Extracting correctly aligned segments from unclean parallel data using character n-gram matching

Maja Popović, Alberto Poncelas

ADAPT Centre, Dublin City University, Ireland
name.surname@adaptcentre.ie

Abstract

Training of Neural Machine Translation systems is a time- and resource-demanding task, especially when large amounts of parallel texts are used. In addition, it is sensitive to unclean parallel data. In this work, we explore a data cleaning method based on character n-gram matching. The method is particularly convenient for closely related language since the n-gram matching scores can be calculated directly on the source and the target parts of the training corpus. For more distant languages, a translation step is needed and then the MT output is compared with the corresponding original part. We show that the proposed method not only reduces the amount of training corpus, but also can increase the system's performance.

1. Introduction

Training of Neural Machine Translation (NMT) systems is a time- and resource-demanding task, especially when large amounts of parallel texts are used. In addition, NMT models are sensitive to unclean parallel data – they tend to perform better with more training data, however it is better to provide less data of better quality. In this work, we explore a data cleaning method based on character n-gram matching. The method is particularly convenient for closely related languages since the n-gram matching scores can be calculated directly on the source and the target parts of the training corpus. For more distant languages, a translation step is needed and then the MT output is compared with the corresponding original part. We show that our method can not only reduce the amount of training data and speed up the training process, but also can lead to better translation quality.

1.1. Related work

As the data has a big impact on the performance of the MT, several publications explored how to identify and remove sentences that could hurt the translation quality of an MT system. Usually, these are noisy parallel sentences where the meaning of the source sentence is different from the meaning of the target sentence. This may often happen in corpora where sentences are automatically crawled. In other cases, although the meaning is the same, the source and target parts are substantially diverging from a literal translation. For example, in the case of subtitling, due to the time limitation, the translations are sometimes simplified in order to be adapted to a particular time frame.

Usual approaches for filtering noisy sentence pairs include comparing characteristics of the sentences such as similarity of lengths, e.g. length difference (Taghipour et al., 2010) or length proportion (Khadivi and Ney, 2005), checking amounts of punctuation or name entities, or evaluating whether the alignment of words in the sentences are well distributed (Taghipour et al., 2010).

Many approaches also involve comparing sentences to their MT-translated version. First, an MT system is trained on the clean data, then it is used to translate the non-English part into English, and then mono-lingual matching methods are used to compare it against the original English side of the parallel corpus. Different matching metrics were used: METEOR (Erdmann and Gwinnup, 2019), Levenshtein distance (Sen et al., 2019), or BLEU (Parcheta et al., 2019) or TER (Freitag et al., 2017).

Other approaches use the word alignments (Bouamor and Sajjad, 2018; Kurfalı and Östling, 2019) or teacher-student frameworks (Freitag et al., 2017). Junczys-Dowmunt (2018) proposed Dual Conditional Cross-Entropy Filtering, where sentence pairs are scored based on the disagreement between MT-translated sentences in both directions. Axelrod et al. (2019) expanded this approach by iteratively improving the MT models with good-quality sentence pairs selected from the comparable set. Since 2018, a shared task on parallel corpus filtering[1] is included in the conference on machine transation (WMT) (Koehn et al., 2018; Koehn et al., 2019), where some of these methods were proposed and tested

While all these methods can be applied to similar language pairs, too, the method proposed in this work is much faster and less resource-intensive.

2. Goal of the paper

The goal of our experiments is to improve a parallel corpus used for training a NMT system and reduce its size by removing misaligned segments without significant loss in translation quality. The results show that the translation quality can even be improved. We carried out systematic experiments on two groups of closely related languages: Slovenian, Serbian and Croatian, as well as Spanish and Portuguese. The criterion for keeping or removing a parallel segment is character n-gram matching score between the source and the target languages. Different thresholds of the score were explored in order to find out the optimal one. For South Slavic languages, the experiments were carried out with Slovenian on one side and Serbian or Croatian on another side in all translation directions. For Spanish and Portuguese, we explored both translation directions.

[1] https://www.statmt.org/wmt18/parallel-corpus-filtering.html

In addition, a preliminary experiment with a distant language pair, English–Serbian, has been carried out. In this case, the source and the target part of the corpus cannot be compared directly, so MT is used as an intermediate step, and then the MT output of the training corpus is compared with the corresponding original part. This type of filtering has already been investigated (Freitag et al., 2017; Erdmann and Gwinnup, 2019; Sen et al., 2019; Parcheta et al., 2019), however using another matching metrics.

3. Cleaning method

Our cleaning method is based on the character n-gram based F-score CHRF[2] (Popović, 2015) which is usually used for MT evaluation to compare MT output with a human reference translation. The score is based on matched character n-grams between the two texts, with n-gram lengths n ranging from 1 to 6. When combining precision and recall into F-score, twice more weight goes to recall, since it is shown to be the best option for MT evaluation (Popović, 2016). It is fast, language-independent, tokenisation-independent, it can be used directly without training and parameter optimisation, and it shows good correlations with human judgments. In 2018, it became one of the baseline automatic metrics in the WMT shared task.[3] Recently, in a study encompassing a number of automatic MT metrics (Mathur et al., 2020). the CHRF score is recommended as a replacement for the widely used BLEU and TER scores.

Closely related langugages While designed for comparing two translations in the same language, the CHRF score can provide useful information about the degree of matching between two texts written in closely related languages. The intuitive idea is that the properly aligned segments will have higher CHRF scores than misaligned ones.

Therefore, for the Slovenian–Croatian/Serbian and Spanish–Portuguese corpora, we calculate the CHRF score for each parallel segment in the training corpora. Examples from the Slovenian–Croatian/Serbian corpora can be seen in Table 3., ordered according their CHRF scores (from highest to lowest). Due to similarity between the languages, the scores between the properly aligned segments are indeed relatively high (top part of the table) in contrast to misaligned segments (10,11,12). Nevertheless, the languages are sufficiently different so that some properly aligned short segments (words) have low scores (8, 9) and might be removed in the process. Still, if those words appear in longer sentences (“Nihče" in 4 and “Veš" in 5), these sentences will get high scores and will not be removed.

Similar tendencies can be found in the Spanish–Portuguese corpus.

Distant languages While the CHRF score can be used for more distant languages, too, the method becomes complicated. The segments cannot be compared directly, so that translation of one part of the training corpus is necessary, which imposes certain challenges. First, the process is not fast anymore since translation of a training corpus requires its time. Second, the translation process introduces more divergences so that more properly aligned segments might end up with lower scores. Third, quality of the MT system used for cleaning has to be taken into account. Generally, better systems result in higher automatic evaluation scores, so the optimal threshold for one MT system is not necessarily the same as for another system. Also, it should be investigated whether lower-quality systems can perform well by increasing the threshold, or the best option would be to use high quality system. Finally, translation direction might play an important role, too, because the performances of the systems in two translation directions usually differ.

Taking all this into account, we carried out a preliminary experiment on the English→Serbian translation. We analysed a system trained on a small training corpus (2M segments), and we cleaned the data using a better system trained on large training corpus (30M segments). The English part was first translated by the “cleaner" system, and then the CHRF scores were calculated between the translated and the original Serbian corpus.

Examples for English–Serbian are shown in Table 3.. It can be noted that the correlation between the CHRF score and the alignment quality became more blurred. Properly aligned segments often yield high scores (1, 2) whereas misaligned segments often yield lower scores (6, 8). Nevertheless, less literal translations with more shifts and divergences from the source text such as segments 5 and 7 also exhibit low scores. Another unnecessarily low score can be seen in a properly aligned segment 9 where the MT system left the word untranslated.

4. Experimental set-up

Data One of the largest sources of publicly available parallel data are *OpenSubtitles* from the OPUS[4] data collection (Tiedemann, 2012). Before the training of the baseline systems, 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. The statistics of the remaining subtitles together with the development and test sets is shown in Table 3.

The development and test data for the closely related language pairs were provided by organisers of the WMT 2020 shared task on translation between similar languages.[5]

For the experiment with English–Serbian, the development test was separated from the training data and consists mainly of subtitles with a small portion of news. The test set consists of publicly available *IMDb* movie reviews[6] used in Lohar et al. (2019).

MT systems All NMT systems are built using the Sockeye implementation (Hieber et al., 2017) of the Transformer architecture (Vaswani et al., 2017). The systems operate on sub-word units generated by byte-pair encoding (BPE) (Sennrich et al., 2016). We set the number of BPE merging operations at 32000. We use shared vocabularies

[2] https://github.com/m-popovic/chrF

[3] http://www.statmt.org/wmt18/

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

[5] http://www.statmt.org/wmt20/similar.html

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

	properly aligned	Slovenian	Croatian/Serbian	CHRF score
1)	yes	Hvala.	Hvala.	100
2)	yes	Pa njegova moč?	A njegova moć?	63.34
3)	yes	Zahtevam, da me izpustite.	Zahtijevam da me pustite.	50.29
4)	yes	Nihče ne govori o iztrebljanju.	Nitko ne govori o istrebljenju.	47.74
5)	yes	Veš, na kaj mislim.	Znaš na što mislim.	37.51
6)	yes	Kaj je narobe s tem norim mačkom?	Što je s tim ludim mačkom?	34.10
7)	yes	Bi rad vedel, kje je?	Zanima te gdje je?	20.51
8)	yes	Veš.	Znaš.	13.14
9)	yes	Nihče!	Nitko!	8.89
10)	no	Seveda je bilo nekaj občudovanja vrednih kotičkov ob Temzi.	Poznat je i kao kavez.	7.54
11)	no	Comprenez vous?	Razumiješ li?	6.13
12)	no	Uh, uh, pogumno, naprej!	Kak' ste, šefe?	2.58

Table 1: Examples from the corpus with closely-related languages: Slovenian segments (left), Serbian/Croatian segments (middle), and character n-gram matching scores between them (right). Correlation between alignment quality and the score can be observed.

	properly aligned	English original	Serbian MT from English	Serbian original	CHRF score
1)	yes	The interior of Texas at that time was completely unknown to the outside world.	Unutrašnjost Teksasa u to vreme bila je potpuno nepoznata spoljnom svetu.	Unutrašnjost Teksasa u to vreme je bila potpuno nepoznata spoljnom svetu.	90.44
2)	yes	But they found a strange and evil country,	Ali našli su čudnu i zlu zemlju,	Ali su naišli na čudnu i zlu zemlju.	63.87
3)	yes	This is a cattail.	Ovo je mačji rep.	Ovo je rogoz.	27.62
4)	no	I don't know whether she's genuinely interested in protecting her husband, or the money.	Ne znam da li je zaista zainteresovana da zaštiti svog muža, ili novac.	Moram doma. Treba me ujak Martin. Ne mogu te tako pustiti.	15.75
5)	yes	It was easy pickings.	To je bio lak plen.	To su mogli lako da nađu.	12.53
6)	no	I don't wanna know.	Ne želim da znam.	Voliš li me?	11.47
7)	yes	Now, those are minerals.	To su minerali.	A sad evo ruda.	9.56
8)	no	How did you find out about the apartment?	Kako si saznao za stan?	Ne bi ga dala zaručniku.	8.51
9)	yes	(THUNDER)	(THUNDER)	(GRMLJAVINA)	5.84

Table 2: Examples from the corpus with distant languages: English original segments (left), MT outputs generated from them (middle left), Serbian original segments (middle right), and character n-gram matching scores between original Serbian and MT outputs from original English (right). Correlation between alignment quality and the score is not obvious, and also might depend on the performance of the used MT system.

for translation systems between the similar languages, and separated vocabularies for English–Serbian systems.

For translation between the South Slavic languages, we tested the data cleaning on multisource and multitarget systems as proposed in Johnson et al. (2017). Multisource setting was used for translation from Serbian/Croatian into Slovenian, and multitarget setting for the other translation direction. We decided to use multilingual systems (with joint Serbian and Croatian data) as baselines because they perform better than bilingual systems, which can be seen in Table 6. For translation between Spanish and Portuguese, only bilingual systems were possible to build. For all pairs of similar languages, we investigated CHRF thresholds in the range from 10 to 50 under the assumption that the threshold of 10 will probably retain too many misaligned segments while the threshold of 50 will remove too many proper segment pairs resulting in too small parallel corpus, so that the optimal threshold lies something between these two values.

For translation between English and Serbian, the system used for cleaning is trained on all publicly available corpora consisting mainly of subtitles (about 30 million segments) as well as a small portion of news (200k segments) and other sources (100k segments). For the preliminary experiment described in this work, we took 2 million segments from subtitles as a starting point and explored the clean-

(a) training corpora

languages		domain	# sentences
close	sl–hr	subtitles	11 213 386
	sl–sr	subtitles	11 780 062
	pt–es	subtitles	22 317 343
distant	en–sr	subtitles	2 000 000
	“cleaner"	mainly subtitles	30 002 892

(b) development and test sets

languages		set	domain	# sentences
close	sl–hr	dev	PR publications	2457
		test	PR publications	2582
	sl–sr	dev	PR publications	1259
		test	PR publications	1260
	pt–es	dev	web and technical	1283
		test	web and technical	1495
distant	en–sr	dev	subtitles+news	1000
		test	movie reviews	485

Table 3: Corpora statistics.

ing of this corpus. The English part was translated by the system trained on the full corpus, and the CHRF scores are calculated between the translation and the original Serbian part. For this language pair, we investigated three CHRF thresholds, 30, 40 and 50 under the assumption that translations below the threshold of 30 are not of sufficient quality to be relied on. This assumption is based on a qualitative manual inspection of the corpus.

Evaluation We evaluate the outputs of our systems using the following three automatic overall evaluation scores: word n-gram precision metric BLEU (Post, 2018), character n-gram F-score chrF (used for cleaning) (Popović, 2015), and character-level translation edit rate characTER (Wang et al., 2016). The BLEU score is used because of the long tradition. The two character level scores are shown to correlate much better with human assessments (Bojar et al., 2017; Ma et al., 2018), especially for morphologically rich languages.

5. Results

5.1. Closely related languages

Table 4 shows results for the translation between the South Slavic languages. It can be seen that the initial hypothesis is true: overall, the best threshold is chrF=20, lying between 10 and 50. It reduced the training corpus by almost half, and surpassed the scores of the system trained on the full data. Larger thresholds remove too many useful segments therefore deteriorating the scores. Threshold of 10, while still yielding better automatic scores than the baseline system, performs worse than the system with the threshold 20.

Four examples of translation from Slovenian into Serbian can be seen in Table 5. For the sentence #1, setting the threshold to 40 results in a mistranslation, while CHRF-10 and CHRF-20 achieve the same performance as the baseline system. The sentence #2 is fully correctly translated only by the CHRF-10 system. The only problem with the CHRF-20 system are two named entitites which, howvever, do not change the meaning of the sentence. Named entities are also corrupted by the baseline system, one is unnecessarily transcribed and another one unnecessarily translated, thus deteriorating adequacy. The CHRF-40 system, apart from an error in the named entity, generates a mistranslation, a local word order error, as well as deletion of a phrase which significantly changes the meaning. In the third sentence, the worst performance can again be observed for the baseline and CHRF-40 systems. Other two systems contain errors which do not change the meaning: the CHRF-10 system deletes a temporal adverb, while the CHRF-20 system makes a sub-optimal lexical choice. For the fourth sentence, the baseline and the CHRF-20 systems generate a correct output while CHRF-10 and CHRF-20 systems generate a mistranslation.

Same tendencies, although less clearly, can be observed for the translation between Spanish and Portuguese presented in Table 6.

5.2. Distant languages

The results for English->Serbian are shown in Table 5.2.. For the sake of completeness, we also show the results of the “cleaner" system trained on the full uncleaned data (row 1). The cleaning is performed on the 2M subset (row 2). For the in-domain test set, removing 0.5M segments with chrF<30 yields better BLEU and characTER and comparable CHRF than when all 2M segments are used. For the movie reviews test set, however, it does not surpass the original performance, but there is no substantial loss.

These results confirm the findings reported in (Freitag et al., 2017), namely that knowledge distillation using a “big" translation system and an automatic evaluation metric is a promising concept for reducing resources necessary for NMT training. Nevertheless, more systematic experiments should be carried out in the future, which should include different cleaning systems with different CHRF thresholds.

6. Conclusions

We propose a method for cleaning parallel training data between closely related languages based on character n-gram matching. The results on South Slavic and Spanish–Portuguese language pairs show that our method can not only reduce the amount of training data and speed up the training process, but it can also lead to better translation quality. Overall, the best option is to remove all segments with the chrF score below 20.

Our preliminary experiment on a more distant language pair, English–Serbian, which also requires an additional translation step, also shows promising results. However, more systematic experiments should be carried out in this direction, such as using different MT systems with different CHRF thresholds for cleaning, using the other translation direction for cleaning, as well as cleaning larger amounts of training data.

Acknowledgments

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

(a) Croatian→Slovenian

training system	size	dev, hr→sl BLEU↑	chrF↑	chrTER↓	test, hr→sl BLEU↑	chrF↑	chrTER↓
HR→SL	11.2M	38.5	65.7	29.4	/	/	/
HR+SR→SL	20.2M	38.8	65.9	29.5	34.7	62.2	34.5
+CHRF-50	1.4M	25.2	54.3	37.3	24.3	53.7	36.6
+CHRF-40	2.8M	31.4	60.4	32.1	30.1	59.5	**32.5**
+CHRF-30	5.6M	36.6	64.3	**29.0**	**35.6**	**63.6**	**29.5**
+CHRF-20	10.7M	**39.7**	**66.5**	**27.0**	**37.1**	**65.2**	**28.2**
+CHRF-10	17.2M	***39.3***	***66.1***	***28.4***	**37.1**	***65.0***	***28.8***

(b) Serbian→Slovenian

training system	size	dev, sr→sl BLEU↑	chrF↑	chrTER↓	test, sr→sl BLEU↑	chrF↑	chrTER↓
SR→SL	11.8M	40.6	67.2	30.3	/	/	/
HR+SR→SL	20.2M	42.1	68.3	28.5	37.7	64.1	33.5
+CHRF-50	1.4M	25.7	54.8	38.5	26.5	55.8	34.9
+CHRF-40	2.8M	33.9	62.0	32.2	33.0	62.2	**30.4**
+CHRF-30	5.6M	39.2	66.1	29.0	**37.9**	**66.2**	**27.4**
+CHRF-20	10.7M	**42.2**	**68.6**	**26.9**	**41.2**	**68.1**	**26.5**
+CHRF-10	17.2M	**42.5**	**68.9**	***28.0***	***40.0***	***67.5***	***27.8***

(c) Slovenian→Croatian

training system	size	dev, sl→hr BLEU↑	chrF↑	chrTER↓	test, sl→hr BLEU↑	chrF↑	chrTER↓
SL→HR	11.2M	33.4	62.6	33.0	/	/	/
SL→HR+SR	20.2M	36.0	63.8	32.6	30.3	58.9	40.0
+CHRF-50	1.4M	25.5	54.9	35.9	24.2	53.8	**36.0**
+CHRF-40	2.8M	31.2	61.1	**30.8**	**30.5**	**60.7**	**30.9**
+CHRF-30	5.6M	35.3	**64.4**	**29.1**	**34.4**	**64.0**	**29.2**
+CHRF-20	10.7M	**36.9**	**65.2**	**28.6**	**35.7**	**64.4**	**28.8**
+CHRF-10	17.2M	***36.3***	***64.8***	**28.6**	***34.8***	***64.2***	***29.1***

(d) Slovenian→Serbian

training system	size	dev, sl→sr BLEU↑	chrF↑	chrTER↓	test, sl→sr BLEU	chrF	chrTER↓
SL→SR	11.8M	33.3	62.3	34.3	/	/	/
SL→HR+SR	20.2M	34.8	63.4	33.4	32.0	60.0	36.4
+CHRF-50	1.4M	22.9	53.5	38.7	26.0	55.2	**34.2**
+CHRF-40	2.8M	29.2	59.9	**32.9**	30.2	**60.6**	**30.2**
+CHRF-30	5.6M	34.6	**63.4**	**31.7**	**36.4**	**64.6**	**27.8**
+CHRF-20	10.7M	**35.5**	**64.2**	**31.5**	**37.0**	**65.1**	***28.2***
+CHRF-10	17.2M	***35.1***	***63.8***	***31.9***	**37.7**	**65.6**	**27.4**

Table 4: Results for translation between South Slavic languages: Croatian→Slovenian (a), Serbian→Slovenian (b), Slovenian→Croatian (c) and Slovenian→Serbian: corpus size and the three automatic MT evaluation scores (BLEU, chrF and characTER).

the European Regional Development Fund (ERDF) through Grant 13/RC/2106. This research was partly supported by the European Association for Machine Translation under its programme "2019 Sponsorship of Activities".

7. References

Amittai Axelrod, Anish Kumar, and Steve Sloto. 2019. Dual monolingual cross-entropy delta filtering of noisy parallel data. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 245–251, Florence, Italy.

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.

Houda Bouamor and Hassan Sajjad. 2018. H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proc. Work-*

source #1	Strojno učenje omogoča analizo nepredstavljive količine podatkov hitreje kot kdaj koli prej.
SL→SR(+HR)	Mašinsko učenje omogućava analizu nezamislive količine podataka brže nego ikada *ranije*.
+CHRF-10	Mašinsko učenje omogućava analizu nezamislive količine podataka brže nego ikada *pre*.
+CHRF-20	Mašinsko učenje omogućava analizu nezamislive količine podataka brže nego ikada.
+CHRF-40	**Strujno** učenje omogućava analizu nezamislive količine podataka brže nego ikad.
source #2	IBM-ov super računalnik Deep Blue je že več kot 10 let nazaj premagal najboljšega šahista Garryja Kasparova.
SL→SR(+HR)	IBMov super kompjuter, **Dip Blu**, je pobedio najboljeg šahista Gerija **Iskra** pre više od 10 godina.
+CHRF-10	IBMov super kompjuter "Deep Blue" je pobedio najboljeg šahista Gerija Kasparova pre više od 10 godina.
+CHRF-20	**IBM** superkompjuter Deep Blue je pobedio najboljeg šahistu **Geri** Kasparova pre više od 10 godina.
+CHRF-40	IBMov super kompjuter Deep Blue **već je prebio** najboljeg šahista **Garrya** Kasparova.
source #3	Trenutno je največ govora o avtonomni vožnji – vožnji avtomobilov brez vpletenosti voznika.
SL→SR(+HR)	Trenutno se najviše **radi** o autonomnoj vožnji... vožnji automobila bez **ičije pomoći** vozača.
+CHRF-10	**XXX** Najviše se priča o autonomnoj vožnji auta bez uključivanja vozača.
+CHRF-20	Trenutno se najviše priča o autonomnoj vožnji automobila bez **umešanosti** vozača.
+CHRF-40	Trenutno **ima najviše govora** o autonomnoj vožnji. **Voziranje** automobila bez **umešanosti** vozača.
source #4	Toda Google je pri tem šel še korak dalje.
SL→SR(+HR)	Ali Google je otišao korak dalje.
CHRF-10	Ali Google je **nastavio** dalje.
CHRF-20	Ali Google je otišao korak dalje.
CHRF-40	Ali Google je **prolazio** i korak dalje.

Table 5: Four examples of translation from Slovenian into Serbian: baseline system and systems with three different thresholds: chrF=10, chrF=20 and chrF=40. Errors are marked in bold ("XXX" denotes a missing word).

(a) Spanish→Portuguese

training		dev, es→pt			test, es→pt		
system	size	BLEU↑	chrF↑	chrTER↓	BLEU↑	chrF↑	chrTER↓
ES→PT	22.3M	30.4	59.1	37.2	36.9	60.8	35.7
+CHRF-50	2.1M	26.8	56.2	40.5	31.3	56.8	39.4
+CHRF-40	4.5M	29.4	58.4	38.3	34.8	59.5	36.8
+CHRF-30	8.9M	30.1	58.8	37.6	35.5	60.1	36.2
+CHRF-20	14.8M	**30.8**	**59.3**	**36.8**	*36.7*	*60.7*	**35.7**
+CHRF-10	20.4M	*30.4*	58.9	***37.1***	*36.7*	60.6	**35.4**

(b) Portuguese→Spanish

training		dev, pt→es			test, pt→es		
system	size	BLEU↑	chrF↑	chrTER↓	BLEU↑	chrF↑	chrTER↓
PT→ES	22.3M	35.7	62.0	34.3	40.5	63.4	33.1
+CHRF-50	2.1M	30.6	58.6	37.7	35.3	60.2	35.8
+CHRF-40	4.5M	34.4	61.2	35.2	38.8	63.0	33.2
+CHRF-30	8.9M	35.3	*61.9*	*34.4*	**40.7**	**64.0**	**32.4**
+CHRF-20	14.8M	**35.8**	**62.0**	34.7	**40.7**	***63.8***	**32.3**
+CHRF-10	20.4M	*35.7*	**62.1**	*34.4*	*40.5*	***63.6***	***32.8***

Table 6: Results for translation between Western Romance languages: Spanish→Portuguese (a) and Portuguese→Spanish (b): corpus size and the three automatic MT evaluation scores (BLEU, chrF and characTER).

shop on Building and Using Comparable Corpora, pages 43–47, Miyazaki, Japan.

Grant Erdmann and Jeremy Gwinnup. 2019. Quality and coverage: The AFRL submission to the WMT19 parallel corpus filtering for low-resource conditions task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 267–270, Florence, Italy, August.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Mac-

English->Serbian		open subtitles + news			IMDb movie reviews		
system	corpus size	BLEU↑	chrF↑	cTER↓	BLEU↑	chrF↑	cTER↓
FULL ("cleaner")	30M	35.5	57.6	46.1	31.5	56.9	39.3
OPENSUB-2M	2M	19.2	43.1	55.0	**17.1**	**43.4**	**52.9**
+CHRF-50	943k	17.9	40.6	56.7	15.5	40.0	55.7
+CHRF-40	1.2M	19.1	42.2	55.4	16.1	41.6	54.6
+CHRF-30	1.5M	**20.3**	**43.0**	**54.6**	17.0	42.5	53.2

Table 7: Results for translation between distant languages: corpus size and the three automatic MT evaluation scores (BLEU, chrF and characTER) for translation from English into Serbian on in-domain (mainly subtitles) and out-of-domain (movie reviews) test sets.

duff 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.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Brussels, Belgium.

Shahram Khadivi and Hermann Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *International Conference on Application of Natural Language to Information Systems*, pages 263–274.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Belgium, Brussels, October.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the wmt 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 56–74, Florence, Italy, August.

Murathan Kurfalı and Robert Östling. 2019. Noisy parallel corpus filtering through projected word embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 277–281, Florence, Italy.

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*, pages 105–113, Florence, Italy.

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 (ACL 2020)*, pages 4984–4997, Online, July.

Zuzanna Parcheta, Germán Sanchis-Trilles, and Francisco Casacuberta. 2019. Filtering of noisy parallel corpora based on hypothesis generation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 282–288, Florence, Italy, August.

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

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany, August.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Parallel corpus filtering based on fuzzy string matching. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 289–293, Florence, Italy, August.

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

Kaveh Taghipour, Nasim Afhami, Shahram Khadivi, and Saeed Shiry. 2010. A discriminative approach to filter out noisy sentence pairs from bilingual corpora. In *Proceedings of 5th International Symposium on Telecommunications (IST 2010)*, pages 537–541, Tehran, Iran.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Eight International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218, Istanbul, Turkey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, USA.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany.