is a kind of predicative verb).

Conclusion

The success ratio of the translation achieved by our system (71.4% for the first Czech to Polish experiment using the rather strict TRADOS' evaluation metrics) justifies the hypothesis that word-for-word translation might be a solution for MT of really closely related languages.

In the near future, we will concentrate on such improvements that promise the biggest improvement (based on frequency of errors): nominal groups (word order, gender agreement), preposition "valency" (case change), and addressing the reader. In parallel, new Czech word sense disambiguation module will be tested, and improvements in the preprocessing of the terminological dictionary are planned, once we are able to get data for training a good Polish tagger. Eventually, of course, we would like to add other Slavic languages as well.

Acknowledgements

This project was supported by the grant GAČR 405/96/K214, and partially supported by the grant GAČR 201/99/0236.

References

Cikhart, P., & Hajič, J. (1999). Word Sense Disambiguation for Czech Texts. In: Proceedings of Text, Speech, Dialogue. Brno 1999. p. 109-114

Hajič, J. (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, J. (2001). Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Charles University Press, Prague.

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

Hajič, J., Hric, J. & Kuboň, V. (2000): Machine Translation of Very Close Languages. In: Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, Washington, USA, April 2000, pp. 7-12

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