Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 361–367

Collaborative development of a rule-based machine translator between Croatian and Serbian

Filip KLUBIČKA[1], Gema RAMÍREZ-SÁNCHEZ[2], Nikola LJUBEŠIĆ[1,3]

[1] University of Zagreb, Ivana Lučića 3, HR-10000 Zagreb, Croatia
[2] Prompsit Language Engineering, Avenida de la Universidad s/n, ES-03202 Elche, Spain
[3] Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

fklubick@ffzg.hr, gramirez@prompsit.com, nljubesi@ffzg.hr

Abstract. This paper describes the development and current state of a bidirectional Croatian-Serbian machine translation system based on the open-source Apertium platform. It has been created inside the Abu-MaTran project with the aims of creating free linguistic resources as well as having non-experts and experts work together. We describe the collaborative way of collecting the necessary data to build our system, which outperforms other available systems.

Keywords: machine translation, collaboration, Apertium, open-source, Croatian, Serbian

1 Introduction

Croatian and Serbian are language varieties and official registers of the pluricentric Bosnian-Croatian-Montenegrin-Serbian (BCMS) language. Although mutually intelligible, the national varieties are standardised differently, and both communities have a high interest to produce documentation that adheres to these standards, if for no other reason, then for the sake of producing standard documents for Serbian, the official language of an EU candidate state. Thus it is sensible to make use of a related language of a recent member state and employ machine translation between these two language varieties to meet this aim.

Creating machine translation (MT) systems for South-Slavic languages, both between themselves and other languages, is also the aim of the Abu-MaTran project.[4] In the first phase of the project, the focus was on MT between English and Croatian, while MT between South-Slavic is the focus of the second phase. The system presented in this paper will be used within the project to increase the amount of English - Serbian parallel data by translating the Croatian side of English-Croatian parallel data to Serbian. It will also be added to another by-product of the Abu-MaTran project - AltLang - a service for translating between language varieties.[5]

[4] http://abumatran.eu
[5] http://www.altlang.net

2 Related work

Forcada et al. (2011) and their open-source Apertium platform have shown that, when doing machine translation between language variants or closely-related languages like Spanish and Catalan, a rule-based shallow transfer approach is often sufficient to produce good quality translations. Indeed, work has already been done in building rule based translators from BCMS into Macedonian and Slovene (Peradin et al, 2014). To our knowledge, however, no similar work has been done for the Croatian-Serbian language pair specifically. The only accessible state of the art system for this pair is *Google Translate*,[6] which reaches a BLEU score of 82.27 in the Serbian-Croatian direction. However, the statistical approach that Google uses, which has also been explored in (Popović et al, 2014) but only using small corpora, is not a feasible option for us, as there are not enough parallel corpora available to train SMT systems that can deal with the minute differences between the two languages without introducing additional noise.

Nonetheless, some free linguistic resources were initially available to us: the HBS monolingual dictionary[7] built for other Apertium language pairs like HBS-Macedonian (Peradin and Tyers, 2012) and HBS-Slovene (Peradin et al, 2014), the SETimes news corpora of both Croatian and Serbian[8] and the hrWaC and srWaC web corpora (Ljubešić and Klubička, 2014). This is always an advantage, as both monolingual and bilingual corpora are extensively used to semiautomatically extract knowledge for Apertium such as frequent non-covered entries, bilingual correspondences, rules, development and test sets, and data needed to train statistical part-of-speech taggers.

Considering the amount of available data, coupled with the fact that differences between Croatian and Serbian occur mostly at the level of orthography and lexicon, with only a bit of syntax (limited only to specific structures and verbal tenses), a rule-based approach makes the most sense. We expect a high quality and a more controlled output from such a system, reproducing other Apertium-based success stories such as the Norwegian Nynorsk-Norwegian Bokmål (Unhammer et al, 2006) or Spanish and Aragonese (Cortés et al, 2012) language pairs.

3 Apertium language pair

The structure of the Croatian-Serbian language pair is based on the same structure shared by other Apertium language pairs. This esentially includes two monolingual dictionaries (source and target) which are used as morphological anlysers/generators, one set of morphological tags for the part-of-speech tagger (currently shared by the two languages involved), and two sets of structural transfer rules (one for each translation direction). However, because there is significant overlap in the lexemes of the languages, instead of two separate monolingual dictionaries, there is only one. In addition to pairing lemmas with inflectional paradigms, this monolingual dictionary - called the metadix - explicitly encodes differences between the three language varieties (Bosnian, Croatian, Serbian) with regards to variant-specific lexemes and the reflex of the vowel

[6] http://translate.google.com
[7] HBS is the ISO 639-3 code for the macrolanguage covering the three languages in question
[8] http://nlp.ffzg.hr/resources/corpora/setimes/

yat.[9] Furthermore, the language pair includes a bilingual dictionary which explicates lexical differences as one-to-one translations, a shared Hidden Markov Model (HMM) tagger, a transfer module for each translation direction and a transliterator for the cyrilic and latin alphabets.

The basic system on which we started making improvements was produced in only a couple of weeks by extracting relevant components from existing language pairs. In other words, we took the dictionaries from HBS-Slovene, the tagger from the HBS module, a bilingual dictionary created from monolingual entries and the transfer rules for agreement between basic noun phrases. Additionally, the work presented in this paper also kicked off the efforts to enrich the HBS monolingual dictionary, which ran in parallel with our construction of the Croatian-Serbian language pair, and resulted in Apertium's largest lexicon to date, with 97,437 lemmas (Ljubešić et al, 2016).

4 Development

Even though there is considerable overlap, the biggest source of differences between Croatian and Serbian is still the differing lexicon. Thus it was important to construct a large, high-coverage bilingual dictionary. Additionally, transfer rules needed to be defined to account for the few syntactic differences between the languages. Each of these tasks was tackled in two phases - at hands-on Abu-MaTran workshops held in Zagreb and within a course held during the winter semester of 2015/2016 at the University of Zagreb, titled *Selected chapters in Natural Language Processing*.[10]

The approach to including non-experts in the process consisted of creating very focused tasks for data which is needed for each of the Apertium modules based on materials created beforehand, e.g. in the form of precomputed bilingual entries that they had to assess. When possible, user-friendly interfaces or very simple spreadsheets were used to lower the technical barrier. After each task, the contributors were able to see the impact of their collaborative work in the translator's performance almost real-time, which proved to be very motivating. While larger groups could work on dictionary entries (as this is an easy task), only a reduced group worked on writing transfer rules (as this requires an advanced level of technical knowledge).

4.1 Adding bilingual entries

First phase: The first workshop was focused on monolingual and bilingual dictionaries.[11] We automatically produced bilingual candidates from comparable corpora - hrWaC and srWaC (Ljubešić and Klubička, 2014) - by identifying lexemes from the Serbian corpus that, given their frequency in the Croatian corpus, were occurring much more frequently than by chance. The workshop participants validated the candidates and

[9] For example, the following lexical entry extracted from the metadix produces either the surface form 'pjevačica' or 'pevačica', depending on the chosen language variant: <e lm="pjevačica">p<par n="e_je__yat"/>vačic<par n="vodnic/a__n"/></e>

[10] Within this course, students were taught about machine translation and the Apertium framework, among other things.

[11] Materials available at http://www.abumatran.eu/?p=292

added additional linguistic information, such as pointing out parts of speech, morphological differences and translation direction details.[12] This workshop resulted in the addition of approximately 485 new entries in a single day. These entries were additionally checked by experts later on.

Second phase: During a one-semester course, our students collected bilingual data and produced many new entries for the bilingual dictionary using several methods, ranging from running texts of their choice through our translator and filling the bilingual dictionary with the untranslated lexemes, to validating and adding bilingual candidates extracted by using the output of a distributional similarity tool (Fišer and Ljubešić, 2011) applied to texts from the Croatian and Serbian Wikipedia. By the end of the course, the dictionary contained 1694 bilingual entries, which is also its current size.

4.2 Adding rules

First phase: Our second workshop focused on transfer rules[13] from Serbian to Croatian. We automatically extracted rules for Serbian to Croatian (Sánchez et al., 2015) and our workshop participants validated them based on actual examples of these rules, answering the simple question "Is this a valid translation?"[14] We taught them how to formalise the rules and presented them with 100 rules to be validated. The implementation of the rules was done by experts after the workshop. In 1 week we implemented 25 new Serbian to Croatian rules and 10 basic Croatian to Serbian rules.

Second phase: Nearing the end of the course, after adding sufficient bilingual entries, the students were taught about shallow transfer rules. They once again looked into the outputs of the texts they ran through the translator and annotated the syntactic mistakes occurring in the translations. Some of the rules that could be fixed via shallow transfer were added during the course for demonstration purposes, but most were formalised and implemeted as a result of the joint work between a language expert and an Apertium expert during a secondment at Prompsit Language Engineering. At the end of this stage, the number of Serbian to Croatian rules was extended to 99 rules, and Croatian to Serbian to 82, which is the current state of the system. Most of the rules implemented cover a bit of syntax via short-distance word shifting (e.g. there are several verbal constructions involving the *da* particle which differ between the languages in regards to word order and whether the *da* particle is present or not)[15], as well as agreement rules (e.g. if the head noun of a noun phrase changes gender in translation, the premodifying adjectives need to change gender as well).[16]

---

[12] Participants would point out whether the translation of a given lexeme is bidirectional (like *direktorica-direktorka*), just from Croatian to Serbian (like *zabava-žurka*), or just from Serbian to Croatian (like *kasnije-docnije*)

[13] Materials available at http://www.abumatran.eu/?p=418

[14] [SR] Zemlje jugoistočne Evrope trebale bi da suraduju
[HR] Zemlje jugoistočne Europe trebale bi suradivati

[15] [SR] da li možeš
[HR] možeš li

[16] [SR] naš brzi računar (masculine)
[HR] naše brzo računalo (neuter)

4.3 Tagger training

Additional insight gained during the workshops and coursework was that the HBS tagger was in serious need of improvement. The tagger we had at the beginning of the described process was using a constraint grammar, and it was producing many errors, which very palpably hindered the translation process. Fortunately, by the time the bilingual lexicon and transfer rules were extended, we had the newly created hr500k Croatian training corpus (Ljubešić et al, 2016) at our disposal, so we decided to train a statistical tagger[17] based on Hidden Markov Models (Rabiner, 1989) with the tools provided in Apertium. A necessary preprocessing step was to transfer the tags in the training corpus from the MULTEXT-East Morphosyntactic Specifications, revised Version 4[18] to Apertium's notation. This was done by automatically mapping the hr500k training corpus to the Apertium tagset, retaining only sentences with full coverage and splitting this dataset into training and test data. This left us with 145,626 tokens (9,465 sentences) of training data and 7,682 tokens (500 sentences) of test data.

Additionally, a tagset file with ambiguity classes was defined so as to narrow down the tagset as much as possible. This step makes learning the morphological disambiguation process feasible as the amount of training data that would be necessary to observe all the possible sequences of full tags, given the rich morphology of the languages, is many orders higher than the amount of data currently available.

We performed a comparative intrinsic evaluation of both the constraint grammar and statistical tagger on the 500-sentence test dataset. We evaluated both taggers via token-level accuracy. In this setting, the improvement in accuracy was quite substantial: while the old constraint grammar-based tagger had an accuracy of 76%, the new HMM tagger achieved an accuracy of 90.19%.

5 Evaluation

Finally, we perform a comparative evaluation of our system, but we present an evaluation of only the Serbian to Croatian direction as this direction was the initial focus of the development and the other direction was still under development at the moment of presenting these results. We compare our system to the output of *Google Translate*,[19] as this is the current state of the art system. For our baseline we assume that the output is identical to the input, a setup which yields the lowest evaluation scores. Our SMT baseline was constructed by training a phrase-based Moses system on 200k segments from the SETimes parallel corpus, with an additional 2 thousand segments of development data, while we use hrWaC2.0 for building the language model (Ljubešić and Klubička, 2014).

For the evaluation we use a test set consisting of 351 Serbian sentences gathered from newspaper texts that were manually translated into Croatian by students. We evaluate the system with BLEU (Papineni et al, 2006) and TER (Snover et al, 2006). Table 1 shows the results of the evaluation process.

[17] It should be noted that even though using the same tagger for both Croatian and Serbian is not ideal, previous experiments (Agić et al., 2013) have shown that only a minor drop in accuracy should be expected from this setting.

[18] https://github.com/ffnlp/sethr/blob/master/mte4r-upos.mapping

[19] Output retrieved on 2016-01-27

	BLEU	TER
baseline	72.66	0.1300
SMT	73.54	0.1255
Google	82.27	0.0873
Apertium	82.97	**0.0782**

Table 1. Results of the MT evaluation. Statistically significantly better results are in bold.

When compared to our baseline systems, the evaluation scores are decidedly positive. When compared to Google's system, we also improve, but the question is whether this improvement is statistically significant. To calculate this we use approximate randomisation with 1000 iterations, and while the reported 0.7 point improvement in BLEU yields a p-value of 0.384, which is too high to prove statistical significance, the improvement in TER by -0.0091 is in fact statistically significant, with a p-value of 0.018. Given that BLEU is known to favour statistical machine translation in its evaluation, it is safe to claim that our system outperforms that of Google.

6 Conclusion

In this paper we present a bidirectional machine translation system between Croatian and Serbian, which was collaboratively developed between the University of Zagreb and Prompsit Language Engineering in the framework of the Abu-MaTran project. To achieve this, we combine Apertium's resources with the University of Zagreb's manpower and resources, taking advantage of our researcher's employment and secondments, as well as hands-on workshops organised as part of our Abu-MaTran activities to get other interested parties to help with the creation of additional necessary linguistic resources.

The result of this work is a system that has been developed in a total of approximately 6 person months (including experiments for semi-automatic extraction of vocabulary and data, work in dictionaries, HMM and implementation of rules, workshop and course materials, training of non-experts and evaluation) and which outperforms the current state of the art. The contribution of this work for the wider community is the release of numerous freely available linguistic tools and resources, as well as the considerable transfer of knowledge between all participating institutions. Additionally, this system opens up the possibility of smoothing the way towards translating official EU documents that are and will be published in Croatian[20] into Serbian, the language of an EU candidate state.

Future work will go into extending the system and further evaluating both translation directions, creating combinations with Bosnian, using it to create synthetic training data, and adding it to AltLang to offer a commercial service that uses the current system to customise content to a specific language variant.

[20] E.g. the *acquis communautaire*; EU parallel corpora and translation memories such as DGT (https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory); the Special Edition of the EU Official Journal (http://eur-lex.europa.eu/eu-enlargement/hr/special.html)

Acknowledgements

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) and the Swiss National Science Foundation grant IZ74Z0_160501 (ReLDI).

References

Agić, Ž., Ljubešić, N., Merkler, D. (2013). Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the Fourth Biennial International Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria.

Cortés Martínez, J.P., O'Regan, J., Tyers, F.M. (2012). Free/Open Source Shallow-Transfer Based Machine Translation for Spanish and Aragonese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Fišer, D., Ljubešić, N. (2011). Bilingual Lexicon Extraction from Comparable Corpora for Closely Related Languages. In *Proceedings of the Recent Advances in Natural Langugage Processing Conference*, Hissar, Bulgaria.

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

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

Ljubešić, N., Klubička, F., Agić, Ž., Jazbec, I. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania.

Peradin, H., Tyers, F. (2012). *A rule-based machine translation system from Serbo-Croatian to Macedonian*. Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012).

Peradin, H., Petkovski, F., Tyers, F. (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*, Reykjavik, Iceland.

Popović, M., Ljubešić N. Exploring cross-language statistical machine translation for closely related South Slavic languages. *Language Technology for Closely Related Languages and Language Variants (LT4CloseLang)*, Doha, Qatar.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*

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

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*

Unhammer, K., Trosterud, T. (2009). Reuse of free resources in machine translation between Nynorsk and Bokmål. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, Alicante, Spain.

Received May 2, 2016 , accepted May 18, 2016