Neural Machine Translation between similar South-Slavic languages

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

Abstract

This paper describes the ADAPT-DCU machine translation systems built for the WMT 2020 shared task on Similar Language Translation. We explored several set-ups for NMT for Croatian–Slovenian and Serbian–Slovenian language pairs in both translation directions. Our experiments focus on different amounts and types of training data: we first apply basic filtering on the *OpenSubtitles* training corpora, then we perform additional cleaning of remaining misaligned segments based on character n-gram matching. Finally, we make use of additional monolingual data by creating synthetic parallel data through back-translation. Automatic evaluation shows that multilingual systems with joint Serbian and Croatian data are better than bilingual, as well as that character-based cleaning leads to improved scores while using less data. The results also confirm once more that adding back-translated data further improves the performance, especially when the synthetic data is similar to the desired domain of the development and test set. This, however, might come at a price of prolonged training time, especially for multitarget systems.

1 Introduction

Machine translation (MT) between closely related languages is, in principle, less challenging than translation between distantly related languages, but it is still far from being solved. While MT between closely related South-Western Slavic languages, Croatian, Slovenian and Serbian based on the rule-based (RBMT) and the phrase-based (PB-SMT) approaches has been investigated in the last years (Etchegoyhen et al., 2014; Petkovski et al., 2014; Klubička et al., 2016; Arčan et al., 2016; Popović et al., 2016a), to the best of our knowledge, the new state-of-the-art neural machine translation (NMT) has not been investigated yet for these languages.

In this work, we first compare bilingual and multilingual systems in order to determine whether joining Serbian and Croatian data is useful. Afterwards, we investigate additional cleaning of remaining misaligned segments by using character n-gram matching scores (Popović, 2015). The beauty of the method for similar languages is that it can be applied directly to the given training corpus providing matching scores for each pair of the source-target segments. For distant languages, translation of one side of the training corpus would be required. Finally, we make use of monolingual data in each of the three languages by creating additional synthetic parallel training sets via back-translation (Sennrich et al., 2016a; Poncelas et al., 2018; Burlot and Yvon, 2018).

2 Language properties

Common properties All three languages, Croatian, Serbian and Slovenian, belong to the South-Western Slavic branch. As Slavic languages, they have a very rich inflectional morphology for all word classes: six cases and three genders for all nouns, pronouns, adjectives and determiners. For verbs, person and many tenses are expressed by the suffix so that the subject pronoun is often omitted. There are two verb aspects, so that many verbs have perfective and imperfective form(s) depending on the duration of the described action. As for syntax, all three languages have quite a free word order, and neither language uses articles, either definite or indefinite. In addition to this, multiple negation is always used.

Croatian and Serbian Croatian and Serbian exhibit a large overlap in vocabulary and a strong morpho-syntactic similarity so that the speakers can understand each other without difficulties. Nev-

ertheless, there is a number of small but notable and also frequently occurring differences between them. The largest differences between the two languages are in vocabulary: some words are completely different, some however differ only by one or two letters. Apart from lexical differences, there are also structural differences mainly concerning verbs: modal verb constructions, future tense, as well as conditional.

Slovenian Even though Slovenian is very closely related to Croatian and Serbian, and the languages share a large degree of mutual intelligibility, a number of Croatian/Serbian speakers may have difficulties with Slovenian and the other way round. The nature of the lexical differences is similar to the one between Croatian and Serbian, namely a number of words is completely different and a number only differs by one or two letters. However, the amount of different words is much larger. In addition to that, the set of overlapping words includes a number of false friends (e.g. *brati* means *to pluck* in Croatian and Serbian but *to read* in Slovenian).

The amount of grammatical differences is also larger and includes local word order, verb mood and/or tense formation, question structure, usage of cases, structural properties for certain conjunctions, as well as some other structural differences. Another important difference is the Slovenian dual grammatical number which refers to two entities (apart from singular for one and plural for more than two). It requires additional set of pronouns, as well as additional sets for noun, adjective and verb inflexion rules not existing either in Croatian or in Serbian.

3 Data

For training, we used publicly available OPUS[1] parallel corpora (Tiedemann, 2012) indicated by the workshop organisers. *OpenSubtitles* is indicated for all translation directions. For Croatian–Slovenian, other corpora are indicated too, but they are either not sentence-aligned (*JW300*) or are extremely noisy (*DGT, MultiParaCrawl*). Therefore, we decided to use only *OpenSubtitles* for all translation directions.

It is worth noting that the organisers also indicated the *SETIMES News* parallel Croatian–Serbian corpus. Developing an additional Croatian–Serbian MT system for converting Serbian data into Croatian and vice versa was shown to be helpful for the PBSMT approach (Popović and Ljubešić, 2014; Popović et al., 2016b). However, our preliminary experiments in this direction indicated that this technique is not helpful for the NMT approach.

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

lang.	set	domain	# sentences
sl-hr	train	*Subtitles*	11 213 386
	dev	*PR publications*	2457
	test	*PR publications*	2582
sl-sr	train	*Subtitles*	11 780 062
	dev	*PR publications*	1259
	test	*PR publications*	1260

Table 1: Corpus statistics.

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 1. The development and test sets were provided by the organisers and originate from Public Relations publications of a business intelligence company.

3.1 Additional cleaning of *OpenSubtitles*

While a large number of noisy parts and misaligned segments was removed from *OpenSubtitles* by the basic filtering procedure, a number of misaligned segments still remained. In order to remove these, we applied additional cleaning based on the character n-gram F-score chrF usually used for MT evaluation (Popović, 2015). For the purpose of cleaning, the chrF score is calculated for each pair of segments in the training data. Due to similarity between the languages, the scores between the properly aligned segments are higher than the scores of misaligned segments. Nevertheless, the languages are sufficiently different so that some properly aligned short segments (or single words) can have low scores, too. Still, if those words also appear in longer sentences, they will not be removed. Preliminary experiments with different thresholds showed that keeping the segments with the chrF score equal or greater than 20 is the best option.

3.2 Using monolingual data

In addition to the parallel *OpenSubtitles* corpora, we also used the monolingual data in each of the three languages which were indicated by the organisers, namely the mixed-domain data collected from Web, *hrWac*, *slWac* and *hrWac* (Ljubešić and Erjavec, 2011; Ljubešić and Klubička, 2014). As a first step, we removed too long and too short sentences, keeping those between 5 and 60 words. Then, we removed sentences with more than 1/3 of non-alphanumeric characters, sentences with URLs, as well as duplicate sentences.

Then, we wanted to rank these sentences according to the relevance for our experiments, namely according to their similarity to the development corpus. For this purpose, we used Feature Decay Algorithm (FDA) (Biçici and Yuret, 2011). This method iteratively 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, after a sentence is selected, its n-grams suffer a penalisation so that they are less likely to be selected in the following iterations. The default FDA system halves the score of an n-gram each time it is selected. Therefore the score of a sentence s is computed as in Equation (1):

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

where Sel is the set of sentences that have been selected and $C_Sel(ngr)$ is the count of occurrences of the *n*-gram ngr. At the end, the set S is converted into the set Sel containing the same sentences, but ranked according to their relevance.

For our experiments, the *hrWac*, *slWac* and *srWac* corpora represented the sets S, and the development sets in the corresponding target language were used as $Seed$.

Back-translated synthetic parallel corpora After ranking the monolingual corpora by FDA, back-translation was applied in order to create additional parallel training corpora. For each translation direction, the first two million best ranked sentences in the target language were translated into the source language by the corresponding NMT system.

Translation from Slovenian: The first two million best ranked Serbian sentences and the first two million best ranked Croatian sentences were translated into Slovenian.

Translation into Slovenian: Slovenian is the target language for two translation directions, and we wanted to have equally relevant Slovenian sentences for both directions. Therefore, we did not take the first two million sentences for one source language and the second two million for the other, because the Slovenian sentences for the first source language would be more relevant than those for the second source language. Instead, we took the first four million best ranked Slovenian sentences, and then translated every odd sentence into Serbian and every even sentence into Croatian.

4 MT systems

All our systems are built using the Sockeye implementation (Hieber et al., 2018) of the Transformer architecture (Vaswani et al., 2017). The systems operate on sub-word units generated by byte-pair encoding (BPE) (Sennrich et al., 2016b). We set the number of BPE merging operations at 32000. We use shared vocabularies between the languages because they are similar. Multilingual 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. We investigated the following set-ups:

1. Systems trained on *OpenSubtitles*

   The four bilingual systems, HR→SL, SR→SL, SL→HR and SL→SR, are trained separately for each language pair and each translation direction on about 11M parallel segments.

   The multisource system HR+SR→SL is trained for translation into Slovenian by joining Serbian and Croatian sources and removing duplicates, thus resulting in 20.2M parallel segments.

   The multitarget system SL→HR+SR is trained for translation from Slovenian on the reversed corpus of 20.2M segments with target language identificators "SR" and "HR" added to the source side.

2. Systems trained on cleaned *OpenSubtitles*

   Two multilingual systems HR+SR→SL_CLEAN and SL→HR+SL_CLEAN are trained on joint *OpenSubtitles* corpora additionally cleaned by the chrF score. The

cleaned corpus consists of 10.8M segments (instead of 20.2M).

3. Systems trained on cleaned *OpenSubtitles* and synthetic back-translated parallel *Wac* data

   Two multilingual systems HR+SR→SL_CLEAN+BT and SL→HR+SR_CLEAN+BT are trained on joint cleaned *OpenSubtitles* corpora together with the corresponding synthetic back-translated data selected from *hrWac*, *slWac* and *srWac*. The monolingual data was back-translated by the corresponding systems trained on cleaned *OpenSubtitles*. The training corpora consist of 14.8M segments.

5 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 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. In addition, the chrF score is recommended as a replacement for BLEU in a recent detailed study encompassing a number of automatic MT metrics (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 and the training time.

Table 2 shows the results both on the development and on the test set for each of the four translation directions. First of all, it can be seen that the automatic scores are relatively low given the similarity of the languages. One reason is domain/genre discrepance between the training and the development/test sets. Another possible reason is the nature of the *OpenSubtitles* corpus. The majority of non-English texts in *OpenSubtitles* are namely human translations from English originals. Therefore, for translation from English, the source language is the original one and the target language is its human translation.[2] On the other hand, for translation not involving English, 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). These effects should be investigated in future work.

[2] And other way round for translation into English.

Results on the development set For the systems trained on *OpenSubtitles*, it can be seen that for each translation direction, multilingual systems yield better automatic scores than bilingual systems at the cost of slightly prolonged training time (from about 3 days to 3-4 days). Therefore we choose the two multilingual systems HR+SR→SL and SL→HR+SR as the baselines and we did not keep the bilingual systems for further experiments.

The chrF cleaning of *OpenSubtitles* reduces the size of the corpus and the training time while slightly improving automatic scores. The reduction in time is slightly smaller for the multitarget translation from Slovenian (down to 2-3 days) than for the multisource translation into Slovenian (down to less than 2 days).

Adding the back-translated data from *Wac* improves the automatic scores for more than 10 points for multisource translation (into Slovenian) and for 5 to 10 points for multitarget translation (from Slovenian). This could be expected, especially since the monolingual data was chosen to be similar to the development data. Nevertheless, this large improvement comes at a price. Although the increase of the corpus is not very large, from 10.8M to 14.8M, the training time increases to (more than) 3 days. It can be noted that for some set-ups, the multitarget system needs more training time. The probable reason is the diversity of the target part of the training corpus – the system has to deal with two target languages, and when synthetic data is added, also with two different domains/genres for each of them.

Results on the test set Based on the results on the development set, we submitted the outputs of the systems with back-translated data (HR+SR→SL_CLEAN+BT, SL→HR+SR_CLEAN+BT) as primary submissions. The outputs of the systems trained on cleaned data (HR+SR→SL_CLEAN, SL→HR+SR_CLEAN) were submitted as first contrastive, and the outputs of the baseline multilingual systems (HR+SR→SL, SL→HR+SR) as second contrastive submissions. The test sets were not at all translated by the initial bilingual systems, therefore the results are not available.

It can be seen that the tendencies for the test set are almost the same as for the development set. The only difference is the larger improvement obtained by cleaning *OpenSubtitles* with the chrF scores. Further detailed analysis involving manual inspec-

(a) Croatian→Slovenian

training			dev, hr→sl			test, hr→sl		
system	size	time	BLEU	chrF	chrTER	BLEU	chrF	chrTER
HR→SL	11.2M	∼3 days	38.5	65.7	29.4	/	/	/
HR+SR→SL	20.2M	3-4 days	38.8	65.9	29.5	34.7	62.2	34.5
HR+SR→SL_CLEAN	10.8M	<2 days	39.7	66.5	27.0	*37.1*	*65.2*	*28.2*
HR+SR→SL_CLEAN+BT	14.8M	∼3 days	**53.9**	**77.7**	**18.9**	**51.9**	**76.4**	**20.0**

(b) Serbian→Slovenian

training			dev, sr→sl			test, sr→sl		
system	size	time	BLEU	chrF	chrTER	BLEU	chrF	chrTER
SR→SL	11.8M	∼3 days	40.6	67.2	30.3	/	/	/
HR+SR→SL	20.2M	3-4 days	42.1	68.3	28.5	37.7	64.1	33.5
HR+SR→SL_CLEAN	10.8M	<2 days	42.2	68.6	26.9	*41.2*	*68.1*	*26.5*
HR+SR→SL_CLEAN+BT	14.8M	∼3 days	**58.0**	**80.4**	**18.5**	**55.2**	**78.4**	**19.1**

(c) Slovenian→Croatian

training			dev, sl→hr			test, sl→hr		
system	size	time	BLEU	chrF	chrTER	BLEU	chrF	chrTER
SL→HR	11.2M	∼3 days	33.4	62.6	33.0	/	/	/
SL→HR+SR	20.2M	3-4 days	36.0	63.8	32.6	30.3	58.9	40.0
SL→HR+SR_CLEAN	10.8M	2-3 days	36.9	65.2	28.6	*35.7*	*64.4*	*28.8*
SL→HR+SR_CLEAN+BT	14.8M	>3 days	**46.1**	**72.7**	**22.8**	**45.1**	**72.3**	**23.3**

(d) Slovenian→Serbian

training			dev, sl→sr			test, sl→sr		
system	size	time	BLEU	chrF	chrTER	BLEU	chrF	chrTER
SL→SR	11.8M	∼3 days	33.3	62.3	34.3	/	/	/
SL→HR+SR	20.2M	3-4 days	34.8	63.4	33.4	32.0	60.0	36.4
SL→HR+SR_CLEAN	10.8M	2-3 days	35.5	64.2	31.5	*37.0*	*65.1*	*28.2*
SL→HR+SR_CLEAN+BT	14.8M	>3 days	**45.5**	**73.3**	**23.4**	**47.6**	**73.6**	**22.1**

Table 2: Results: Croatian→Slovenian (a), Serbian→Slovenian (b), Slovenian→Croatian (c) and Slovenian→Serbian: corpus size, training time, and the three automatic MT evaluation scores (BLEU, chrF and characTER).

tion is needed to better understand this difference.

6 Summary and outlook

This work investigates different set-ups for training NMT systems for translation between three closely related South-Slavic languages: Slovenian on one side, and Serbian and Croatian on the other side. We explore different sizes and types of training corpora, as well as bilingual and multilingual systems. Our results show that for all translation directions, multilingual systems with joint Croatian and Serbian data perform better than bilingual systems. The results also show that cleaning misaligned segments using character n-gram matching (chrF score) represents a fast and useful method for closely related languages, which improved the evaluation scores while reducing corpus size and training time. Finally, we confirm that adding back-translated synthetic data, which is the usual practice in neural machine translation, can yield large improvements of evaluation scores also for these languages. Nevertheless, for multitarget translation, it might result in a prolonged training time due to increased variety of the target language side.

Future work should include more genres and domains, as well as detailed analysis of errors and problems in order to further improve the performance of NMT between South Slavic languages.

Acknowledgments

The ADAPT SFI Centre for Digital Content Technology (www.adaptcentre.ie) 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.

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.

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.

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.

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.

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.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 46–53, Reykjavik, Iceland.

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 (Volume 1: Research Papers)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

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.

Filip Klubička, Gema Ramírez-Sánchez, and Nikola Ljubešić. 2016. Collaborative development of a rule-based machine translator between Croatian and Serbian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2016)*, pages 361–367.

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.

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. 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.

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.

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.

Filip Petkovski, Francis Tyers, and Hrvoje Peradin. 2014. Shallow-transfer rule-based machine translation for the western group of south slavic languages. In *Proceedings of the 9th Workshop on Free/open-Source Language Resources for the Machine Translation of Less-Resourced Languages (SaLTMil 2014)*, pages 25–30, Reykjavik, Iceland.

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. 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. Association for Computational Linguistics.

Maja Popović, Mihael Arčan, and Filip Klubička. 2016a. Language related issues for machine translation between closely related south Slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52, Osaka, Japan.

Maja Popović, Kostadin Cholakov, Valia Kordoni, and Nikola Ljubešić. 2016b. 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.

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.

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, Brussels, Belgium.

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.

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.

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.

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.

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.

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.

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. Association for Computational Linguistics.