A method to restrict the blow-up of hypotheses of a non-disambiguated shallow machine translation system

Jernej Vičič
University of Primorska
Koper, Slovenia
jernej.vicic@upr.si

Petr Homola
Charles university
Prague, Czech republic
homola@ufal.mff.cuni.cz

Vladislav Kuboň
Charles university
Prague, Czech republic
vk@ufal.mff.cuni.cz

Abstract

The article presents two automatic methods that reduce the complexity of the ambiguous space introduced by the omission of the part of speech tagger from the architecture of a shallow machine translation system. The methods were implemented in a fully functional translation system for related languages. The language pair chosen for the experiments was Slovenian-Serbian as these languages are highly inflectional with morphologically ambiguous forms. The empirical evaluations show an improvement over the original system.

Keywords

MT, RBMT, SMT, related languages, ranking, Apertium.

1 Introduction

The article presents two automatic methods that reduce the complexity of the ambiguous space introduced by the omission of the part of speech tagger from the architecture of a shallow transfer machine translation system.
The shallow parsing machine translation architecture is suitable for the translation systems for related languages as shown in [5, 15]. Authors [12] and the authors of translation systems for related languages Apertium [5] and Česílko [11] suggest using an architecture similar to the one presented in figure 1. This architecture employs statistical part of speech (POS) tagger for disambiguation of the morphological analysis of the source language. The quality level of the tagging process of today's state-of-the-art POS taggers for highly inflectional languages like [10] and [8] is relatively low, comparing to the quality of POS taggers for the analytical languages like the English language, and also comparing to the overall quality of the translation systems for related languages.
[14] proposes a change of the basic architecture, the omission of the statistical POS tagger from the early stages of the translation process and the introduction of a ranking mechanism in the post-processing phase. The proposed architecture is presented in the figure 2. The ranking mechanism is based on the statistical trigram target language model and models the most probable sentence in the target language. The ranker selects the best translation among all available translation candidates.
The rest of the article is organized as follows: the section 2 presents the current accomplishments in the field of the machine translation for related languages, section 3 presents the basic motives that led the authors to this set of experiments. The section 4 shows the motivation that led to the experiment presented in the paper. the section 5 presents the main method and the section presents the evaluation methodology with the results. The article concludes with a discussion.

2 State of the art

The majority of the translation systems for related languages uses the shallow parsing machine translation architecture as shown in [15] and in the overview of the translation systems presented in 2.1. The figure 1 shows the architecture of the of the most known translation systems for related languages [5] and Česílko [11] and used with variations in the majority of the systems presented in 2.1.

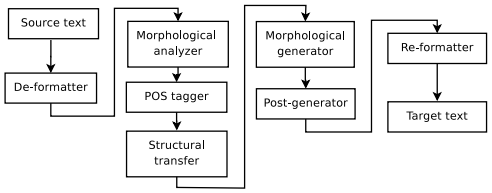

Fig. 1: *The modules of a typical shallow transfer translation system. The systems [5, 11, 17, 19] follow this design.*

2.1 Available MT systems for related languages

A few experiments in the domain of machine translation for related languages have led to the construction of more or less functional translation systems. The systems are ordered alphabetically:

- [2] for Turkic languages.
- Apertium, [5], for Romance languages.
- [7, 3, 1] for Scandinavian languages.

International Conference RANLP 2009 - Borovets, Bulgaria, pages 460–464

- Ruslan and Česílko, [9, 12], for highly inflectional Slavic languages, mostly language pairs with Czech language.
- [18] for Gaelic languages; Irish (Gaeilge) and Scottish Gaelic (G'aidhlig).
- [19] for the North Sámi to Lule Sámi language pair.
- Guat, [20], for highly inflectional Slavic languages, mostly language pairs with Slovenian language.

The experiments presented in this paper are based on technologies presented by [5] and [12].

3 Motivation

The most important motivative reasons for this research are:

- The production of a new POS tagger, especially a good quality tagger, is not a simple task. One of the easiest methods is training of a stochastic tagger based on HMM algorithm [21]. Some parts of this task can be automatized using unsupervised learning methods or supervised learning methods like [4], but it still involves the selection of a new tag set, the production of a tagged training corpus, testing of the corpus and at the end the basic learning process.
- The quality level of the tagging process of today's state-of-the-art POS taggers for highly inflectional languages like [10] and [8] is relatively low, comparing to the quality of POS taggers for the analytical languages like the English language, and also comparing to the overall quality of the translation systems for related languages.
- According to the today's most used designs for translation systems for related languages, the shallow transfer translation systems, the disambiguation module follows the source language morphological analysis at the beginning of the translation process. This design is shown on figure 1. Such design is adopted by Apertium [5] and Česílko [11]. Errors produced at the early stages of the translation process usually cause bigger problems than errors introduced at latest phases as later phases of the translation rely on the output of the preceding phases.
- Multiple translation candidates allow selection of the best candidates in the final phase when all available data for the translation has been accumulated. The most common translation errors are fluency errors of the target language and not adequacy errors. These errors commonly do not interfere with the meaning of the translation but rather on the grammatical correctness of the translation. They are mostly caused by the errors in morphological analysis or morphological syntheses.

The omission of the tagger and introduction of a ranking scheme based on target language statistical model as suggested in [13] yields better translation results as suggested in the same paper. The newly proposed architecture is shown on figure 2. This method is further described in the section 4. The introduction of multiple translation candidates generated from all possible morphological ambiguities as suggested in [13] leads to an exponential growth of the number of possible translation candidates.

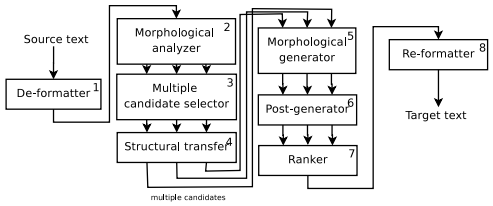

Fig. 2: *The newly proposed architecture of a shallow transfer translation system.*

The output of the morphological analysis is a set of all possible morphological tags describing each word. Every word with more than only one tag can be observed as a set of possible ambiguities. In the case of highly inflectional languages like the pair presented in this paper the number of ambiguous possibilities increases. The set of all possible translation candidates is constructed as the vector product of all ambiguous sets. The number of possible translation candidates grows exponentially with the length of the sentence, the equation 1 shows the upper limit of the number of possible translation candidates.

$$|TC| = \prod_{i=0}^{|S_{max}|} x_{max} \tag{1}$$

where TC is the set of possible translation candidates for the longest sentence S_{max} and x_{max} is the biggest number of ambiguities for a word. Although the equation 2, which shows the average number of possible translation candidates, presents much lower numbers, the complexity of the problem still remains exponential.

$$|TC| = \prod_{i=0}^{|\bar{S}|} \bar{x} \tag{2}$$

Equations 3 and 4 show empirical values for an example source sentence and typical numbers collected from a corpus test-set.

$$\begin{gathered} |\bar{S}| = 40 \\ \bar{x} = 15 \\ |TC| = \textstyle\prod_{i=0}^{40} 15 = 110,573323209e+45 \end{gathered} \tag{3}$$

$$\begin{gathered} |S| = 15 \\ x = 3 \\ |TC| = \textstyle\prod_{i=0}^{15} 3 = 14,348,907 \end{gathered} \tag{4}$$

4 Using a statistical post-processor instead of a POS tagger

A statistical language model assigns a probability to a sequence of words by means of a probability distribution. Using a language model, we assume that the training corpus used in the experiments is a good enough representation of the observed language.
The authors of the newest version of *Česílko*, the platform for MT for related languages [14], suggest a modified architecture for the translation system with the omission of the POS tagger, shown of figure 2. The experiment shown in [13] even shows a better translation performance of a similar architecture change on a different translation platform, the Apertium [5]. The experiment was conducted on a Romance language pair Portuguese - Spanish.
The exclusion of the tagger from the system has to be compensated somewhere else in the translation process since the output of the MT system is supposed to be one sentence.
An essential part of the whole MT system is the statistical post processor. The main problem with our simple MT process described in the previous sections is that both morphological analyser and transfer introduce a huge number of ambiguities into the translation. It would be very complicated (if possible at all) to resolve this kind of ambiguity by hand-written rules. That is why we have implemented a stochastic post-processor which aims to select one particular sentence that suits best the given context.

We use a simple language model based on trigrams (trained on word forms without any morphological annotation) which is intended to sort out "wrong" target sentences (these include grammatically ill-formed sentences as well as inappropriate lexical mapping). The current model has been trained on a corpus of 9 million words which have been randomly chosen from the Serbian Wikipedia[1].

Let us present an example of how this component of the system works. In the source text we had the following Slovenian segment (matrix sentence):

(5) *Kozarec* *na* *zeleni*
glass-N,M,SG,NOM on green-ADJ,F,SG,NOM
mizi ...
table-N,F,SG,NOM
"A glass on a green table ..."

The rule-based part of the system has generated four target segments: *Čaša na zelenom stolu*, *Čaša na zelenim stolu*. *Čaša na zelenim stolovima*, *Čaša na zelenom stolovima*.
The words *zelen* and *miza* are (fem.sg.loc and fem.du.nom). According to the language model, the ranker has (correctly) chosen the first sentence as the most probable result. There are also many homonym word forms that result in different lemmas in the target languages. For example, the word *bleda* means both "spinach beet-n.f.sg.nom" or "Beta vulgaris var cicla" and "pale-a.f.sg.nom". The ranker is supposed to sort out the contextually wrong meaning in such cases.

[1] http://sr.wikipedia.org

5 Method description

The authors of [13] suggest using a shallow parser to reduce the number of the translation candidates. The parser has to be hand-crafted. We propose a combination of two automatic methods to reduce the number of possible translation candidates:

- An automatically constructed set of local-agreement rules to discard improbable translation candidates.
- A statistical ranker of the translation candidates based on the POS tags of the source language to rank and select the remaining translation candidates from first method.

The *Guat* [20], a translation system for related languages based on *Apertium* [5], was used as the reference system and also as the sandbox for the implementation of proposed methods. Guat is automatically constructed so there is still room for improvement mainly through data correction tasks. The basic architecture of the system follows the architecture of apertium [5] and is presented in figure 1. The exponential complexity of the problem was addressed with a combination of two methods, each is presented in a separate section. The local agreement rules were used to discover and discard improbable translation candidates, the method is described in section 5.1. A statistical ranker [14] trained on the POS tags of a hand checked corpus [6] was used to select an n-best set of translation candidates, the method is presented in 5.2. The results and further discussion are presented in sections 7 and 8.

5.1 Discarding improbable translation candidates using local agreement rules

In this experiment we tried to use the agreement of morphological descriptors of adjacent words as a criterion to discard improbable translation candidates. The agreement of morphological descriptors can be modelled using rules based on regular expressions. The rules are described in section 5.1.1. The same format of rules as defined in the Apertium framework was used as it was powerful enough for the chosen task and it was already based on the same technology. The automatic induction of such rules is presented in the section 5.1.1. All the applicable rules, whose pattern regular expression describes part of the translation candidate, were applied on the translation candidate. If a rule changes part of the translation candidate, the candidate is discarded.

5.1.1 Automatic induction of local agreement rules

The automatic induction of the local agreement rules produces the same format of the rules as used by Apertium transfer module, but the method is limited to the discovery of local agreement. The requirements for the method are much simpler, just a morphologically annotated corpus of the source language.
Trigrams and bigrams with morphological descriptions

were extracted from source language part of the corpus. The corpus used as training data was [6], which was hand checked for errors in morphosyntactic tags. The source language used in our experiment was Slovenian language although the same method could be used for other languages as the method is not language specific. Each bigram and trigram was checked for agreement among tags of different words, the tags and their positions were free. If any agreements were found, a candidate for a rule was stored. The POS tags of the source bigram or trigram present the pattern part of the rule, the action part of the rule is constructed from all the morphosyntactic tags with agreement information. The rule candidates were grouped according to the pattern and action definitions, each group with a predefined number of candidates was chosen as a valid rule.

Although many improbable rules were discarded, longer sentences still yielded unmanageable number of translation candidates. Basically the growth of the problem was still exponential although the base was reduced. There is no guarantee on the upper limit of the number of translation candidates using this method. Another method was devised that coped with this problem, the ranking of translation candidates, described in section 5.2.

5.2 Ranking of the translation candidates

The empirical results in the table 2, the system named rules, show an improvement of the method that discards the improbable translation candidates using automatically induced rule, presented in section 5.1, over the original system and even over the system using all possible translation candidates. The problem of the number of possible translation candidates still rests to be resolved. The number of possible translation candidates, non-discarded by the first method, depends on the induced rules and in the worst case it is still exponential. A mechanism for ranking the remaining translation candidates and selecting a manageable subset had to be applied. A variation of the mechanism described in the section 4 was used in this experiment.
The main part of the MSD descriptors, the Category, was extracted from the source part of the training corpus [6] presenting a list of Category descriptors instead of the original words. The ranker was trained in the same way as explained in the 5.2 thus learning the most probable POS tag sequences of the source language.
Translation candidates obtained from the morphoogical analyser can be scored using this ranking scheme. A scoring scheme enables the selection of an n-best set of possible translation candidates thus reducing the problem to a fixed upper limit ensuring a limit to the worst-case translation time. The n can be an arbitrary number although lower numbers yield a translation quality penalty as shown in the ranker part of the table 2.

6 Evaluation method

The edit-distance [16] was used to count the number of edits needed to produce a correct target sentence from automatically translated sentence. The metric counts the number of deletions, insertions and substitutions that need to be performed among the observed sequences. This procedure shows how much work has to be done to produce a good translation. The metric roughly reflects the complexity of post-editing task.
The evaluation comprised of selecting 57 sentences from testing data, translating these sentences using the translation system and manually correcting the output of the system to a suitable translation. By suitable translation we mean a translation that is syntactically correct and expresses the same meaning as the source sentence. The sentences were chosen by length (shorter sentences than 15 words). This limitation enabled a fair comparison of the translation quality of all the systems, same test-set, as same systems used the full set of possible translation candidates. The complexity of each sentence was arbitrary, there was no special selection of the sentences using this criteria although shorter sentences are usually simpler in structure.
Two variants of the edit-distance [16] were calculated for each set of examples, the edit distance based on words and edit distance based on characters.
The results of this evaluation can be compared to results of the same metric used on similar systems; [14] and [20]. Language pair's properties and similarities of our system in comparison to [14] make the comparison feasible. The evaluation was conducted as a test on a low number of test translations due to the time and space limitations.

7 Results

Table 1: *The number of translation candidates for each system.*

System	All	All cand.	Used
original	57	57	57
all	44 million	44 million	34,526
rules	44 million	934,326	4,284
ranker	44 million	1,325,216	6,569
rules+ranker	44 million	437,123	4,041

Description of the table 1:
System - Name of the tested system, each system is presented in this section.
All - the number of all candidates produced from tested sentences.
All cand. - the number of all translation candidates entering the last translation phase, the ranking phase.
Used - the number of unique candidates entering the last translation phase, the ranking phase.

Description of the table 2: The figures in table 2 show the difference between 1 and the weighed edit distance, meaning higher values show better results.
System - Name of the tested system, each system is presented in this section.

Table 2: *Translation quality for each system.*

System	E. D. char.	E. D. word
original	0.848	0.778
all	0.892	0.826
rules	0.896	0.829
ranker	0.888	0.824
rules+ranker	0.896	0.829

E. D. char. - the edit distance based on characters showing the percentage of characters that were not changed.
E. D. word. - the edit distance based on words showing the percentage of words that were not changed.

Description of the systems presented on tables 1 and 2:
original - the translation system [20] based on original Apertium [5] architecture;
all - the *original* system with the omission of the statistical tagger. All possible ambiguous translation candidates were used.
rules - the *all* system with the introduction of the method based on automatically induced local agreement rules
ranker - the *all* system with the introduction of the method based on the ranking mechanism. The threshold was set at 100.000 translation candidates.
rules+ranker - the *all* system with the introduction of the method based on automatically induced local agreement rules with the ranking mechanism as a back off in case of too many translation candidates.

8 Discussion

The first experiment, the introduction of the proposed method by [14] to a new language pair, Slovenian - Serbian, showed an even bigger improvement of the proposed method as the original experiment.
The empirical evaluation showed that the introduction of the proposed methods increased the translation quality of the overall system by a statistically significant margin. The proposed methods successfully limit the number of possible translation candidates.
The empirical evaluation was conducted on a relatively small test sample due to time and resources limitations, the experiment should be evaluated on a bigger test-set. Further tests should be performed in the discovery of the best ranker threshold limit.
Some of the deterministic possibilities to restrict the blow-up of hypotheses have been explored, namely rule-based or heuristic removal of contextually inappropriate hypotheses (e.g. parser [14], tag sequences presented in this paper). Exploring different representations of the data as (multi)graph with multisets of edges and later compacting of the graph (contraction of fully/morphologically identical edges). The later approach should provide interesting results, especially for languages with high case syncretism.

References

[1] L. Ahrenberg and M. Holmqvist. Back to the future? the case for english-swedish direct machine translation. In *Proceedings of Recent Advances in Scandinavian Machine Translation*, 2005.

[2] K. Altintas and I. Cicekli. A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*, 2002.

[3] E. Bick and L. Nygaard. Using danish as a cg interlingua: A wide-coverage norwegian-english machine translation system. In *Proceedings of NODALIDA, Tartu*, 2007.

[4] T. Brants. Tnt – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*. Seattle, WA, 2000.

[5] A. M. Corbi-Bellot, M. L. Forcada, S. Ortiz-Rojas, J. A. Prez-Ortiz, G. Ramirez-Sanchez, F. Sanchez-Martinez, I. Alegria, A. Mayor, and K. Sarasola. An open-source shallow-transfer machine translation engine for the romance languages of spain. In *Proceedings of the Tenth Conference of the European Association for Machine Translation*, pages 79–86, May 2005.

[6] L. Dimitrova, N. Ide, V. Petkevič, T. Erjavec, H. J. Kaalep, and D. Tufis. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In *COLING-ACL*, pages 315–319, 1998.

[7] H. Dyvik. Exploiting structural similarities in machine translation. *Computers and Humanities*, 28:225–245, 1995.

[8] T. Erjavec. Multilingual tokenisation, tagging, and lemmatisation with totale. In *Proceedings of the 9th INTEX/NOOJ Conference*, 2006.

[9] J. Hajič. An mt system between closely related languages. In *Proceedings of the third conference of the European Chapter of the Association for Computational Linguistics*, 1987.

[10] J. Hajič. Morphological tagging: data vs. dictionaries. In *Proceedings of the North American chapter of the Association for Computational Linguistics conference*, 2000.

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

[12] J. Hajič, J. Hric, and V. Kuboň. Machine translation of very close languages. In *Proceedings of the 6th Applied Natural Language Processing Conference*, 2000.

[13] P. Homola and V. Kuboň. Improving machine translation between closely related romance languages. In *Proceedings of EAMT*, pages 72 – 77, 2008.

[14] P. Homola and V. Kuboň. A method of hybrid mt for related languages. In *Proceedings of IIS*, 2008.

[15] P. Homola, V. Kuboň, and J. Vičič. Shallow transfer between slavic languages. In *Recent Advances in Intelligent Information Systems*, page 219–232. Academic publishing house EXIT, Warsaw, 2009.

[16] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk*, pages 845–848, 1965.

[17] K. P. Scannell. Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, 2006.

[18] K. P. Scannell. Machine translation for closely related language pairs. *unknown*, 2008.

[19] F. M. Tyers, L. Wiechetek, and T. Trosterud. Developing prototypes for machine translation between two sámi languages. In *Proceedings of EAMT*, 2009.

[20] J. Vičič. Rapid development of data for shallow transfer rbmt translation systems for highly inflective languages. In *Jezikovne tehnologije, language technologies : zbornik konference : proceedings of the conference*, pages 98–103, 2008.

[21] L. R. Welch. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4):1–14, 2003.