On Machine Translation of User Reviews

Maja Popović, Andy Way
ADAPT Centre
School of Computing
DCU, Ireland
name.surname@adaptcentre.ie

Alberto Poncelas
Rakuten Asia
Singapore
aponcelas2@hotmail.com

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

Abstract

This work investigates neural machine translation (NMT) systems for translating English user reviews into Croatian and Serbian, two similar morphologically complex languages. Two types of reviews are used for testing the systems: *IMDb* movie reviews and *Amazon* product reviews.

Two types of training data are explored: large out-of-domain bilingual parallel corpora, as well as small synthetic in-domain parallel corpus obtained by machine translation of monolingual English *Amazon* reviews into the target languages. Both automatic scores and human evaluation show that using the synthetic in-domain corpus together with a selected subset of out-of-domain data is the best option.

Separated results on *IMDb* and *Amazon* reviews indicate that MT systems perform differently on different review types so that user reviews generally should not be considered as a homogeneous genre. Nevertheless, more detailed research on larger amount of different reviews covering different domains/topics is needed to fully understand these differences.

1 Introduction

Machine translation (MT) has evolved very rapidly since the emergence of neural approaches in 2015, and it is being used for different genres and domains. Every year, evaluation campaigns which include both human and automatic evaluation are carried out with the goal of advancing the state of the art. The most well-known is the WMT shared task[1] which focuses on news articles and (since 2016) on biomedical texts, and both can be considered as instances of "formal written text". The IWSLT evaluation campaign[2], on the other hand, focuses on the translation of TED talks, and some European projects (TraMOOC, transLectures) investigated the translation of online lectures. In both cases, the text can be considered to be "formal speech", with the challenges of dealing with characteristics of spoken language and speech recognition output.

Recently, interest in the translation of user-generated content in the form of "informal written text" has been increasing. For example, JSALT 2019 workshop[3] focused on translation of very noisy text content originating from sources like WhatsApp, Twitter and Reddit.

In this work, we focus on a different type of written user-generated content, namely user reviews. While the style is not as colloquial and noisy as that of Twitter or of other similar sources, it certainly is much less formal than news texts or other sources that have been investigated traditionally in the MT community. There are also important applications for focusing on this kind of data, both from commercial and from user perspective. More and more companies are expanding into multinational markets, and user reviews of products have become an important asset for online transactions and a feature that many customers expect to find. And in the era of always-available internet connectivity, many individuals rely on experiences of other people not only for guiding purchasing decisions, but also for entertainment options like choosing movies, books, restaurants, etc. In this work, we focus on both kinds of user reviews, namely product reviews from *Amazon* and movie reviews from *IMDb*.

Translating user reviews can increase and improve its reach and utility. The main issue for human translation is the fact that there is way too

[1] http://www.statmt.org/wmt20/

[2] http://workshop2019.iwslt.org/index.php

[3] https://www.clsp.jhu.edu/workshops/19-workshop/improving-translation-of-informal-language/

Proceedings of Recent Advances in Natural Language Processing, pages 1109–1118
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_124

much content to be translated. Therefore, MT is very helpful for this kind of content. However, the genre introduces several important challenges, such as informal language, spelling errors, a large number of domains/topics, and lack of in-domain parallel (bilingual) data.

In this work, we compare two approaches for building MT systems for translating user reviews: training on large parallel out-of-domain data and training on small synthetic in-domain data. We also compare MT performance on two types of user reviews: *IMDb* movies and *Amazon* products.

We investigate Croatian and Serbian as target languages, as a case involving mid-size less-resourced morphologically rich European languages. For these languages, a reasonable amount of out-of-domain parallel data is publicly available to train an NMT system, however still much lower than for "major" European languages (such as German, French, Spanish).

All our experiments were carried out on publicly available data sets. We used OPUS[4] parallel data for out-of-domain training and a selected set of Amazon reviews[5] for in-domain training. For development, we used the publicly available texts[6] consisting of a selected set of English *IMDb* reviews[7] and their Croatian and Serbian human translations. For testing, we used another selected set of *IMDb* reviews as well as a selected set of *Amazon* reviews. Neither of the test reviews has been investigated yet, and they will also be made publicly available.

1.1 Related work

A considerable amount of work in the Computational Linguistics/Natural Language Processing community has been done on processing user-generated content, mostly on sentiment analysis, but also on different aspects of machine translation (MT). Some papers investigate translating social media texts in order to map widely available English sentiment labels to a less supported target language (Balahur and Turchi, 2012, 2014).

Several researchers attempted to build parallel corpora for user-generated content in different language pairs in order to facilitate MT (Jehl et al., 2012; Ling et al., 2013; San Vicente et al., 2016), while (Banerjee et al., 2012) explored methods for domain adaptation. A recent JSALT Workshop[8] dealt with improving MT for messages (Messenger, WhatsApp), social media (Facebook, Instagram, Twitter), and discussion forums (Reddit). Evaluating MT outputs of user-generated content was the topic of several publications, too. Two important measures of overall quality, comprehensibility and fidelity, were investigated in (Roturier and Bensadoun, 2011) in order to compare different English-to-German and English-to-French MT systems for technical support forums, and automatic estimation of these two measures for English-to-French MT was investigated in (Rubino et al., 2013). Maintaining sentiment polarity in German-to-English MT of Twitter posts was explored in (Lohar et al., 2017, 2018). However, none of these publications explored translation of user reviews.

The first publication about MT for user reviews (Lohar et al., 2019) explored translating English *IMDb* reviews into Croatian and Serbian and reported results of both automatic and human evaluation. However, all the systems were trained on very small amounts of parallel data so that the reported performance was rather low. More experiments on the same *IMDb* reviews were carried out (Popović et al., 2020), however, still only small amounts of training data were used. Also, no results of any kind of human evaluation were reported.

In this work, different sizes of the training corpora were explored, including a large corpus consisting of all publicly available parallel data for the two language pairs. Two types of reviews are explored, *IMDb* and *Amazon*, and both automatic scores as well as results of human evaluation are reported. In addition, differences between the two types of reviews are examined in order to see whether all user reviews can be considered as a homogeneous genre.

2 Building 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). Previous work on the given two target languages (Popović et al., 2020) reported that multilingual sys-

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

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

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

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

[8] https://www.clsp.jhu.edu/workshops/19-workshop/improving-translation-of-informal-language/

tem which translates into both languages performs better than two separated bilingual systems. Therefore, all our systems are multilingual, built using the same technique as (Johnson et al., 2017; Aharoni et al., 2019), namely adding a target language label "SR" or "HR" to each source sentence. The amount of Croatian and Serbian data is balanced in all set-ups in order to achieve optimal performance for both target languages.

The systems operate on sub-word units generated by byte-pair encoding (BPE) (Sennrich et al., 2016b) with 32000 BPE merge operations both for the source and for the target language texts. We do not use shared vocabularies between the source and the target languages because they are distinct. On the other hand, we built a joint vocabulary for the two target languages because they are very similar.

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.

"Teacher/student" model As a first step, we built a system trained on all publicly available parallel data consisting of about 55 million sentences. These data, however, do not contain any user reviews. On the other hand, there is a vast amount of monolingual English user reviews, and in order to get use of it, we created a synthetic in-domain parallel corpus which is a widely used practice in NMT (Sennrich et al., 2016a; Zhang and Zong, 2016; Burlot and Yvon, 2018; Poncelas et al., 2018). We selected a set of about four million sentences from *Amazon* reviews originating from 14 different topics, and translated them by the system trained on out-of-domain data. In this way, we applied so-called "teacher/student" model, or "knowledge distillation" (Saleh et al., 2020; Chen et al., 2017; Kim and Rush, 2016). Knowledge distillation is the training of a smaller network (student) who learns from an already trained network (teacher). The idea is that the student will be performing much faster and hopefully approximately well as the teacher. The method is often used for reducing the amount of training data, to speed up the process, as well as for domain adaptation.

In our set-up, knowledge distillation is used for domain adaptation: the teacher model is the system trained on a large amount of out-of-domain parallel data. This system is used to create a small synthetic in-domain corpus, which is then used to train the student model.

"Advanced student" model The best option for using synthetic training corpora for NMT is not to use them alone, but to enrich "natural" parallel corpora. However, we do not have any natural in-domain parallel corpora. Yet, some parts of the large out-of-domain corpora might be more useful for translating reviews than others, especially subtitles which are usually informal spoken language. To explore this potential, we ranked out-of-domain sentences according to their similarity to user reviews, and extracted the most similar ones to combine them with the synthetic parallel corpus and train an "advanced student" model.

The details about all data sets and data selection are presented in the next section.

3 Data sets

3.1 User reviews

***IMDb* movie reviews**[9] (Maas et al., 2011) consist of about 10 sentences and 230 words on average. Each review is labelled with a score: negative reviews have a score<4 out of 10, positive reviews have a score>7 out of 10, and the reviews with more neutral ratings are not included.

In our experiments, *IMDb* reviews were used for development and testing, but not for training.

***Amazon* product reviews**[10] (McAuley et al., 2015) are generally shorter, consisting of 5 sentences and 93 words on average. Each review is labelled with a rating from 1 (worst) to 5 (best). The reviews are divided into 24 categories/topics/domains, and we used the reviews from the following 14 topics: "Beauty", "Books", "CDs and Vinyl", "Cell Phones and Accessories", "Grocery and Gourmet Food", "Health and Personal Care", "Home and Kitchen", "Movies and TV", "Musical Instruments", "Patio, Lawn and Garden", "Pet Supplies", "Sports and Outdoors", "Toys and Games", and "Video Games".

For our systems, *Amazon* reviews were used both for training as well as for testing, however

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

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

not for development. In order to obtain a balanced multi-target training corpus, half of the selected reviews from each of the topics were translated into Serbian and another half into Croatian.

3.2 Out-of-domain data

We used the publicly available OPUS[11] parallel data (Tiedemann, 2012) as out-of-domain data. The vast majority of these resources for the desired language pairs consists of *OpenSubtitles*, and there are also *SETIMES News*, *Bible*, *Tilde*, *EU-bookshop*, *QED*, and *Tatoeba* corpora. In addition, we used *GlobalVoices* for Serbian, and *hrenWac*, *TED* and *Wikimedia* for Croatian. In total, the corpus is well balanced over the two target languages.

3.3 Selected out-of-domain data

As mentioned in Section 2, we extracted a set of sentences from the out-of-domain subtitles according to their similarity to *Amazon* reviews. The subtitles were ranked using the Feature Decay Algorithm (FDA) (Biçici and Yuret, 2011, 2015; Poncelas et al., 2018; Poncelas, 2019). FDA selects sentences from a 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 diversity, the n-grams are penalised proportionally to the number of instances already present in Sel. During the execution of FDA, candidate sentences from the set S are selected one by one according to the following score:

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

The sentence s 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 each step, and stopping after enough sentences have been extracted.

For our experiment, the out-of-domain subtitles represent the set S, and the *Amazon* reviews are $Seed$. From the 4 million English review sentences selected for training, we selected 140,000 sentences as seed (about 10,000 from each of the topics). We then used this seed to extract the similar sentence pairs from English-Croatian and English-Serbian subtitles. For each target language, we selected the top 9 million sentence pairs, thus 18M balanced sentence pairs in total.

Table 1 shows number of sentences, running words and distinct words (vocabulary) in training, development and test sets, as well as contributions of each of the review types.

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

4 Experimental set-up

In order to systematically explore influence of different sizes and natures of training data, we built the following MT systems:

– GENERAL (teacher model): system trained on all publicly available out-of-domain parallel data.

– REVIEWS (student model): system trained on in-domain synthetic corpus consisting of original English *Amazon* reviews and their translations generated by the GENERAL system.

– REVIEWS+SELECTED (advanced student): system trained on combination of synthetic in-domain data and selected natural out-of-domain data. We investigated different amounts of selected data:

- REVIEWS+6M: adding 6 million selected out-of-domain sentences (3M for each target language)
- REVIEWS+12M: adding 12 million selected out-of-domain sentences (6M for each target language)
- REVIEWS+18M: adding all 18 million selected out-of-domain sentences (9M for each target language)

5 Results

5.1 Comparing MT systems

In order to get a quick feedback about each of our systems, we first evaluated them using the following three automatic overall evaluation scores: sacreBLEU (Post, 2018), chrF (Popović, 2015) and characTER (Wang et al., 2016).

The two best systems according to automatic scores, the “teacher" system GENERAL and the “advanced student" system REVIEWS+18M, were also evaluated by human annotators. The evaluators marked all words considered as adequacy errors, as described in (Popović, 2020), on a sub-set of about 200 sentences per system.

(a) training sets

training	general corpus	*Amazon* reviews	selected subtitles
sentences	55,556,238	3,956,785	18,000,000
running words en	564,781,595	69,737,751	217,466,330
hr+sr	468,039,263	/	180,782,847
vocabulary en	1,264,079	671,196	587,784
hr+sr	2,843,079	/	1,537,498

(b) development and test sets

	development			test		
	IMDb reviews			*Amazon* + *IMDb* reviews		
	en	hr	sr	en	hr	sr
sentences	485			1,170		
running words	8,530	7,510	7,607	16,861	14,594	14,985
vocabulary	2,456	3,193	3,201	3,807	5,400	5,366

(c) percentage of different reviews in different sets

review type	% of sentences in		
	train	dev	test
IMDb movies	0	100	29.7
Amazon products	100	0	70.3

Table 1: Data statistics: number of sentences, running words and distinct words (vocabulary) in training (a), development and test sets (b), and contribution (% of segments) of *IMDb* and *Amazon* reviews in training, development and test sets (c).

The results are presented in Table 2, and the tendencies are same for both target languages. As expected, the small synthetic in-domain corpus alone (REVIEWS) cannot achieve the same performance as the large out-of-domain corpus (GENERAL), however the difference in scores is not so large as could be expected considering the difference in the sizes (55M vs 4M) as well as the fact that the target part of the in-domain corpus is machine translated. Adding 6M of selected parallel sentences (REVIEWS+6M) slightly improves the performance, while additional 6M selected sentences (REVIEWS+12M) yield (and even slightly improve) the performance of the GENERAL "teacher" system. Adding 18M selected sentences (REVIEWS+18M) only slightly improves over the REVIEWS+12M system, and definitely outperforms the GENERAL "teacher" system. Since the improvements from 12M to 18M are rather small, we did not experiment with larger selected corpora.

We also present the scores for two on-line MT systems, AMAZON and GOOGLE, and it can be seen that our best two systems outperform both of them. Although their automatic scores are notably lower than the two best systems, they were also evaluated by human annotators in order to gather more annotations for comparing two different types of reviews which will be described in the next section.

Before moving to that, we will present a set of translation examples for the two best systems in Table 3. The first four sentences represent examples where the review-oriented "advanced student" system REVIEWS+18M performs better. In the sentence (1), the GENERAL system completely mistranslated the noun phrase "reddish brown hair", and in the sentences (2) and (3) it choose incorrect variant of ambiguous source words "characters" and "care". In the sentence (4), the word order is not optimal.

In sentences (5) and (6), REVIEWS+18M performed better on the first part of the sentence while GENERAL performed better on the second part. GENERAL failed to properly rephrase the first part of the sentence (5) and generated overly literal translation. In sentence (6), it choose incorrect variant of the ambiguous source word "great". On the other hand, REVIEWS+18M failed to properly

(a) English→Croatian

en→hr		development (*IMDb*)			test (*Amazon+IMDb*)			
system	size	BLEU ↑	chrF ↑	cTER ↓	BLEU ↑	chrF ↑	cTER ↓	human ↓
GENERAL	55M	31.6	57.4	39.1	30.6	57.0	39.9	14.2
REVIEWS	4M	26.2	53.7	42.7	26.3	54.2	41.3	/
REVIEWS+6M	10M	26.3	53.9	42.4	26.4	54.6	41.4	/
REVIEWS+12M	16M	31.7	58.0	39.2	30.7	57.2	39.5	/
REVIEWS+18M	22M	**32.1**	**58.2**	**39.0**	**31.4**	**57.8**	**38.8**	**12.6**
AMAZON	n.a.	30.9	57.6	**38.9**	29.7	56.7	39.0	18.3
GOOGLE	n.a.	28.6	55.7	40.6	26.6	53.0	43.8	17.4

(b) English→Serbian

en→sr		development (*IMDb*)			test (*Amazon+IMDb*)			
system	size	BLEU ↑	chrF ↑	cTER ↓	BLEU ↑	chrF ↑	cTER ↓	human ↓
GENERAL	55M	32.1	57.3	39.0	29.8	55.2	40.4	14.2
REVIEWS	4M	26.6	53.6	42.4	26.1	52.8	42.1	/
REVIEWS+6M	10M	27.2	54.0	42.2	26.2	52.9	42.3	/
REVIEWS+12M	16M	**31.9**	**57.6**	**38.2**	29.7	55.5	40.1	/
REVIEWS+18M	22M	**31.9**	**57.6**	38.4	**29.9**	**55.6**	**40.0**	**13.5**
AMAZON	n.a.	26.7	54.6	40.8	25.2	52.4	42.5	25.6
GOOGLE	n.a.	26.4	54.2	40.9	25.4	52.8	41.9	24.0

Table 2: Comparison of English→Croatian (a) and English→Serbian (b) systems trained on different texts by automatic evaluation scores: BLEU, chrF and characTER as well as by percentage of words marked as adequacy errors by human evaluators ("human").

disambiguate the word "review" in sentence (5) and omitted the preposition "of" in sentence (6). For sentences (7), (8) and (9), GENERAL performed well while REVIEWS+18M produced errors. In (7) and (8), it failed to rephrase properly, and in (9) to generate the correct variant of the ambiguous word "bean". Finally, both systems failed in translating noun phrases in sentences (10) and (11), although in different ways. In sentence (10), GENERAL generated a noun phrase with changed meaning (animals are cruel instead of someone being cruel to them) and REVIEWS+18M even left the word "cruelty" untranslated. In sentence (11), REVIEWS+18M failed in disambiguation of the word "poor", while GENERAL changed the meaning of the entire noun phrase into "charger with cell phones of poor quality".

5.2 Comparing *Amazon* and *IMDb* reviews

In order to compare the MT performance of two types of reviews, separated scores for joint target languages are presented in Table 4. The *reviews+18M* system shows the best results for both types of reviews, which means that the "knowledge distillation" in form of forward translation of *Amazon* reviews by the *general* system was helpful for both review types.

Furthermore, for all systems, automatic scores are notably better for *Amazon* product reviews than for *IMDb* movie reviews, indicating that *IMDb* is more difficult for machine translation. However, the tendencies of human scores are different, except for GOOGLE. For other systems (our two and AMAZON), the evaluators found less errors in *IMDb* than in *Amazon* reviews. Also, it has to be taken into account that *IMDb* reviewers were not used for training, only *Amazon* reviews, which can influence the results. More experiments with equal distributions in training and test sets should be carried out in future work.

After looking into errors marked by human evaluators in order to identify the most prominent error types (Popović, 2021), we found out that there are some differences in frequencies of certain error types, presented in Table 5. The largest difference can be seen for named entities, which are generally more frequent in *IMDb* reviews. Some types of errors are, however, more frequent in *Amazon* reviews, such as ambiguous words (words with different meanings in different contexts), gender er-

(1) source	Do not buy this unless you purposely want **reddish brown hair**.
reference	Ne kupujte ovo osim ako ciljano ne želite crvenkasto smeđu kosu.
GENERAL$^-$	Ne kupujte ovo, osim ako ne želite **rashlađenu** kosu **Reddisha**.
REVIEWS+18M$^+$	Ne kupujte ovo osim ako ne želite crvenosmeđu kosu.
(2) source	Boring **Characters**
reference	Dosadni likovi
GENERAL$^-$	Dosadni **karakteri**
REVIEWS+18M$^+$	Dosadni likovi
(3) source	Wonderful Skin **Care**
reference	Odlična nega kože
GENERAL$^-$	Predivna **briga za kožu**.
REVIEWS+18M$^+$	Predivna nega kože.
(4) source	This was a pretty dull movie, **actually**.
reference	Ovo je zapravo bio poprilično dosadan film.
GENERAL$^-$	Ovo je bio prilično dosadan film, **zapravo**.
REVIEWS+18M$^+$	Ovo je zapravo bio prilično dosadan film.
(5) source	**I had high hopes** for this product after reading all the wonderful **reviews**.
reference	Veliku nadu sam polagao u ovaj proizvod nakon čitanja svih tih divnih recenzija.
GENERAL^{-+}	**Imao sam velike nade za** ovaj proizvod nakon čitanja svih prekrasnih recenzija.
REVIEWS+18M^{+-}	Polagao sam velike nade u ovaj proizvod nakon čitanja svih divnih **kritika**.
(6) source	A **Great** Story. The Most Amazing Tale **of** Human Ingenuity and Creativity!
reference	Sjajna priča. Najneverovatnija pripovetka o ljudskoj dovitljivosti i kreativnosti!
GENERAL^{-+}	**Velika** priča. Najneverovatnija priča o ljudskoj genijalnosti i kreativnosti!
REVIEWS+18M^{+-}	Sjajna priča. Najneverovatnija priča **X ljudske** genijalnosti i kreativnosti!
(7) source	i don't like this kind of films, i feel like somebody is trying to **pull my leg**.
reference	ne volim ovakve filmove, osjećam se kao da me netko pokušava prevariti.
GENERAL$^+$	ne volim ovakve filmove, osjećam se kao da me netko pokušava prevariti.
REVIEWS+18M$^-$	ne sviđa mi se ova vrsta filmova, osjećam se kao da me netko pokušava **povući za nogu**.
(8) source	My sense is that it depends **to a large degree on** the dog.
reference	Imam utisak da dosta zavisi od samog psa.
GENERAL$^+$	Moj osećaj je da to mnogo zavisi od psa.
REVIEWS+18M$^-$	Moj osećaj je da to zavisi **od velikog stepena na** psa.
(9) source	I only recently discovered **vanilla bean** paste.
reference	Tek sam skoro otkrio pastu od zrna vanile.
GENERAL$^+$	Nedavno sam otkrio pastu od **X** vanile.
REVIEWS+18M$^-$	Nedavno sam otkrio pastu od vanile **i pasulja**.
(10) source	Horrifying **Animal Cruelty**
reference	Užasavajuća okrutnost prema životinjama
GENERAL$^-$	Zastrašujuća **životinjska okrutnost**
REVIEWS+18M$^-$	Užasna **Životinjska Cruelty**
(11) source	**Poor Quality Cell Phone Charger**
reference	Punjač mobitela loše kvalitete
GENERAL$^-$	Punjač **s lošim kvalitetnim mobilnim telefonima**
REVIEWS+18M$^-$	**Siromašni** punjač za mobitel

Table 3: Translation examples for the two best systems, GENERAL and REVIEWS+18M. Errors together with the corresponding English parts are marked in bold. For the first four sentences, REVIEWS+18M is better; for (5) and (6), the two systems exhibit errors in different parts of the sentence; for (7), (8) and (9), GENERAL is better; for (10) and (11), both systems fail at the same part of the sentence.

en→hr+sr review type	system	BLEU ↑	chrF ↑	cTER ↓	% of errors (human) ↓
Amazon products	GENERAL	57.8	68.7	26.5	15.1
	REVIEWS+18M	**58.2**	**69.2**	**26.1**	**14.0**
	AMAZON	56.6	67.9	26.8	21.4
	GOOGLE	56.5	67.6	27.7	19.4
IMDb movies	GENERAL	**49.1**	63.3	31.6	12.9
	REVIEWS+18M	48.9	**63.6**	**31.0**	**11.8**
	AMAZON	46.7	61.5	33.6	20.5
	GOOGLE	44.2	58.2	38.0	23.6

Table 4: Comparison of automatic scores and human evaluation for two different types of reviews: *Amazon* products and *IMDb* movies. The scores are calculated on the joint test set for both target languages. All automatic scores are better for *Amazon* product reviews than for *IMDb* movie reviews, while the situation is different for human evaluation.

error type (%)	*IMDb*	*Amazon*
named entity	6.7	**2.8**
ambiguous word	**10.9**	12.9
gender	**1.8**	3.4
untranslated	**0.9**	2.5
non-existing word	**0.7**	1.6

Table 5: Different error types in *IMDb* and *Amazon* user reviews; the largest difference can be noted for named entity errors, which are especially frequent in *IMDb*.

rors, untranslated words (English words copied into translation) as well as non-existing words (which do not exist either in the source or in the target language).

All these results indicate that there are differences between different types of reviews so that user reviews generally do not represent a homogeneous genre. However, the analysis is carried out on relatively small amount of data, especially human evaluation, so that it is not yet possible to draw any conclusions about the nature of these differences. Further analysis on more data as well as detailed analysis of different review topics including more review types (such as hotel reviews from Trip Advisor) should be carried out in future work.

6 Summary and outlook

This work investigates machine translation of two types of user reviews, *IMDb* movie reviews and *Amazon* product reviews, from English into Serbian and Croatian.

Since one of the main challenges for MT of user reviews is lack of parallel in-domain training data, we explored a possibility to make use of large out-of-domain bilingual parallel corpora as well as monolingual in-domain English corpora. We trained a general “teacher" system on all out-of-domain data and then used this system to create a small synthetic in-domain parallel corpus by translating English *Amazon* reviews into the target languages. Both automatic scores and human evaluation show that using this synthetic in-domain corpus together with a selected sub-set of out-of-domain data is the best option.

The results on separated *IMDb* and *Amazon* reviews indicate that MT systems perform differently on different review types so that user reviews generally should not be considered as a homogeneous genre. However, evaluating and training on larger amount of different reviews covering different domains/topics is needed to identify the nature of differences between different types of reviews, and also influence of different topics. Another direction of future work should include using more in-domain data, as well as other techniques for domain adaptation.

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.

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

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

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

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.

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.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 18)*, pages 1925–1935, Vancouver, Canada.

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.

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

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.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 16)*, pages 1317–1327, Austin, Texas.

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.

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

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

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

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.

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

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

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)*, pages 249–258, 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 (WMT 2015)*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online).

Maja Popović. 2021. On nature and causes of observed mt errors. In *Proceedings of the MT Summit 2021*, Online.

Maja Popović, Alberto Poncelas, Marija Brkić, and Andy Way. 2020. Neural machine translation for translating into Croatian and Serbian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2020)*, pages 102–113, Barcelona, Spain (Online).

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.

Johann Roturier and Anthony Bensadoun. 2011. Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the MT Summit XIII*, Xiamen, China.

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

Fahimeh Saleh, Wray Buntine, and Gholamreza Haffari. 2020. Collective wisdom: Improving low-resource neural machine translation using adaptive knowledge distillation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 20)*, Barcelona, Spain (Online).

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

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.

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.