АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic building of Machine Translation

PIAP- GA-2012-324414

D4.1a MT systems for the first development cycle

Dissemination level	Public
Delivery date	2013/08/31
Status and version	Final, 1.0
Authors and affiliation	Antonio Toral (DCU), Santiago Cortés-Vaíllo (Prompsit), Sergio Ortiz-Rojas (Prompsit), Gema Ramírez-Sánchez (Prompsit), Mikel L. Forcada (UA)

Project funded by the European Community under the Seventh Framework Programme for Research and Technological Development

Table of Contents

Executive Summary

One of the goals of project Abu-MaTran is to rapidly build a machine translation system from English to Croatian. Indeed, after collecting the necessary corpora (described in deliverable D3.1a) a number of systems were built and the best one (according to the evaluation described in deliverable D5.1) was made available through http://translator.abumatran.eu/ on July 1, simultaneously with the accession of Croatia to the European Union.

To build the systems, two different avenues have been taken:

- Building a statistical machine translation system using the limited amount of parallel corpora available for English and Croatian
- Building a hybrid machine translation system using, on the one hand, the larger amount of parallel text available for English and Slovene (most of it EU material), and, on the other hand, a Slovenian to Croatian rule-based system built as part of the Apertium free/open-source machine translation project (Forcada et al. 2011; http://www.apertium.org).

This deliverable describes the English–Croatian systems built in the first development cycle.

1 Introduction

This deliverable describes the English–Croatian systems built in the first development cycle. To build the systems, two different avenues have been taken:

- Building a statistical machine translation system using the limited amount of parallel corpora available for English and Croatian.
- Building a hybrid machine translation system using, on the one hand, the larger amount of parallel text available for English and Slovene (most of it EU material), and, on the other hand, a Slovenian to Croatian rule-based system built as part of the Apertium free/open-source machine translation project (Forcada et al., 2011).

Details on the evaluation of these systems are given in deliverable D5.1a.

2 Direct English→Croatian statistical systems

Two different English→Croatian (EN→HR) systems have been built, using the English–Croatian corpora described in Deliverable D3.1.

To build these systems, we have used MGIZA (Gao and Vogel., 2008) 0.7.3[1] to align the corpora, Moses (Koehn et al., 2007) 1.0[2] to train a regular (non-hierarchical) phrase-based statistical machine translation system with a maximum phrase length of 7 words, and IRSTLM (Federico et al., 2008) 5.80.01[3] to train the target-language model with a maximum *n*-gram size of 5 and the modified Knesser-Ney discounting.

The systems are as follows:

1. **EN-HR1**: trained on the concatenation of the SETimes, hrenWaC and TED Talks corpora, using the Croatian side of the training sets as target-language model.
2. **EN-HR2**: as EN-HR1 but training the target language model on the Croatian side of the training sets and the hrWaC monolingual corpus (described in Deliverable D3.1a).

Regarding the other translation direction, Croatian→English (HR→EN), two different systems have been built, as follows:

1. **HR-EN1**: trained on the concatenation of the SETimes, hrenWaC and TED Talks corpora, using the English side of the training sets as target-language model.
2. **HR-EN2**: as HR-EN1 but training the target language model on the English side of the training sets and the WMT'13 monolingual corpora (described in Deliverable D3.1a).

3 Indirect English→Slovene→Croatian systems

In view of the small amount of parallel data publicly available for the English–Croatian language pair, an alternative avenue was explored to build the machine translation system, which involved building a hybrid machine translation system using, on the one hand, the larger amount of parallel text available for English and Slovene (most of it EU material) to build an English→Slovene MT system, and, on the other hand, a rule-based Slovene→Croatian system built as part of the Apertium

1 http://www.kyloo.net/software/doku.php/mgiza:overview
2 http://www.statmt.org/moses/
3 http://sourceforge.net/projects/irstlm/

free/open-source machine translation project[4] (Forcada et al. 2011).

3.1 English→Slovene statistical machine translation systems

Using the same settings as in the previous section, an English→Slovene statistical machine translation systems was built. This system is trained on the concatenation of three parallel sources (Europarl, DGT-TM and EU bookshop) while the language model is built on the target-side of these parallel sources and the slWaC monolingual corpus. The corpora used are described in Deliverable D3.1a.

3.2 Slovene→Croatian rule-based machine translation system

A rule-based Slovene–Croatian (actually Slovene–Croatian/Bosnian/Serbian) system[5] was built by Hrvoje Peradin, Francis Tyers and Filip Petkovski based on the Apertium platform. The system contains around 13,000 lexical entries in each language, around 16,000 bilingual correspondences and around 1,000 morphological inflection paradigms. Part-of-speech tagging is mainly dealt with by constraint grammar rules (30 from Slovene to Croatian, 200 from Croatian to Slovene). The system also contains around 50 structural transfer rules in each direction to deal with the most important morphosyntactic divergences. The coverage of dictionaries is 85% on the Croatian side of SETimes and 95% on the Slovene side of Europarl. The system is more developed for the Croatian→Slovene direction than for Slovene→Croatian.

3.3 Hybrid English→Croatian systems

The following configurations were built and tried:

1. **Pivot**: the *n* best outputs of one of the English→Slovene statistical MT system described in Section 3.1 are directly fed into the Slovene→Croatian rule-based MT system described in Section 3.2 (Toral, 2012). The best of the *n* translations produced by the rule-based system is output. This system is noted as EN-SL-HR. We have two systems:
    1. EN-SL-HR-1 where "1" stands for *n* equals 1. This is a two stage system where the Slovene→Croatian system translates the top translation produced by the English→Slovene system.
    2. EN-SL-HR-o where "o" stands for oracle. In this system *n* equals 2000, i.e. the top 2000 translations from the English→Slovene system are fed to the Slovene→Croatian system. The top translation in Croatian (according to an oracle that selects the translation with highest sentence-level BLEU score) is output.
2. **Synthetic corpus**: the Slovene side of the training corpus is translated with the Slovene→Croatian rule-based MT system described in Section 3.2; then, the resulting English–"Croatian" corpus is used to train a SMT system as described in Section 2. This has been done in two different ways:
    - using the complete English–"Croatian" corpus for training
    - using only those sentences in which the rule-based system did not find any unknown word (words that were not in the dictionary of the rule-based system). Systems using these subset of sentence pairs are referred as "f" (filtered).

4 http://www.apertium.org
5 http://sourceforge.net/projects/apertium/files/apertium-hbs-slv/

3. **Direct + Indirect**: a direct system is used primarily (EN-HR2 for EN→HR and HR-EN2 for HR→EN) and an indirect system built with the synthetic corpus is used as back-off.[6] Direct data (which is smaller but arguably "cleaner" than the synthetic data) is given priority, but indirect data is also used, as being much larger, we hypothesise it should help to improve the coverage of the overall system. These systems are noted as EN-HR2-b, EN-HR2-bf, HR-EN2-b and HR-EN2-bf, where "b" stands for backoff and "f" stands for filtered.

Note that while systems Direct + Indirect operate in both directions (EN→HR and HR→EN), the pivot system is used only in the EN→HR direction.

4 Conclusions

This document has described the machine translation systems that have been built for the English–Croatian language pairs (in both directions). We have built statistical direct systems that use parallel data for the English—Croatian language pair (described in D3.1a) and also indirect systems pivoting via Slovene, given (a) the availability of large amounts of parallel data for English—Slovene (compared to the amount of parallel data available for English—Croatian) and (b) the high level of similarity between Slovene and Croatian. The best of these systems (according to the evaluation described in D5.1a), was made available on July 1, 2013 (the EU accession date for Croatia) through http://translator.abumatran.eu/.

6 Scoring is carried out with either phrase table and the second table is used for unknown n-grams up to size 7. More details at http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc23

Bibliography

Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M. (2011), "Apertium: a free/open-source platform for rule-based machine translation", Machine Translation, (Special Issue on Free/Open-Source Machine Translation) 25:2, 127-144.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo (2008), “IRSTLM: an open source toolkit for handling large scale language models”. In INTERSPEECH, pages 1618–1621. ISCA.

Qin Gao and Stephan Vogel. 2008. “Parallel implementations of word alignment tool. In Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pages 49–57. Association for Computational Linguistics .

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst (2007), “Moses: open source toolkit for statistical machine translation”. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 177–180.

Antonio Toral (2012), “Pivot-based Machine Translation between Statistical and Black Box systems”, in Proceedings of the 15th Annual Conference of the European Association for Machine Translation. Trento (Italy). May 2012.