Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 354–360

Dealing with Data Sparseness in SMT with Factored Models and Morphological Expansion: a Case Study on Croatian

Víctor M. SÁNCHEZ-CARTAGENA[1], Nikola LJUBEŠIĆ[2,3], Filip KLUBIČKA[3]

[1] Prompsit Language Engineering, Av. Universitat, s/n. Edifici Quorum III, E-03202, Elx, Spain
[2] Dept. of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 32, SI-1000, Ljubljana, Slovenia
[3] Dept. of Information and Communication Sciences, University of Zagreb

vmsanchez@prompsit.com, nljubesi@ffzg.hr, fklubicka@gmail.com

Abstract. This paper describes our experience using available linguistic resources for Croatian in order to address data sparseness when building an English-to-Croatian general domain phrase-based statistical machine translation system. We report the results obtained with factored translation models and morphological expansion, highlight the impact of the algorithm used for tagging the corpora, and show that the improvement brought by these methods is compatible with the application of data selection on out-of-domain parallel corpora.

Keywords: data sparseness, factored translation models, morphological expansion

1 Introduction

Data sparseness is a well-known problem that phrase-based statistical machine translation (SMT) systems suffer from when dealing with highly inflected languages, especially when the highly inflected one is the target language (TL). In these kinds of languages, a single word (lemma) can have dozens of different inflected forms. Translation performance is hampered because it is difficult to observe all the forms of a given word (in the different contexts relevant for translation) in the training corpus.

This paper is part of the Abu-MaTran project, where we aim to provide machine translation support to Croatian, as the official language of a new EU country. Croatian is a highly inflected language and hence it is affected by data sparseness. For instance, adjectives inflect for 3 genders, 2 numbers and 7 cases and the *hrLex* Croatian inflectional lexicon (Ljubešić et al, 2016b) contains 939 unique morphosyntactic description tags. In this paper, we show how we dealt with that problem in a general-domain English-to-Croatian phrase-based SMT system by leveraging a Croatian inflectional lexicon and adapting existing solutions in the literature, namely factored translation models

(Koehn and Hoang, 2007) and morphological expansion (Turchi and Ehrmann, 2011). Sections 2 and 3 respectively describe these solutions, while Section 4 shows that they can be successfully combined with a data selection strategy. The paper ends with a brief description of related approaches and some concluding remarks and future directions.

2 Factored translation models

Factored translation models (Koehn and Hoang, 2007) split the translation of words in the translation of different factors (surface forms, lemmas, lexical categories, morphosyntactic information, etc.). Among the different ways these factors can be combined, we opted for producing a surface form factor and a morphosyntactic description (MSD) factor for each word in the output, and used two different language models (LMs), one operating on surface forms and another one on MSDs. This setup has been reported to be effective (and efficient in terms of decoding time) when the TL is highly inflected but the SL is not (Skadinš et al., 2010), since it helps the decoder to produce grammatically correct phrases that have not been observed in the training corpus. We considered the following aspects when building our factored phrase-based SMT system.

- Order of the MSD LM. The order of surface-form based LMs is usually set to 5. As the number of different MSDs is several orders of magnitude lower than the number of different surface forms, a greater order can also be considered.
- Corpora tagging algorithm. In order to obtain the MSD factor of the TL side of the parallel corpus and the TL monolingual corpus, a part-of-speech (PoS) tagger is needed. We tested the effect of lexicon constraining in PoS tagging.

2.1 Constrained part of speech tagging

The best performing tagger available for Croatian is a CRF-based tagger (Ljubešić et al, 2016b). While the tagger makes use of the *hrLex* inflectional lexicon (Ljubešić et al, 2016b), the corresponding lexicon entries are not used for constraining the tagger to the potential tags, but just as features. As a result, the number of translations for each SL phrase in a factored system grows when compared to a non-factored phrase-based SMT system: in the system setups described in Section 2.2, the average number of translations per SL phrase in the phrase table of our non-factored system is 2.091. This value grows to 2.119 in the factored system that uses the CRF tagger. Moreover, we observed that the increase in the number of translation options caused by unconstrained tagging is more relevant in frequent SL phrases. For instance, the Croatian surface form *kuća* (*house*) can only be analysed as a feminine, singular, nominative, common noun or as a feminine, plural, genitive, common noun according to the lexicon. However, in our factored system with unconstrained tagging (Section 2.2), we can find 8 phrase table entries whose SL word is *house* and whose TL surface form factor is *kuća*. Additional MSDs include cardinal number, adjective, or proper noun. Since according to some studies (Ling et al., 2012), the presence of redundant phrase translations can hurt translation quality, we also tested a modified version of the tagger in which it only selects the MSDs present in the lexicon.[4] As a result, the average number of translations per SL phrase was 2.111.

[4] We post-processed the output of the CRF tagger: for each word tagged with an MSD not present in the lexicon, we replaced it with the most likely one from the lexicon according to a 3-gram

We compared the original and the constrained tagger on the test set traditionally used to evaluate taggers on Croatian: 300 sentences (6306 tokens) from news, general web and Wikipedia domains (Agić and Ljubešić, 2014): constraining the tagger slightly reduces accuracy from 0.9253 to 0.9232.

2.2 Experiments and results

We built our phrase-based SMT system from corpora crawled from the web: we consider them to be the most suitable ones for an open-domain system. In particular, we used *hrenWaC* as parallel corpus (Ljubešić et al., 2016a) and *hrWaC* (Ljubešić and Klubička, 2014) as TL monolingual corpus. The parallel corpus contains 1 166 732 sentences, 32 908 281 English words and 29 199 856 Croatian words. The size of the vocabularies is 605 929 (English) and 888 405 (Croatian): the ratio between them is 1.47, which gives us an idea of the morphological richness of Croatian as compared with English. The monolingual corpus contains 67 403 231 sentences and 1 404 303 868 words. We used Moses[5] with the MIRA tuning algorithm (Watanabe et al., 2007). We estimated a 5-gram surface-form LM from the TL monolingual corpus. Our factored system contains an additional MSD LM estimated from the same monolingual corpus. We experimented with orders 3, 5 and 7 and the two tagging alternatives discussed in the previous section. We used KenLM and Knesser-Ney discounting.[6]

Table 1. Results of the evaluation of factored models. A score in bold means that the system outperforms the plain baseline by a statistically significant margin according to paired bootstrap resampling (Koehn, 2004) ($p = 0.05$, 1 000 iterations).

MSD LM order	constrained tagging	BLEU	TER	METEOR
baseline	-	0.2356	0.6351	0.2119
3	N	**0.2429**	**0.6296**	**0.2152**
5	N	0.2408	0.6327	**0.2142**
7	N	0.2373	0.6352	0.2125
3	Y	**0.2458**	**0.6226**	**0.2167**
5	Y	**0.2432**	**0.6256**	**0.2161**
7	Y	0.2413	**0.6280**	**0.2154**

We tuned the systems with *newstest2012* and evaluated them with *newstest2013*, as Pirinen et al. (2016) did. Table 1 shows the values of the BLEU, TER and METEOR evaluation metrics for the basic phrase-based SMT system and the different factored alternatives. Results show that constraining the tagging brings a consistent improvement (for all evaluation metrics and MSD LM orders), thus confirming the observations by

LM of MSDs estimated from the same annotated corpus from which the CRF tagger was trained (Ljubešić et al, 2016b). Words not found in the lexicon were assigned the special tag `UNK` (this was not done when evaluating tagging accuracy in order to perform a fair comparison with the unconstrained tagger).

[5] http://www.statmt.org/moses/. Corpora were normalized, tokenized and truecased with the tools provided with Moses. Parallel sentences with more than 80 tokens were removed.

[6] https://kheafield.com/code/kenlm/

Ling et al. (2012). Concerning the order of the MSD LM, the best translation quality is obtained for order 3. We observed that, as we increase the value of the order, short-distance agreement deteriorates, but long-distance agreement does not improve. A couple of examples can be found in Table 2. This is probably caused by the fact that the order of the constituents of the sentence is relatively free in Croatian. Thus, it is difficult to predict MSDs in the TL with n-grams that cross constituent boundaries.

Table 2. Example sentences illustrating the difference in local agreement between different orders of the MSD LM. In the first example, the phrase *obične smrt* should be in the genitive case, but it is nominative in the order 7 alternative; in the second example, the adjective *egipatske* should be in neuter gender in order to agree with *društva*, but it is feminine in the order 7 alternative.

	Example 1	Example 2
source	The courage of ordinary death.	[...] respect for the other elements of Egyptian society.
order 3	hrabrost **obične smrti**.	[...] poštovanje za druge elemente **egipatskog društva**.
order 7	hrabrost **obična smrt**.	[...] poštovanja za druge elemente **egipatske društva**.
ref.	hrabrost **obične smrti**.	[...] poštovanja ostalih elemenata **egipatskog društva**.

3 Morphological expansion

Factored systems with an additional MSD LM cannot produce surface forms in Croatian that have not been observed in the training corpus. In order to further mitigate the data sparseness problem, we enhanced our system with morphological expansion. It consists of creating new phrase table entries from existing ones by means of changing values of morphological inflection attributes and inflecting words accordingly. We followed a strategy[7] inspired by Turchi and Ehrmann (2011) but we restricted the process with linguistically motivated rules so as to avoid the need to optimise filtering thresholds. We created new phrase pairs by changing only the TL side of existing phrase pairs, and only for those phrase pairs whose TL side is a single word or a grammatically meaningful phrase. We select, among others, TL phrases that contain a noun, a noun phrase, an adjective, a verb, etc. Then, we generate new phrases with all the possible values of the morphological inflection features not present in English.[8] We added the generated phrase pairs to a new phrase table which is combined with the original one at decoding time by means of independent decoding paths (Koehn and Schroeder, 2007).

We added morphological expansion to the best factored system described in Section 2.2 and repeated the evaluation. In view of the positive results of reducing translation alternatives by constraining PoS tagging, we also evaluated an alternative morphological expansion strategy in which only noun phrases were expanded. Our expansion rules generate only 3 alternatives for them (for nominative, accusative and instrumental cases), while the number of generated entries is higher for verbal phrases and adjectives. Results displayed in Table 3 show that there is not a clear difference between both expansion

[7] Implementation is available at: https://github.com/vitaka/morph-xpand-smt

[8] The file with the expansion rules used and some comments can be found at https://github.com/vitaka/morph-xpand-smt/blob/master/tags_29-1-2016-somecases.

strategies, and that morphological expansion is not able to outperform the factored system (the difference is not statistically significant). We performed a manual analysis on the agreement errors found in 40 sentences randomly selected from the test set and found that for only 25% of all the agreement errors proper word forms were not present in the phrase table, out of which only 23% were generated by morphological expansion. Most needed words were not generated because they were not present in the lexicon.

Table 3. Results of the evaluation of morphological expansion.

System	BLEU	TER	METEOR
best factored	0.2458	0.6226	0.2167
+ morph. expansion noun phrases	0.2470	0.6232	0.2174
+ morph. expansion all	0.2460	0.6235	0.2179

4 Combination with data selection

A different way of dealing with data sparseness is to augment the training corpus by selecting the most suitable sentences from out-of-domain data. Pirinen et al. (2016; Section 2.6) followed that strategy and obtained a significant improvement over a plain phrase-based SMT system trained on *hrenWaC* (Table 4 shows the result of evaluating their approach; the evaluation setup defined in Section 2.2 was followed). In order to make the most of the available resources for English–Croatian we combined both approaches: we built a system with factored translation models (following the best setup in Section 2.2) on the parallel corpora obtained as a result of data selection. Results, which are also depicted in Table 4, confirm that both approaches can be successfully combined, allowing us to reach the state-of-the-art system, *Google Translate*.[9]

Table 4. Results of the evaluation of the combination of data selection and the best factored setup. There are not statistically significant differences between that combination and *Google Translate*.

System	BLEU	TER	METEOR
hrenWaC + factored	0.2458	0.6226	0.2167
data selection	0.2576	0.6060	0.2264
Google Translate	0.2673	0.5946	0.2321
data selection + factored	0.2700	0.5963	0.2338

5 Related work

Factored models have been deeply studied by Tamchyna and Bojar (2013), who concluded that automatically searching for the best factored model architecture in a given

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

language pair is not feasible. Successful application of factored models to different language pairs has been already reported by other authors, like Bojar (2007) and Koehn et al. (2010). Regarding morphological expansion, to the best of our knowledge, the approach by Turchi and Ehrmann (2011) is the only one that addresses the expansion of the TL side of the phrase table. Concerning other ways of adding linguistic information to an SMT system, we refer the reader to the survey by Costa-Jussà and Farrús (2014).

6 Conclusions and future work

In this paper, we presented a set of strategies on how to leverage existing Croatian linguistic resources to address data sparseness in a general-domain English-to-Croatian SMT system. Applying factored models showed to be successful. We observed that accuracy of PoS tagging and translation performance are not correlated and that increasing the order of the MSD LM is counterproductive. A combination of factored models and data selection allowed us to build a system that reaches state-of-the-art commercial tools. Improvement obtained with morphological expansion was negligible.

Since most of the agreement errors were not caused by lack of inflected forms in the phrase table, and the best results were obtained with a low-order MSD LM because of the free constituent order in Croatian, hybridisation with an RBMT system that performs full syntactic analysis (Labaka et al., 2014) could further improve the results.

Acknowledgements

Research funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

References

Agić, Ž., Ljubešić, N. (2014). The SETimes.HR Linguistically Annotated Corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Ondřej Bojar (2007). English-to-Czech factored machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Costa-Jussà, M. R., Farrús, M. (2014). Statistical machine translation enhancements through linguistic levels: A survey. *ACM Computing Surveys* 46:3.

Koehn. P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Koehn, P., Haddow, B., Williams, P., Hoang, H. (2010). More Linguistic Annotation for Statistical Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden.

Koehn, P., H. Hoang (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* , Prague, Czech Republic.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Labaka, G., España-Bonet, C., Màrquez, L., Sarasola, K. (2014). A hybrid machine translation architecture guided by syntax. *Machine Translation*, 28(2):91–125.

Ling, W., Graça, J., Trancoso, I., Black, A. (2012). Entropy-based Pruning for Phrase-based Machine Translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea.

Ljubešić, N., Esplà-Gomis, M., Toral, A., Klubička, F. (2016). Producing Monolingual Web Corpora and Bitext at the Same Time - SpiderLing and Bitextor's Love Affair. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

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. 29–35.

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.

Pirinen, T., Rubino, R., Sánchez-Cartagena, V.M., Klubička, F., Toral, A. (2016). *D5.1c Evaluation of the MT systems deployed in the third development cycle*. Automatic building of Machine Translation. FP7-PEOPLE-2012-IAPP project deliverable (http://www.abumatran.eu/?page_id=59).

Skadinš, R., Goba, K., Šics, V. (2010). Improving SMT for Baltic Languages with Factored Models. In *Proceedings of the Fourth International Conference Baltic HLT, Frontiers in Artificial Intelligence and Applications*, Riga, Latvia.

Tamchyna, A., Bojar, O. (2013). No Free Lunch in Factored Phrase-Based Machine Translation. In *Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, Samos, Greece.

Turchi, M., Ehrmann, M. (2011). Knowledge Expansion of a Statistical Machine Translation System using Morphological Resources. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics*, Tokyo, Japan.

Watanabe, T., Suzuki, J., Tsukada, H., Isozaki, H. (2007). Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic.

Received May 2, 2016 , accepted May 16, 2016