Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 368–375

Re-assessing the Impact of SMT Techniques with Human Evaluation: a Case Study on English↔Croatian

Antonio TORAL[1], Raphael RUBINO[2], Gema RAMÍREZ-SÁNCHEZ[3]

[1] ADAPT Centre, School of Computing, Dublin City University, Ireland
[2] Universität des Saarlandes, 66123 Saarbrücken, Germany
[3] Prompsit Language Engineering, Avenida de la Universidad s/n, ES-03202 Elche, Spain

atoral@computing.dcu.ie, raphael.rubino@uni-saarland.de, gramirez@prompsit.com

Abstract. We re-assess the impact brought by a set of widely-used SMT models and techniques by means of human evaluation. These include different types of development sets (crowdsourced vs translated professionally), reordering, operation sequence and bilingual neural language models as well as common approaches to data selection and combination. In some cases our results corroborate previous findings found in the literature, when those approaches were evaluated in terms of automatic metrics, but in some other cases they do not.

Keywords: human evaluation, operation sequence model, bilingual neural language model, data selection

1 Introduction

In the field of statistical machine translation (SMT), when new models and techniques are introduced, it is rather common to assess their performance in terms of automatic metrics solely for just one or a few language pairs. Upon showing significant improvement and being implemented as free/open-source software, some of these techniques become then widely used in the community. In this paper we select a relevant set of such techniques and evaluate the impact they bring by means of a human evaluation.

This paper is part of a wider activity whose goal is to rapidly provide machine translation (MT) for under-resourced languages along with a better insight on how do they work and perform in a way that is meaningful not just for researchers but also for industrial adopters of MT. Our case study is on Croatian given its strategic importance to the EU as the official language of a recent member state. In this work we aim to assess the impact of the components of several SMT systems (English–Croatian in both directions) built for this purpose. To meet this aim, we evaluate, both automatically and manually, each component one at a time. Namely:

1. We assess the impact of using different development sets, produced by professional and amateur translators.
2. Compare the use of three reordering models (word-, phrase-based and hierarchical).
3. Measure the impact of using additional models recently introduced in the SMT pipeline. Specifically, the operation sequence model (OSM) and bilingual neural language models (BiNLM).
4. Assess the impact of different ways to select and combine data sets.
5. Compare our best systems to widely-used commercial systems.

2 Experimental Setting

2.1 MT Systems

SMT systems are trained with Moses 3.0,[4] using default settings unless mentioned otherwise, and tuned with MIRA (Cherry and Foster, 2012). Language models are of order 5 with Kneser-Ney modified smoothing.

The following publicly available parallel corpora are used for training: HrEnWaC 2.0,[5] the DGT Translation Memory,[6] the JRC Acquis,[7] SETIMES(Agić and Ljubešić, 2014), TED talks,[8] OpenSubtitles 2013 cleaned (Esplà-Gomis et al., 2014) and SrEnWaC.[9] The last one is a parallel corpus for Serbian–English. The Serbian side is translated to Croatian with a rule-based system in order to get more English-Croatian parallel text.[10] Language models are trained on the hrWaC corpus[11] for Croatian, and on all the available English data for the translation task at WMT15.[12]

The development set consists of multiple translations into Croatian of the first 1,011 sentences from the English side of the WMT2012 test set. Namely, we have 1 translation produced by a professional translator and 2 by amateur translators. The test set consists of the first 1,000 sentences from the English side of the WMT2013 test set, translated into Croatian by a native speaker.

2.2 Evaluation

Each experiment is evaluated both automatically and manually (except the one in Section 3.4). We used the widely used automatic metrics BLEU (Papineni et al., 2002) and

[4] https://github.com/moses-smt/mosesdecoder/tree/RELEASE-3.0
[5] http://hdl.handle.net/11356/1058
[6] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory
[7] http://tinyurl.com/CroatianAcquis
[8] http://nlp.ffzg.hr/resources/corpora/ted-talks/
[9] http://hdl.handle.net/11356/1059
[10] https://svn.code.sf.net/p/apertium/svn/staging/apertium-hbs_HR-hbs_SR/
[11] http://nlp.ffzg.hr/resources/corpora/hrwac/
[12] http://www.statmt.org/wmt15/translation-task.html

TER (Snover et al., 2006). Statistical significance is calculated on BLEU scores with paired bootstrap resampling (1,000 iterations and $p = 0.95$).

The human evaluation consists of ranking MT outputs with Appraise.[13] For each experiment 100 randomly selected segments were ranked. All the annotations were carried out by 2 native Croatian speakers with an advanced level of English. The following guidelines were provided to the annotators:

```
Given translations by more than two MT systems, the task is to rank them:
- Rank system A higher (rank1) than B (rank2), if the output of the first
is better than the output of the second.
- Rank both systems equally, A rank1 and B rank1, if the outputs are of
the same quality
- Use the highest rank possible, e.g. if you've three systems A, B and C,
and the quality of A and B is equivalent and both are better than C, then
do: A=rank1, B=rank1, C=rank2. Do NOT use lower rankings, e.g.: A=rank2,
B=rank2, C=rank4.
```

We then derive a human score for each system with the TrueSkill method adapted to MT evaluation (Sakaguchi et al., 2014) following its usage at WMT15.[14] Namely, we run 1,000 iterations of rankings followed by clustering ($p = 0.95$). If two systems are placed in different clusters (column "range" in results' tables) then the one with lower range is considered significantly better.

3 Experiments

3.1 Development Sets

In this experiment we aim to assess the impact of using a development set obtained by professional versus amateur translators. While professional translations should lead to a higher quality data set, its cost is in our case close to an order of magnitude (both in terms of price and time) higher than crowdsourcing. In this sense, Zbib et al. (2013) built MT systems for Arabic–English using development sets that were professionally translated and crowdsourced. They compared tuning with one reference (either professional or crowdsourced) and using both together as multiple references. The latter setup led to the best results. In our experiments we aim to corroborate these results for English-to-Croatian[15] and also compare the use of professional and crowdsourced translations for tuning. Results are shown in Table 1.

Following Zbib et al. (2013), we would expect the combination of both professional and crowdsourced translations to perform better than professional ones, which in turn would be expected to perform better than crowdsourced translations. An interesting question would be whether two crowdsourced translations would perform better than

[13] https://github.com/cfedermann/Appraise

[14] https://github.com/mjpost/wmt15

[15] This experiment is run only into Croatian as it is for this language that we have multiple references for the development set.

Table 1. Results (different development sets) for English→Croatian.

System	English→Croatian			
	BLEU	**TER**	**Human**	**Range**
crowd1	0.2362	0.6250	0.110	1-4
crowd2	0.2344	**0.6242**	0.027	1-4
crowd1+2	0.2348	0.6287	-0.109	2-5
prof	0.2273	0.6457	**0.225**	1-4
prof+crowd1+2	**0.2363**	0.6303	-0.253	3-5

one professional translation (Zbib et al. (2013) had only 1 translation of each type) as two crowdsourced translations, in our setup, are 5 times cheaper than one professional translation. The results are mixed: there is no clear winner neither for automatic metrics nor for human evaluation. This may indicate that the impact of the development set in our setup is too small to be of importance.

3.2 Reordering Models

We compare using a word-based reordering model solely (the default in the Moses MT toolkit) to adding two additional models: 1) phrase-based (with the same three orientations as the word-based model: monotone, swap and discontinuous) and 2) hierarchical (with four orientations: non merged, discontinuous, left and right). This combination has been shown to yield the best performance in terms of automatic metrics for English–Chinese and English–Arabic (Galley and Manning, 2008). Here we evaluate it for another language and not only automatically but also manually. Results are shown in Table 2.

Table 2. Results using different reordering models. Best results shown in bold.

System	English→Croatian				Croatian→English			
	BLEU	**TER**	**Human**	**Range**	**BLEU**	**TER**	**Human**	**Range**
Word-based	**0.2363**	**0.6303**	-0.117	1-2	0.3392	0.5211	**0.168**	1
Three	0.2355	0.6336	**0.117**	1-2	**0.3404**	**0.5202**	-0.168	2

The results are mixed. In terms of automatic metrics, the differences are very small and not significant. According to the human evaluation, the word-based model alone leads to significantly better results than using three models for Croatian-to-English. Although the usual word order is the same in English and Croatian (subject-verb-object, the rich case structure of Croatian makes it a rather free word order language. This may explain the non-conclusive results. The remaining experiments use three reordering models.

3.3 Additional Components

In this experiment we assess the impact brought by two additional components, OSM (Durrani et al., 2011) and BiNLM (Devlin et al., 2014), used both alone and jointly. Results are shown in Table 3.

Table 3. Results using additional components (OSM and BiNLM). Best results shown in bold. † indicates that a system is significantly better than the other ones ($p = 0.05$).

System	English→Croatian				Croatian→English			
	BLEU	TER	Human	Range	BLEU	TER	Human	Range
None	0.2355	0.6336	-0.330	3-4	0.3404	0.5202	-0.373	3-4
OSM	0.2408	0.6265	0.231	1-3	0.3460	**0.5088**	**0.207**	1-3
BiNLM	0.2379	0.6310	-0.352	3-4	0.3471	0.5138	0.039	1-4
OSM+BiNLM	**0.2457**†	**0.6198**	**0.452**	1-2	**0.3499**	0.5090	0.127	1-4

OSM results in gains over the baseline consistently, but this is not the case for BiNLM (lower human score into Croatian). The joint use of OSM and BiNLM leads to the best BLEU scores for both directions. Human ranges put OSM+BiNLM and OSM alone on top into Croatian while differences are not significant into English.

3.4 Data Selection and Combination

While the previous experiments used only HrEnWaC as training data, here we consider all available parallel corpora[16] (cf. Section 2.1) and experiment with data selection and combination.

For data selection, we concatenate the parallel corpora and rank their parallel sentences according to the bilingual cross-entropy difference heuristic (Moore and Lewis, 2010; Axelrod et al., 2011). This method was shown to reach state-of-the-art performance for domain adaptation in SMT (Rubino et al., 2014; Banerjee et al., 2015). As in-domain language model we consider SETimes and as out-of-domain a random set of sentences from the concatenated dataset of equal size. We split the ranked sentences into two groups, the top 25% (row "top" in Table 4) and the bottom 75% (rows "bottom"). We then experiment with applying the vocabulary saturation filter (Lewis and Eetemadi, 2013) to the bottom data set (rows "vsf"). Specifically, we drop sentences for which all its words have been seen already at least 10 times (Rubino et al., 2014).

As for combining data sets, we consider concatenation (rows "concat") and linear interpolation (rows "tmc") (Sennrich, 2012) of selected datasets (rows "tmc top bottom") and phrase tables built on the individual parallel corpora (row "7pt").

The systems evaluated in this experiment are built on different amounts of parallel data. Therefore, on top of providing evaluation metrics, we detail the size of each system, in terms of number of sentences in the training parallel corpora (column "%

[16] We keep the RMs, OSM and BiNLM trained on HrEnWaC as in this experiment we aim to measure the impact of additional entries in the phrase table.

Sent.")[17] and in terms of number of tokens (measured in millions in column "# token"). Results are shown in Table 4.

Table 4. Results applying data selection and combination. Best results shown in bold.

System	English→Croatian		Croatian→English		% Sent.	# tokens (M)	
	BLEU	TER	BLEU	TER		en	hr
concat	0.2409	0.6178	0.3636	0.4942	100	215.1	176.2
top	0.2487	**0.6115**	0.3679	0.4993	25	81.4	69.7
concat vsf	0.2423	0.6177	0.3640	0.4913	41	117.2	99.2
tmc top bottom	0.2480	0.6120	0.2974	0.5731	100	215.1	176.2
tmc top bottom vsf	**0.2497**	0.6118	0.3646	0.4906	41	117.2	99.2
tmc 7pt	0.2445	0.6147	**0.3721**	**0.4878**	100	215.1	176.2

While most systems outperform the baseline (concatenation of all the corpora), there is no clear winner among them. If one considers a trade-off between translation quality and parallel data size, then systems "top" (trained on 25% of the sentence pairs) and "tmc top bottom vsf" (trained on 41% of the sentence pairs) seem the best choices.

3.5 Comparison to Commercial Systems

Finally, we evaluate our best system[18] against commercial systems available online by Yandex, Microsoft and Google.[19]

Results are shown in Table 5. Google comes on top for both directions, but it is not significantly better than our system (according to the human evaluation), which on its turn is significantly better than both Microsoft and Yandex for English-to-Croatian, and than Yandex only for Croatian-to-English.

Table 5. Results comparing to commercial systems. Best results shown in bold. † indicates that a system is significantly better than the other ones ($p = 0.05$).

System	English→Croatian				Croatian→English			
	BLEU	TER	Human	Range	BLEU	TER	Human	Range
Google	**0.2673**†	**0.5946**	**0.899**	1-2	**0.4099**†	**0.4635**	**0.672**	1-2
Ours	0.2544	0.6081	0.515	1-2	0.3852	0.4819	0.291	1-3
Microsoft	0.2281	0.6263	-0.650	3-4	0.3658	0.5199	0.104	2-3
Yandex	0.2030	0.6801	-0.793	3-4	0.3463	0.5311	-1.306	4

[17] This is measured as the percentage of sentences used (taking 100% as the total used in the concatenation).

[18] I.e. system "top" from the previous experiment with reordering models, OSM and BiNLM trained on the "top" dataset.

[19] https://translate.yandex.com/, http://bing.com/translator/ and https://translate.google.com/, respectively. The translations were obtained on December 22nd, 2015.

4 Conclusions

In this paper we have re-assessed the impact brought to SMT systems by a set of widely-used techniques by means of individual experiments for a not very common language pair in the literature: English–Croatian. Namely, we have performed human evaluations on different types of development sets (crowdsourced vs translated professionally), reordering models, operation sequence and bilingual neural language models as well as common approaches to data selection and combination. In some cases our results have corroborated previous findings found in the literature, when those approaches were evaluated solely in terms of automatic metrics, but in some other cases they did not.

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).

References

Željko Agić and Nikola Ljubešić. The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of LREC*, 2014. ISBN 978-2-9517408-8-4.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*, pages 355–362, 2011.

Pratyush Banerjee, Raphael Rubino, Johann Roturier, and Josef van Genabith. Quality estimation-guided supplementary data selection for domain adaptation of statistical machine translation. *Machine Translation*, 29(2):77–100, 2015.

Colin Cherry and George Foster. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of NAACL*, pages 427–436, 2012. ISBN 978-1-937284-20-6.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of ACL*, pages 1370–1380, 2014.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. A joint sequence translation model with integrated reordering. In *Proceedings of ACL*, pages 1045–1054, 2011.

Miquel Esplà-Gomis, Mikel L. Forcada, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Sergio Ortiz-Rojas, Tommi Pirinen, Raphaël Rubino, and Antonio Toral. Deliverable D3.1b Acquisition for cycle 2. Technical report, Abu-MaTran project, 2014.

Michel Galley and Christopher D Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of EMNLP*, pages 848–856, 2008.

William D Lewis and Sauleh Eetemadi. Dramatically reducing training data size through vocabulary saturation. In *Proceedings of WMT*, pages 281–291, 2013.

Robert C. Moore and William Lewis. Intelligent Selection of Language Model Training Data. In *Proceedings of ACL*, pages 220–224, 2010.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002.

Raphael Rubino, Antonio Toral, Victor M Sánchez-Cartagena, Jorge Ferrández-Tordera, Sergio Ortiz-Rojas, Gema Ramırez-Sánchez, Felipe Sánchez-Martınez, and Andy Way. Abu-MaTran at WMT 2014 translation task: Two-step data selection and RBMT-style synthetic rules. In *Proceedings of WMT*, 2014.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. In *Proceedings of WMT*, pages 1–11, 2014.

Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of ACL*, pages 539–549, 2012.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231, 2006.

Rabih Zbib, Gretchen Markiewicz, Spyros Matsoukas, Richard M. Schwartz, and John Makhoul. Systematic Comparison of Professional and Crowdsourced Reference Translations for Machine Translation. In *Proceedings of NAACL*, pages 612–616, 2013.

Received May 3, 2016 , accepted May 18, 2016