[*Translating and the Computer* 29, November 2007]

Rapid development of RBMT systems for related languages

Jernej Vicic
University of Primorska
Cankarjeva 5, 6000 Koper, Slovenia
jernej.vicic@upr.si

Abstract

The article describes a new 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 (Systran, 2007)(Promt, 2007) 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.

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.

The system is based on Apertium (Armentano-Oller et al., 2007) (Corbí-Bellot et al., 2005), an open-source RBMT toolkit. 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 three categories: monolingual dictionaries used by morphological analysis and morphological generation, bilingual dictionaries used in lexical transfer and structural transfer rules used in structural transfer.

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 automatic structural shallow-transfer rule construction (Sánchez-Martínez et al., 2006) was used to construct a set of structural transfer rules.

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 Slovene language are (reasonable or even good quality): part of speech tagger (Erjavec et al., 2000), lemmatizer (Erjavec et al., 2004), stemmer, 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. Evaluation was conducted on the functional machine translation system and the results presenting coverage using referential corpus and selected evaluation metrics are shown. Objective and subjective evaluation methods were used as only a correct mixture of methods minimizes evaluation bias. Translation quality evaluation was conducted using subjective evaluation methods where a set of native speakers scored translations. Automatic objective measures NIST and BLEU (Papineni et al., 2001) were used to ensure wider coverage. Bilingual corpus was used in both automatic evaluations. Conclusions present strong and weak points of this approach and explore grounds for further work.

1. Introduction

Slovene and Serbian language belong to the group of southern Slavic languages that are spoken mostly on the territory of former Yugoslavia. Slovenian language is mostly spoken in Slovenia, Serbian language is mostly spoken in Serbia. 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 both countries are closely connected. Younger generations, the post-yugoslavia breakage generations, have difficulties in mutual communication, *so* there is quite big interest in construction of an automatic machine translation system for this language pair.

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.

All methods and materials discussed in this paper were tested on a fully functional machine translation system based on Apertium (Armentano-Oller et al., 2006) and (Corbí-Bellot et al., 2005), 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 (Day, 2007) 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. Following section describes results and evaluation methods. The last section describes discussion and further work.

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.

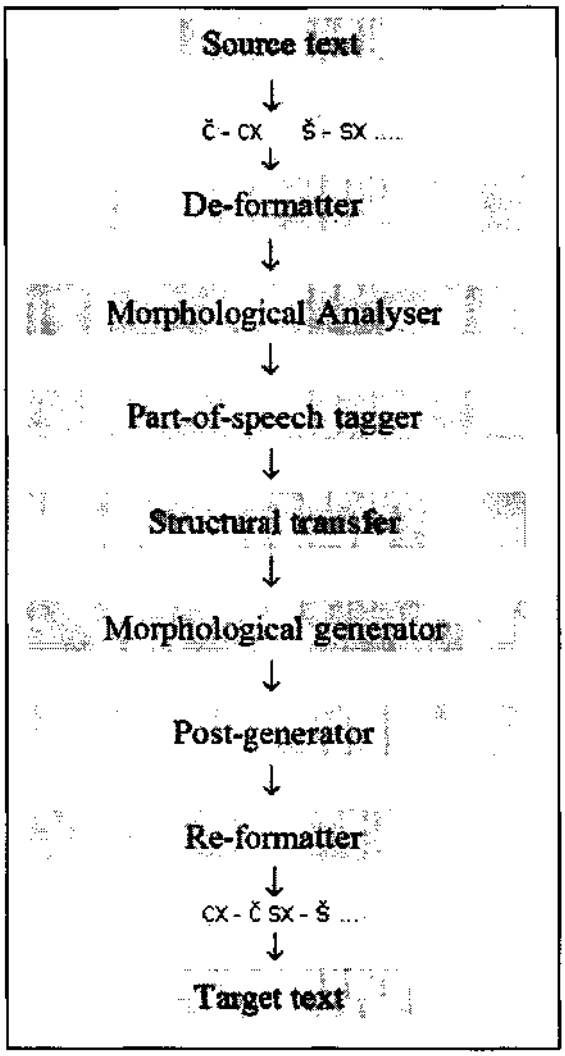

Figure 1: Modules of a standard Apertium system

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 1, where the specially addressed modules are marked with a new colour 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 automatic structural shallow-transfer rule construction (Sánchez-Martínez et al., 2006) will be used to construct a set of structural transfer rules.

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 (Erjavec, 2006) and (Brants, 2000), lemmatizer (Erjavec, 2006) and (Erjavec et al., 2004), stemmer (Popovič et al., 1992) and (Vilar et al., 2000), 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 (Erjavec, 2004) that was used mostly for evaluation purposes. This research focuses mostly on lexical level mainly for these reasons:

- Lexical level presents the starting ground for written text translation.
- Related languages, particularly the language pair we based our study upon, usually share the same sentence structure. Most of the translation takes place on lexical level.
- Unlike some well-known languages, like English, southern Slavic languages express most of the meaning by inflecting words and less by word order.

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. Still some basic structural transfer rules were constructed to greatly enhance translation performance at a small cost in expert hours.

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 (Toporišič, 2000).

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

Totale toolkit (Erjavec, 2006) was used to POS tag (Brants, 2000) and lemmatize (Erjavec et al., 2004) words in the bilingual word list; POS tagger was also used in automatic paradigm classifying, see chapter 3.3.1 for further description.

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

POS tagger from Totale (Erjavec 2006) 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.

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 improbable two-character

Ћ = Ć = Cx	ћ = ć = cx	Ч = Č = Cy	ч = č = cy
Ђ = Đ = Dx	ђ = đ = dx	Џ = Dž = Dy	џ = dž = dy
Ш = Š = Sx	ш = š = sx	Ж = Ž = Zx	ж = ž = zx
Љ = Lj = Lx	љ = lj = lx	Њ = Nj = Nx	њ = nj = nx

Figure 2: Special characters were converted into impossible two-character pairs

combinations following AURORA coding (Vitas, 1979) like shown on Figure 2. First module, the coder, was inserted at the beginning of the translation pipeline; the decoder was inserted at the end.

3.3 *Paradigm tagging*

During this case study we developed two methods to group words into pre-prepared paradigm classes (tag paradigms to words). An example paradigm description is shown in Figure 3. The methods were developed with available materials and tools that we could use. The first method relies on POS tagger and the second method relies on a big monolingual corpus.

```
<pardef n="korak_n">
    <e><p><l/><r><s n="n"/><s n="m"/>< n="sg"/>< n="nominative"/></r></p></e>
    <e><p><l>a</l><r><s n="n"/><s  n="m"/><s n="sg"/><s n="genitive"/></r></p></e>
    <e><p><l>u</l><r><s n="n"/><s n="ra"/><s n="sg"/><s n="dative"/></r></p></e>
    <e><p><l/><r><s n="n"/><s n="m"/><s n="sg"/><s n="acusative"/></r></p></e>
    <e><p><l>u</l><r><s n="n"/><s n="m"/><s n="sg"/><s n="locative"/></r></p></e>
    <e><p><l>om</l><r><s n="n"/><s n="m"/><s n="sg"/><s n="instrumental"/></r></p></e>
    <e><p><l>a</l><r><s n="n"/><s n="m"/><s n="sistdv"/><s n="nominative"/></r></p></e>
    <e><p><l>ov</l><r><s n="n"/><s n="m"/><s n="sistdv"/><s  n="genitive"/></r></p></e>
    <e><p><l>oma</l><r><s n="n"/><s n="m"/><s n="sistdv"/><s  n="dative"/></r></p></e>
    <e><p><l>a</l><r><s n="n"/><s  n="m"/><s n="sistdv"/><s n="acusative"/></r></p></e>
    <e><p><l>ih</l><r><s n="n"/><s  n="m"/><s n="sistdv"/><s n="locative"/></r></p></e>
    <e><p><l>i</l><r><s n="n"/><s  n="m"/><s n="sistdv"/><s  n="instrumental"/></r></p></e>
    <e><p><l>i</l><r><s n="n"/><s n="m"/><s n="pl"/><s n="nominative"/></r></p></e>
    <e><p><l>ov</l><r><s n="n"/><s n="m"/><s n="pl"/><s n="genitive"/></r></p></e>
    <e><p><l>om</l><r><s n="n"/><s  n="m"/><s n="pl"/><s n="dative"/></r></p></e>
    <e><p><l>e</l><r><s n="n"/><s  n="m"/><s n="pl"/><s n="acusative"/></r></p></e>
    <e><p><l>ih</l><r><s n="n"/><s n="m"/><s n="pl"/><s n="locative"/></r></p></e>
    <e><p><l>i</l><r><s n="n"/><s  n="m"/><s n="pl"/><s n="instrumental"/></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 (Erjavec, 2006) was available for Slovenian language. Words were tagged using full MSD descriptions (Erjavec, 2004) and grouped into classes with same descriptions (words that had the same POS tag were grouped together). This process produced 312 classes in Slovene and 274 classes in Serbian language; see Table 1 for details. A linguist manually tagged the classes to paradigms.

The TNT tagger (Brants, 2000), 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*

Each language part of the bilingual word list was treated independently using the same method, but obviously different corpus. Each word from bilingual word list was stemmed using a modified version of (Popovič, 1992) 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 of words. This multiset is searched in monolingual referential corpus, in our case (Erjavec et al., 1998) and (Serbian, 2007), 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.

4. Results

This chapter presents the motivation, methodology and finally the results of the evaluation process.

4.1 Motivation and methodology

Our research focused on rapid construction of lexical data, mainly paradigm tagging. Primary evaluation goal was evaluation of paradigm tagging methods; evaluation of the complete translation system was also carried out.

Paradigm tagging evaluation task consisted of manual evaluation of paradigm tagging process on selected words. The whole bilingual word-list was paradigm tagged using the method presented in chapter 3.3.1. A statistically meaningful and objectively still feasible number of words (600) were hand-checked for errors in paradigm tags. The words were divided into five classes: perfect tag, wrong

paradigm (tagging method error), wrong stem (stemmer error), wrong POS (POS tagger error), unknown word (mostly OCR error).

Evaluation of the translations was performed in two parts:

- Automatic objective evaluation of using BLEU (Papineni, 2001) metric.
- Non-automatic subjective evaluation.

Bilingual parallel corpus (Erjavec, 2004) was used in automatic (BLEU) evaluation of translations. This corpus consists of 8600 sentences that were not used in translation system construction. Subjective evaluation was performed after first poor BLEU results triggered some distrust. Many authors agree that BLEU metric systematically penalizes RBMT systems (Callison-Burch et al., 2006) and it is not suited for highly inflexible languages. Authors of METEOR (Banerjee et al., 2005), (Lavie, 2007) state that their system fixes most of the problems encountered using BLEU metric; they state that METEOR correlates highly with human judgement. Unfortunately METEOR does not support our language pair, we hope to change this in the near future, see further work

Subjective manual evaluation of translation quality was performed according to the annual NIST Machine Translation Evaluation Workshop by the Linguistic Data Consortium guidelines. The most widely used methodology when manually evaluating MT is to assign values from two five-point scales representing fluency and adequacy. These scales were developed for the annual NIST Machine Translation Evaluation Workshop by the Linguistic Data Consortium (LDC, 2005).

The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a hypothesis translation:

5 = All
4 = Most
3 = Much
2 = Little
1 = None

The second five-point scale indicates how fluent the translation is. When translating into Serb the values correspond to:

5 = Flawless Serb
4 = Good Serb
3 = Non-native Serb
2 = Disfluent Serb
1 = Incomprehensible

Separate scales for fluency and adequacy were developed under the assumption that a translation might be disfluent but contain all the information from the source.

Four independent evaluators (two native speakers) evaluated sets of 100 sentences using this methodology.

4.2 Results

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

Objective and subjective evaluation methods were used in final testing as only a correct mixture of methods minimizes evaluation bias. Translation quality evaluation was conducted using subjective evaluation methods; where a group of native and near-native speakers scored translations. Automatic objective measures NIST and BLEU (Papineni, 2001) were used to ensure wider coverage. Bilingual corpus (Erjavec, 2004) was use d in all evaluation processes.

Table 1: Preliminary values describing translational data

number of lemmata	74584
number of paradigms sl	58
number of paradigms sr	59
number of classes sl	312
number of classes sr**	274
% of wrong paradigm tags	18.4

**the number of sr classes is smaller due to finer POS tag definition

4.2.1 Paradigm tagging method evaluation

Evaluation of paradigm tagging quality was quite simple in organization and methodology, but quite resource consuming: 600 randomly selected words from monolingual dictionary were manually checked. Each word was assigned a value from this set:

- Perfect tag (word was correctly tagged)
- Wrong paradigm (the method erroneously assigned the tag)
- Wrong stem (stemmer error)
- Wrong POS (POS tagger error)
- Unknown word (error in bilingual wordlist)

Figure 4 shows values for Slovene words and Figure 5 shows values for Serbian words.

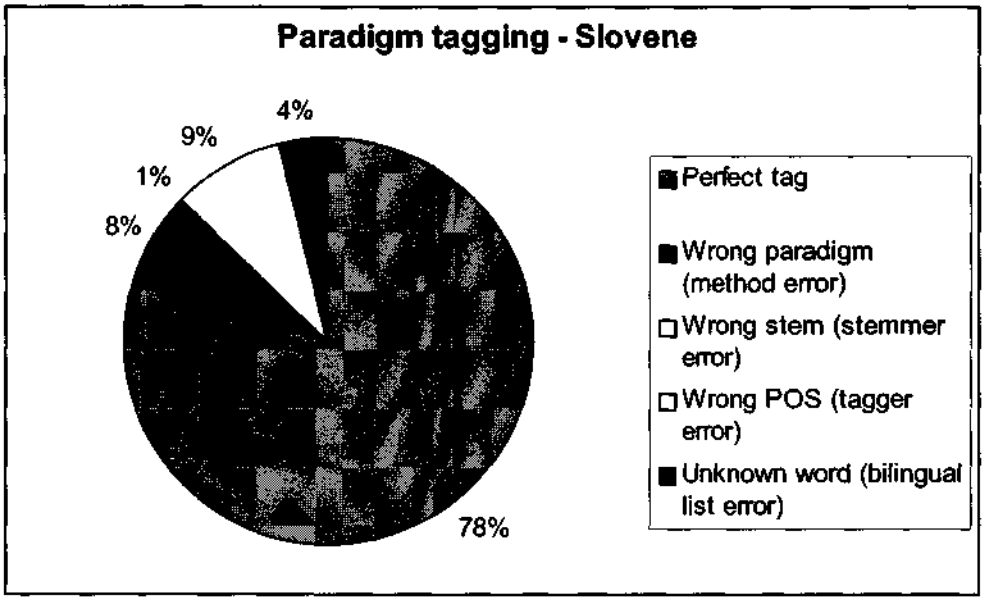

Figure 4: Paradigm tagging evaluation, Slovene part

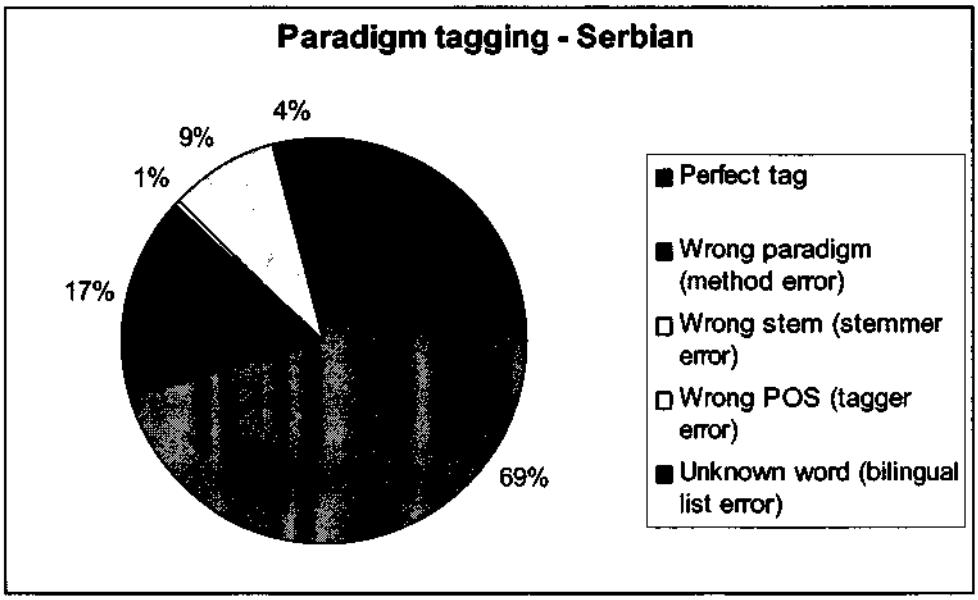

Figure 5: Paradigm tagging evaluation, Serbian part

4.2.2 Translation evaluation

A bilingual parallel corpus (Erjavec, 2004) was used for automatic evaluation using BLEU metric (Papineni et al., 2001). The results are presented in Table 2. The values are quite low, partly due to reasons explained in (Callison-Burch et al., 2006), partly due to unknown words in test corpus.

Table 2: automatic evaluation

Number of test sentences	6.669
Bleu value	0,07

Figure 6 shows results of evaluation of translation quality using subjective measures using methodology (LDC, 2005). The methodology is explained in chapter 4.1. Four independent evaluators (two native speakers) evaluated sets of 100 sentences using this methodology. Standard deviation shows that evaluators agreed with their scores.

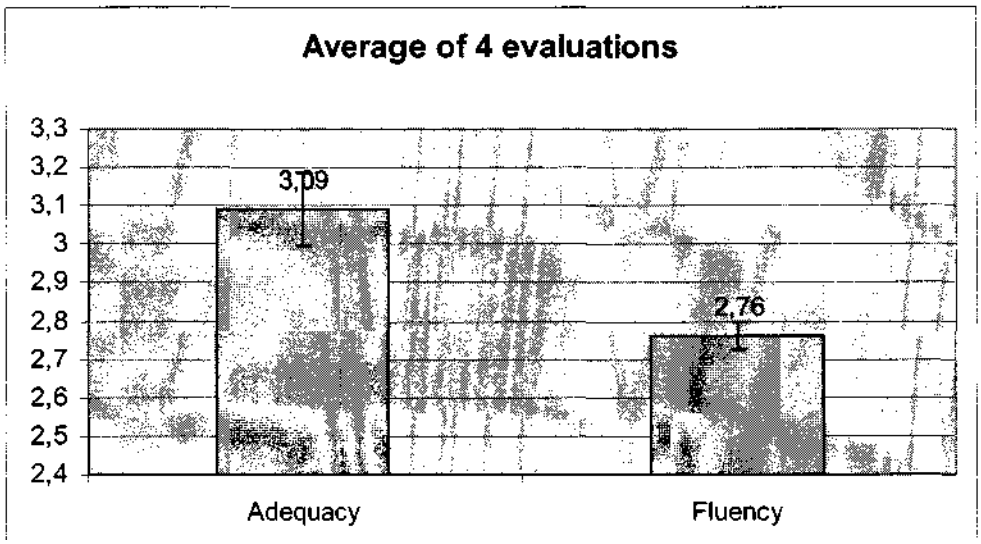

Figure 6: Paradigm tagging evaluation, Serbian part

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 in 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.

Automatic paradigm construction from tagged corpora presents a completely independent task that would abstract the need for an expert of the field, thus lowering the cost and time of production of a new machine translation system.

Automatic evaluation metrics present a big problem for RBMT systems and even more for translation pairs as discussed in chapter 4.1. A change of available metric and thorough testing of results will provide a confirmation of refusal of the hypothesis.

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 and Mikel L. Forcada for their generous support and feedback during the design and implementation of the core system.

6. References

Armentano-Oller Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramirez-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

Banerjee, S. Lavie A. (2005), "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan

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.

Callison-Burch Chris, Osborne Miles, and Koehn Philipp. (2006). "Re-evaluating the role of Bleu in machine translation research". In Proceedings of EACL.

Callison-Burch Chris, Fordyce Cameron, Koehn Philipp, Monz Christof, Schroeder Josh (2007). "(Meta-) Evaluation of Machine Translation". Proceedings of ACL-2007 Workshop on Statistical Machine Translation.

Corbí-Bellot Antonio M., Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Sánchez-Ramirez, 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),

Day Robert A. (2007): How to Write and Publish a Scientific Paper, http://www-math.science.unitn.it/LRM3D2/report.htm

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

Erjavec Tomaž (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

Erjavec, Tomaž, Dzeroski Sasa (2004): Machine Learning of Language Structure: Lemmatising Unknown Slovene Words, Applied Artificial Intelligence, 18

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

Lavie, A., Agarwal. A. (2007): "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments", Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007), Prague

LDC. 2005. Linguistic data annotation specification: "Assessment of fluency and adequacy in translations". Revision 1.5.

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.

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

Promt (2007): http://www.e-promt.com/

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)

Serbian monolingual corpus, (2007): http://korpus.matf.ba.ac.vu/

Systran (2007): http://www.systran.co.uk/

Toporišič J., 2000. Slovenska slovnica, Založba Obzorja, 2000, Maribor

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

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