Abu-MaTran

Automatic building of Machine Translation

PIAP- GA-2012-324414

D3.1a Acquisition 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), Gema Ramírez-Sánchez (Prompsit), Mikel L. Forcada (UA), Nikola Ljubešić (UZ)

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

Table of Contents

Executive Summary

For the fast building of machine translation systems containing corpus-based components (such as statistical machine translation systems or hybrid systems including rule-based components), one needs to collect data for the languages involved; in particular, one needs sentence-aligned parallel corpora (that is, a large amount of sentences and their translations) and, optionally, large amounts of target-language text.

One of the goals of project Abu-MaTran is to rapidly build a machine translation system from English to Croatian. Indeed, a number of systems were built (described in deliverable D4.1a) and the best one (according to the evaluation described in deliverable D5.1a) was made available through http://translator.abumatran.eu/ on July 1st, 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).

This deliverable describes the English–Croatian and English–Slovene corpora collected from different sources (including the criteria used to select them) to build such systems.

1 Introduction

The purpose of this deliverable is to describe the corpora acquired in the first development cycle to build a set of English–Croatian machine translation systems, which are described in deliverable D4.1a and evaluated in deliverable D5.1a.

2 Parallel corpora

2.1 Parallel corpora for English–Croatian

Table 2.1 shows the available corpora that have been used for the English–Croatian language pair, as well as their size (number of sentence pairs) and a brief description.

Corpus	Size	Description
SETimes	201,910	Content published on the SETimes.com news portal.
hrenWaC	99,001	Crawled from web pages.
TedTalks	86,348	Transcribed TED talks.
Total	387,259	

Table 2.1: parallel corpora used for English–Croatian.

The SETimes.com corpus, initiated by Tyers and Alperen (2010), and improved by Nikola Ljubešić[1] is now available through the Opus website.[2] This second version has been used.

The hrenWac (Croatian–English web-as-a-corpus), compiled by partner UZ,[3] is available through Opus.[4]

Transcribed TED Talks[5] were translated to Croatian and made available by Željko Agić at partner UZ.

2.2 Parallel corpora for English–Slovenian

Table 2.2 shows the available corpora that have been used for the English–Slovenian language pair, as well as their size (number of sentence pairs) and a brief description.

The Europarl corpus (Koehn 2005)[6] contains the transcription of European Parliament[7] speeches.

The DGT-TM[8] contains documents from the Acquis Communautaire (EU laws).

The EUbookshop corpus has been described above in Section 2.1.

1 http://www.nljubesic.net/resources/corpora/setimes/
2 http://opus.lingfil.uu.se/
3 http://nlp.ffzg.hr/resources/corpora/hrenwac/
4 http://opus.lingfil.uu.se/hrenWaC.php
5 https://wit3.fbk.eu/
6 http://statmt.org/europarl/
7 http://www.europarl.europa.eu/
8 http://ipsc.jrc.ec.europa.eu/index.php?id=197

Corpus	Size	Description
Europarl	623,490	Transcriptions from European Parliament.
DGT-TM	2,663,301	Translation memories from European Commission's Directorate-General for Translation.
EUbookshop	426,641	Publications from European institutions.
Total	3,713,432	

Table 2.2: parallel corpora used for English–Slovenian

2.3 Reasons to select these particular parallel corpora

Although several other corpora are publicly available for these language pairs, they have not been considered for different reasons:

- Belonging to a very restricted specific domain, e.g. KDE4 and PHP (software documentation), EMEA (medical), ECB (banking).
- Containing noisy data and spoken language, e.g. Open subtitles.
- Small size, e.g. Tatoeba contains less than 1,000 sentence pairs for English–Croatian (but is used in deliverable D5.1a for evaluation).

While it is true that some of the corpora considered could be deemed as domain-specific, they cover different subjects. E.g. Europarl contains parliamentary speeches in a variety of topics.

3 Monolingual corpora

To build the target-language models in statistical machine translation, it is convenient to use a target-language corpus which is larger than the target-side of the parallel corpus, in particular when the parallel corpus is not very large.

To translate from English to Croatian, the hrWaC corpus[9] was used for this purpose. For the opposite language direction (Croatian to English), we use additional resources from the WMT'13 shared task,[10] namely the English side of several parallel corpora (News commentary, Europarl, United Nations, Giga French–English and Common Crawl) and English monolingual news sources (News crawl 2007 to 2012). Finally, the system that translates trough Slovenian uses slWaC[11] for language modelling purposes.

4 Conclusions

This document has described the parallel and monolingual corpora used to build the statistical English–Croatian and English–Slovene machine translation systems described used to translate between English and Croatian (described in Deliverable D4.1a), the best of which (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/.

9 http://nlp.ffzg.hr/resources/corpora/hrwac/
10 http://www.statmt.org/wmt13/translation-task.html
11 http://nlp.ffzg.hr/resources/corpora/slwac/

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.

Koehn, P. (2005) Europarl: A Parallel Corpus for Statistical Machine Translation, in Proceedings of Machine Translation Summit X (Phuket, Thailand, September 2005), pp. 79–86.

Tyers, Francis M. and Murat Serdar Alperen (2010), South-East European Times: A parallel corpus of the Balkan languages. In Proceedings of the Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages at LREC 2010, pp. 49–53 (http://www.lrec-conf.org/proceedings/lrec2010/workshops/W22.pdf)