A Fast Implementation of Rules Based Machine Translation Systems for Similar Natural Languages

Jernej Vičič
Faculty of Mathematics, Natural Sciences and Information Technologies
University of Primorska
E-mail: jernej.vicic@upr.si
http://www.jt.upr.si/doktoratjernej/thesis/final/

Thesis Summary

Keywords: machine translation, machine translation of related languages, shallow transfer RBMT, RBMT

Received: March 25, 2013

This paper is and extended abstract of the doctoral thesis [1]. It presents an overview of the systems and methods for the natural language machine translation. It focuses primarily on systems and methods for shallow transfer rule based machine translation which are better suited for the translation of related languages. The major problem of the rule-based translation systems is costly manual production of dictionaries and translation rules in the case of a classical approach to building such systems. The work provides an overview over the collection of selected and new methods designed for automatic production of materials for the installation of systems based on translation rules.

Povzetek: Pričujoče delo je razširjen povzetek doktorske disertacije [1]. Predstavlja pregled strojnega prevajanja naravnih jezikov, osredotoča se predvsem na sisteme in metode za prevajanje na osnovi pravil plitkega prenosa, ki so najprimernejše za sorodne naravne jezike. Največja težava sistemov, ki temeljijo na pravilih, je dolgotrajna in draga ročna izdelava slovarjev ter prevajalnih pravil v primeru klasičnega pristopa h gradnji prevajalnih sistemov na osnovi pravil. Delo ponuja pregled zbirke izbranih in na novo zasnovanih metod samodejne izdelave gradiv za postavitev prevajalnih sistemov na osnovi pravil.

1 Introduction and problem statement

The paper presents an attempt to automate all data creation processes of a rule-based shallow-transfer machine translation system and its background. Several methods that automate some parts of the shallow transfer Rule Based Machine Translation (RBMT) system construction have been presented and are even used as part of the construction toolkits like Apertium [2], which is a widely used open source toolkit for creating machine translation systems between related languages.

Parts of the creation process have been addressed by several authors, some of these technologies have been used in our experiments along with newly developed methods. All methods and materials discussed in this paper were tested on a fully functional machine translation system based on Apertium. The system uses an architecture similar to the one presented in Figure 1.

Although it seems that Statistical Machine Translation (SMT) would be a perfect choice as some of the best performing machine translation systems are based on the SMT technologies, the stochastic approach has a couple of drawbacks that cannot be ignored; the SMT systems, to be successful, require huge amounts of parallel texts.

Another reason for choosing the RBMT approach is the nature of the languages involved in our experiments (Slovenian paired with Serbian, Czech, English and Estonian language). These are languages with rich inflectional morphology and as such they present a big problem for SMT.

Last but not least reason for using an RBMT machine translation system is the chance for the linguistic experts to further refine the results of the automatically produced data and thus to be able to improve the system in a controlled way.

2 Methodology

The modules presented in Figure 1 and numbered with numbers 1 through 5 require linguistic data (monolingual dictionaries, bilingual dictionaries, translation rules, etc.). Each module was examined and a method for linguistic data creation was designed.

The following types of data are needed for all modules of the system: the monolingual source dictionary with morphological information for source language parsing, monolingual target dictionary with morphological information for target language generation, bilingual translation dictionary, finite-state rules for shallow transfer and local agree-

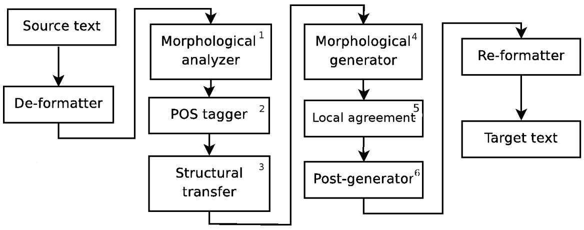

Figure 1: The modules of a typical shallow transfer translation system. The system [2] follows this design. An addition of the original architecture is the local agreement module tagged as number 6.

ment, statistical target language model, modeled source language tags.

3 Evaluation methodology and results

The evaluation focused only on the translation quality; the translation speed and responsiveness of the system, user-friendliness and other features of the translation systems are not presented. Were used the following methods: the automatic objective evaluation using the METEOR [3] metric, the non-automatic evaluation using weighted Levenshtein edit-distance [4] on a human corrected output of the translation system, the non-automatic subjective evaluation following [5] guidelines. The translation system was constructed according to the methodology presented in Section 2 using the selected training set. The evaluated values in each fold and the average final values are presented.

4 Discussion and further work

The agreement among all three evaluation methods is quite high, which shows that the results of the evaluation process are valid. The translation quality of the Slovenian-Serbian translation system is higher than the systems for distant language pairs. This can be attributed to the fact that the similarity of the first language pair is bigger.

The automatically generated linguistic data is far from perfect and additional manual labor will have to be executed in order to obtain better translation quality.

References

[1] J. Vičič, "Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike," Ph.D. dissertation, Univerza v Ljubljani, 2012. [Online]. Available: http://eprints.fri.uni-lj.si/1778/

[2] S. A. M. Corbi-Bellot, M. L. Forcada, Ortiz-Rojas, "An open-source shallow-transfer machine translation engine for the Romance languages of Spain," in *EAMT*, 2005, pp. 79–86.

[3] A. Lavie and M. J. Denkowski, "The Meteor metric for automatic evaluation of machine translation," *Machine Translation*, vol. 23, no. 2-3, pp. 105–115, Sep. 2009.

[4] K. S. Fu, *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1982.

[5] LDC, "Linguistic data annotation specification: Assessment of fluency and adequacy in translations," LDC, Tech. Rep., 2005.