Simplify-then-Translate: Automatic Preprocessing for Black-Box Translation

Sneha Mehta[1], **Bahareh Azarnoush**[2], **Boris Chen**[2], **Avneesh Saluja**[2], **Vinith Misra**[2], **Ballav Bihani**[2], **Ritwik Kumar**[2]
[1]Department of Computer Science, Virginia Tech, VA
[2]Netflix Inc., CA
snehamehta@vt.edu, {bazarnoush, bchen, asaluja, vmisra, bbihani, ritwikk}@netflix.com

Abstract

Black-box machine translation systems have proven incredibly useful for a variety of applications yet by design are hard to adapt, tune to a specific domain, or build on top of. In this work, we introduce a method to improve such systems via automatic pre-processing (APP) using sentence simplification. We first propose a method to automatically generate a large in-domain paraphrase corpus through back-translation with a black-box MT system, which is used to train a paraphrase model that "simplifies" the original sentence to be more conducive for translation. The model is used to preprocess source sentences of multiple low-resource language pairs. We show that this preprocessing leads to better translation performance as compared to non-preprocessed source sentences. We further perform side-by-side human evaluation to verify that translations of the simplified sentences are better than the original ones. Finally, we provide some guidance on recommended language pairs for generating the simplification model corpora by investigating the relationship between ease of translation of a language pair (as measured by BLEU) and quality of the resulting simplification model from back-translations of this language pair (as measured by SARI), and tie this into the downstream task of low-resource translation.

1 Introduction

Modern translation systems built on top of a sequence transduction approach (Sutskever, Vinyals, and Le, 2014; Bahdanau, Cho, and Bengio, 2014) have greatly advanced the state and quality of machine translation (MT). These systems generally rely on the availability of large-scale parallel corpora, and while unsupervised (Lample et al., 2017) or semi-supervised (Saluja et al., 2014) approaches are a popular area of research, production-grade translation systems still primarily leverage bitexts when training. Efforts such as WMT[1] provide such corpora for select language pairs, which has enabled neural MT systems to achieve state-of-the-art performance on those pairs. However, low resource MT (language pairs for which parallel data is scarce) remains a challenge.

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

In this work, we focus on improving translation quality for low-resource translation (i.e., from English *into* a low-resource language) in the *black-box MT* (BBMT) setting - namely a system which has been trained and tuned *a priori* and for which we cannot access the model parameters or training data for fine-tuning or improvements. Examples of such systems include those provided by commercial vendors e.g., Google Translate[2] or Microsoft Translator[3]. While some provide the option of fine-tuning on domain-specific data under certain conditions, how to improve the performance of such black-box systems on domain-specific translation tasks remains an open question. We investigate methods to leverage the BBMT system to preprocess input source sentences in a way that preserves meaning and improves translation in the target language. Specifically, a large-scale parallel corpus for English simplification is obtained by back-translating (Sennrich, Haddow, and Birch, 2016a) the reference translations of several high-resource target languages. The resulting parallel corpus is used to train an Automatic Preprocessing model (APP) (§ 2), which transforms source sentence into a form that preserves meaning while also being easier to translate into a low-resource language. In effect, the APP model attacks the longstanding problem of handling complex idiomatic and non-compositional phrases (Lin, 1999; Sag et al., 2002), and simplifies these expressions to more literal, compositional ones that we hypothesize are easier to translate.

We use the APP model to simplify the source sentences of a variety of low-resource language pairs and compare the performance of the black-box MT system on the original and simplified sentences. Note that only one APP model needs to be trained per source language and this model can be applied to a variety of low-resource language pairs as long as the source language is the same. In our study we focus on the domain of conversational language as used in dialogues of TV shows. We picked this domain since here language tends to be colloquial and idiomatic. We empirically show improvement in translation quality in this domain across a variety of low resource target languages (§3). This improvement is further corroborated with side-by-side human

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

[3]https://www.bing.com/translator

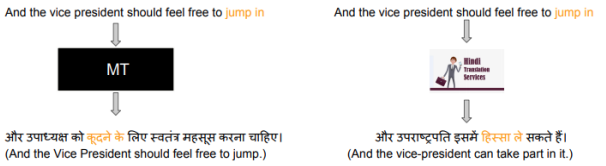
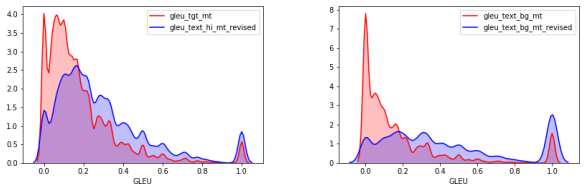


Figure 1: Machine vs. Human Translation



(a) English-Hindi (b) English-Bulgarian

Figure 2: Sentence-level GLEU for direct translations and back-translations. The red density curve represents the distribution of GLEU scores obtained by the direction translation direction and the blue density curve represents the distribution of GLEU scores obtained by using back-translations.

evaluations (§4.2) and evaluating on post-editing efficiency (§4.1). Lastly, we perform an empirical analysis to probe further into which high-resource language pairs should be selected to obtain a good quality simplification corpus for a given language, before discussing connections with related work (§5) and concluding.

2 APP with Back Translations

Consider the example in Fig. 1. The source sentence "The vice president should feel free to jump in" has been translated by Google Translate [4] incorrectly to Hindi as "Vice President should feel free to jump inside". The system was unable to correctly translate the idiomatic and non-compositional phrase "jump in", which in this context means "intervene" or "get involved", and instead translated it literally. An expert human Hindi translator would take these idiomatic expressions into account when generating the Hindi translation. Indeed, when back-translating the reference translation into English we obtain "The Vice President should feel free to take part" (in the conversation). Such instances where MT systems incorrectly translate sentences containing phrases, idioms or complex words for low-resources (i.e. with smaller training sets) languages pairs are fairly common.

In other words, the back-translation is different in meaning than the natural source sentence. This problem is prevalent even when these BBMT systems are fine-tuned on domain-specific data and is exacerbated when dealing with low-resource language pairs, simply because the paucity of data does not allow the MT models to infer the translations of the myriad of phrases and complex words. For instance, in the example above 'jump in' was interpreted compositionally as 'jump' + 'in'. To generate better quality or even acceptable translations, it is imperative to simplify such complex sentences into simpler forms while still preserving meaning.

Using automated models to simplify such sentences is a well-studied problem. Though, when it comes to the ultimate task of domain-specific translation, it is not entirely clear what data is best suited to train such simplification models. Open source datasets like WikiLarge (Zhang and Lapata, 2017) or Simple PPDB (Zhao et al., 2018) are good options to explore, but the domain mismatch and dataset size may pose a challenge. In particular, WikiLarge dataset contains 296K sentence pairs of descriptive text while our domain of interest in this study is conversational dialogues from TV Shows and movies. Collecting a large amount of domain-specific simplification data could be prohibitive, forcing one to consider alternatives when constructing their simplification models.

To address this problem we make use of the observation that translating back-translations is easier than translating naturally occurring source sentences, which has been corroborated by numerous studies (Graham, Haddow, and Koehn, 2019; Zhang and Toral, 2019). Consider a set of around 30K uniformly-sampled sentence pairs from the English-Bulgarian (En-Bg) subtitles corpus appearing on TV shows and movies from a subscription video-on-demand provider[5]. The BLEU score in the natural or direct direction (En $\rightarrow$ Bg) is only 10.20, whereas when following the reverse direction (Bg $\rightarrow$ En $\rightarrow$ Bg) and translating back-translations instead of original source sentences, the BLEU score dramatically improves to 33.39. Probing a little deeper, Fig. 2 shows the distributions of sentence-level GLEU scores (Wu and others, 2016) for two language pairs: English-Bulgarian (right) and English-Hindi (left). We observe the trend that GLEU scores have generally improved when using back-translations. The area where the blue curve dominates the red curve can be considered as the 'scope-of-simplification'.

Thus it seems that human reference translations when back-translated to the original language (in this case English) are a rich source of simplifications (e.g. "jump in" is simplified to "take part"). This observation leads to two immediate corollaries - 1) by back-translating the ground truth human translations to the source language we obtain a (perhaps noisy) simplified version of the original source, and 2) we can learn a function to map the source sentences to their simplified forms by training a sequence-to-sequence (S2S) model from the aforementioned generated parallel corpus. We term the resulting simplification model an **Automated Preprocessing** (or APP) model.

We formalize our APP model as a S2S model from source sentences in one language to back-translations of ground truth sentences i.e., *translationese* targets in the same language. The *translationese* targets can be obtained from multiple high-resource language pairs using black-box MT sys-

[4] This specific translation is observed in translate.google.com as of September 5, 2019.

[5] www.netflix.com

tems and the trained APP model can be applied to a variety of low resource language pairs. Let (X^i, Y^i) be the i^{th} training bitext corpus (X is source, Y is target) with source language s_i and target language t_i for $i \in \{1, \ldots, M\}$, where M is the number of training language pairs, and let $j \in \{M+1, \ldots, M+N\}$ refer to the N test bitext corpora. Note that for the experiments in this paper, s_i is fixed to English $\forall i \in \{1, \ldots, M+N\}$ so we simply refer to it as s. The APP procedure is as follows:

- Obtain back-translations of target train sets Y^i for $i = 1$ to M to language s_i given by $T^1, T^2, \ldots, T^M$ using BBMT models $MT_{t_i \to s} \forall i$.
- Train an APP simplification model f_{APP} on the combined parallel corpus $\bigcup_{i=1}^{M} \{(X^i, T^i)\}$.
- At test time, preprocess the source X_s^j for each test language pair j using the trained APP model to obtain the simplified source X^{j*}, where

$$X^{j*} = f_{APP}(X^j) \tag{1}$$

- Translate the simplified source using the BBMT model for the j^{th} test language pair.

$$\hat{Y}^{j*} = MT_{s \to t_j}(X^{j*}) \tag{2}$$

APP provides in-domain simplification bitext at scale and from the same BBMT system that we eventually use to translate into low-resource languages, thus providing a more flexible solution than using precompiled simplification corpora. In the next sections, we compare the performance of BBMT system outputs with and without APP simplifications i.e. $\hat{Y}^j* \; \& \; \hat{Y}^j$ respectively. Further, we compare the APP models trained on in-domain vs out-of-domain corpora.

3 Evaluation

We first compare the in-domain APP model to a S2S model trained on the WikiLarge corpus, by evaluating downstream translation performance on low-resource languages, followed by an evaluation based on human judgements and post-editing efficiency. These experiments are conducted on subtitles dataset from a subscription video-on-demand provider. We also validate the approach on another subtitles dataset and verify that the improvements we see in the first set of experiments generalize to other corpora in the same domain. In all of our experiments we used the Google Translate BBMT system.

3.1 Datasets and Metrics

FIGS This dataset comes from subtitles appearing on 12,301 TV shows and movies from a subscription video-on-demand provider. Titles include but are not limited to: "How to Get Away with Murder", "Star Trek: Deep Space Nine", and "Full Metal Alchemist". We take four high-resource language pairs namely: English-**F**rench (1.3 million parallel subtitle events i.e., sentence pairs), English-**I**talian (1.0M), English-**G**erman (1.2M) and English-**S**panish (6.5M). We collectively refer to this dataset as FIGS. We use the APP simplification procedure to obtain English simplification parallel corpora resulting in 9.5M subtitle events i.e., sentence pairs. This dataset contains short sentences with an average length of 7. For evaluation, we pick 7 low-resource language pairs namely: English-Hungarian (En-Hu), English-Ukrainian (En-Uk), English-Czech (En-Cs), English-Romanian (En-Ro), English-Bulgarian (En-Bg), English-Hindi (En-Hi), and English-Malay (En-Ms). The test set statistics for each dataset is given in Table 1. We refer to the APP model trained on this dataset as 'FigsAPP'.

Table 1: FIGS test set statistics

Language	#sentences
En-Hu	27,393
En-Uk	30,761
En-Cs	35,505
En-Ro	47,054
En-Bg	30,714
En-Hi	23,496
En-Ms	11,713

Wikilarge The WikiLarge dataset (Zhang and Lapata, 2017) was compiled by using sentence alignments from other Wikipedia-based datasets (Zhu, Bernhard, and Gurevych, 2010; Woodsend and Lapata, 2011; Kauchak, 2013), and contains 296K instances of 1-to-1 and 1-to-many alignments. This is a widely used benchmark for test simplification tasks. The train split contains 296K sentence pairs and the validation split contains 992 sentence pairs. We call the simplification model trained on this dataset as 'Wiki-APP'.

Open Subtitles The Open Subtitles dataset (Lison and Tiedemann, 2016) is a collection of translated movie subtitles obtained from opensubtitles.org[6]. It contains 62 languages and 1,782 bitexts. We first train two MT models using two high resource language pairs (Es-En and Ru-En) using the Transformer architecture (Vaswani et al., 2017). Then using the MT models above we train two APP models obtained from the same language pairs English-Spanish (En-Es) and English-Russian (En-Ru). We sample 5M sentence pairs each for training MT and APP models from the corresponding Open Subtitles corpora and filter out short ($length < 3$) and long ($length > 50$) sentence pairs. Note that training sets for both MT and APP models are disjoint. For evaluation, we pick the following six language pairs: three randomly picked pairs English-Armenian (En-Hy), English-Ukrainian(En-Uk) and English-Bulgarian (En-Bg) and three pairs in which the target language is similar to Spanish including English-Catalan (En-Ca), English-Portuguese (En-Pt) and English-Romanian (En-Ro). We sample 50,000 pairs from the low-resource test bitexts and filter out pairs with length less than 3 and greater than 50. We call the APP model obtained from En-Es dataset as 'OS-EsAPP' and from En-Ru dataset as 'OSRuAPP'.

Turk and PWKP To test the performance of APP models obtained from different language pairs (§ 4.3) we pick two

[6] http://www.opensubtitles.org/

open source simplification datasets. The first dataset is the Turk (Xu et al., 2016) dataset which contains 1-to-1 alignments focused on paraphrasing transformations, and multiple (8) simplification references per original sentence (collected through Amazon Mechanical Turk). To evaluate the performance on this dataset we use the SARI metric they introduced. The next dataset we use is the test set of the PWKP dataset (Zhu, Bernhard, and Gurevych, 2010). This dataset contains only 1-to-1 mapping between source and reference and hence we use the BLEU metric to evaluate the simplification performance.

Metrics We evaluate the translation performance of the BBMT system using the commonly used BLEU metric (Papineni et al., 2002). For subtitle generation, expert linguists post-edit subtitles at the event or dialog level, hence it is useful to look at the impact that simplification brings at the sentence-level, motivating the choice of sentence-level GLEU (Wu and others, 2016), which has been shown to be better correlated with human judgements at the sentence-level as compared to sentence-level BLEU. Furthermore, we compute the normalized edit (Levenshtein) distance between a translation output and human reference translation also known as translation error rate (TER), which has been shown to correlate well with the amount of post-editing effort required by a human (Snover et al., 2006). This metric provides yet another way to evaluate the quality of translation and completes our comprehensive suite of metrics.

3.2 Implementation

For training the APP simplification model we use the Transformer architecture (Vaswani et al., 2017) through the tensor2tensor[7] library. We also evaluate BLEU using the implementation in that library and report the uncased version of the metric. For computing the TER score we use the implementation provided by (Snover et al., 2006) [8].

All experiments are based on the transformer base architecture with 6 blocks in the encoder and decoder. We use the same hyper-parameters for all experiments, i.e., word representations of size 512 and feed-forward layers with inner dimension 4096. Dropout is set to 0.2 and we use 8 attention heads. Models are optimized with Adam (Kingma and Ba, 2014) using $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e - 9$, with the same learning rate schedule as Vaswani et al. (2017). We use 50,000 warmup steps. All models use label smoothing of 0.1 with a uniform prior distribution over the vocabulary. We run all experiments using machines with 4 Nvidia V100 GPUs. We use a sub-word vocabulary of size 32K implemented using the word-piece algorithm (Sennrich, Haddow, and Birch, 2016a) to deal with out-of-vocabulary words and the open vocabulary problem in S2S language models.

4 Results

Table 2 compares the performance of APP models trained on FIGS (in-domain) dataset and the WikiLarge (out-of-domain) datasets on the FIGS test set using the BBMT system. The values in columns 1-3 indicate the BLEU score after translating the original sentence and after simplifying using the FigsAPP and WikiAPP models respectively. There is uniform improvement across all languages when using the FigsAPP model, ranging from 3.5% (relative) for En-Uk to 11.6% for En-Bg. On the other hand, performance degrades significantly on all languages when simplified using a model trained on the WikiLarge dataset. Fig. 3 shows the distribution of sentence-level GLEU for each target language in the FIGS test set. Mean GLEU increases for En-Hu, En-Ro, En-Ms and En-Bg.

Table 2: In-domain vs out-of-domain simplification performance. It is evident that APP models trained on an out-of-domain simplification corpus (WikiLarge) degrades performance, whereas in-domain simplification corpora (FIGS) boosts performance.

language pair	original	FigsAPP	WikiApp
En-Hu	17.69	**18.92**	11.86
En-Uk	17.57	**18.18**	12.51
En-Cs	21.62	**22.35**	16.71
En-Ro	26.08	**27.98**	21.56
En-Bg	14.9	**16.63**	12.39
En-Hi	14.45	**15.53**	11.39
En-Ms	19.14	**20.37**	13.00

Table 3: Translation performance (BLEU) before and after of OSEsAPP and OSRuAPP models on six low resource target language pairs from the Open Subtitles corpus.

language pair	original	OSEsAPP	OSRuAPP
En-Ca	27.25	**27.84**	23.36
En-Hu	**7.05**	6.28	5.79
En-Pt	25.11	**25.5**	22.28
En-Ro	**26.18**	25.03	22.40
En-Uk	11.73	**11.77**	11.61
En-Bg	23.71	**24.68**	23.90

Table 3 shows the results of APP on the Open Subtitles dataset. It can be noted that the performance improves for Catalan (ca) and Portuguese (pt) which are languages similar to the language used for training the simplification corpus. Additionally for Bulgarian an improvement of 4.1% can be observed. Moreover, the performance of OSEsAPP is better than the performance of OSRuAPP. This can be attributed to the fact that En-Ru is a harder language pair to translate than En-Es and hence the simplification model obtained from En-Ru is of worse quality than En-Es. We further elaborate this point in section § 4.3.

4.1 Post-Editing Efficiency

We also observe that TER decreases for all languages, which is intuitive to understand because the APP simplification brings the sentences closer to their literal human translation. Table 4 shows TER score for the FIGS test corpora for translating into seven low resource languages, before and after applying the APP simplification. As can be seen, TER decreases for all languages after simplification with a reduction of 6.9%, 6.9% and 7.2% for target languages Hungarian,

[7]https://github.com/tensorflow/tensor2tensor

[8]http://www.cs.umd.edu/ snover/tercom/

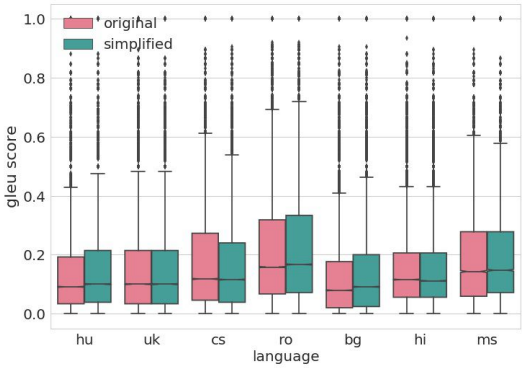


Figure 3: Sentence-level GLEU scores

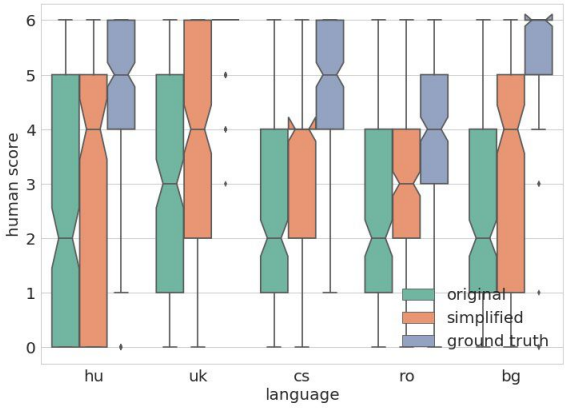


Figure 4: Human Evaluation Results

Table 4: Translation Edit Rate (TER) score of translations before and after applying the APP simplification for test language-pairs from the FIGS dataset. 'original' and 'simple' columns show the TER for translations before and after APP where the last column indicates the percentage decrease in TER.

language	original	simple	%Δ
En-Hu	0.86	0.80	-7%
En-Uk	0.74	0.70	-5.4%
En-Cs	0.78	0.77	-1.3%
En-Ro	0.72	0.67	-6.9%
En-Bg	0.83	0.77	-7.3%
En-Hi	0.62	0.60	-3.3%
En-Ms	0.76	0.72	-5.3%

Romanian and Bulgarian respectively. The reduction in TER correspondingly translates to a reduction in post-editing effort required by translators using the BBMT system as an assistive tool.

4.2 Human Evaluation

Simplified sentences with worse GLEU than their baseline non-simplified counterparts might not necessarily be of worse quality; rather, they may just be phrased differently than the reference sentence. We thus perform a side-by-side human evaluation to verify if APP-simplified translations improve MT quality, which allows us to assess via human evaluation these supposedly worse translations and validate translations with GLEU improvements at the same time. For this purpose, we restricted evaluation to five languages from the FIGS test set (hu, uk, cs, ro, and bg) and sampled 100 sentences from the fraction of sentences for which $\Delta GLEU > 0.4$ and 100 sentences from the sentences for which $\Delta GLEU < 0$, where $\Delta GLEU$ for one sentence pair (x, y) is defined as;

$$\Delta GLEU = GLEU(MT(x^*), y) - GLEU(MT(x), y) \quad (3)$$

and x^* is the simplified source sentence. For each language we show the source sentence, the BBMT output of the original sentence, the BBMT output of the simplified sentence, and the ground truth human translation per language side-by-side to expert linguists. We ask them to rate the quality of the three translations (original BBMT, simplified BBMT and human translation) according to the scale used by Wu and others (2016) described in Table 5. Since translation is generally easier for shorter sentences, in order to get a representative sample of challenging sentences we only selected sentences with more than 4 tokens for this study.

Table 5: Human evaluation ratings and their description

Rating	Description
0	completely nonsense translation
2	the sentence preserves some of the meaning of the source sentence but misses significant parts
4	the sentence retains most of the meaning of the source sentence, but may have some grammar mistakes
6	perfect translation: the meaning of the translation is completely consistent with the source, and the grammar is correct

Fig 4 displays results from the human evaluation, with corresponding values stated in Table 6. The green box represents scores from translations of original sentences, the orange box represents scores from translations from simplified sentences and the blue box represents the ground truth human translation. As expected the human translation has the highest score. Also worth noting is the jump in score intervals between original and simplified translations. For Hungarian, Czech, and Bulgarian median scores jump from $2 \rightarrow 4$, with improvements for Ukrainian ($3 \rightarrow 4$) and Romanian ($2 \rightarrow 3$) as well. Thus at least for these languages we can conclude that APP simplification results in improved translation output, as determined by expert human translators. This concurs with the initial observation e.g., for Bulgarian in Fig. 2, where the number of sentences with score of 0.0 are squashed and their scores shifted higher. This

means simplification can improve upon erroneous or bad-quality translations. Table 6 shows the mean human scores for original, simplified and reference translations as well as percentage of sentences for which human score improved, worsened and remained the same after simplification for all 200 sentences per language.

Table 6: Human Evaluation statistics for the FIGS test language pairs. 'Original Mean', 'Simple Mean' and 'Human Mean' represent the mean human scores before, after APP and ground truth human scores respectively. The next three columns indicates the percentage of samples with improved (% +ve), worse (% -ve), and same (% same) performance after simplification.

lang	Original Mean	Simple Mean	Human Mean	% +ve	% -ve	% same
Hu	2.52	3.11	4.45	38.5%	18.5%	43%
Uk	3.02	3.61	5.8	43%	15%	42%
Cs	2.43	3.12	4.77	49%	16.5%	34.5%
Ro	2.56	3.0	3.91	41.5%	25.5%	33%
Bg	2.34	3.33	5.365	51%	22%	27%

4.3 Ease of Translation vs Simplification Quality

We further seek to investigate the relationship between the translation language pair used to generate the simplification corpus and the quality of the resulting simplification model. To this end, we train translation systems on two language pairs: English-Spanish (En-Es) and English-Russian (En-Ru). The En-Es pair is an *easier* language pair to translate than En-Ru as reflected by the BLEU scores of the SOTA MT systems on these pairs. We pick an En-Es system trained to reach a BLEU of 42.7 and an En-Ru system trained to a BLEU of 34.23. We hypothesize that the APP models resulting from an easier language pair will be of better quality because it is easier to generate a good quality parallel simplification corpus. To test this hypothesis, we train simplification models obtained from the En-Es and En-Ru translation pairs and test the simplification performance on two standard simplification test sets.

Table 7 presents the results of this experiment, specifically the performance of the simplification models trained via automatically-generated simplification corpora obtained from the En-Es and En-Ru OpenSubtitles models. We also test the performance of an in-domain simplification model trained on the WikiLarge dataset. As expected, the performance of the in-domain model exceeds the performance of the models trained on En-Es and En-Ru datasets. It is also interesting to note that the En-Es model outperforms the En-Ru models on both datasets, underlining our hypothesis that easier language pairs result in better APP models. This idea can be used to inform the high-resource language pair(s) to pick for training an APP model for a target language pair. For instance, to simplify English it would be better to pick a high-resource pair which is easier to translate from English, e.g. Spanish. While simplifying Catalan(ca), it would be good to pick a translation pair like Catalan-Spanish which would be easier to translate than Catalan-English.

Table 7: Simplification performance of APP models trained on corpora generated by En-Es and En-Ru datasets on Turk and PWKP open source datasets.

(Metric) Dataset	Wiki	En-Es	En-Ru
(SARI) Turk	31.9	26.7	23.2
(BLEU) PWKP	53.5	32.7	23.9

Table 8: Positive and negative qualitative examples of simplifications brought about by APP.

Original:	I still, think **you're nuts**, but not as nuts as I thought
Simple:	I still think **you're crazy**, but not as crazy as I thought
Original:	in a town **only five miles** from Kabul.
Simple:	in a city **eight kilometers** from Kabul.
Original:	This case is **so far over your head**, it'd make your nose bleed.
Simple:	This case is **so complicated** that it would bleed your nose.
Original:	When I was **marooned** here, my first meal was a pheasant.
Simple:	When I was **stranded** here, my first meal was a pheasant.
Original:	*She* jumped from the window of Room 180.
Simple:	*He* jumped out of the window of Room 180.

4.4 Qualitative Analysis

Here we provide qualitative examples of simplifications generated by the APP approach and how it helps in improving BBMT performance. Table 8 shows some example simplifications from the FIGS test datasets. Phrases highlighted in bold were converted to 'simpler' phrases. In the first example "you're nuts" was replaced by a non-idiomatic/simpler phrase "you're crazy". In the second example, distance of five miles was almost accurately converted from imperial to the metric system whereas in the third example non-compositional phrase "so far over your head" was translated into the compositional phrase "so complicated". In the next example, an infrequent word "marooned" was replaced by its more frequent counterpart "stranded". Finally, the last example shows the kinds of errors introduced by the APP model where it occasionally replaces pronouns like 'it', 'she' by 'he'.

Table 9 gives examples of how APP simplifications can help the BBMT systems make fewer errors. The first example shows a sample from English-Romanian translation pair. Direct translation of the source makes the BBMT system incorrectly translate "fixating" as "fixing" (as observed in the backtranslation of the BBMT output) where as simplifying "fixating on" as "thinking about" produces a more meaningful translation. Similarly in the second example from the English-Catalan pair of the Open Subtitles corpus, APP replaces the colloquial word 'swell' by its meaning 'great' and hence results in a translation that is identical to the reference.

5 Related Work

Automatic text simplification (ATS) systems aim to transform original texts into their lexically and syntactically simpler variants. In theory, they could also simplify texts at the

Table 9: Qualitative examples of how APP simplification can help mitigate BBMT errors on the FIGS and Open Subtitles datasets. Here x is the input source.

x:	If only we can **stop fixating** on the days
BBMT(x):	Dacă numai putem opri fixarea pe zile
APP(x):	If only we could **stop thinking** about the days
BBMT(APP(x)):	Dacă am putea să nu ne mai gândim la zile
Reference:	Trebuie să nu ne mai gândim la zile
x:	Another **swell** party, Jay.
BBMT(x):	Una altra festa de l'onatge, Jay.
APP(x):	Another **great** party, Jay.
BBMT(APP(x)):	Una altra gran festa, Jay.
Reference:	Una altra gran festa, Jay.

discourse level, but most systems still operate only on the sentence level. The motivation for building the first ATS systems was to improve the performance of machine translation systems and other text processing tasks, e.g. parsing, information retrieval, and summarization (Chandrasekar, Doran, and Srinivas, 1996). It was argued that simplified sentences (which have simpler sentential structures and reduced ambiguity) could lead to improvements in the quality of machine translation (Chandrasekar, Doran, and Srinivas, 1996) A large body of work, since then has investigated text simplification for machine translation and found that this approach can improve fluency of the translation output and reduce technical post-editing effort (Štajner and Popovic, 2016).

Researchers have attempted to build simplification systems for different languages such as English, Spanish, (Saggion et al., 2015), and Portuguese (Aluísio et al., 2008). Wubben, van den Bosch, and Krahmer (2012) use phrase-based machine translation for sentence simplification based on the PWKP dataset (Zhu, Bernhard, and Gurevych, 2010). However, these systems are modular, rule-based (Poornima, Dhanalakshmi, and Soman, 2011), limited by data or language specific.

End-to-end simplification which is more similar to our work also has been studied by applying RNN-based sequence-to-sequence models to the PWKP dataset or transformer based models that are integrated with paraphrase rules (Zhao et al., 2018) and trained on English to English parallel simplification corpora and are current state-of-the-art. These methods are limited by the availability of parallel simplification corpora and especially the ones that can adapt to new domains. We propose a general framework that can be used to collect large-scale data for any language to train in-domain end-to-end data-driven lexical simplification systems.

Our work capitalizes on the observation that synthetically-generated source sentences resulting from reversing the translation direction on a parallel corpus yield better translations. These "back-translations" (Sennrich, Haddow, and Birch, 2016b; Poncelas et al., 2018) can augment relatively scarce parallel data with by translating the plentiful target monolingual data to the source language.

Various methods have been explored to improve low-resource translation. (Zoph et al., 2016) transfer parameters from an MT model trained on a high-resource language pair to low-resource language pairs and observe performance improvement. To improve performance on spoken language domain, researchers have finetuned state-of-the-art models trained on domains in which data is more abundant (Luong and Manning) whereas others have used data augmentation techniques (Fadaee, Bisazza, and Monz, 2017) to bring improvements in low-resource translation. Above approaches assume access to the underlying MT system whereas we assume a black-box scenario.

6 Conclusion

In this work we introduced a framework for generating a large-scale parallel corpus for sentence simplification, and demonstrated how the corpus can be used to improve the performance of black-box MT systems (especially on low-resource language pairs) and increase the post-editing efficiency at the subtitle-event i.e., sentence level. Moreover, we perform thorough empirical analysis to give insights into language pairs to select for simplifying a given language. Our results suggest that easier a language pair to translate, the better the simplification model that will result.

It should be noted that even though this work mainly focuses on simplification of English, our method is general and can be used to automatically generate simplification parallel corpora and thus data-driven simplification models using state-of-the-art architectures for any given language. Moreover, it accommodates collecting multiple reference simplifications for a given source sentence by leveraging open-source multilingual corpora. Using the insight that translating multiword expressions and non-compositional phrases is hard and simplifying these expressions before translating helps, our work merges two important sub-fields of NLP (machine translation and sentence simplification) and paves the path for future research in both of these fields.

References

Aluísio, S. M.; Specia, L.; Pardo, T. A.; Maziero, E. G.; and Fortes, R. P. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the Eighth ACM Symposium on Document Engineering*, DocEng '08, 240–248. New York, NY, USA: ACM.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Chandrasekar, R.; Doran, C.; and Srinivas, B. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, 1041–1044. Association for Computational Linguistics.

Fadaee, M.; Bisazza, A.; and Monz, C. 2017. Data augmentation for low-resource neural machine translation. In *ACL*, 567–573. Vancouver, Canada: Association for Computational Linguistics.

Graham, Y.; Haddow, B.; and Koehn, P. 2019. Translationese in machine translation evaluation. *CoRR* abs/1906.09833.

Kauchak, D. 2013. Improving text simplification language modeling using unsimplified text data. In *ACL*, 1537–1546.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lin, D. 1999. Automatic identification of non-compositional phrases. In *ACL*, 317–324.

Lison, P., and Tiedemann, J. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. *European Language Resources Association*.

Luong, M.-T., and Manning, C. D. Stanford neural machine translation systems for spoken language domains.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318. Association for Computational Linguistics.

Poncelas, A.; Shterionov, D.; Way, A.; Wenniger, G. M. d. B.; and Passban, P. 2018. Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*.

Poornima, C.; Dhanalakshmi, V.; and Soman, K. 2011. Rule based sentence simplification for english to tamil machine translation system. *International Journal of Computer Applications* 25(8):38–42.

Sag, I. A.; Baldwin, T.; Bond, F.; Copestake, A. A.; and Flickinger, D. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, 1–15.

Saggion, H.; Štajner, S.; Bott, S.; Mille, S.; Rello, L.; and Drndarevic, B. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Trans. Access. Comput.* 6(4):14:1–14:36.

Saluja, A.; Hassan, H.; Toutanova, K.; and Quirk, C. 2014. Graph-based semi-supervised learning of translation models from monolingual data. In *ACL*. Baltimore, Maryland: Association for Computational Linguistics.

Sennrich, R.; Haddow, B.; and Birch, A. 2016a. Improving neural machine translation models with monolingual data. In *ACL*, 86–96. Berlin, Germany: Association for Computational Linguistics.

Sennrich, R.; Haddow, B.; and Birch, A. 2016b. Neural machine translation of rare words with subword units. In *ACL*, 1715–1725. Berlin, Germany: Association for Computational Linguistics.

Snover, M.; Dorr, B. J.; Schwartz, R.; Micciulla, L.; and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas* 223 – 231.

Štajner, S., and Popovic, M. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, 230–242.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *NeurIPS*. Curran Associates, Inc. 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Woodsend, K., and Lapata, M. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *EMNLP*, 409–420. Stroudsburg, PA, USA: Association for Computational Linguistics.

Wu, Y., et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wubben, S.; van den Bosch, A.; and Krahmer, E. 2012. Sentence simplification by monolingual machine translation. In *ACL*, ACL '12, 1015–1024. Stroudsburg, PA, USA: Association for Computational Linguistics.

Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; and Callison-Burch, C. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics* 4:401–415.

Zhang, X., and Lapata, M. 2017. Sentence simplification with deep reinforcement learning. In *EMNLP*, 584–594. Copenhagen, Denmark: Association for Computational Linguistics.

Zhang, M., and Toral, A. 2019. The effect of translationese in machine translation test sets. *CoRR* abs/1906.08069.

Zhao, S.; Meng, R.; He, D.; Saptono, A.; and Parmanto, B. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *EMNLP*, 3164–3173. Brussels, Belgium: Association for Computational Linguistics.

Zhu, Z.; Bernhard, D.; and Gurevych, I. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 1353–1361. Stroudsburg, PA, USA: Association for Computational Linguistics.

Zoph, B.; Yuret, D.; May, J.; and Knight, K. 2016. Transfer learning for low-resource neural machine translation. In *EMNLP*, 1568–1575. Austin, Texas: Association for Computational Linguistics.