Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 185-188 recibido 29-03-2015 revisado 27-04-2015 aceptado 08-06-2015

Automatic Acquisition of Machine Translation Resources in the Abu-MaTran Project *

Adquisición automática de recursos para traducción automática en el proyecto Abu-MaTran

Antonio Toral, Tommi Pirinen, Andy Way,
ADAPT Centre, School of Computing, Dublin City University, Ireland

Raphaël Rubino, Gema Ramírez-Sánchez, Sergio Ortiz-Rojas, Víctor Sánchez-Cartagena, Jorge Ferrández-Tordera,
Prompsit Language Engineering, S.L., Elx, Spain

Mikel Forcada, Miquel Esplà-Gomis,
Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain

Nikola Ljubešić, Filip Klubička,
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Prokopis Prokopidis and Vassilis Papavassiliou
Institute for Language and Speech Processing, Athens, Greece
info@abumatran.eu

Resumen: Este artículo presenta una panorámica de las actividades de investigación y desarrollo destinadas a aliviar el cuello de botella que supone la falta de recursos lingüísticos en el campo de la traducción automática que se han llevado a cabo en el ámbito del proyecto Abu-MaTran. Hemos desarrollado un conjunto de herramientas para la adquisición de los principales recursos requeridos por las dos aproximaciones más comunes a la traducción automática, modelos estadísticos (corpus) y basados en reglas (diccionarios y reglas). Todas estas herramientas han sido publicadas con licencias libres y han sido desarrolladas con el objetivo de ser útiles para ser explotadas en el ámbito comercial.
Palabras clave: Traducción automática, adquisición de recursos lingüísticos, cooperación entre universidad y empresa

Abstract: This paper provides an overview of the research and development activities carried out to alleviate the language resources' bottleneck in machine translation within the Abu-MaTran project. We have developed a range of tools for the acquisition of the main resources required by the two most popular approaches to machine translation, i.e. statistical (corpora) and rule-based models (dictionaries and rules). All these tools have been released under open-source licenses and have been developed with the aim of being useful for industrial exploitation.
Keywords: machine translation, acquisition of language resources, industry-academia cooperation

1 Introduction

Abu-MaTran (Automatic bulding of Machine Translation)[1] is a four-year EU Marie-Curie IAPP (Industry-Academia Partnerships and Patways) project (2013–2016) that seeks to enhance industry-academia cooperation as a key aspect to tackle one of Europe's biggest challenges: multilinguality. The consortium is made up of four research institutions (Dublin City University, Universitat d'Alacant, University of Zagreb and the Institute for Language and Speech Processing in Athens) and one industry partner (Prompsit Language Engineering).

We aim to increase the hitherto low indus-

* The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

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

ISSN 1135-5948

trial adoption of machine translation (MT) by identifying crucial research techniques, preparing them to be suitable for commercial exploitation and finally transferring this knowledge to industry. On the opposite direction, we transfer back to academia the know-how of industry regarding management, processes, etc. to make research products more robust. The project exploits the open-source business model, all the resources produced are released as free/open-source software, resulting in effective knowledge transfer beyond the consortium.

While MT is nowadays a rather mature technology, it is still far from being widely adopted in industry. We argue that this has to do with the lack of required language resources (LRs). For example, if we look at the level of MT support for European languages, out of 30 languages, only 3 languages are considered to count with moderate to good support (English, Spanish and French), while the level of support for the remaining 27 languages ranges from fragmentary to weak or even none (Rehm and Uszkoreit, 2013).

An important strand of the Abu-MaTran project aims to alleviate this so-called LR bottleneck by providing ready-to-use tools for the automatic acquisition of the LRs required by MT systems. This paper provides an overview of the research and development actions carried out in the project in this respect. We also detail the resources that have been acquired.

While the tools we develop aim to be generic, we have a specific case study. This case study has been selected according to its strategic interest in the European context. We acquire the required resources to provide MT for the official language of a new member state of the EU (Croatian) and then extend to related South-Slavic languages official in candidate member states, such as Serbian and Bosnian. It should be noted that all these languages are considered to be under-resourced (Rehm and Uszkoreit, 2013).

The rest of the paper is organised as follows. Section 2 deals with the acquisition of resources for statistical MT (SMT) systems, namely corpora. Next, Section 3 regards the acquisition of resources for rule-based MT (RBMT) systems, namely dictionaries and rules. Finally, Section 4 derives conclusions and outlines future work directions.

2 Corpora

This section covers the acquisition of corpora, both monolingual (Section 2.1) and parallel (Section 2.2), as well as the cleaning of noisy parallel corpora (Section 2.3).

2.1 Acquisition of Monolingual Corpora

Monolingual corpora constitute a cheap (in comparison to parallel corpora) and important resource for SMT systems as they can be used to build language models for the target language.

We propose to crawl top-level domains (e.g. `.hr` for Croatia) in order to acquire vast amounts of monolingual data. The procedure has been used to crawl data for Croatian (1.9 billion tokens), Bosnian (429 million tokens) and Serbian (894 million tokens) (Ljubešić and Klubička, 2014) as well as for Catalan (779 million tokens) (Ljubešić and Toral, 2014).

While the previous approach yields general-domain data, we have also developed a novel tool to crawl tweets, given the growing importance of social media. This tool, TweetCat (Ljubešić, Fišer, and Erjavec, 2014), has been used to acquire tweets for Croatian, Serbian and other similar languages (235 million words) as well as for Slovene (38 million words).

2.2 Acquisition of Parallel Corpora

Compared to monolingual corpora, the acquisition of parallel corpora is considerably more complex. While the main aim for building these corpora is to train SMT systems, we envisage other purposes too such as assisting translators (Rubino et al., 2015).

We have built upon two parallel crawlers previously developed by project partners for research purposes, ILSP Focused Crawler (Papavassiliou, Prokopidis, and Thurmair, 2013)[2] and Bitextor (Esplà-Gomis and Forcada, 2010).[3] In Abu-MaTran we have prepared them to be ready for commercial exploitation. As a result, it is now straightforward to use these tools for crawling parallel data (Papavassiliou et al., 2014). In the project we have used these crawlers to acquire parallel corpora for the

[2] `http://nlp.ilsp.gr/redmine/projects/ilsp-fc`

[3] `http://sourceforge.net/projects/bitextor/`

tourism domain (Esplà-Gomis et al., 2014) for the language pair Croatian–English (140 thousand sentence pairs).

2.3 Cleaning of Noisy Parallel Corpora

There are vast amounts of publicly available parallel data that are not clean enough to be usable to be used for training MT systems. We have proposed a cleaning procedure so that these corpora can be useful to train MT systems (Forcada et al., 2014a). We have applied this procedure to OpenSubtitles,[4] a set of corpora made of open-domain subtitles available for several language pairs. The cleaning procedure fixes some recoverable errors and removes noisy sentence pairs (e.g. misaligned pairs). We have evaluated our procedure on the OpenSubtitles corpus for English–Croatian. An SMT system built on the clean version outperforms a system built on the original corpus by approximately 10 BLEU points absolute (Forcada et al., 2014b).

3 RBMT Resources

Part of our research is focused on RBMT systems. These systems have proven to be a sensible choice when translating between related languages, which is the case of the South-Slavic languages covered in Abu-MaTran. One of the weakest points of RBMT is that developing such systems may result expensive, since linguists have to manually encode the translation rules and dictionaries used by these systems. Our research focuses then on the automatic and semi-automatic acquisition of the main resources used by RBMT: dictionaries (Section 3.1) and rules (Section 3.2).

3.1 Dictionaries

We have proposed a novel approach to assist non-expert users to add new words to the morphological dictionaries used in RBMT systems (Esplà-Gomis et al., 2014). Our method helps the user to add unknown words and find their correct morphological paradigm by asking the user about the possible derivations of the word. A hidden Markov model is used to minimise the amount of necessary questions.

[4] `http://opus.lingfil.uu.se/OpenSubtitles.php`

3.2 Transfer Rules

Transfer rules encode the information needed to deal with the grammatical divergences between languages and they are usually developed by linguists. In order to enable the rapid and cheap building of RBMT systems we have developed a novel approach that learns shallow-transfer MT rules from very small amounts (a few hundreds of sentences) of parallel corpora (Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez, 2015).

Experiments on five language pairs have shown that the translation quality significantly improves that obtained with previous approaches (Sánchez-Martínez and Forcada, 2009) and is close to that obtained with handcrafted rules.

4 Conclusion and Future Work

We have provided an overview of the research and development activities carried out in the Abu-MaTran project to alleviate the LR bottleneck in MT. To this end, we have tackled the acquisition of resources needed to build SMT (corpora) and RBMT systems (dictionaries and rules).

Regarding SMT resources, we have established a robust pipeline to crawl monolingual and parallel corpora that is ready for commercial exploitation. We have also devised a novel procedure to clean publicly available corpora that are not usable for MT as they are.

As for RBMT resources, we have proposed 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.

In the remaining two years of the project, we will continue our work on acquisition as follows. Regarding corpora, we plan to combine the approaches for crawling of top-level domains and parallel crawling in a single tool. This will allow users to crawl both monolingual and parallel data for any language that is associated to a top-level domain by issuing a single command. As for linguistic resources, we will apply the tools presented for the acquisition of dictionaries and rules to bootstrap the development of a rule-based MT system for the pair of closely-related languages Croatian–Serbian.

References

Esplà-Gomis, M. and M. L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *Prague Bull. Math. Linguistics*, 93:77–86.

Esplà-Gomis, M., V. M. Sánchez-Cartagena, J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. L. Forcada, and R. C. Carrasco. 2014. An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation Translation*, pages 19–26, Dubrovnik, Croatia, June.

Esplà-Gomis, M., F. Klubička, N. Ljubešić, S. Ortiz-Rojas, V. Papavassiliou, and P. Prokopidis. 2014. Comparing two acquisition systems for automatically building an english-croatian parallel corpus from multilingual websites. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Forcada, M. L., S. Ortiz-Rojas, T. Pirinen, R. Rubino, and A. Toral. 2014a. Abu-MaTran deliverable D4.1b MT systems for the second development cycle. `http://www.abumatran.eu/?page_id=59`.

Forcada, M. L., T. Pirinen, R. Rubino, and A. Toral. 2014b. Abu-MaTran deliverable D5.1b Evaluation of the MT systems deployed in the second development cycle. `http://www.abumatran.eu/?page_id=59`.

Ljubešić, N., D. Fišer, and T. Erjavec. 2014. TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Ljubešić, N. and F. Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden.

Ljubešić, N. and A. Toral. 2014. cawac - a web corpus of catalan and its application to language modeling and machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Papavassiliou, V., P. Prokopidis, M. Esplà-Gomis, and S. Ortiz. 2014. Abu-MaTran deliverable D3.2. Corpora Acquisition Software. `http://www.abumatran.eu/?page_id=59`.

Papavassiliou, V., P. Prokopidis, and G. Thurmair. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria, August.

Rehm, G. and H. Uszkoreit. 2013. META-NET Strategic Research Agenda for Multilingual Europe 2020. `http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf`. [Online; accessed 27 March 2015].

Rubino, R., M. Esplà-Gomi, A. Toral, V. Papavasiliou, and P. Prokopidis. 2015. DIY Domain Specific Parallel Corpora for Translators. In *To appear in Proceedings of the IV International Conference on Corpus Use and Learning to Translate (CULT)*, Alacant, Spain.

Sánchez-Cartagena, V. M., J. A. Pérez-Ortiz, and F. Sánchez-Martínez. 2015. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech and Language*, 32(1):46–90.

Sánchez-Martínez, F. and M. L. Forcada. 2009. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34:605–635.