Building English-to-Serbian Machine Translation System for IMDb Movie Reviews

Pintu Lohar, Maja Popović and Andy Way
ADAPT Centre
Dublin City University
Ireland
name.surname@adaptcentre.ie

Abstract

This paper reports the results of the first experiment dealing with the challenges of building a machine translation system for user-generated content involving a complex South Slavic language. We focus on translation of English IMDb user movie reviews into Serbian, in a low-resource scenario. We explore potentials and limits of (i) phrase-based and neural machine translation systems trained on out-of-domain clean parallel data from news articles (ii) creating additional synthetic in-domain parallel corpus by machine-translating the English IMDb corpus into Serbian. Our main findings are that morphology and syntax are better handled by the neural approach than by the phrase-based approach even in this low-resource mismatched domain scenario, however the situation is different for the lexical aspect, especially for person names. This finding also indicates that in general, machine translation of person names into Slavic languages (especially those which require/allow transcription) should be investigated more systematically.

1 Introduction

Social media platforms have become hugely popular web-sites where Internet users can communicate and spread information worldwide. Social media texts, such as user reviews and micro-blogs, are often short, informal, and noisy in terms of linguistic norms. Usually, this noise does not pose problems for human understanding, but it can be challenging for NLP applications such as sentiment analysis or machine translation (MT). Additional challenge for MT is sparseness of bilingual (translated) user-generated texts, especially for neural machine translation (NMT). The NMT approach has emerged in recent years and already replaced statistical phrase-based (PBMT) approach as state-of-the-art. However, NMT is even more sensitive to the low-resource settings and domain mismatch (Koehn and Knowles, 2017). Therefore, the challenge of translating user-generated texts is threefold, and if the target language is complex, then fourfold.

In this work, we focus on neural machine translation of English IMDb movie reviews into Serbian, a morpho-syntactically complex South Slavic language. To the best of our knowledge, this is the first experiment dealing with machine translation of user-generated content involving a South Slavic language. The main questions of our research described in this work are (i) What performance can be expected of an English-to-Serbian machine translation system trained on news articles and applied to movie reviews? (ii) Can this performance be improved by translating the monolingual English movie reviews into Serbian thus creating additional synthetic in-domain bilingual data? (iii) What are the main issues and what are the most important directions for the next experiments?

In order to answer these questions, we build a neural (NMT) machine system on the publicly available clean out-of-domain news corpus, and a phrase-based (PBMT) system trained on the same data in order to compare the two approaches in this specific scenario. After that, we use these two systems to generate synthetic Serbian movie reviews thus creating additional in-domain bilingual data. We then compare five different set-ups in terms of corpus statistics, overall automatic scores, and error analysis.

All our experiments were carried out on publicly available data sets. In order to encourage further research on the topic, all Serbian human translations of IMDb reviews produced for purposes of this research are made publicly available, too[1].

[1] https://github.com/m-popovic/imdb-corpus-for-MT

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 105–113,
Florence, Italy, 2 August 2019.

2 Related Work

A considerable amount of work has been done on social media analysis, mostly on the sentiment analysis of user-generated texts, but many publications deal with different aspects of translation of user-generated content. Some papers investigate translating social media texts in order to map widely available English sentiment labels to a less supported target language and thus be able to perform the sentiment analysis in this language (Balahur and Turchi, 2012, 2014). Several researchers attempted to build parallel corpora for user-generated content in order to facilitate MT. For example, translation of Twitter microblog messages by using a translation-based cross-lingual information retrieval system is applied in (Jehl et al., 2012) on Arabic and English Twitter posts. (Ling et al., 2013) crawled a considerable amount of Chinese-English parallel segments from micro-blogs and released the data publicly. Another publicly available corpus, TweetMT (naki San Vicente et al., 2016), consists of Spanish, Basque, Galician, Catalan and Portuguese tweets and has been created by automatic collection and crowd-sourcing approaches. (Banerjee et al., 2012) investigated domain adaptation and reduction of out-of-vocabulary words for English-to-German and English-to-French translation of web forum content. Estimation of comprehensibility and fidelity of machine-translated user-generated content from English to French is investigated in (Rubino et al., 2013), whereas (Lohar et al., 2017) and (Lohar et al., 2018) explore maintaining sentiment polarity in German-to-English machine translation of Twitter posts.

Whereas South Slavic languages are generally less supported in the NLP, they have been investigated in terms of user-generated content. For example, sentiment classification of Croatian Game reviews and Tweets is investigated in (Rotim and Šnajder, 2017), and (Ljubešić et al., 2017) proposes adapting a standard-text Slovenian POS tagger to tweets, forum posts, and user comments on blog posts and news articles. These languages have been dealt with in machine translation research as well. (Maučec and Brest, 2017) gives an overview of Slavic languages and PBMT, and (Popović and Ljubešić, 2014) explores similarities and differences between Serbian and Croatian in terms of PBMT. Linguistic characteristics of South Slavic languages which are problematic for PBMT were investigated in (Popović and Arčan, 2015), and (Popović, 2018) compares linguistically motivated issues for PBMT with those of the recently emerged NMT.

However, to the best of our knowledge, MT of user-generated texts involving South Slavic languages has not been investigated so far. In this work, we present the first results of translating English IMDb movie reviews into Serbian.

3 Data Sets

We carried out our experiments using the publicly available "Large Movie Review Dataset"[2](Maas et al., 2011) which contains $50,000$ IMDb user movie reviews in English. The data set is mainly intended for sentiment analysis research, so each review is associated with its binary sentiment polarity label "positive" or "negative". Negative reviews have a score $\leq$4 out of 10, positive reviews have a score $\geq$7 out of 10 and the reviews with more neutral ratings are not included. The overall distribution of labels is balanced, namely $25k$ positive and $25k$ negative reviews. In the entire collection, no more than 30 reviews are allowed for any particular movie.

For our experiments, we kept 200 reviews (100 positive and 100 negative) containing about $2,500$ sentences for testing purposes, and used the remaining $49,800$ reviews (about $500k$ sentences) for training. Human translation of the test set into Serbian, which is necessary for fast automatic evaluation of MT outputs, is currently in progress, and at the time of our first experiment described in this work, Serbian reference translations were available for 33 test reviews (17 negative and 16 positive) containing 485 sentences (208 negative and 277 positive).

For the baseline out-of-domain training, we used the South-east European Times (SEtimes) news corpus (Tyers and Alperen, 2010) consisting of about $200k$ parallel sentences from the news articles. In order to be able to compare the results with the in-domain scenario, the development set is extracted from the SETimes corpus, too.

4 Expanding English IMDb Reviews into a Bilingual Training Corpus

The Serbian language is generally not very well supported in terms of NLP resources. The English-Serbian publicly available parallel OPUS

[2]http://ai.stanford.edu/ amaas/data/sentiment/

data[3] consists mostly of subtitles, which are rather noisy. The only really clean parallel corpus there is "SEtimes", which is the reason why we used it for the baseline system in our first experiments – we wanted to avoid any effects of noisy data. To the best of our knowledge, there are no publicly available parallel corpora containing user-generated texts in Serbian.

Therefore, we created synthetic IMDb parallel corpus by translating English IMDb reviews into Serbian using our baseline systems. This technique is shown to be very helpful for NMT systems (Sennrich et al., 2016; Poncelas et al., 2018; Burlot and Yvon, 2018) and has become a common practice in the development of NMT systems. It is usually called "back-translation", because the monolingual in-domain data is normally written in the target language and then translated into the source language. In this way, the synthetic corpus consists of noisy source and clean natural target language texts. In our case, however, we are interested in translating into Serbian but we do not have any movie reviews in Serbian, only in English (the source language). Therefore, we actually applied the "forward-translation" technique, which is also shown to be helpful, albeit less than back-translation (Park et al., 2017; Burlot and Yvon, 2018).

In our case, we expected it to be even more sub-optimal than for some other language pairs, because our target language is more complex than the source language in several aspects. The Serbian language, as other Slavic languages, is morphologically rich and has a rather free word order. Furthermore, unlike other Slavic languages, it is bi-alphabetical (with both Latin and Cyrillic scripts) so attention should be payed in order not to mix the two scripts in one corpus. Another possible inconsistency in corpora is different handling of person names – in Cyrillic, only transcription is possible, whereas in Latin both transcription as well as leaving the original are allowed. Apart from this, all person names are declined, as in other Slavic languages.

Usually, back- and/or forward-translation is performed by an NMT system in order to improve the performance of a baseline NMT system. Recently, a comparison between NMT and PBMT back-translation (Burlot and Yvon, 2018) shown that using a PBMT system for synthetic data can lead to comparable improvement of the baseline NMT system with a lower training cost. Therefore, we decided to use and compare both approaches for improving our baseline NMT system.

[3] http://opus.nlpl.eu/

5 Experimental Set-up

For our experiment, we have built one PBMT English-to-Serbian system using Moses toolkit (Koehn et al., 2007) and four English-to-Serbian NMT models using OpenNMT (Klein et al., 2017) in the following way:

- Train an out-of-domain PBMT system on the SEtimes corpus.
- Train a baseline out-of-domain NMT system on the SEtimes corpus.
- Translate the English IMDb training corpus into Serbian using the PBMT system, thus generating a synthetic parallel corpus IMDb$_{pbmt}$.
- Translate the English IMDb training corpus into Serbian using the baseline NMT system, thus generating a synthetic parallel corpus IMDb$_{nmt}$.
- Train a new NMT system on the SEtimes corpus enriched with the IMDb$_{pbmt}$ corpus.
- Train another NMT system using SEtimes corpus enriched with the IMDb$_{nmt}$ corpus.
- Train one more NMT system using SEtimes corpus enriched with both IMDb$_{pbmt}$ and IMDb$_{nmt}$ corpora (IMDb$_{joint}$).

Table 1 shows the statistics for each of the three training corpora (SEtimes, IMDb$_{pbmt}$ and IMDb$_{nmt}$), for the development set, as well as for the test set. First, it can be noticed that the IMDb training corpus contains more than twice segments and running words than the English part of the SEtimes corpus, and it has a much larger vocabulary. Another fact is that, due to the rich morphology, the Serbian SEtimes vocabulary is almost twice as large as the English one. Nevertheless, this is not the case for the synthetic IMDb data, where the Serbian vocabulary is only barely larger or even comparable to the English one. This confirms the intuition about sub-optimal forward translation mentioned in the previous section – machine translated data generally exhibit less lexical and syntactic variety than natural data (Burlot and Yvon,

train	reviews	segments	words (en)	voc (en)	words (sr)	voc (sr)
SEtimes (natural)	/	224167	4675549	81064	4439280	155447
IMDb (natural)	49800	536433	11313315	223972	/	/
IMDb$_{pbmt}$	49800	536433	/	/	12012734	236272
IMDb$_{nmt}$	49800	536433	/	/	11077566	195912

dev (SEtimes)	/	1000	20338	4757	19244	6806
OOV rate [%]		SEtimes	0.25	5.6	0.48	7.9
		IMDb	1.29	19.9	/	/
		IMDb$_{pbmt}$	/	/	2.21	29.0
		IMDb$_{nmt}$	/	/	2.18	29.0

test (IMDb)	33	485	8530	2548	7630	3220
OOV rate [%]		SEtimes	1.16	17.5	1.83	22.2
		IMDb	0.24	4.2	/	/
		IMDb$_{pbmt}$	/	/	2.39	27.4
		IMDb$_{nmt}$	/	/	2.76	32.3

Table 1: Corpus statistics

2018), and here we are additionally dealing with a scarce out-of-domain MT system translating into a more complex language.

For the development set, as intuitively expected, out-of-vocabulary rates are smaller for the in-domain SEtimes corpus, and for the less morphologically complex English language. As for the test set, the English part behaves in the same way, namely the OOV rates are smaller when compared to the in-domain IMDb training corpus. However, for the synthetic Serbian data, the OOV rates are comparable with those of the out-of-domain development corpus and much higher than for development corpus when compared to its in-domain the SEtimes corpus, which again illustrates the effects of sub-optimal synthetic data.

6 Results

6.1 Overall Automatic Evaluation

We first evaluated all translation outputs using the following overall automatic MT evaluation metrics: BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009), TER (Snover et al., 2006), chrF (Popović, 2015) and characTER (Wang et al., 2016). BLEU, METEOR and TER are word-level metrics whereas chrF and characTER are character-based metrics. BLEU, METEOR and chrF are based on precision and/or recall, whereas TER and characTER are based on edit distance. The results both for the development as well as for the test set can be seen in Table 2.

The results for the development set are as it could intuitively be expected: the best option is to use a NMT system trained on the in-domain data (baseline), and using any kind of additional out-of-domain data deteriorates all scores.

As for the test set, it could be expected that the scores will be worse than for the development set. However, several interesting tendencies can be observed. First of all, the baseline NMT system outperforms the baseline PBMT system despite the scarcity of the training corpus and domain mismatch (Koehn and Knowles, 2017), however only in terms of word-level scores – both character-level scores are better for the PBMT system. Furthermore, adding IMDb$_{pbmt}$ data deteriorates all word-level scores and improves both character-level scores. On the other hand, adding IMDb$_{nmt}$ data improves all baseline scores, but the improvements of the character-based scores are smaller than those yielded by adding the IMDb$_{pbmt}$ corpus. Finally, using all synthetic data IMDb$_{joint}$ improves all scores (except BLEU) over the baseline, however the improvements are smaller than the improvements of each individual synthetic data sets (IMDb$_{nmt}$ for word-level scores and IMDb$_{pbmt}$ for character-level scores).

6.2 Automatic Error Analysis

In order to better understand the character-metrics preference for the PBMT-based systems, we car-

(a) Overall automatic evaluation scores for the development set (SEtimes)

development set (SEtimes)						
system	training corpus	BLEU↑	METEOR↑	TER↓	chrF↑	chrTER↓
PBMT	SEtimes	33.1	29.4	48.9	61.2	41.5
NMT	SEtimes	**39.2**	**32.2**	**42.6**	**62.7**	**39.1**
	SEtimes+IMDb$_{pbmt}$	36.2	30.8	44.7	61.1	41.0
	SEtimes+IMDb$_{nmt}$	38.1	31.7	43.0	61.6	40.1
	SEtimes+IMDb$_{joint}$	35.1	30.2	45.5	59.8	41.9

(b) Overall automatic evaluation scores for the test set (IMDb)

test set (IMDb)						
system	training corpus	BLEU↑	METEOR↑	TER↓	chrF↑	chrTER↓
PBMT	SEtimes	10.8	18.6	69.1	40.5	56.3
NMT	SEtimes	13.7	19.2	65.8	37.4	61.4
	SEtimes+IMDb$_{pbmt}$	11.6	19.0	66.9	**40.7**	**55.3**
	SEtimes+IMDb$_{nmt}$	**14.7**	**20.4**	**63.2**	38.8	60.2
	SEtimes+IMDb$_{joint}$	13.3	19.7	64.8	40.6	55.5

Table 2: Overall word-level and character-level automatic evaluation scores for the development (SEtimes) and the test (IMDb) corpus.

ried out a more detailed evaluation in the form of error classification. Automatic error classification of all translation outputs is performed by the open source tool Hjerson (Popović, 2011). The tool is based on combination of edit distance, precision and recall, and distinguishes five error categories: inflectional error, word order, omission, addition and mistranslation. Following the set-up used for a large evaluation involving many language pairs and translation outputs in order to compare the PBMT and NMT approaches in (Toral and Sánchez-Cartagena, 2017), we group omissions, additions and mistranslations into a unique category called lexical errors. The results for both development and for the test set can be seen in Table 3 in the form of error rates (raw error count normalised over the total number of words in the translation output).

Again, the findings for the in-domain development set could be intuitively expected, and are in line with the findings of (Toral and Sánchez-Cartagena, 2017): the NMT system better handles grammatical features (morphology and word order) than the PBMT system, whereas there is no difference regarding lexical aspect.

The tendencies for the inflectional errors are same for the test set. The lowest inflectional error rate can be observed for the baseline NMT system, and it is slightly increased when the IMDb$_{nmt}$ corpus is added. Other three systems, involving the PBMT approach, exhibit much more inflectional errors. For the other two error categories, the situation is slightly different. Word order is also better for the baseline NMT system than for the PBMT system, however adding the IMDb$_{nmt}$ corpus does not improve it whereas the IMDb$_{pbmt}$ corpus does. Possible reason is the free word order in the Serbian language, so that the system trained on IMDb$_{pbmt}$ data simply generated the word order closest to the one in the reference translation. As for the lexical errors, it can be seen that the lexical error rate is much higher for the baseline NMT system than for the baseline PBMT system, which corresponds to the domain-mismatch challenge for NMT (Koehn and Knowles, 2017). Furthermore, the highest reduction of this error type is achieved when the IMDb$_{pbmt}$ corpus is added.

6.3 Manual Inspection of Lexical Errors

In order to further explore the increase of the lexical errors in systems involving the NMT model, we carried out a qualitative manual inspection of three translation outputs: from the baseline NMT system, from the NMT system with additional IMDb$_{pbmt}$ corpus, and from the NMT system with additional IMDb$_{nmt}$ corpus.

We found out that in general, there are many person names (actors, directors, etc., as well as characters) in the IMDb corpus. As mentioned in Section 4, Serbian (Latin) allows both transcrip-

(a) Error rates (%) for the development set (SEtimes)

development set (SEtimes)				
system	training corpus	inflection	word order	lexical
PBMT	SEtimes	15.4	5.3	**36.1**
NMT	SEtimes	**11.8**	**4.0**	**36.1**
	SEtimes+IMDb$_{pbmt}$	12.5	4.4	37.2
	SEtimes+IMDb$_{nmt}$	11.8	4.1	36.6
	SEtimes+IMDb$_{joint}$	12.6	4.4	38.0

(b) Error rates (%) for the test set (IMDb)

test set (IMDb)				
system	training corpus	inflection	word order	lexical
PBMT	SEtimes	14.2	5.1	54.1
NMT	SEtimes	**10.0**	4.9	60.1
	SEtimes+IMDb$_{pbmt}$	14.4	**4.6**	**53.7**
	SEtimes+IMDb$_{nmt}$	10.4	5.0	57.3
	SEtimes+IMDb$_{joint}$	13.4	4.7	53.8

Table 3: Results of automatic error analysis including three error categories for the development (SEtimes) and test (IMDb) corpus.

tion as well as leaving the original names, but it should be consistent in a text. Whereas in the test reference translation the names were left in the original, neither of the MT systems handled the names in a consistent manner. Both PBMT and NMT-based systems generated originals, transcriptions and sometimes unnecessary translations of the names in a rather random way, and in addition, NMT-based systems often omitted or repeated (the parts of) the names.

This finding could explain both the increase of the lexical error rates as well as decrease of the character-level overall scores for the NMT-based systems. Several examples can be seen in Table 4, and for each example, the best version of the given name is shown in bold. The names on the left were problematic for the baseline NMT system and then improved (albeit not always in the perfect way) by adding the IMDb$_{pbmt}$ corpus, but not improved (or even worsened) by adding the IMDb$_{nmt}$ corpus. The names on the right were treated properly both by the baseline NMT system as well as by the IMDb$_{nmt}$ system, however the IMDb$_{pbmt}$ system transcribed the first name thus making it more distant from the reference, and unnecessarily translated the second name as though it were a common noun.

This finding, together with the facts described in Section 4, indicate that Serbian, as well as other Slavic person names and other name entities should be further investigated in the context of machine translation, not only for movie reviews or other types of user-generated context, but in general.

7 Summary and Outlook

In this work, we focused on the task of building an English-to-Serbian machine translation system for IMDb reviews. We first trained a phrase-based and a neural model on out-of-domain clean parallel data and used it as baselines. We then generated additional synthetic in-domain parallel data by translating the English IMDb reviews into Serbian using the two baseline machine translation systems. This "forward-translation" technique improved the baseline results, although "back-translation" (translating natural Serbian texts into English) would be more helpful. Further analysis shown that morphology and syntax are better handled by the neural approach than by the phrase-based approach, whereas the situation is different for the lexical aspect, especially for person names. This finding also indicates that in general, machine translation of person names into Slavic languages (especially those which require/allow transcription) should be investigated more systematically.

The most important directions for the future work on user-generated texts are finding appropriate Serbian texts (for example, movie review ar-

	IMDb$_{pbmt}$ is better	IMDb$_{nmt}$ is better
source	best Clark Kent	to watch Patrick Duffy
reference	najbolji **Clark Kent**	gledati **Patricka Duffyja**
SEtimes	best Kent	pratiti **Patrick Duffy**
SEtimes+IMDb$_{pbmt}$	najbolji **Klark Kentu**	da gledaju Patrik Dafi
SEtimes+IMDb$_{nmt}$	best Kent Kent	pratiti **Patrick Duffy**
source	the Richard Donner Cut	Kate Winslet (as Rose)
reference	verziju **Richarda Donnera**	Kate Winslet (kao **Rose**)
SEtimes	odlaska Richard Cut	Winslet (kao Jack)
SEtimes+IMDb$_{pbmt}$	**Ričard Donner** smanji	Kate Winslet (kao ruža)
SEtimes+IMDb$_{nmt}$	Richard Cut Cut	Kate Winslet (kao **Rose**)
source	Lester's Superman II	
reference	**Lesterov** Supermen II	
SEtimes	's Superman II	
SEtimes+IMDb$_{pbmt}$	**Lestera** u Superman II	
SEtimes+IMDb$_{nmt}$	's Superman II	
source	scriptwriter Tony Morphett	
reference	scenarista **Tony Morphett**	
SEtimes	scenarista Tony Tony	
SEtimes+IMDb$_{pbmt}$	scenarista **Toni Morphett**	
SEtimes+IMDb$_{nmt}$	scenarista Tony Tony	

Table 4: Examples of different name entities (person names)

ticles in the news) and using them for enlarging the in-domain part of the training corpus by back-translation, as well as enlarging out-of-domain data by cleaning the subtitles corpora, and by back-translating monolingual Serbian news articles. In addition, more IMDb reviews should be evaluated in future experiments. Apart from this, future work should involve other types of user-generated content, such as product or hotel reviews and micro-blog posts, as well as other (South) Slavic languages.

Acknowledgments

This research was supported by the ADAPT Centre for Digital Content Technology at Dublin City University, funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund.

References

Alexandra Balahur and Marco Turchi. 2012. Multilingual Sentiment Analysis using Machine Translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Korea.

Alexandra Balahur and Marco Turchi. 2014. Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. *Computer Speech and Language*, 28(1):56–75.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2012. Domain Adaptation in SMT of User-Generated Forum Content Guided byOOV Word Reduction: Normalization and/or Supplementary Dat. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, pages 169–176, Trento, Italy.

Franck Burlot and François Yvon. 2018. Using Monolingual Data in Neural Machine Translation: a Systematic Study. In *Proceedings of the 3rd Conference on Machine Translation (WMT 2018)*, pages 144–155, Belgium, Brussels.

Laura Jehl, Felix Hieber, and Stefan Riezler. 2012. Twitter Translation Using Translation-based Cross-lingual Retrieval. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT 2012)*, pages 410–421.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for SMT. In *Proceedings of 45th annual meeting of the ACL on interactive poster & demonstration sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver.

Alon Lavie and Michael J. Denkowski. 2009. The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation*, 23(2-3):105–115.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as Parallel Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 176–186, Sofia, Bulgaria.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2017. Adapting a State-of-the-Art Tagger for South Slavic Languages to Non-Standard Text. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, pages 60–68, Valencia, Spain.

Pintu Lohar, Haithem Afli, and Andy Way. 2017. Maintaining Sentiment Polarity of Translated User Generated Content. *The Prague Bulletin of Mathematical Linguistics*, 108(1):73–84.

Pintu Lohar, Haithem Afli, and Andy Way. 2018. Balancing Translation Quality and Sentiment Preservation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 2018)*, pages 81–88, Boston, MA.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics and Human Language Technologies (ACL-HLT 2011)*, pages 142–150, Portland, Oregon, USA.

Mirjam Sepesy Maučec and Janez Brest. 2017. Slavic languages in phrase-based statistical machine translation: a survey. *Artificial Intelligence Review*, 51(1):77–117.

Kishore Papineni, Salim Roukos, Todd Ward, and Wie-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, PA.

Jaehong Park, Jongyoon Song, and Sungroh Yoon. 2017. Building a Neural Machine Translation System Using Only Synthetic Parallel Data. *CoRR*.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating Back translation in Neural Machine Translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, Alicante, Spain.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *Prague Bulletin of Mathematical Linguistics*, 96:59–68.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT 2015)*, pages 392–395, Lisbon, Portugal.

Maja Popović. 2018. Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation*, 32(3):237–253.

Maja Popović and Mihael Arčan. 2015. Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, pages 97–104, Antalya, Turkey.

Maja Popović and Nikola Ljubešić. 2014. Exploring cross-language statistical machine translation for closely related South Slavic languages. In *Proceedings of the EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 76–84, Doha, Qatar.

Leon Rotim and Jan Šnajder. 2017. Comparison of Short-Text Sentiment Analysis Methods for Croatian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, pages 69–75, Valencia, Spain.

Raphael Rubino, Jennifer Foster, Rasoul Samad Zadeh Kaljahi, Johann Roturier, and Fred Hollowood. 2013. Estimating the Quality of Translated User-Generated Content. In *Proceedings of 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 1167–1173, Nagoya, Japan.

Iñaki San Vicente, Iñaki Alegria, Cristina Espa na Bonet, Pablo Gamallo, Hugo Goncalo Oliveira, Eva Martinez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. 2016. TweetMT: A Parallel Microblog Corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 86–96, Berlin, Germany.

Matthew Snover, Bonnie J. Dorr, Richard M. Schwartz, and Linnea Micciulla. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, MA.

Antonio Toral and Víctor Manuel Sánchez-Cartagena. 2017. A Multifaceted Evaluation of Neural versus Statistical Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, Valencia, Spain.

Francis M Tyers and Murat Serdar Alperen. 2010. South-east European Times: A parallel corpus of Balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, Valetta, Malta.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation Edit Rate on Character Level. In *Proceedings of the 1st Conference on Machine Translation (WMT 2016)*, pages 505–510, Berlin, Germany.