АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic Building of Machine Translation

PIAP- GA-2012-324414

D2.5 End-user workshop

Dissemination level	Public
Delivery date	2016/12/31
Status and version	Final, v1.0
Authors and affiliation	Víctor M. Sánchez-Cartagena (Prompsit) and Antonio Toral (DCU)

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

Contents

Executive Summary

This deliverable corresponds to task T.2.5. (End-user Workshop) within work package 2 (Dissemination and Outreach) aiming at the dissemination of the project for different audiences (academia and general public, in this case). Two workshops covering the same topics were organized by the industrial partner, Prompsit Language Engineering, during the secondments of Víctor M. Sánchez-Cartagena at Dublin City University (DCU) in November 2016 and at University of Alicante (UA) in December 2016. The objective of both workshops was to publicise a set of tools developed by project partners that make understanding statistical machine translation easier for students, translators, etc.

The first workshop was held under the supervision of Prof. Andy Way and the support of DCU's technical and administrative staff. The workshop took place at DCU's School of Computing premises on 17th November 2016 with the participation of 6 people from the School of Applied Language and Intercultural Studies at DCU. The attendees were mainly researchers and lecturers from the Translation Studies field. The workshop lasted 3 hours and combined theory and hands-on activities. All workshop materials that were produced as part of this task (participant's guide and slides) have been made available to the general public through the Abu-MaTran website. They have been published under a free license.

The second workshop was held under the supervision of Dr. Felipe Sánchez-Martínez and the support of UA's technical and administrative staff. The workshop took place at UA's Escuela Politécnica Superior premises on 16th December 2016 with the participation of 11 people. The attendees were mainly students (some of them already graduated) of a translation degree at UA and the workshop lasted 4 hours. The contents were broadly the same ones as in the first workshop, but they were adapted to the audience: students instead of teachers. As in the first workshop, all materials have been made available to the general public through the Abu-MaTran website.

Both workshops received positive feedback. The participant's guide is likely to be used by some DCU workshop attendees in the Machine Translation courses they teach, while the students who attended the workshop held at UA understood the impact of the different resources in the training of statistical machine translation systems thanks to the hands-on activities.

1 Workshop description

The *End-user Workshop*, which was finally named *Workshop on the Abu-MaTran project: tools for teaching machine translation* when it was held at DCU and *Workshop on statistical machine translation for curious translators* when it was held at UA, aimed at summarizing the achievements of the project and presenting tools developed by project partners that make understanding statistical machine translation easier for students, translators, etc.

In particular, two tools were presented: Bicrawler,[1] a service that allows users to easily obtain bilingual data from multilingual websites; and MTradumàtica, a web interface with which anyone can train and test phrase-based statistical machine translation systems in a couple of clicks.

The main achievements of the Abu-MaTran project were presented to the participants by means of a talk supported by slides. The linguistic resources acquired and MT systems built were described, as well as the main challenges found during the process. The talk also focused on the successful participation of the project partners in different MT shared tasks and on the dissemination activities carried out. This talk was shortened in the workshop held at UA, as researchers are more likely to be interested in the project achievements than students.

The tool Bicrawler was introduced by means of a short talk, in which the approach followed by web crawling tools in order to produce a parallel corpus from a multilingual website was outlined, stressing the linguistic resources used in each step:

1. Download web pages (documents)
2. Extract text and remove HTML tags
3. Detect language of documents
4. Identify documents that are mutual translation
5. Extract parallel sentences from each document pair

Afterwards, attendees worked on a hands-on session in which they learnt how to use Bicrawler and the format of the parallel corpora it produces.

[1] `http://bicrawler.com`

The tool MTradumàtica was also introduced in a talk, in which the principles behind statistical machine translation were explained with the help of slides. This talk was extended in the workshop held at UA because many students were not familiar at all with statistical machine translation. In the workshop held at DCU, the attendees followed the participant's guide in order to install the tool on their own computers (so that they can use it later for preparing their teaching materials). In the workshop held at UA, an online version of the tool was prepared and students did not need to install the tool. In both workshops, attendees used MTradumàtica to train statistical machine translation systems with different parallel and monolingual corpora and study the impact of each resource by examining the phrase table and the language model. In particular, they learnt how to use MTradumàtica to:

- Train an statistical machine translation system (including a language model) from a parallel corpus
- Translate sentences using a trained system
- Inspect the phrase table and check why some words from the source-language sentence were not translated
- Check the perplexity of different translation hypotheses according to the language model
- Train statistical machine translation systems from the concatenation of different parallel and monolingual corpora
- Tune statistical machine translation systems

The guide as well as the slides were distributed to all participants and are available under a free/open-source creative commons license through the Abu-MaTran website at: `http://www.abumatran.eu/?p=474` and `http://www.abumatran.eu/?p=479`. A copy of both materials is also attached to this deliverable.

2 Conclusions

We can highlight some very positive aspects that arose from the organization of these workshops:

- Attendees: although the number of participants in the workshop held at DCU was not large (6 people), all of them were researchers in the linguistics field and 4 of them were Machine Translation lecturers. This will boost the dissemination of the tools presented. The number of participants in the workshop held at UA was larger (11 people).

- Useful content: the workshop guide allows users who could not attend the workshop to access, install, and understand the tools presented. Additionally, the attendees who experimented problems with the installation of MTradumàtica during the workshop held at DCU were able to finish all the proposed activities on their own after the workshop thanks to the availability of the guide and the software. All the attendees to the workshop held at UA were able to finish the proposed activities before the end of the workshop.

- Feedback: we received positive feedback about the workshop from the participants. Lecturers who attended the DCU workshop thought it was a very useful teaching tool, while students who attended the UA workshop understood how statistical machine translation works and the impact of the different resources because they trained and tested statistical machine translation systems with their own hands.

- Re-usability: the section about the MTradumàtica tool of participant's guide is likely to be used by some DCU workshop attendees in the Machine Translation courses they teach.

A Workshop Materials

The written materials used at the workshops (slides and guide) are delivered together with this document in PDF format. The fonts for slides (LibreOffice Open Document format) and guide (LaTeX format) will be delivered on demand and under a Creative Commons Attribution-ShareAlike 3.0 license.