АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic building of Machine Translation

PIAP- GA-2012-324414

D3.1b. Acquisition for cycle 2

Dissemination level	Public
Delivery date	2014/12/31
Status and version	Final, v1.0
Authors and affiliation	Miquel Esplà-Gomis (UA), Mikel L. Forcada (UA), Nikola Ljubešić (UZ), Vassilis Papavassiliou (ILSP), Prokopis Prokopidis (ILSP), Sergio Ortiz-Rojas (Prompsit), Raphaël Rubino (Prompsit), Víctor Sánchez-Cartagena (Prompsit), Antonio Toral (DCU)

PEOPLE

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

Table of Contents

Executive Summary

This deliverable (D3.1b) describes the activities carried out within the project during the period between the first milestone (or "first development cycle", month 7) and the second milestone period (or "second development cycle", month 24), regarding the acquisition of translation resources to be added to those reported in public deliverable D3.1a (Toral et al. 2013a). These resources are needed to build improved statistical and rule-based machine translation (MT) components. The acquired resources are, on the one hand, monolingual and parallel text corpora for statistical MT, harvested using the improved web crawlers developed in the project, and, on the other hand, dictionaries and translation rules, for rule-based MT, using semiautomatic techniques for dictionary enrichment and fully automatic techniques to learn rules from parallel corpora, both developed in the project. Work has encompassed a variety of languages but has focussed on the Croatian–English language pair, and two domains: the tourist domain and an independent news domain. The deliverable refers to papers describing the acquisition techniques and the resulting resources, presented at main conferences and in a mainstream journal.

1 Introduction

This deliverable reports on the activities carried out within the project regarding acquisition during the second milestone period (M7 to M24). The main aim of the deliverable is to report on the automatic acquisition of additional resources to those used in the first milestone (cf. D3.1a, Toral et al. 2013a) in order to improve the performance of machine translation, focusing on the Croatian–English language pair.

The rest of this deliverable is organised as follows. In Section 2 we report on our efforts to automatically acquire monolingual corpora. We cover both the acquisition of generic data (Section 2.1) as well as user-generated content (Section 2.2). In Section 3 we report on our lines of work for the acquisition of parallel corpora, which have concerned a tourism corpus for English–Croatian (Section 3.1) and a set of parallel corpora for independent news involving 15 languages (Section 3.2). To close this section, we present a method we have created to derive synthetic parallel data for under-resourced languages (Section 3.3). Subsequently, Sections 4 and 5 deal with the acquisition of linguistic resources that are used in rule-based machine translation, dictionaries and rules, respectively.

2 Monolingual Corpora

2.1 Acquisition of Generic Monolingual Corpora

Monolingual corpora are a cheap (in comparison to parallel corpora) and important resource for statistical machine translation. In the first milestone of the project we have already used the hrWaC v1.0 web corpus of Croatian (Ljubešić and Erjavec, 2011) for producing language models of the target language. Given the recent improvements of the tools for building large corpora from the web and the constantly increasing amount of data available, we have decided to recrawl the Croatian (.hr) top-level domain and additionally crawl top-level domains of Bosnia (.ba) and Serbia (.rs). All three crawls were run for 21 days on 16 server-grade cores. The Bosnian and Serbian web corpora will be especially useful in the third year of the project (2015) when we will extend our machine translation technologies to these languages.

Ljubešić and Klubička (2014) describe the corpora construction process which yielded a Croatian corpus of 1.9 billion tokens, a corpus of Bosnian containing 429 million tokens and a corpus of Serbian containing 894 million tokens. The constructed corpora represent the largest corpora available for all the three languages. During the construction process special emphasis was put on discriminating between these three languages via a simple, but data-heavy token-level unigram language model since all three languages are used on all three top-level domains and standard language identification algorithms cannot cope with this task. All three corpora were then linguistically annotated with the tools developed inside the project (Agić and Ljubešić, 2014).

The usefulness of the corpora produced goes beyond language modeling for statistical machine translation. First, they can be used for assessing the frequency of a specific form which is already used in guessing the inflective morphological paradigms in a semi-automatic approach to extending morphological dictionaries, which are a vital resource for rule-based machine translation. Secondly, by performing multilingual crawls of top-level domains, multilingual hotspots can be identified, which are good candidates for in-depth crawling with the complementary goal of identifying parallel data.

To further assess the potential of this approach, we crawled a web corpus for the top level domain of Catalan (.cat). The only human intervention in the process had to do with (i) compiling a seed list of URLs to start the crawling process and (ii) providing a clean sample of Catalan data to perform distance-based language identification where each document being over a distance threshold to this language sample is discarded. Crawling lasted for 21 days and yielded a corpus of 779 million tokens. We assessed the usefulness of this corpus on two tasks: language modeling and machine

translation. For both tasks, our crawled corpus brings substantial improvements on performance. Ljubešić and Toral (2014) describe the process in more detail.

2.2 Acquisition of Monolingual Corpora of User Generated Content

Recent interest in the automatic processing of user-generated content can be easily explained with the advent of social networks and its huge potential for both research and industry. Machine translating such content has its appeal as well, and the most straightforward way of adapting a statistical translator to such content is by providing it with a target-language language model built from user-generated content.

One of the social networks most open to the idea of using its data either for research or for building products is Twitter, which offers convenient APIs that allow users to easily gather data for analysis and modelling. On the other side, metadata on the language the tweet is written in is very unreliable which makes the process of collecting tweets written in a medium-density language in no way straight-forward. This is why we developed the TweetCaT tool[1] (Ljubešić et al. 2014), allowing users, by defining seed terms specific for a language, to identify users that tweet in that language and collect all their tweets during a longer period of time. By collecting Twitter corpora of Croatian, Serbian and Slovene we have shown that even for such low-density languages, corpora of a few million tweets can be built within a time period as short as one week.

During the 235-day long collection procedure (which is still running) two Twitter corpora were produced: one for Slovene, 38 million words in size, and another one for Croatian, Serbian and other similar languages, containing 235 million words. Discriminating between Twitter users of the latter language group will be dealt with in the second part of the project.

First preliminary experiments on using the collected Twitter collections for building language models for statistical machine translation to translate tweets show a significant increase in the BLEU score in comparison to using generic language models. We will continue with these experiments and report in detail about them in Year 3 deliverables.

3 Parallel Corpora

3.1 Acquisition of Parallel Corpora for Tourism

While on the first milestone of the project our focus was to build a generic MT system for Croatian–English to be ready on the date Croatia joined the EU (1st July 2013), for the second milestone we consider also specific domains. Specifically, we are interested in using fully automatic crawling procedures to acquire domain-specific parallel data that will then be used to train domain-specific systems. Our motivation is to use this fully automated approach to provide high-performance MT systems for strategic domains. Given the importance of tourism to the Croatian economy (15.4% of its GDP in 2012), we decided to use this domain as our use case for the current milestone.

In this task, we have acquired bilingual Croatian–English (HR–EN) datasets that will be used in the MT systems developed in WP4 (Development and deployment of MT). To this end, we used two open-source crawlers, Bitextor and the ILSP Focused Crawler, which were extended and enhanced during the project (Abu-MaTran Deliverable 3.2, Papavassiliou et al., 2013), in order to crawl a set of 23 websites from the domain of tourism.

Preliminary work towards data acquisition for the tourism domain (Espla-Gomis et al., 2014) showed: i) the usefulness of combining these two crawlers in order to maximise the amount of data harvested, and ii) the need of homogenizing the output of each crawler. Following these conclusions, we used six configurations of Bitextor and ILSP-FC to harvest the 23 websites and

[1] https://github.com/nljubesi/tweetcat

construct six sentence-aligned corpora. Then, we designed a post-processing workflow to ensure the comparability of the data obtained by both crawlers. This process yielded a translation memory comprising 139,938 aligned segments. It is worth mentioning that the delivered corpus contains useful information about each aligned segment including the web page it was acquired from, the crawl configuration and the alignment confidence. A detailed description of the entire corpus acquisition workflow is included in Section 3.1 of Toral et al. (submitted).

3.2 Acquisition of Parallel Corpora for Independent News

In this task we have acquired corpora from the domain of independent news. The corpora acquired are both monolingual (for 15 languages) and parallel (for all the combinations of those 15 languages). It should be noted that these are the first publicly available corpora for the domain of independent news, and thus they open the possibility for doing further research on these type of text, e.g. corpus studies and domain-specific machine translation.

Toral (2014) describes this task in further detail.

3.3 Quality Estimation for Parallel Data Generation

Synthetic parallel data are an alternative to bilingual parallel corpora which are usually manually translated or post-edited from automatic translation. To extend our work on pivot-based MT evaluated during Milestone 1 of the project (cf. deliverable D4.1a, Toral et al. 2013b), we make use of the available English–Slovene parallel data and the Slovene to Croatian rule-based MT system (Peradin et al. 2014) based on Apertium (Forcada et al., 2011), using the Slovene language as a pivot to translate from English to Croatian.

However, the low quality of the English–Croatian parallel corpus generated remains an issue. In Rubino et al. (2014) we describe a new approach to filter synthetic parallel sentences based on MT quality estimation (QE). Estimating the quality of MT output automatically is the ability to judge the correctness of a translation without any translation reference. These experiments are to be considered as a first step towards the generation of reliable synthetic parallel data for under-resourced languages. The results of this study show significant improvement in terms of automatic metrics obtained on two test sets using our approach compared to a random selection of synthetic parallel data.

4 Dictionaries

Part of our research is focused on rule-based machine translation systems (RBMT). These systems have proven to be a good 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. The objective of our research is to develop methods which enable non-expert users to improve a RBMT system. In this section, we describe the method proposed to help a non-expert user to add new words to the morphological dictionaries used in RBMT.

Our approach is described by Esplà-Gomis et al. (2014b). We assume an scenario in which a user is using the machine translation system to translate a text, but one word is not recognised by the system and, therefore, it remains untranslated. In this case, the user may want to add this word and its translation to the morphological dictionaries of the RBMT system. To do so, it is necessary to determine the stem of the word as well as its inflection paradigm. Our approach helps the user in this task in two steps:

- First, all those stem–paradigm pairs fitting the unknown word form proposed are detected.

- Then, the user is asked about several inflected word forms resulting from applying the different inflection paradigms to the unknown word. In this way, the user only has to validate a small number of word forms proposed by the system to find the best fitting paradigm.

The first step of this process is carried out by using a suffix tree which contains all the suffixes generated by each paradigm in the dictionary, as well as the information about which paradigms generate them. In this way, the stem/paradigm candidates can be efficiently retrieved by just checking which suffixes fit the word form to be added. For choosing the words to be shown to the user, an ID3 decision tree is used. The objective of this approach is to obtain an optimal solution, that is, one that shows the fewest number of words to be validated by the user. This tree is built in such a way that in each node a binary decision is taken, which splits the set of categories thus maximising the entropy. In our case, the decision to be made in each node corresponds to the validation of a given word form. If the word form is validated, those pairs stem/paradigm which do not produce it are discarded. Conversely, if the word form is rejected, those pairs generating it are discarded. The word forms are proposed to the user following the structure of the tree until only one stem/paradigm candidate is accepted. In a first approximation, all the pairs stem/paradigm are equiprobable, but given that the word to be added appears in a context, we also tried an approach using a hidden Markov model (HMM) to obtain probabilities for the different paradigm candidates. This second approach resulted in a reduction of the number of word forms that the user had to validate.

5 Rules

RBMT systems need translation rules, in addition to dictionaries, in order to properly carry out the translation task. Since rules encode the information needed in order to deal with the grammatical divergences between languages, they can only be developed by expert linguists. In order to enable the use of RBMT when the cost and slow development cycles of the development of the rules by trained linguists cannot be afforded, we have developed a new approach with which to automatically learn shallow-transfer MT rules (like those used by the Apertium RBMT platform, Forcada et at., 2011) from very small parallel corpora: barely a few hundreds of sentences.

Our new strategy is able to achieve a high degree of generalisation over the linguistic phenomena observed in the training corpus thanks to the fact that it is the first approach in which conflicts between rules are resolved by choosing the most appropriate ones according to a global minimisation function rather than proceeding in a pairwise greedy fashion. In addition, our approach is able to select the proper subset of rules which ensure the most appropriate segmentation of the input sentences to be translated with them. Experiments conducted using five different language pairs with the free/open-source rule-based MT platform Apertium show that translation quality significantly improves when compared to previous approaches (that have less generalisation power and produce rules that cause a deficient segmentation), and is close to that obtained using handcrafted rules. Moreover, the resulting number of rules is considerably smaller, which eases human revision and maintenance. A detailed description of this rule inference approach is given by Sánchez-Cartagena et al. (2014b).

The new rule inference algorithm can also be used outside RBMT. In particular, it can be applied, together with existing dictionaries, to build a hybrid MT system that consists of an SMT system whose translation model is enriched with the existing dictionaries and a set of transfer rules inferred from the training parallel corpus. The resulting hybrid system is thus able to generalise the translation knowledge contained in the parallel corpus to sequences of words that have not been observed in it, as long as they share lexical category and/or morphological inflection features with the sequences of words present in the corpus. A system built according to these principles took part in the WMT 2014 shared translation task (Sánchez-Cartagena et al., 2014). Although the hybrid system did not obtain a statistically significant improvement over an SMT baseline, the factors that

prevent the achievement of a higher translation quality have been identified (mainly, the poor quality of the dictionaries), and they will be overcome in forthcoming editions of the WMT shared translation task.

6 Conclusions

This deliverable has covered the work done in the area of acquisition (work package WP3) during the period of the second milestone of the project (M7–M24). We have worked on the acquisition of two types of resources (corpora and linguistic data), which can be loosely identified with the different MT paradigms that will make use of them (statistical and rule-based, respectively).

Regarding corpora, we have covered the acquisition of both monolingual and parallel corpora. For monolingual corpora we have proposed methodologies to crawl generic and user-generated data, and used them to acquire corpora for Croatian and other related languages. With respect to parallel corpora, we have improved two crawlers (Bitextor and the ILSP Focused Crawler) to crawl domain-specific parallel corpora. We have then use these crawlers to acquired a parallel corpus for Croatian–English in the domain of tourism. Still concerning parallel corpora, we have acquired the first set of corpora in the literature for the domain of independent news.

Moving on to linguistic resources, we have worked on the acquisition of dictionaries and rules. For dictionaries, we have proposed a methodology that allow non-expert users to improve the coverage of the dictionaries used by rule-based MT systems. As regards rules, on the one hand, a new approach that learns translation rules from very small, morphologically analysed parallel corpora, through the global minimization of a loss function, and shows a performance which is very close to that obtained with handcrafted rules. On the other hand, these rules may be used, together with dictionaries, to extend the phrase table of a statistical MT system; results are still preliminary and do not show statistically significant improvements over the baseline.

Bibliography

Toral, A., Cortés-Vaíllo, S., Ramírez-Sánchez, G., Forcada, M.L., Ljubešić, N. (2013a) "Abu-MaTran Deliverable D3.1a: Acquisition for the first development cycle", version 1.0, available from http://www.abumatran.eu/?page_id=59

Toral, A., Cortés-Vaíllo, S., Ortiz-Rojas, S., Ramírez-Sánchez, G., Forcada, M.L. (2013b) "Abu-MaTran Deliverable D4.1a: MT systems for the first development cycle", version 1.0, available from http://www.abumatran.eu/?page_id=59

Toral, A. (2014) "TLAXCALA: A Multilingual Corpus of Independent News". In Proceedings of the 9th Language Resources and Evaluation Conference (LREC), pp. 3689–3692.

Toral, A., Esplà-Gomis, M., Ljubešić, N., Papavassiliou, V., Prokopidis, P., Rubino, R., and Way, A. (submitted) "Crawl and Crowd to Bring Machine Translation to Under-resourced Languages". Submitted to the Language Resources and Evaluation journal.

Esplà-Gomis, M., Klubička, F., Ljubešić, N., Ortiz-Rojas, S., Papavassiliou, V., Prokopidis, P. (2014) "Comparing Two Acquisition Systems for Automatically Building an English-Croatian Parallel Corpus from Multilingual Websites". In Proceedings of the 9th Language Resources and Evaluation Conference (LREC), pp. 1252–1258.

Esplà-Gomis, M., Sánchez-Cartagena, V.M., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Forcada, M.L., Carrasco, R.C. (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 16th Annual Conference of the European Association for Machine Translation (EAMT), pp. 19–26.

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. (2011) "Apertium: a free/open-source platform for rule-based machine translation", *Machine Translation* 25:2, 127-144

Koehn, P. (2005) "Europarl: A parallel corpus for statistical machine translation", Proceedings of Machine Translation Summit X, pp. 79–86.

Ljubešić, N., Klubička, F. (2014), "{bs,hr,sr}WaC — Web Corpora of Bosnian, Croatian and Serbian". In Proceedings of the 9th Workshop Web as Corpus (WaC).

Ljubešić, N., Fišer, D., Erjavec, T. (2014), "TweetCaT: A Tool for Building Twitter Corpora of Smaller Languages". In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), pp. 2279–2283.

Ljubešić, N., Toral, A. (2014), "caWaC – A Web Corpus of Catalan and its Application to Language Modeling and Machine Translation". In Proceedings of the 9th Language Resources and Evaluation Conference (LREC), pp. 1728–1732.

Lui, M., Baldwin, T. (2012), "langid.py: An Off-the-shelf Language Identification Tool", In Proceedings of The 50th Annual Meeting of the Associaton for Computational Linguistics, proceedings of System Demonstrations, pp. 25-30

Papavassiliou, V., Prokopidis, P., Thurmair, G. (2013), "A Modular Open-source Focused Crawler for Mining Monolingual and Bilingual Corpora from the Web". In Proceedings of the 6th Workshop on Building and Using Comparable Corpora (BUCC), pp. 43–51.

Papavassiliou, V., Prokopidis, P., Esplà-Gomis, M., Ortiz, S. (2014). "Abu-MaTran Deliverable 3.2: Corpora Acquisition Software", version 1.0, available from http://www.abumatran.eu/?page_id=59

Peradin, H., Petkovsky, F., Tyers, F.M. (2014) "Shallow-transfer rule-based machine translation for the Western group of South Slavic languages", in Proceedings of the 9th SaLTMiL workshop on Free/open-Source Language Resources for the Machine Translation of Less-Resourced Languages at the 9th Language Resources and Evaluation Conference (LREC), pp. 25–30.

Rubino, R., Toral, A., Ljubešić, N., Ramírez-Sánchez. G. (2014), "Quality Estimation for Synthetic Parallel Data Generation". In Proceedings of the 9th Language Resources and Evaluation Conference (LREC), pp. 1843–1849.

Sánchez-Cartagena, V.M., Pérez-Ortiz, J.A., Sánchez-Martínez, F. (2014) "The UA-Prompsit hybrid machine translation system for the 2014 Workshop on Statistical Machine Translation". In Proceedings of the 9th Workshop on Statistical Machine Translation (WMT).

Sánchez-Cartagena, V.M., Pérez-Ortiz, J.A. Sánchez-Martínez, F. (2014) "A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora". *Computer Speech & Language* (Special Issue on Hybrid Machine Translation). Article in press. DOI: http://dx.doi.org/10.1016/j.csl.2014.10.003