Rapid development of RBMT systems for related languages, a case study on language pair Slovenian-Serbian

Jernej Vičič

University of Primorska
Glagoljaska 8, Koper
E-mail: jernej.vicic@upr.si

Abstract

The article describes novel way of constructing rule-based machine translation systems (RBMT). RBMT systems are currently among the best performing machine translation systems. Most of the "big named" machine translation systems [5] and [6] belong to this category, but these systems have a big drawback; construction of such systems demands a great amount of time and resources, thus resulting very expensive.
The article describes methods that automate parts of the construction process. The methods were evaluated on a case study: construction of a fully functional machine translation system of closely related language pair Slovene – Serb.
The system is based on Apertium [1] and [3], an open-source RBMT toolkit.
Evaluation was conducted on a fully functional machine translation system.

1 Introduction

Slovene and Serbian language belong to the group of southern Slavic languages that were spoken mostly in former Yugoslavia. Slovenian language is mostly spoken in Slovenia, Serbian language is mostly spoken in Serbia and in Montenegro. The languages share common roots and even more importantly they share common recent historical environment, these languages were spoken in the same country, even taught in schools as languages of the surroundings.
Economies of all three states are closely connected and younger generations, the post-yugoslavia breakage generations, have difficulties in mutual communication, so there is quite big interest in construction of such translation system.
Both languages belong to the southern Slavic language group; they are highly inflective and morphologically and derivationally rich languages and defer greatly from mostly used languages in electronic materials like English, Arabic, Chinese, Spanish and French. This means that most of the data and translation methods must be at least revisited or even worse rewritten. This language pair is closely related lexicographically and syntactically which simplifies most of the normal translation system production steps.
The machine translation system is based on Apertium [1] and [3], an open-source RBMT toolkit.
Apertium is an open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English–Catalan). The platform provides a language-independent machine translation engine tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs.
All these properties make Apertium a perfect choice in a cost effective machine translation system development.
The rest of the article is organized according to [2] as follows:
Apertium, the open-source MT platform that was used as basis in the case study, is described in the first section following the introduction. Materials and methods describe already available language processing tools and materials, mainly corpora. The newly developed methods are described in the same section. Results and evaluation methods are described in the last section.

2 The Apertium open-source MT platform

Apertium uses a shallow-transfer machine translation engine which processes the input text in stages, as in an assembly line: de-formatting, morphological analysis, part-of-speech disambiguation, shallow structural transfer, lexical transfer, morphological generation, and re-formatting.
The data needed by the presented stages can be grouped into four categories: monolingual dictionaries used by morphological analysis and morphological generation, bilingual dictionaries used in lexical transfer, structural transfer rules used in structural transfer and Part Of Speech (POS) tagging used in disambiguation.
The modules are shown on Figure 3, where the specially addressed modules are marked with a new color and the two newly added modules are inserted.
Each group's data creation was addressed by a particular method; monolingual dictionaries were constructed using bilingual dictionary data and applying automatic paradigm tagging techniques; bilingual dictionary was constructed using available bilingual word-list but a few methods for automatic bilingual dictionary construction were investigated; a method for

? = ? = Cx
? = ? = cx
*** = ? = Cy**

automatic structural shallow-transfer rule construction [7] will be used to construct a set of structural transfer rules.

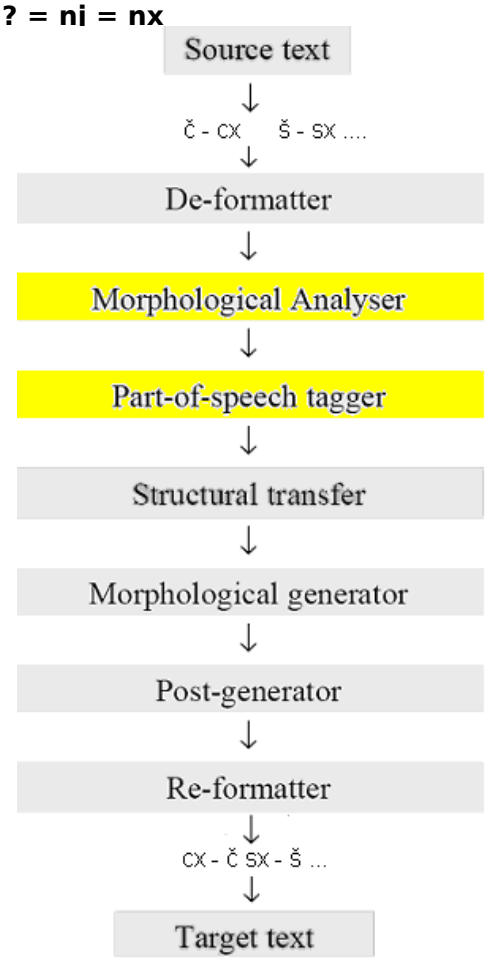

Figure 3: All modules os a standard Apertium system

3 Materials and methods

A research of already available and accessible language processing tools and materials, mostly corpora, revealed that there is a reasonably big amount of work already done for Slovenian language, less for Serbian. The tools for Slovenian language are (reasonable or even good quality): part of speech tagger [8] and [9], lemmatizer [8] and [10], stemmer [11] and [13], none of these tools exists for Serbian language. Both languages have solid monolingual reference corpora (going into hundreds of millions) and a small bilingual corpus that was used mostly for evaluation purposes.

Only lexicographic modules were taken into consideration in this case study as the work on the project is still in progress. We concentrated the research on preceding modules, the lexicographic modules, as they present the basis for all translation stages.

1 The system is still in development phase

3.1 Automating data creation using available tools and materials

Monolingual and bilingual dictionaries were constructed using a large bilingual word list of unchecked quality. Paradigms were hand-written according to [12].

Some paradigms such as numbers, abbreviations and punctuation were taken from preexisting materials, mostly from Spanish-Catalan and English-German Apertium data modules.

Totale toolkit [8] was used to POS tag [9] and lemmatize [10] words in the bilingual word list; POS tagger was also used in automatic paradigm classifying, see chapter for further description.

Some post-processing was necessary due to errors in bilingual word list and unsuccessful paradigm tagging.

POS tagger from Totale [8] was also used as the disambiguation module instead of the original apertium tagger.

Structural transfer rules were simply copied from existing data, exactly from Spanish-Catalan translation system. We acknowledge that this is far from being ideal but the system is built in modules that allow gradual construction of a new system thus allowing us to deal with structural transfer in second phase.

A small demo system implementation for research purposes showed that with a few adaptations that would address properties uncommon with starting translation system like inflectional variety in both languages and special number, the dual, in Slovenian language, the starting rules would mostly suffice.

3.2 Overcoming Apertium limitations

Apertium was built as a machine translation system for related romance languages and some properties still reflect the first design, like fixed codepage. All modules are still fixed to Latin-1 codepage, which is not suitable for Slavic languages that mostly share Latin-2 codepage.

Figure 3: Special characters were converted into impossible two-character pairs

The modules are being rewritten to support Unicode standard, but at the moment we had to use available tools and deal with this problem. There are 8 special characters in the new language pair and we constructed two simple modules that translate these characters into impossible two-character combinations following AURORA coding [16] like shown on Figure 3. First module, the coder, was inserted at the beginning of the

```
<pardef n="korak__n">
 <e><p>
  <l/>
  <r>
   <s n="n"/><s n="m"/>
   <s n="sg"/>
   <s n="genitive"/>
  </r>
 </p></e>
 <e><p>
  <l>a</l>
  <r>
   <s n="n"/>
   <s n="m"/>
   <s n="sg"/>
   <s n="acusative"/>
  </r>
 </p></e>
</pardef>
```

Figure 3: Paradigm example, Noun, masculine 1. paradigm (korak)

3.3.1 Paradigm tagging using POS tagger

An already trained and tested POS tagger [8] was available for Slovenian language. Words were tagged using full MSD [14] descriptions and grouped into classes with same descriptions (words that had the same POS tag were grouped together). This process produced 141 classes in Slovene and 274 classes in Serbian language; see Table 1 for details. A linguist manually tagged the classes to paradigms. The difference in number of classes in mostly due to finer MSD descriptions in Serbian language.

The TNT tagger [9], which was used in the process, relies heavily on context to disambiguate ambiguities. In a word list each word is treated separately, there is no context, so the word tagging quality is lower than the values on running text.

3.3.2 Paradigm tagging using monolingual referential corpus

Bilingual word list was treated for each language separately using the same method, but obviously different corpus. Each word from bilingual word list was stemmed using a modified version of [11] algorithm that takes into consideration only extensions that were present in paradigms. This means that each word is shortened of the longest possible extension producing word's stem. All extensions are attached to the stem producing a multiset[2] of words. This multiset is searched in monolingual referential corpus, in our case [15] and [17], all words that are found in corpus present a list of possible extensions, thus reducing the number of all extensions to a moderate number.

The multiset of possible extensions is compared to groups of extensions retrieved from paradigm descriptions; the paradigm that has most matches in this comparison is selected as the most likely paradigm from the word, i.e. the word is tagged with this paradigm. Paradigms are selected or tagged only if a predefined value of matched postfix if found. The words that are not selected by this method can be tagged manually or tagged with a paradigm that is most likely.

2 A multiset is a generalization of a set. A member of a multiset can have more than one membership.

4 Results

The translation quality of the overall system still leaves to be desired, the bleu value was below 0.05, so translation quality tests were conducted just to test the capabilities and methods.

Table 1 presents some preliminary values describing the most important translation data properties.

Objective and subjective evaluation methods will be used in final testing as only a correct mixture of methods minimizes evaluation bias. Translation quality evaluation will be conducted using subjective evaluation methods; where a group of native speakers will score translations. Automatic objective measures NIST and BLEU [4] will be used to ensure wider coverage. Bilingual corpus [14] will be used in all evaluation processes.

number of lemmata:	74584
number of paradigms sl:	38
number of paradigms sr*:	34
number of classes sl:	141
number of classes sr**:	274
% of wrong paradigm tags	18.4
*the number of sl classes is bigger due to unfinished work	
**the number of sr classes is bigger due to finer POS tag definition	

Table 1: Preliminary values describing translational data

5 Discussion and further work

The system is still under heavy development; we still have to improve translation data quality through improvement of automatic methods but unfortunately also through manual correction. Parallel we will modify the existing structural transfer rules.

The bilingual word list will be changed due to licensing problems as we expect to release the translation system as part of Apertium bundle under open-source licensing.

The problems that we encountered this case study and promising results led us to the idea of a toolset that would automate most of the steps (possibly all steps) of a standard translation system creation process.

6 Acknowledgements

This research was partially funded by the Vice-rectorate for Research, Development and Innovation of the Universitat d'Alacant. Special thanks go to Sergio Ortiz-Rojas for helping me with the implementation.

References

[1] Armentano-Oller Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe

Sánchez-Martínez, Miriam A. Scalco 2006: Open-source Portuguese-Spanish machine translation, In Lecture Notes in Computer Science 3960 (Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006, Springer-Verlag 2006, p. 50-59

[2] Robert A. Day: *How to Write and Publish a Scientific Paper, http*://www-math.science.unitn.it/LRM3D2/report.htm

[3] Corbí-Bellot Antonio M., Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Sánchez-Ramírez, Felipe Sánchez-Martínez, Iñaki Alegria, Aingeru Mayor, Kepa Sarasola 2005: An open-source shallow-transfer machine translation engine for the romance languages of Spain, Proceedings of the European Association for Machine Translation, 10th Annual Conference (Budapest, Hungary, 30-31.05.2005),

[4] Papineni Kishore , Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2001: Bleu: a method for automatic evaluation of machine translation. Technical Report, RC22176, IBM, 2001.

[5] Promt: http://www.e-promt.com/

[6] Systran: http://www.systran.co.uk/

[7] Sánchez-Martínez, Felipe and Ney, Hermann, 2006:
Using Alignment Templates to Infer Shallow-Transfer Machine Translation Rules, Advances in Natural Language Processing, Proceedings of 5th International Conference on Natural Language Processing {FinTAL}, 756--767, 2006, (Copyright Springer-Verlag)

[8] Tomaž Erjavec, 2006: Multilingual tokenisation, tagging, and lemmatisation with totale. V: 9th INTEX/NOOJ Conference, Belgrade, Serbia, June 1-3, 2006

[9] Brants, Thorsten, 2000: *TnT -- a statistical part-of-speech tagger*, In Proceedings of the 6th Applied NLP Conference, ANLP-2000, April 29 -- May 3, 2000, Seattle, WA.

[10] Erjavec, Tomaz, Dzeroski Sasa, 2004: Machine Learning of Language Structure: Lemmatising Unknown Slovene Words, Applied Artificial Intelligence, 18

[11] Popovič, M., Willett, P. 1992. *The effectiveness of stemming for natural language access to Slovene textual data*. Journal of the American Society for Information Science, 43(5), 384-390

[12] Toporišič J., 2000. *Slovenska slovnica*, Založba Obzorja, 2000, Maribor

[13] Vilar P., Dimec J., 2000. *Krnjenje kot osnova nekaterih nekonvencionalnih metod poizvedovanja*, Knjižnica, Ljubljana, 44(2000)48

[14] Tomaz Erjavec, 2004. MULTEXT-East Version 3: *Multilingual Morphosyntactic Specifications*, Lexicons and Corpora. In M. T. Lino and M. F. Xavier (ur.) Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04. Paris: ELRA

[15] Erjavec, T., Gorjanc, V., Stabej, M.: *Korpus FIDA* International Multi-Conference Information Society - IS'98, 6 - 7 October 1998, [Ljubljana]. - Ljubljana : Institut Jožef Stefan, 1998

[16] Vitas, D.; 1979: *Prikaz jednog sistema za automatsku obradu teksta*, Zbornik radova XIV jugoslovenskog medunarodnog simpozijuma o obradi podataka In-formatica 79, Bled, Slovensko drustvo INFORMATIKA, Ljubljana, p. 7 101

[17] Serbian monolingual corpus, 2007: http://korpus.matf.bg.ac.yu/