Neural Machine Translation for translating into Croatian and Serbian

Maja Popović, Alberto Poncelas, Andy Way
ADAPT Centre, School of Computing
Dublin City University, Ireland
name.surname@adaptcentre.ie

Marija Brkić Bakarić
Department of Informatics
University of Rijeka, Croatia
mbrkic@uniri.hr

Abstract

In this work, we systematically investigate different set-ups for training of neural machine translation (NMT) systems for translation into Croatian and Serbian, two closely related South Slavic languages. We explore English and German as source languages, different sizes and types of training corpora, as well as bilingual and multilingual systems. We also explore translation of English IMDb user movie reviews, a domain/genre where only monolingual data are available.

First, our results confirm that multilingual systems with joint target languages perform better. Furthermore, translation performance from English is much better than from German, partly because German is morphologically more complex and partly because the corpus consists mostly of parallel human translations instead of original text and its human translation. The translation from German should be further investigated systematically.

For translating user reviews, creating synthetic in-domain parallel data through back- and forward-translation and adding them to a small out-of-domain parallel corpus can yield performance comparable with a system trained on a full out-of-domain corpus. However, it is still not clear what is the optimal size of synthetic in-domain data, especially for forward-translated data where the target language is machine translated. More detailed research including manual evaluation and analysis is needed in this direction.

1 Introduction

Whereas South Slavic languages are generally less supported and investigated in natural language processing, they have been explored in the field of machine translation (MT). Nevertheless, a large part of the work deals with the previous state-of-the-art approach, namely phrase-based statistical machine translation (PBSMT) (Popović and Ljubešić, 2014; Toral et al., 2014; Popović and Arčan, 2015; Arčan et al., 2016; Popović et al., 2016; Sánchez-Cartagena et al., 2016; Maučec and Brest, 2017), while much less work can be found about the new state-of-the-art, neural machine translation (NMT) (Lakew et al., 2018; Lohar et al., 2019).

In this work, we focus on NMT into Croatian and Serbian, two very closely related South Slavic languages. We explore two source languages, English and German. The main goals of our research are to explore two source languages, each of them with different sizes and types of training corpora, as well as to test our systems on translating English user reviews, a challenging domain/genre where no parallel training data are available. For these purposes, we train bilingual and multilingual NMT systems on different publicly available parallel training corpora. For translating English user reviews, we also explore different types of synthetic parallel in-domain data (Sennrich et al., 2016a; Zhang and Zong, 2016; Burlot and Yvon, 2018; Poncelas et al., 2018a), which is a widely used practice in NMT. We explored two types of synthetic data: back-translated (BT) and forward-translated (FT). BT data consist of in-domain target language data and their machine translations into English, whereas FT data consist of English data and their machine translations into Serbian and Croatian. All our experiments were carried out on publicly available data sets.

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 102–113
Barcelona, Spain (Online), December 13, 2020

For all our systems, we present three distinct automatic MT evaluation scores, training corpus size, percentage of the particular target language in the training corpus, as well as training time. In order to get a broader picture about the current state of NMT into these two languages, we also present automatic scores for MT outputs[1] of two on-line systems: Google Translate[2] and Amazon Translate.[3]

1.1 Related work

As already mentioned, a large number of publications deal with the PBSMT approach for South Slavic languages. An overview of Slavic languages and PBSMT is given in (Maučec and Brest, 2017). Translating from Croatian into English for tourism domain is presented in (Toral et al., 2014), while factor models for the same language pair were explored in (Sánchez-Cartagena et al., 2016). Similarities and differences between Serbian and Croatian in terms of building PBMT systems were investigated in (Popović and Ljubešić, 2014) for news domain, as well as in (Popović et al., 2016) for the educational domain. Linguistic phenomena posing problems for PBSMT systems between Serbian and Slovenian on one side and English and German on the other side were investigated in (Popović and Arčan, 2015). Publicly available on-line PBSMT systems for translating between English on one side and Croatian, Slovenian and Serbian on another are described in (Arčan et al., 2016).

With the emergence of NMT, several publications have compared its performance with PBSMT. Translation errors from English into Croatian are analysed in (Klubička et al., 2018) and discourse phenomena for the same language pair in (Šoštarić et al., 2018), while (Popović, 2018) explores linguistically motivated issues for the English–Serbian language pair.

Different training set-ups for NMT from English into Serbian and Croatian were first investigated in (Lakew et al., 2018). They explored individual bilingual systems for each language, systems built on merged unlabelled data, as well as multilingual systems built on merged labelled data, namely with a language identifier in each source sentence. In addition, they investigated combinations of unlabelled and labelled data. Their results showed that the multilingual approach with a language identifier is the most promising. However, they carried out the experiments only on a very small TED corpus consisting of about 100k segments for each target language. In our work, we systematically investigate different corpora with sizes in the range of 300k to 40M segments.

NMT for translating English user reviews into Serbian has been addressed in (Lohar et al., 2019). They compared PBSMT and NMT systems trained on out-of-domain data, and then further investigated the NMT system with additional synthetic data. However, they explored only Serbian as a target language, and their baseline system was built on a very small News corpus of 200k segments. In this work, we also include Croatian, and we build the systems on more data, both out-of-domain as well as in-domain.

2 Data sets

2.1 OPUS parallel data

For all our systems, we use the publicly available OPUS[4] parallel data (Tiedemann, 2012). The vast majority of these resources for the desired language combinations consists of *OpenSubtitles*. For English and both target languages, we also used *SETIMES News*, *Bible*, *Tilde*, *EU-bookshop*, *QED*, and *Tatoeba* corpora. In addition, we used *GlobalVoices* for Serbian, and *hrenWac*, *TED* and *Wikimedia* for Croatian. For German, we only used *OpenSubtitles* because other corpora are rather sparse. However, including these corpora might be interesting for future work.

The original parallel data were filtered in order to eliminate noisy parts: too long segments (more than 100 words), segment pairs with disproportional sentence lengths, segments with more than 1/3 of non-alphanumeric characters, as well as duplicate segment pairs were removed. Table 1 shows the number of remaining segments which were used for training and testing the systems. For testing English systems,

[1] The outputs were generated at the beginning of August 2020.

[2] https://translate.google.com/

[3] https://aws.amazon.com/translate/

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

	en-hr		en-sr		de-hr		de-sr	
	# of sentences		# of sentences		# of sentences		# of sentences	
domain	training	test	training	test	training	test	training	test
Subtitles	23 956 612	800	29 445 286	800	9 816 820	1044	10 620 905	1062
News	202 133	200	221 109	200	/	/	/	/
Other	856 594	/	92 357	/	/	/	/	/

Table 1: Statistics of the parallel OPUS data: English–Croatian (en-hr), English–Serbian (en-sr), German–Croatian (de-hr) and German–Serbian (de-sr).

we used 800 segments from *Subtitles* and 200 segments from *News*. For testing German systems, we used only *Subtitles*.

2.2 Movie reviews

Translation from English is also tested on publicly available texts[5] consisting of English user movie reviews from IMDb and their human translations into Serbian and Croatian. This test set, also used in (Lohar et al., 2019), was compiled from the publicly available IMDb corpus[6] created for sentiment analysis. For this test set, we explored the use of synthetic data obtained from the following monolingual in-domain data:

Movies: Croatian and Serbian texts collected from web sites dedicated to movie overviews and cinema programmes. This corpus is very small, consisting of about 100k Serbian segments and 5k Croatian segments.

Selected: Extracted from the mixed Croatian and Serbian web data *hrWac* and *srWac* (Ljubešić and Erjavec, 2011; Ljubešić and Klubička, 2014) using the Feature Decay Algorithm (FDA) (Biçici and Yuret, 2011; Biçici and Yuret, 2015; Poncelas et al., 2018b; Poncelas, 2019). FDA selects sentences from an initial set S based on the number of n-grams which overlap with an in-domain text $Seed$ and adds these sentences to a selected set Sel. In addition, in order to promote a diversity, the n-grams are penalized proportionally to the number of instances present in Sel. During the execution of FDA, candidate sentences from the set S are selected one by one according to the score defined in (1):

$$score(s, Seed, Sel) = \frac{\sum\limits_{ngr \in \{s \bigcap Seed\}} 0.5^{C_{Sel}(ngr)}}{\text{length(s)}} \tag{1}$$

The sentence with the highest score is removed from S and added to Sel. The count of occurrences of n-gram ngr in the selected set Sel, $C_{Sel}(ngr)$, is updated so that in the following iterations this n-gram contributes less to the scoring of one sentence. The process is executed iteratively, adding a single sentence from the set S to the selected set Sel at a time, and it stops after enough sentences have been extracted.

For our experiment, the *hrWac* and *srWac* corpora represented the set S, and the *Movies* data were used as $Seed$. Before applying FDA, the undesirable sentences were removed from S: those containing URLs, those with more than 1/3 of non-alphanumeric characters, duplicates, as well as too long (more than 60 words) and too short (less than 5 words) sentences. In this way, we obtained about 500k selected segments for each target language.

IMDb: English data consisting of about 600k IMDb movie reviews which were not used as a test set.

Amazon: English Amazon movie reviews from the Amazon product review collection[7] (McAuley et al., 2015). We used the first 1M segments (from about 15M in total) for the experiments described in this work. Using more of these data could be an interesting direction for future work.

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

[6] https://ai.stanford.edu/~amaas/data/sentiment/

[7] http://jmcauley.ucsd.edu/data/amazon/

	en→hr			en→sr		
		# of sentences			# of sentences	
corpus	languages	training	test	languages	training	test
Movies	hr-en(BT)	5 357	/	sr-en(BT)	117 768	/
Selected	hr-en(BT)	449 388	/	sr-en(BT)	549 756	/
IMDb	en-hr(FT)	306 874	485	en-sr(FT)	306 874	485
Amazon	en-hr(FT)	500 000	/	en-sr(FT)	500 000	/

Table 2: Statistics of the movie reviews data.

BT and FT synthetic parallel corpora The Serbian and Croatian data *Movies* and *Selected* were translated into English by an NMT system in order to create BT synthetic parallel corpora. The English *IMDb* and *Amazon* data were translated into Serbian and Croatian by an NMT system thus providing FT synthetic parallel corpora. In order to obtain a balanced corpus in terms of the two target languages, we translated half of the *IMDb* (about 300k segments) and half of the *Amazon* (about 500k segments) corpora into Croatian, and the other two halves into Serbian. More details about the NMT systems used for BT and FT can be found in the next section. Detailed statistics for all movie reviews can be seen in Table 2.

3 NMT systems

All our systems are based on the Transformer architecture (Vaswani et al., 2017) and built using the Sockeye implementation (Hieber et al., 2018). The systems operate on sub-word units generated by byte-pair encoding (BPE) (Sennrich et al., 2016b). We set the number of BPE merge operations at 32000 both for the source and for the target language texts. We do not use shared vocabularies between the source (English, German) and the target (Serbian, Croatian) languages because they are distinct. For multilingual systems, on the other hand, we build a joint vocabulary for the two target languages (Serbian and Croatian) because they are very similar. These systems are built using the same technique as (Johnson et al., 2017) and (Aharoni et al., 2019), namely adding a target language label “SR” or “HR” to each source sentence.

All the systems have Transformer architecture with 6 layers for both the encoder and decoder, model size of 512, feed forward size of 2048, and 8 attention heads. For training, we use Adam optimiser (Kingma and Ba, 2015), initial learning rate of 0.0002, and batch size of 4096 (sub)words. Validation perplexity is calculated after every 4000 batches (at so-called “checkpoints”), and if this perplexity does not improve after 20 checkpoints, the training stops. For set-ups with less than two million segments,[8] following the recommendations for low-resource settings in (Sennrich and Zhang, 2019), we changed the following parameters: training stops after 10 checkpoints instead of 20, initial learning rate is 0.0001 instead of 0.0002, and checkpoint interval is 100 batches instead of 4000.

3.1 Systems for translating from English

The following set-ups were investigated for translating from English into Croatian and Serbian:

Clean data only The bilingual (**EN→HR_CLEAN**, **EN→SR_CLEAN**) and multilingual (**EN→HR+SR_CLEAN**) low-resourced systems are trained only on *News* and *Other* data from Table 1. The segments in these texts are mainly properly aligned, so we refer to it as “clean”. It is worth noting that this corpus is rather unbalanced in the terms of target languages, as can be observed from Table 1: there are only about 300k segments for Serbian and about 1M for Croatian.

***Subtitles* only (first 2M segments)** In order to compare bilingual and multilingual systems in a balanced scenario, a small sub-set of *Subtitles* is used. The bilingual systems **EN→HR_SUBS2M** and

[8] This threshold was chosen intuitively. Systematic experiments with different corpus sizes and different parameters should be carried out in order to determine the exact threshold for the low-resourced scenario.

EN→SR_SUBS2M are trained on the first 2M segments for each target language. For the multilingual system **EN→HR+SR_SUBS2M**, the duplicates were removed after joining the corpora so that 3.6M segments remained.

Cleaning: *Subtitles* contain a number of misaligned segments even after the basic filtering. Therefore, a multisource system in the opposite direction[9] "*hr+sr→en_subs2M*" was trained on the same data and was used for removing segment pairs with the negative log-probability larger than 2 from its own training corpus. The threshold of 2 was chosen after a qualitative manual inspection of parallel segments and their log-probabilities. The opposite translation direction was used for cleaning because of different complexity levels of the languages; probabilities provided by an MT system are usually more reliable when translating from more complex languages into less complex ones. After cleaning, the number of parallel segments is reduced to 2.7M and this cleaned corpus is used to train a cleaned multilingual system **EN→HR+SR_CLEANED**.

***Subtitles* (cleaned first 2M segments) + clean data** The multilingual system **EN→HR+SR_CLEAN/ED** is trained on the cleaned 2.7M segments from *Subtitles* merged with the low-resourced clean data, consisting in total of 4.1M segments. This system is used as a baseline for translating English movie reviews.

A multisource system in the opposite direction "*hr+sr→en_clean/ed*" is trained on the same data and is used for cleaning the full *Subtitles* corpus as well as for back-translation of movie reviews.

Full data As mentioned above, in order to reduce noise in the whole *Subtitles* corpus, the multisource "*hr+sr→en_clean/ed*" system was used with the log-probability threshold of 3 (again, established after a manual qualitative inspection of the parallel segments and their log-probabilities). The multilingual system **EN→HR+SR_FULL_CLEAN/ED** is then built on all clean and all cleaned *Subtitles*. For the sake of completeness, bilingual systems **EN→HR_FULL** and **EN→SR_FULL** are trained on all data from Table 1, without any cleaning.

3.2 Systems for translating from German

The following set-ups were investigated for translating from German into Croatian and Serbian:

***Subtitles* (first 2M segments)** The bilingual systems **DE→HR_SUBS2M** and **DE→SR_SUBS2M** are trained on the first 2M segments from *Subtitles* for each target language. For the multilingual system **DE→HR+SR_SUBS2M**, the duplicates were removed so that 3.5M segments remained.

Cleaning: Analogously to translation from English, a multisource system in the opposite direction "*hr+sr→de_subs2M*" was used for removing misaligned segments from its own training corpus. Although German is morphologically more complex than English, it is still less complex than Serbian and Croatian, so we also used the opposite translation direction for cleaning. After cleaning, the number of segments is reduced to 2.3M, and the cleaned multilingual system **DE→HR+SR_CLEANED** is trained on this corpus.

Multisource German+English: Because the corpora including German are smaller than those including English, we wanted to check the potential of joining German and English corpora in order to build a multisource system. As the first step, we wanted to test it on a more or less balanced corpus in terms of source languages. For this purpose, we built a multilingual system **DE+EN→HR+SR_CLEAN/ED** by merging the cleaned German *Subtitles* with the English clean data and cleaned (first 2M) *Subtitles*.

Full *Subtitles* Again, the multisource "*sr+hr→de_clean*" system was used with the translation score threshold of 3. The multilingual system **DE→HR+SR_FULL_CLEANED** is then built on all cleaned *Subtitles* data. Also, two bilingual systems **DE→HR_FULL** and **DE→SR_FULL** are trained on all *Subtitles* data from Table 1, without any cleaning.

[9] It is worth noting that the multilingual system is better than the bilingual systems also for the opposite translation direction.

3.3 Systems for translating English movie reviews

The main challenge when translating this test set is that there are no readily available parallel corpora of movie reviews which could be used for in-domain training. Therefore, we apply the following strategy: we start from the multilingual EN→HR+SR_CLEAN/ED system trained on 4.1M clean segments as the baseline, and enrich it with different types and amounts of synthetic parallel movie reviews. The following systems are trained:

baseline EN→HR+SR CLEAN/ED: on the clean data together with the cleaned first 2M segments from OpenSubtitles.

+MOVIES-BT: on the baseline system data enriched with back-translated *Movies* data. The system used for back-translation is the same "*hr+sr-en_clean/ed*" system used for cleaning the full *Subtitles* data. This system can be seen as the baseline system in the opposite translation direction.

+SELECTED-BT: the baseline system data enriched with back-translated *Movies* and *Selected* data.

+IMDB-FT: on the data of the system "+SELECTED-BT" enriched with forward-translated *IMDb* data. Forward translation was performed by the system "+SELECTED-BT".

+AMAZON-FT: on the data used for the system "+IMDB-FT" enriched with forward-translated *Amazon* data. Forward translation was performed by the system "+IMDB-FT".

In addition to these mid-resourced partially in-domain systems, we also translate the movie reviews test set by the three systems trained on full out-of-domain OPUS data (EN→HR_FULL, EN→SR_FULL, EN→HR+SR_FULL_CLEAN/ED).

4 Results

We evaluate our systems using the following three automatic overall evaluation scores: sacreBLEU (Post, 2018), chrF (Popović, 2015) and characTER (Wang et al., 2016). The BLEU score is based on word n-gram (n in range from 1 to 4) precision and brevity penalty which should replace recall. The chrF score is based on character n-gram matching (n in range from 1 to 6) instead of word n-gram matching. It is F-score which weights recall two times more than precision. The characTER score is based on edit distance which takes into account not only substitutions, insertions and deletions, but also word sequence reorderings and character sequences in unmatched words. We use the BLEU score because of the long tradition of using it for MT evaluation despite well-known faults, and the two character level scores because they are shown to correlate much better with human assessments (Bojar et al., 2017; Ma et al., 2018). Recently, the chrF score is recommended as a replacement for BLEU (Mathur et al., 2020).

In addition to the automatic MT evaluation scores, for each of the systems we report the size of the training corpus, the percentage of particular target language data in this corpus, as well as the training time in terms of days.

4.1 Translation from English

The results for translation from English are presented in Table 3. As expected, multilingual systems are better than bilingual for all set-ups, even for the unbalanced clean low-resourced corpus. However, the improvements are smaller for Croatian, the language with more data (77.1% of segments). Another observation is that multilingual systems trained on cleaned data, with reduced corpus size and similar or shorter training time, demonstrate better translation performance in terms of three automatic scores than bilingual systems trained on uncleaned data.

Adding the first 2M cleaned segments from *Subtitles* to the small clean data (EN-HR+SR_CLEAN/ED) results in scores which are approaching those obtained by full data, although the corpus size is 6 to 10 times smaller (4.1M vs. 26/30/39M). The training time is more than two times shorter (two days vs. five days). This could be interesting for cases when a trade-off between performance and resources can play a role. Adding German data does not prolong the training time, but it slightly deteriorates all scores.

(a) Translation from English into Croatian

	training			test, en→hr, subs+news		
system	size	%hr	time	BLEU↑	chrF↑	chrTER↓
EN→HR_CLEAN	0.9M	100	<12h	22.3	48.5	52.1
EN→HR+SR_CLEAN	1.3M	77.1	<12h	22.6	48.7	51.8
EN→HR_SUBS2M	2.0M	100	<1d	20.0	44.4	49.9
EN→HR+SR_SUBS2M	3.6M	45.7	<1d	22.1	47.2	46.8
EN→HR+SR_SUBS2M_CLEANED	2.7M	44.8	<12h	20.9	45.6	48.2
EN→HR+SR_CLEAN/ED	4.1M	55.5	<2d	**32.4**	**56.7**	**43.4**
DE+EN→HR+SR_CLEAN/ED	6.4M	56.1	<2d	31.8	56.2	43.7
EN→HR_FULL	26M	100	~5d	33.0	57.4	41.6
EN→HR+SR_FULL_CLEAN/ED	39M	39.1	~5d	**33.7**	**58.2**	**41.0**
Google	*n.a.*	*n.a.*	*n.a.*	*26.9*	*53.4*	*46.2*
Amazon	*n.a.*	*n.a.*	*n.a.*	***31.7***	***57.1***	***42.5***

(b) Translation from English into Serbian

	training			test, en→sr, subs+news		
system	size	%sr	time	BLEU↑	chrF↑	chrTER↓
EN→SR_CLEAN	0.3M	100.0	<12h	19.3	43.6	60.4
EN→HR+SR_CLEAN	1.3M	22.9	<12h	22.8	47.0	58.0
EN→SR_SUBS2M	2.0M	100	<1d	19.2	43.1	55.0
EN→HR+SR_SUBS2M	3.6M	54.3	<1d	22.2	46.5	52.1
EN→HR+SR_SUBS2M_CLEANED	2.7M	55.2	<12h	20.7	45.2	52.4
EN→HR+SR_CLEAN/ED	4.1M	44.5	<2d	**33.8**	**56.6**	**47.6**
DE+EN→HR+SR_CLEAN/ED	6.4M	43.9	<2d	32.6	55.6	48.4
EN→SR_FULL	30M	100	~5d	**35.5**	57.6	46.1
EN→HR+SR_FULL_CLEAN/ED	39M	60.9	~5d	35.2	**57.7**	**45.6**
Google	*n.a.*	*n.a.*	*n.a.*	24.4	50.8	51.5
Amazon	*n.a.*	*n.a.*	*n.a.*	26.5	52.5	51.1

Table 3: Results for translation from English into Croatian (above) and Serbian (below): corpus size, percentage of the target language, training time, and the three automatic MT evaluation scores: BLEU (higher values are better), chrF (higher values are better) and characTER (lower values are better).

Apart from this, it can be seen that only Amazon Translate for Croatian is comparable to our best systems while all other on-line systems are clearly outperformed.[10]

4.2 Translation from German

Table 4 shows the results for translation from German. First of all, it can be noted that, as expected, the scores for translating from German are generally much worse than for translating from English. Due to this discrepancy in performance, the multisource system including both English and German data DE+EN-HR+SR_CLEAN/ED improves the scores for German although it slightly deteriorates the scores for English (Table 3).

One possible reason for the discrepancy is that the German language is morphologically more complex than English, although it has more similarities with the target languages (such as grammatical gender, cases, verb prefixes, etc.). Still, these general similarities often cannot be mapped (such as usage of different cases for the same constructions, different grammatical gender of the same noun, etc.), which is not very helpful for the translation process. Another factor, namely the nature of the corpus, could also play an important role. The majority of non-English texts in *OpenSubtitles* are human translations from English original sentences. Therefore, for translation from English, the source language is the original one and the target language is its human translation. In contrast, for translation from German, both sides are human translations, which can have a strong impact on performance (Kurokawa et al., 2009; Vyas et al., 2018; Zhang and Toral, 2019). A thorough investigation of the data should be carried out in future work in order to better understand these results.

Apart from this, it is once more confirmed that multilingual systems yield better scores than bilingual.

[10] On one hand, our systems were trained on in-domain data. On the other hand, the companies providing on-line systems have access to huge amounts of data and computer resources.

(a) Translation from German into Croatian

	training			test, de→hr, subs		
system	size	%hr	time	BLEU↑	chrF↑	chrTER↓
DE→HR_SUBS2M	2.0M	100.0	<1d	15.8	36.6	60.5
DE→HR+SR_SUBS2M	3.5M	56.6	<1d	18.2	39.1	57.9
DE→HR+SR_SUBS2M_CLEANED	2.3M	57.2	<12h	16.5	37.3	58.9
DE+EN→HR+SR_CLEAN/ED	6.4M	56.1	<2d	**18.6**	**39.4**	**57.6**
DE→HR_FULL	9.8M	100.0	∼3d	20.7	41.9	55.4
DE→HR+SR_FULL_CLEANED	12.6M	55.6	∼3d	**21.4**	**42.4**	**55.3**
Google	*n.a.*	*n.a.*	*n.a.*	*14.8*	*38.2*	*59.7*
Amazon	*n.a.*	*n.a.*	*n.a.*	*18.5*	*40.6*	*57.1*

(b) Translation from German into Serbian

	training			test, de→sr, subs		
system	size	%sr	time	BLEU↑	chrF↑	chrTER↓
DE→SR_SUBS2M	2.0M	100.0	<1d	14.7	34.9	60.8
DE→HR+SR_SUBS2M	3.5M	43.4	<1d	16.1	36.4	59.5
DE→HR+SR_SUBS2M_CLEANED	2.3M	42.8	<12h	14.9	35.4	59.8
DE+EN→HR+SR_CLEAN/ED	6.4M	43.9	<2d	**15.7**	**36.8**	**59.1**
DE→SR_FULL	10.6M	100.0	∼3d	18.2	39.1	56.9
DE→HR+SR_FULL_CLEANED	12.6M	44.4	∼3d	**18.5**	**39.4**	**56.7**
Google	*n.a.*	*n.a.*	*n.a.*	*11.6*	*34.8*	*61.9*
Amazon	*n.a.*	*n.a.*	*n.a.*	*12.4*	*35.6*	*60.5*

Table 4: Results for translation from German into Croatian (above) and Serbian (below): corpus size, percentage of the target language, training time, and the three automatic MT evaluation scores: BLEU (the higher the better), chrF (the higher the better) and characTER (the lower the better).

Furthermore, our systems result in better scores than on-line systems.

4.3 Translation of English movie reviews

Table 5 shows the results for the IMDb test set. The multisource system including German also deteriorates the scores for this test set. As for synthetic parallel data, it can be noted that, as expected, the BT data improve the scores, especially for Serbian because there are many more movie reviews in this language. FT data further improves the scores at the cost of slightly prolonged training. In contrast, simply including all out-domain data results in better scores although there is no in-domain data. In addition, contrary to all other test sets, bilingual system for Croatian results in better automatic scores than the multilingual cleaned system. For all these reasons, it is still hard to draw conclusions about translating user reviews, so further systematic research including manual evaluation and analysis of MT outputs is needed to find an optimal set-up.

As for on-line systems, again only Amazon Translate into Croatian is comparable with our best system. All other on-line systems are outperformed both by our best system as well as by our second and third best systems.

5 Summary and outlook

This work presents a systematic investigation of different set-ups for training NMT systems for translation into Serbian and Croatian, two closely related South Slavic languages. We explore English and German as source languages, different sizes and types of training corpora, as well as bilingual and multilingual systems. We also explore translation of English IMDb user movie reviews, a domain/genre where only monolingual data are available.

Our results confirm that multilingual systems with joint target languages perform better. The performance of translation from English is generally much better than from German, partly because German is morphologically more complex and partly because the corpus consists mostly of parallel human translations instead of original text and its human translation. More research should be carried out on translation from German (as well as on more languages other than English) in order to better understand the poten-

(a) Translation of English IMDb movie reviews into Croatian

	training			test, en→hr, imdb		
system	size	%hr in domain	time	BLEU↑	chrF↑	chrTER↓
EN→HR+SR_CLEAN/ED	4.1M	0	<2d	27.4	53.9	41.9
+MOVIES-BT	4.2M	4.4	<2d	27.4	54.1	41.6
+SELECTED-BT	5.2M	45.0	<2d	27.4	55.0	40.9
+IMDB-FT	5.8M	50.0	<3d	28.2	55.9	40.2
+AMAZON-FT	6.8M	50.0	∼3d	**29.2**	**56.6**	**39.8**
DE+EN→HR+SR_CLEAN/ED	6.4M	0	<2d	26.7	53.4	42.4
EN→HR_FULL	25.5M	0	∼5d	**31.8**	**57.6**	**39.3**
EN→HR+SR_FULL_CLEAN/ED	39M	0	∼5d	30.6	56.7	39.8
Google	*n.a.*	*n.a.*	*n.a.*	*28.6*	*55.7*	*40.6*
Amazon	*n.a.*	*n.a.*	*n.a.*	*30.9*	***57.6***	***38.9***

(b) Translation of English IMDb movie reviews into Serbian

	training			test, en→sr, imdb		
system	size	%sr in domain	time	BLEU↑	chrF↑	chrTER↓
EN→HR+SR_CLEAN/ED	4.1M	0	<2d	26.5	53.3	42.2
+MOVIES-BT	4.2M	95.6	<2d	28.0	54.7	41.1
+SELECTED-BT	5.2M	55.0	<2d	28.8	55.6	40.0
+IMDB-FT	5.8M	50.0	<3d	29.6	56.3	39.4
+AMAZON-FT	6.8M	50.0	∼3d	**30.2**	**56.5**	**39.0**
DE+EN→HR+SR_CLEAN/ED	6.4M	0	<2d	26.5	53.3	42.0
EN→SR_FULL	30M	0	∼5d	31.5	56.9	39.3
EN→HR+SR_FULL_CLEAN/ED	39M	0	∼5d	**31.6**	**57.2**	**39.1**
Google	*n.a.*	*n.a.*	*n.a.*	*26.4*	*54.2*	*40.9*
Amazon	*n.a.*	*n.a.*	*n.a.*	*26.7*	*54.6*	*40.8*

Table 5: Results for translation of English IMDb movie reviews into Croatian (above) and Serbian (below): corpus size, percentage of the target language in the in-domain training corpus, training time, and the three automatic MT evaluation scores (BLEU, chrF and characTER)

tials and limits of the approaches tested here. Still, for both source languages, our best systems perform better than the two online-systems, Google Translate and Amazon Translate, whereby the Amazon systems for translating English into Croatian are comparable with ours.

For translating user reviews, creating synthetic in-domain parallel data through back- and forward-translation and adding them to a small out-of-domain parallel corpus can yield performance comparable with a system trained on a full out-of-domain corpus. Our experiments still leave some important questions open, such as the impact of the size of synthetic FT data and impact of performance of MT system which generated this data. Therefore, more detailed research including manual evaluation and analysis of translated reviews is needed in this direction.

Apart from this, removing misaligned segments by log-probabilities provided by MT systems should be investigated systematically, by comparing different NMT systems for cleaning together with different thresholds for log-probabilities.

Acknowledgements

The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106. This research was partly funded by financial support of the European Association for Machine Translation (EAMT) under its programme “2019 Sponsorship of Activities”.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 3874–3884, Minneapolis, Minnesota, June.

Mihael Arčan, Maja Popović, and Paul Buitelaar. 2016. Asistent – A Machine Translation System for Slovene, Serbian and Croatian. In *Proceedings of the Tenth Conference on Language Technologies and Digital Humanities (JDTH 2016)*, pages 13–20, Ljubljana, Slovenia, September.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 272–283, Edinburgh, Scotland.

Ergun Biçici and Deniz Yuret. 2015. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):339–350.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation (WMT 2017)*, pages 489–513, Copenhagen, Denmark, September.

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, November.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 2018)*, pages 200–207, Boston, MA, March.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, May.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative Fine-grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *Machine Translation*, 32(3):195–215, September.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *In Proceedings of MT Summit XII*, pages 81–88, Ottawa, Canada, August.

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 156–164, Brussels, Belgium, October.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In *Proceedings of the 14 Conference on Text, Speech and Dialogue (TSD 2011)*, Lecture Notes in Computer Science, pages 395–402, Pilsen, Czech Republic, September. Springer.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, April.

Pintu Lohar, Maja Popović, and Andy Way. 2019. Building English-to-Serbian Machine Translation System for IMDb Movie Reviews. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, pages 105–113, Florence, Italy, August.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 671–688, Belgium, Brussels, October.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July.

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

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, pages 43–52, Santiago, Chile.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018a. Investigating Back translation in Neural Machine Translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, pages 249–258, Alicante, Spain, May.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018b. Investigating backtranslation in neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 249–258, Alicante, Spain.

Alberto Poncelas. 2019. *Improving transductive data selection algorithms for machine translation*. Ph.D. thesis, Dublin City University.

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, May.

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 (LT4CloseLang 2014)*, pages 76–84, Doha, Qatar, October. Association for Computational Linguistics.

Maja Popović, Kostadin Cholakov, Valia Kordoni, and Nikola Ljubešić. 2016. Enlarging scarce in-domain English-Croatian corpus for SMT of MOOCs using Serbian. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 97–105, Osaka, Japan, December.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT 2015)*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 186–191, Brussels, Belgium, October.

Victor M. Sánchez-Cartagena, Nikola Ljubešić, and Filip Klubička. 2016. Dealing with data sparseness in SMT with factured models and morphological expansion: a case study on Croatian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2016)*, pages 354–360, Riga, Latvia.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 211–221, Florence, Italy, July.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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, August.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1715–1725, Berlin, Germany, August.

Margita Šoštarić, Christian Hardmeier, and Sara Stymne. 2018. Discourse-related language contrasts in English-Croatian human and machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 36–48, Brussels, Belgium, October.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey, May.

Antonio Toral, Raphael Rubino, Miquel Esplá-Gomis, Tommi Pirinen, Andy Way, and Gema Ramirez-Sanchez. 2014. Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 2014)*, pages 220–224, Dubrovnik, Croatia, June.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008, Long Beach, CA, December.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 1503–1515, New Orleans, Louisiana, June.

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, August.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (WMT 2019)*, pages 73–81, Florence, Italy, August. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 1535–1545, Austin, Texas, November.