Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain*

Antonio Toral†, Raphael Rubino⋆, Miquel Esplà-Gomis‡, Tommi Pirinen†, Andy Way†, Gema Ramírez-Sánchez⋆

† NCLT, School of Computing, Dublin City University, Ireland

{atoral,tpirinen,away}@computing.dcu.ie

⋆ Prompsit Language Engineering, S.L., Elche, Spain

{rrubino,gramirez}@prompsit.com

‡ Dep. Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain

mespla@dlsi.ua.es

Abstract

We present an extrinsic evaluation of crawlers of parallel corpora from multilingual web sites in machine translation (MT). Our case study is on Croatian to English translation in the tourism domain. Given two crawlers, we build phrase-based statistical MT systems on the datasets produced by each crawler using different settings. We also combine the best datasets produced by each crawler (union and intersection) to build additional MT systems. Finally we combine the best of the previous systems (union) with general-domain data. This last system outperforms all the previous systems built on crawled data as well as two baselines (a system built on general-domain data and a well known online MT system).

1 Introduction

Along with the addition of new member states to the European Union (EU), the commitment with multilingualism in the EU is strengthened to give support to new languages. This is the case of Croatia, the last member to join the EU in July 2013, and of the Croatian language, which became then an official language of the EU.

Croatian is the third official South Slavic language in the EU along with Bulgarian and Slovene. Other surrounding languages (e.g. Serbian and Bosnian), although still not official in the EU, belong also to the same language family and are the official languages of candidate member states, thus being also of strategic interest for the EU.

We focus on providing machine translation (MT) support for Croatian and other South Slavic languages using and producing publicly available resources. Following our objectives, we developed a general-domain MT system for Croatian–English and made it available online on the day Croatia joined the EU. It is, to the best of our knowledge, the first available MT system for this language pair based on free/open-source technologies.

New languages in the EU like Croatian can benefit from MT to speed up the flow of information from and into other EU languages. While this is the case for most types of content it is especially true for official documentation and for content in particular strategic sectors.

Tourism is one of the most important economic sectors in Croatia. It represented 15.4% of Croatia's gross domestic product in 2012 (up from 14.4% in 2011).[1] With almost 12 million foreign tourists visiting Croatia annually, the tourism sector results in income of 6.8 billion euro.

The increasing number of tourists in Croatia makes tourism a relevant domain for MT in order to provide them with quick and up-to-date information about the country they are visiting. Although most visitors come from non-English speaking countries,[2] English is frequently used as a lingua franca. This observation led us to our first approach to support the Croatian tourism sec-

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.eubusiness.com/news-eu/croatia-economy.nrl

[2] According to the site croatia.eu, top emitting countries are Germany (24.2%), Slovenia (10.8%), Austria (8.9%), Italy (7.9%), Czech Republic (7.9%), etc.

tor: to provide MT adapted to the tourism domain from Croatian into English. Later, we will provide MT in the visitors' native languages, i.e. German, Slovene, etc.

We take advantage of a recent work that crawled parallel data for Croatian–English in the tourism domain (Esplà-Gomis et al., 2014). Several datasets were acquired by using two systems for crawling parallel data with a number of settings. In this paper we assess these datasets by building MT systems on them and checking the resulting translation performance. Hence, this work can be considered as an extrinsic evaluation of these crawlers (and their settings) in MT.

Besides building MT systems upon the domain-specific crawled data, we study the concurrent exploitation of domain-specific and general-domain data, with the aim of improving the overall performance and coverage of the system. From this perspective, our case study falls in the area of domain adaptation of MT, following previous works in domains such as labour legislation and natural environment for English–French and English–Greek (Pecina et al., 2012) and automotive for German to Italian and French (Läubli et al., 2013).

The rest of the paper is organised as follows. Section 2 presents the crawled datasets used in this study and details the processing undertaken to prepare them for MT. Section 3 details the different MT systems built. Section 4 shows and comments the results obtained. Finally, Section 5 draws conclusions and outlines future lines of work.

2 Crawled Datasets

Datasets were crawled using two crawlers: ILSP Focused Crawler (FC) (Papavassiliou et al., 2013) and Bitextor (Esplà-Gomis et al., 2010). The detection of parallel documents was carried out with two settings for each crawler: 10best and 1best for Bitextor and reliable and all for FC (see (Esplà-Gomis et al., 2014) for further details). It is worth mentioning that reliable and 1best are subsets of all and 10best, respectively. These subsets were obtained with a more strict configuration of each crawler and, therefore, are expected to contain higher quality parallel text. In addition, a set of parallel segments was obtained by aligning only those pairs of documents which were checked manually by two native speakers of Croatian.

Both Bitextor and FC segment the documents aligned by using the HTML tags. These segments were re-segmented in shorter segments and tokenised with the sentence splitter and tokeniser included in the Moses toolkit.[3]

The resulting segments were then aligned with Hunalign (Varga et al., 2005), using the option `realign`, which provides a higher quality alignment by aligning the output of the first alignment. The documents from each website were concatenated prior to aligning them using tags (`<p>`) to mark document boundaries. Aligning multiple documents at once allows Hunalign to build a larger dictionary for alignment while ensuring that only segments belonging to the same document pair are aligned to each other. The resulting pairs of segments were filtered to remove those with a confidence score lower than 0.4.[4]

From the aligned segments coming from manually checked document pairs we remove duplicate segments. We only keep pairs of segments with confidence score higher than 1.[5] These segments are randomised and we keep two sets, one of 825 segmens for the development set and one of 816 segments for the test set.

From the other 4 datasets, those obtained with the different settings of the two crawlers (1best, 10best, all and reliable), duplicate pairs of segments were also removed. Pairs of segments appearing either in the test or development set were also removed. The remaining pairs of segments are kept and will be used for training MT systems.

Apart from the domain-specific crawled data we use additional general-domain (gen) data gathered from several sources of Croatian–English parallel data: hrenWaC,[6] SETimes[7] and TED Talks.[8] These three datasets are concatenated and will be used to build a baseline MT system.

Table 1 presents statistics (number of sentence pairs, number of tokens and number of unique tokens in source (Croatian) and target (English) language) of the previously introduced parallel datasets for Croatian–English. The table shows

[3] `https://github.com/moses-smt/mosesdecoder`

[4] Manual evaluation for English, French and Greek concluded that 0.4 was an adequate threshold for Hunalign's confidence score (Pecina et al., 2012).

[5] While segment pairs with score above 0.4, as shown above, are deemed to be of reasonable quality for training, we raise the threshold to 1 for test and development data.

[6] `http://nlp.ffzg.hr/resources/corpora/hrenwac/`

[7] `http://nlp.ffzg.hr/resources/corpora/setimes/`

[8] `http://zeljko.agic.me/resources/`

Dataset	# s. pairs	# tokens	# uniq t.
dev	825	30,851	10,119
		34,558	7,588
test	816	28,098	9,585
		31,541	7,366
gen	387,259	8,084,110	288,531
		9,015,757	149,430
1best	27,761	592,236	80,958
		680,067	46,671
10best	34,815	760,884	86,391
		864,326	52,660
reliable	23,225	613,804	71,657
		706,227	37,399
all	27,154	719,526	77,291
		819,353	40,095
union	52,097	1,243,142	103,671
		1,418,950	60,956
intersection	5,939	131,569	28,761
		155,432	16,290

Table 1: Statistics of the parallel datasets. For each dataset the first line corresponds to statistics for Croatian and the second to English.

two additional datasets: union and intersection. These are the union and intersection of datasets 10best and reliable.

3 Machine Translation Systems

Phrase-based statistical MT (PB-SMT) systems are built with Moses 2.1 (Koehn et al., 2007). Tuning is carried out on the development set with minimum error rate training (Och, 2003).

All the MT systems use an English language model (LM) from our system for French→English at the WMT-2014 translation shared task (Rubino et al., 2014).[9] We built individual LMs on each dataset provided at WMT-2014 and then interpolated them on a development set of the news domain (news2012).

Most systems are built on a single dataset, hence they have one phrase table and one reordering table. These systems include a baseline built on the general-domain data (gen), four systems built on the crawled datasets (1best, 10best, reliable and all) and two systems built on the union and intersection of the best performing[10] dataset of each crawler: 10best and reliable.

There is also one system (gen+u) built on two datasets, the general-domain (gen) dataset and a domain-specific dataset (union). Phrase tables from the individual systems gen and union are interpolated so that the perplexity on the development set is minimised (Sennrich, 2012).

[9] http://www.statmt.org/wmt14/translation-task.html

[10] According to the BLEU score on the development set.

System	BLEU	METEOR	TER	OOV
gen	0.4092	0.3005	0.5601	9.5
google	0.4382	0.2947	0.5295	-
1best	0.5304	0.3478	0.4848	7.6
10best	0.5176	0.3436	0.5016	7.2
reliable	0.4064	0.2945	0.5755	12.6
all	0.4105	0.2927	0.5756	12.4
union	0.5448	0.3583	0.4726	6.3
inters.	0.3224	0.2456	0.6582	23.1
gen+u	0.5722	0.3767	0.4451	4.1

Table 2: SMT results.

4 Results

The MT systems are evaluated with a set of state-of-the-art evaluation metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Lavie and Denkowski, 2009). For each system we also report the percentage of out-of-vocabulary (OOV) tokens.

Table 2 shows the scores obtained by each MT system. We compare our systems to two baselines: a PB-SMT system built on general-domain data (gen) and an on-line MT system, Google Translate[11] (google).

Systems built solely on in-domain data outperform the baselines (1best and 10best) or obtain similar results (reliable and all). Different crawling parameters of the same crawler (10best vs 1best and reliable vs all) do not seem to have much of an impact. In fact, while the scores by 1best are slightly better than scores by 10best, the latter scored slightly better on the development set (and thus it is used in system union).

The union of data crawled by both Bitextor (10best) and FC (reliable) achieves a further improvement over the top performing system built on data by a single crawler (BLEU 0.5448 vs 0.5304). The system built on the intersection is the least performing system (BLEU 0.3224) but it should be noted that this system is built on a very small amount of data (5,939 sentence pairs, cf. Table 1).

Finally a system built on the interpolation of the systems union and gen obtains the best performance, beating all the other systems for all metrics. In the interpolation procedure system union was weighted around 85% and system gen around 15%. Hence, the data provided by the union of the crawlers, although considerably smaller than the general-domain data (52,097 vs 387,259 sentence pairs), is considered more valuable for translating the domain-specific development set.

[11] http://translate.google.com/

5 Conclusions and Future Work

We have presented an extrinsic evaluation of parallel crawlers in MT. Our case study is on Croatian to English translation in the tourism domain.

Given two crawlers, we have built PB-SMT systems on the datasets produced by each crawler using different settings. We have then combined the best datasets produced by each crawler (both intersection and union) and built additional MT systems. Finally we have combined the best of the previous systems (union) with general-domain data. This last system outperforms all the previous systems built on crawled data as well as two baselines (a PB-SMT system built on general-domain data and a well known on-line MT system).

As future work we plan to build MT systems for other relevant languages. As German, Slovene and Italian account for over 50% of incoming tourists in Croatia, we consider of strategic interest to build systems that translate from Croatian into these languages. Even more as it seems that on-line MT systems covering these pairs do not perform the translation directly but use English as a pivot.

Croatian–Slovene is a pair of closely-related languages, already covered by Apertium.[12] We plan to perform domain adaptation on tourism of this rule-based MT system following previous work in this area (Masselot et al., 2010). For the remaining languages (German and Italian), we plan to build SMT systems with crawled data following the approach presented in this paper.

References

Esplà-Gomis, Miquel, Felipe Sánchez-Martínez, and Mikel L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. *The Prague Bulletin of Mathemathical Linguistics*, 93:77–86.

Esplà-Gomis, Miquel, Filip Klubička, Nikola Ljubešić, Sergio Ortiz-Rojas, Vassilis Papavassiliou, and Prokopis Prokopidis. 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*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

[12] `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-hbs-slv/`

Läubli, Samuel, Mark Fishel, Manuela Weibel, and Martin Volk. 2013. Statistical machine translation for automobile marketing texts. In *Machine Translation Summit XIV: main conference proceedings*, pages 265–272.

Lavie, Alon and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Masselot, François, Petra Ribiczey, and Gema Ramírez-Sánchez. 2010. Using the apertium spanish-brazilian portuguese machine translation system for localisation. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation.*

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167.

Papavassiliou, Vassilis, Prokopis Prokopidis, and Gregor 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.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Pecina, Pavel, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, and Josef van Genabith. 2012. Domain adaptation of statistical machine translation using web-crawled resources: a case study. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 145–152.

Rubino, Raphael, Antonio Toral, Victor M. Sánchez-Cartagena, Jorge Ferrández-Tordera, Sergio Ortiz-Rojas, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Andy Way. 2014. Abu-MaTran at WMT 2014 Translation Task: Two-step Data Selection and RBMT-Style Synthetic Rules. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*.

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP*, pages 590–596.