Final Results of Abu-MaTran (Automatic building of Machine Translation)

Antonio Toral
Faculty of Arts
University of Groningen
NL-9712 EK Groningen
a.toral.ruiz@rug.nl

Víctor Sánchez-Cartagena
Prompsit Language Enginering
E-03202 Elx
vmsanchez@prompsit.com

Mikel L. Forcada
Dept. Lleng. i Sist. Inform.
Universitat d'Alacant
E-03071 St. Vicent del Raspeig
mlf@ua.es

Abstract

We present the final results of Abu-MaTran (http://www.abumatran.eu), a 4-year project (January 2013–December 2016) on rapid development of machine translation for under-resourced languages. It was funded under the Marie Curie's Industry-Academia Partnerships and Pathways 2012 programme. The Abu-Matran consortium had 5 partners (4 academic and 1 industrial) in four different countries.

1 Introduction

Abu-MaTran sought to enhance industry-academia cooperation as a key aspect to tackle one of Europes biggest challenges: multilingualism. We aimed to increase the hitherto low industrial adoption of machine translation (MT) by identifying crucial cutting-edge research techniques, making them suitable for commercial exploitation. We also aimed to transfer back to academia the know-how of industry to make research results more robust. We worked on a case study of strategic interest for Europe: MT for the language of a new member state (Croatian) and for related languages. All the resources produced have been released as free/open-source software, resulting in effective knowledge transfer beyond the funded period.

2 Results

At EAMT 2017 we will present a selection of the final results of the project, including the following:

- **Web crawling:** A novel pipeline to crawl massive amounts of parallel and monolingual data from the Internet's top level domains that is ready for commercial exploitation.
- **Acquisition of language resources** (bilingual dictionaries and transfer rules): We have developed methodologies (i) to enable non-expert users to improve the coverage of morphological dictionaries and (ii) to learn automatically translation rules from very small parallel corpora.
- **Language models:** Implementation of a novel cloud-based language model that allows us to use effectively vast amounts of monolingual data in phrase-based statistical MT.
- **Linguistically-augmented approaches,** including morph-segmentation approaches, to phrase-based and neural MT.
- **Improved data selection** of training data for MT using linguistic information and quality estimation techniques.
- **Collaborative development of MT:** development of state-of-the-art rule-based MT between closely-related languages through a collaborative process.
- **Dissemination:** Workshops on (i) tools for teaching MT and on (ii) methodologies for rapid development of MT for under-resourced languages; and the establishment of a linguistics Olympiad in Spain.

All the tools and data sets developed within the project were released according to free/open-source licenses and can be found at the project's website.[1]

[1] http://www.abumatran.eu/