Customizing Neural Machine Translation for Subtitling

Evgeny Matusov[*], **Patrick Wilken**[*], **Panayota Georgakopoulou**[*]
AppTek
Aachen, Germany
ematusov@apptek.com pwilken@apptek.com yota@athenaconsultancy.eu

Abstract

In this work, we customized a neural machine translation system for translation of subtitles in the domain of entertainment. The neural translation model was adapted to the subtitling content and style and extended by a simple, yet effective technique for utilizing inter-sentence context for short sentences such as dialog turns. The main contribution of the paper is a novel subtitle segmentation algorithm that predicts the end of a subtitle line given the previous word-level context using a recurrent neural network learned from human segmentation decisions. This model is combined with subtitle length and duration constraints established in the subtitling industry. We conducted a thorough human evaluation with two post-editors (English-to-Spanish translation of a documentary and a sitcom). It showed a notable productivity increase of up to 37% as compared to translating from scratch and significant reductions in human translation edit rate in comparison with the post-editing of the baseline non-adapted system without a learned segmentation model.

1 Introduction

In recent years, significant progress was observed in neural machine translation (NMT), with its quality increasing dramatically as compared to the previous generation of statistical phrase-based MT systems. However, user acceptance in the subtitling community has so far been rare. The reason for this, in our opinion, is that the state-of-the-art off-the-shelf NMT systems do not address the issues and challenges of the subtitling process in full.

In this paper, we present a customized NMT system for subtitling, with focus on the entertainment domain. From the user perspective, we show how the quality of translation and subtitle segmentation can improve in such a way that significantly reduced post-editing is required. We believe that such customized systems would lead to greater user acceptance in the subtitling industry and would contribute to the wider adoption of NMT technology with the subsequent benefits the latter brings in terms of productivity gain and time efficiency in subtitling workflows.

The paper is structured as follows. We start with the review of related research in Section 2. Section 3 describes the details of our baseline NMT system and how it compares to NMT systems from previous research. Section 4 presents the details of the changes to the MT system that were necessary to boost its performance on the subtitling tasks for entertainment domain, with a focus on Latin American Spanish as the target language. In Section 5, we present a novel algorithm for automatic subtitle segmentation that is combined with rule-based constraints which are necessary for correct subtitle representation on the screen. Finally, Section 6 describes the automatic and human evaluation of the proposed system, including post-editing experiments and feedback from professional translators.

2 Related Work

Evaluation of post-editing time and efficiency gain was presented by Etchegoyhen et al. (2014) on multiple language pairs and with many post-editors. However, that work only evaluated statistical MT systems, whereas here we evaluate a neural MT system. Also, the aspect of subtitle segmentation was not explicitly considered there; it was not clear what segmentation was used, if at all. Interesting findings on evaluation of statisti-

[*] equal contribution

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 82–93
Florence, Italy, August 1-2, 2019.

cal MT for subtitling in production can be found in the work of Volk et al. (2010), who perform an extensive subjective error analysis of the MT output. Aspects of customizing MT, again statistical, using existing subtitle collections are discussed in (Müller and Volk, 2013).

There is little work on subtitle segmentation, and to the best of our knowledge, no research which targets segmentation of MT output. The work by Álvarez et al. (2017) uses conditional random fields and support vector machines to predict segment boundaries, whereas in this paper we rely on recurrent neural networks. That algorithm is evaluated in terms of monolingual post-editing effort in the work of Álvarez Muniain et al. (2016). Lison and Meena (2016) predict dialog turns in subtitles, which is related to subtitle segmentation, but was beyond the scope of our work. The latest research of Song et al. (2019) deals with predicting sentence-final punctuation within non-punctuated subtitles using a long-short-term memory network (LSTM); that model, and also the punctuation prediction LSTM of Tilk and Alumäe (2015) is related to what we use in our work, but we deal with subtitle segmentation that is more complex and less well-defined than prediction of punctuation, as we show in Section 5.

3 Baseline NMT Architecture

We trained our NMT models using an open-source toolkit (Zeyer et al., 2018) that is based on TensorFlow (Abadi et al., 2015). We trained an attention-based RNN model similar to Bahdanau et al. (2015) with additive attention.

The attention model projects both the source and the target words into a 620-dimensional embedding space. The bidirectional encoder consists of 4 layers, each of which uses LSTM cells with 1000 units. We used a unidirectional decoder with the same number of units. In the initial (sub)epochs, we employed a layer-wise pretraining scheme that resulted in better convergence and faster overall training speed (Zeyer et al., 2018). We also enhanced the computation of attention weights using fertility feedback similar to Tu et al. (2016); Bahar et al. (2017).

The training data was preprocessed using Sentencepiece (Kudo and Richardson, 2018), with 20K and 30K subword units estimated separately for English and Spanish, respectively, without any other tokenization. In training, all our models relied on the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001. We applied a learning rate scheduling according to the Newbob scheme based on the perplexity on the validation set for a few consecutive evaluation checkpoints. We also employed label smoothing of 0.1 (Pereyra et al., 2017). The dropout rate ranged from 0.1 to 0.3.

Our baseline general-domain NMT system is a competitive single system that obtains the case-sensitive BLEU score of 34.4% on the WMT newstest 2013 En-Es set[1].

4 NMT Adaptation

4.1 Domain and style adaptation

Film content covers a large variety of genres, thus it is not easy to characterize the domain of these type of data. However, subtitles typically have shorter sentences than general texts (e.g. news articles), and brief utterances abound in many films. To create a customized system for subtitles, we used the OpenSubtitles parallel data[2], downloaded from the OPUS collection (Lison and Tiedemann, 2016), as the main training corpus. The corpus was filtered by running FastText based language identification (Joulin et al., 2016) and other heuristics (e.g. based on source/target lengths and length ratios in tokens and characters). In addition, we used other conversational corpora, such as GlobalVoices, transcribed TED talks and in-house crawled English-Spanish transcripts of the EU TV as parallel training data. We also added Europarl and News Commentary data to the main training corpus as sources of clean and well-aligned sentence pairs.

Neural MT systems often have problems translating rare words. To mitigate this problem, we developed a novel data augmentation technique. First, we computed word frequency statistics for the main training corpus described above. Then, we defined auxiliary out-of-domain training data from which we wanted to extract only specific sentence pairs. These data included all other publicly available training data, including ParaCrawl, CommonCrawl, EUbookshop, JRC-Acquis, EMEA, and other corpora from the OPUS collection. We computed word frequencies for each of these auxiliary corpora individually. Next,

[1]The BLEU score of a top online MT provider on this set was 35.0% as of July 2018.

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

for each sentence pair in each auxiliary corpus we checked that:

- either the source or the target sentence has at least one word that is *rare* in the main corpus, and
- neither the source sentence, nor the target sentence includes any word that is out-of-vocabulary for the main training corpus and at the same time is *rare* in the auxiliary corpus.

We defined a word to be *rare* if its frequency is less than 50. Finally, we limited the total number of running words we add (counted on the source side) to 100M per auxiliary corpus. This was done to avoid oversampling from large, but noisy corpora such as CommonCrawl.

In practice, for the En-Es training data, 145M words of auxiliary data were added, which is ca. 17% of the auxiliary training data that was available. Overall, we used ca. 39M lines of parallel training data for training, with 447M running words on the English and 453M running words on the Spanish side.

Additional domain adaptation may include fine-tuning of the trained model with a reduced learning rate on in-domain data, as e.g. in the work of Luong and Manning (2015). Since we were aiming at covering all possible film genres, we did not perform this additional fine-tuning in our experiments. We also did not use any back-translated target language monolingual data.

4.2 Handling language variety

Most MT systems do not differentiate between European and Latin American (LA) Spanish as the target language, providing a single system for translation into Spanish. However, significant differences between the two language varieties require the creation of separate subtitles for audiences in Latin America and Spain.

Almost no parallel corpora are available for training NMT systems, in which the target language is explicitly marked as Latin American Spanish, and the majority of the public corpora represent European Spanish (such as proceedings of the European Parliament). However, large portions of the in-domain OpenSubtitles corpus contain Latin American Spanish subtitles. We follow a rule-based approach to label those documents/movies from the OpenSubtitles corpus as translations into LA Spanish. If the plural form of the word "you" is "ustedes" that is used in Latin American Spanish, then we mark the whole document as belonging to this language variety. Since this word is used frequently in movie dialogues, we can label a significant number of documents as belonging to LA Spanish (a total of 192M running words when counted on the Spanish side of the parallel data).

We then train a multilingual system similarly to Firat et al. (2016). We do not change the neural architecture, but add a special token at the beginning of the source sentence to signal LA Spanish output for all training sentence pairs which we labeled as translations into LA Spanish with the rule-based method described above. This is also similar to using tokens for domain control as in the work of Kobus et al. (2016). We used a development set labeled as having translations into LA Spanish to track convergence and for selection of the final training epoch.

An alternative approach that was applied to low-resource language pairs by Neubig and Hu (2018) would have been to pre-train the model on all English-Spanish data, and then continue training on sentence pairs with LA Spanish targets. However, we did not follow this approach to avoid overfitting to the style of the OpenSubtitles corpus instead of adapting to the LA Spanish language variety.

4.3 Towards document-level translation

Subtitles often contain short sentences which, when translated by NMT individually, provide very little context for correct translation of certain words and phrases, such as pronouns. Yet this context is available in preceding sentences. As a step towards document-level translation, we created a training corpus of OpenSubtitles in which we spliced two or more consecutive subtitles from the same film, as well as their translations, until a maximum length of K tokens was reached on the source side. We inserted a special separator symbol between each pair of spliced sentences both on the source and the target side. The idea was that the NMT system can learn to produce these separator symbols and learn not to re-order words across them, so that the original sentence segmentation can be restored. At the same time, because of the nature of the recurrent model, the context of the previous sentences would also be memo-

rized by the system and would affect the translation quality of the current sentence[3].

We created two copies of OpenSubtitles corpus of only spliced sentence pairs with $K = 20$ and $K = 30$, respectively, and used this corpus in training together with all the other data described in Section 4.1. During inference, we also spliced consecutive short sentences from the same film until a threshold of $K = 20$ tokens was reached and then translated the resulting test set. Thus, each sentence was translated only once, either as part of a spliced sentence sequence or as an individual (long) sentence. A possibly better, but more redundant approach would have been to cut out the translation of only the last sentence in a spliced sequence, and then re-send the corresponding source sentence as context for translating the next sentence. However, for time reasons we did not test this approach. In the future, we also plan to expand on the existing research on document-level translation (Miculicich et al., 2018; Wang et al., 2017) and encode the previous inter-sentence context in a separate neural component. Even the first step towards expanding context beyond a single sentence described above led to some improvements in translation, and in particular pronoun disambiguation, as will be seen in Section 6.

5 Subtitle Segmentation

The output of the NMT system has to be formatted in an appropriate way when displayed on the screen. Typically, there exists a fixed character limit per subtitle line, the number of lines should not exceed two, and the text in a subtitle has to be as long as needed to match the user's reading speed, so that it is possible for viewers to read the subtitle and also watch the film at the same time. Beyond that, we want line and subtitle boundaries to occur in places where the flow of reading is harmed as little as possible. While the first two requirements can be implemented as hard rules, optimizing boundaries for readability is more subtle and a lack thereof can easily expose the subtitle as being machine generated, especially when compared to a professionally created one. Punctuation and part-of-speech information can indicate possible segmentation points. However, in general finding good boundaries is not straight-forward and depends on syntax and semantics.

We therefore employ a neural model to predict segment boundaries. It consists of a 128-dimensional word embedding layer and two 256-dimensional bi-directional LSTM layers, followed by a softmax. The output is a binary decision, i.e. we generate two probabilities per input word w_i: the probability $p_{B,i}$ of inserting a segment boundary after position i, and the probability $1 - p_{B,i}$ of the complementary event.

We train the model on the Spanish OpenSubtitles 2018 corpora of the OPUS Project (Lison and Tiedemann, 2016), which we tokenize and convert to lower-case. The data comes in XML format, including annotated sentence boundaries and timestamps for the subtitle units. We use all subtitle boundaries occurring in between words of a sentence as ground truth labels. Training is performed on all sentences containing at least one subtitle boundary, leading to a corpus size of 16.7M sentences.

To enforce the additional requirements mentioned above, we integrate the neural segmentation model into a beam search decoder. The search happens synchronous to the word positions of the input. At each step there are three possible expansions of a partial hypothesis: no boundary, line boundary, or subtitle boundary after the current word. The natural logarithm of the segmentation model probability is used as score (making no distinction between line and subtitle boundaries). Penalties for the following auxiliary features are subtracted:

1. *character limit*: penalty $q_1 = \infty$ if a line is longer than allowed;
2. *number of lines*: penalty q_2 for every line exceeding two in a given subtitle;
3. *similar line lengths*: penalty q_3 per difference in length of subsequent lines within a subtitle, measured in characters;
4. *expected subtitle lengths*: penalty q_4 per deviation from expected subtitle lengths, measured in characters; we expect subtitle lengths to be as in the source language, only scaled according to the difference in sentence length between the source sentence and its translation.

The third feature is supposed to lead to geometrically pleasing line lengths. In particular, it avoids

[3] We came up with this approach on our own, but later found it to be similar to the work of Tiedemann and Scherrer (2017), who include a single previous translation unit with a separator symbol as additional context.

orphans, i.e. lines with very few words in them. The forth feature attempts to keep the translation in sync with the video by keeping the number of characters in a subtitle similar to the source language. This also means that the subtitle duration will be suited for a similar reading speed as the one set in the source file. As a side effect, this feature ensures that we predict the right number of subtitle boundaries for a given sentence.

We use a beam size of 100. The penalties are set to $q_2 = 10$, $q_3 = 0.1$ and $q_4 = 1$. Furthermore, we use a margin of 20% and 30% of the line and subtitle lengths for features 3 and 4, respectively, in which no penalty is applied.

For the baseline approach, we do the segmentation using the four heuristics only, i.e. without the neural segmentation model. This is similar to algorithms used in existing subtitling tools and makes a direct analysis of the effect of the segmentation model possible.

6 Experimental Results

6.1 User experience

To confirm the improvement in quality described in Sections 3 and 4 and the usability of the ensuing output we sought the feedback of professional translators. We selected the language pair US English into LA Spanish for our case study and used video materials of two different genres:

- *Home*[4], a documentary about Earth, composed of aerial shots of our planet and narrated by a single voice over narrator, in a paced manner with well-structured sentences;
- *Lucy: The Bean Queen*[5], an all-time classic sitcom, full of puns and idiomatic language.

We asked an experienced English subtitler to create subtitle files to be used as input for machine translation purposes, with 6.6K running words for *Home* and 2.7K running words for *Lucy*, following well-established subtitling conventions in the source audio language (English). These subtitle files were subsequently machine translated into LA Spanish using both the baseline and the adapted MT systems described in previous sections, the latter including the inter-sentence context for short sentences and the proposed novel subtitle segmentation algorithm.

[4]https://archive.org/details/HOME_English

[5]https://archive.org/details/TLS_Lucy_The_Bean_Queen

We asked two translators to perform a post-editing evaluation of the two MT outputs. Both have between 11-20 years of experience each in all types of subtitling work. PE1 comes from Colombia and PE2 from Argentina. Both have experience with MT of general texts and PE1 had limited prior experience with the use of MT in subtitling. We split the two source files in three roughly equal sections and asked the translators to perform the following tasks:

- Translate Part 1 straight from the template file, without deviating from the set timings, subtitle number and segmentation;
- Post-edit Part 2 using output from the baseline MT system;
- Post-edit Part 3 using output from the adapted MT system.

The translators did not know the output of which system they were post-editing. We asked the translators to work consecutively, as they normally would, taking as few breaks as possible and recording their actual work time to the nearest minute. We asked them to include the time for research they would normally perform as part of their translation task in this measurement and review their work one final time before submitting it, as they would under live working conditions in order to submit a file of publishable quality level. We then asked the translators to answer a survey, which included answers to the demographic information mentioned above, plus a qualitative survey of the machine translation output and the post-editing experience, using a combination of ranking scale scores and free-text questions.

6.2 Translation speed benchmarking

Both translators were asked to translate "from scratch" Part 1 of each of the two template files, totaling 24 minutes/220 subtitles for *Home* and 8 minutes/118 subtitles for *Lucy*, in order to obtain their benchmark speed for each type of material. PE1 with 2.08/2.0 subtitles per minute for Home and Lucy, respectively, turned out to be significantly faster than PE2 (1.18/1.44 subtitles per minute) and maintained similar speed irrespective of the film genre. Their translated files for Part 1 of the templates were used as gold reference for performing automatic MT evaluation, with its results shown in the next section.

Mode	System	BLEU [%]	TER [%]	charTER [%]
W	baseline	52.3	50.3	42.8
	adapted	53.6	49.2	41.2
S	baseline	49.9	58.5	51.8
	adapted	54.7	49.3	41.9
	base segm.	50.8	57.8	52.4
L	baseline	37.2	60.1	53.4
	adapted	44.0	49.3	42.1
	base segm.	38.2	59.4	53.4

Table 1: Case-sensitive MT error measures on part 1 of the *Home* documentary computed in 3 different modes: using full sentences with real words only (W), on the level of subtitles (S), or on the level of subtitles with line breaks within a subtitle marked with a special token `BR` both in MT output and reference translation. The BLEU scores are computed against two human reference translations created "from scratch", other measures against the translation of PE1.

Mode	System	BLEU [%]	TER [%]	charTER [%]
W	baseline	26.3	68.4	61.3
	adapted	30.3	61.5	56.8
	sent-level	30.2	62.8	54.8
S	baseline	26.6	85.6	60.4
	adapted	31.1	76.1	56.4
	sent-level	30.5	77.3	54.9
	base segm.	31.0	78.2	58.8
L	baseline	21.8	85.7	61.6
	adapted	30.4	75.6	56.6
	base segm.	25.7	79.4	59.4

Table 2: Case-sensitive MT error measures on part 1 of the *Lucy: The Bean Queen* documentary computed as in Table 1.

6.3 Automatic evaluation

We computed automatic MT metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and CharacTER (Wang et al., 2016) on the first part of each template for which we now had two independent human reference translations. We computed the scores three times using different evaluation modes. In the mode (W), we computed the scores and error rates on the full sentences; thus, pure MT quality is evaluated, and any segmentation decisions are ignored. In the (S) mode, we compared the subtitles with each other. Thus, any words and phrases wrongly placed in a different (e.g. previous or next) subtitle would count as errors. Finally, in the (L) mode we additionally add a special symbol to represent a line break (in rare cases, two breaks) within a subtitle. Thus, an incorrect line break is an extra token error that directly affects all error metrics. To summarize, the (S) and (L) evaluation modes jointly judge the MT and segmentation quality, whereas the (W) mode only judges the MT quality.

Table 1 shows these results for the *Home* video. We observe an improvement in BLEU from 52.3 to 53.6%, as computed with two reference translations, when comparing the baseline system with the adapted one that uses previous sentence context. This improvement becomes much larger in the (S) and (L) evaluation modes, which confirms the quality of the segmentation algorithm as compared with the baseline heuristics-only segmentation. The other error measures improve similarly with the adapted MT output and the proposed segmentation algorithm. We also show the result of the adapted system, but with the baseline segmentation. The result for this system is slightly better than for the baseline due to the generally better MT quality, but because of the incorrect segmentation it is very far from human references when the evaluation is performed on the level of subtitles.

On part 1 of the *Lucy* sitcom (Table 2), the improvements with the adapted system are more significant when the MT quality alone is evaluated. This is expected, since the style of the input is further away from the general-domain (news) data that was used to train the baseline system. On the other hand, the improvements with the new segmentation algorithm w.r.t baseline segmentation seem to be significant, but less pronounced, since here we are dealing with generally shorter subtitles, many of them one-liners. Nevertheless, the improvement in the (L) evaluation mode, where incorrect line breaks within a subtitle are considered as errors, is as large as 8 BLEU percentage points absolute, from 21.8 to 30.4%. Table 2 also shows the results for the adapted system, but when translating individual sentences without inter-sentence context (lines *sent-level*). We observed only insignificant reduction of the pure MT quality in BLEU and TER; the CharacTER even improved. The test sample was too small to make any conclusions here. Nevertheless, we observed cases where the translation of some words (e.g., pronouns) was better when consecutive short sentences were translated as a single unit as described

Test set/system	HTER [%]		SER [%]	
	PE1	PE2	PE1	PE2
p. 2 baseline	36.7	51.4	88.0	99.0
p. 3 adapted	27.8	44.2	67.2	79.6
p. 2 adapted*	36.2	46.5	83.6	88.7

Table 3: Case-sensitive Human Translation Edit Rate (HTER) and Subtitle Edit Rate (SER) on the post-edited parts 2 and 3 of the *Home* documentary. *The comparison of the adapted NMT on section 2 is against the human post-editing of the baseline NMT output.

Test set/system	HTER [%]		SER [%]	
	PE1	PE2	PE1	PE2
p. 2 baseline	73.8	82.7	89.4	91.7
p. 3 adapted	44.0	59.5	71.9	80.5
p. 2 adapted*	60.6	72.0	87.9	90.9

Table 4: Case-sensitive Human Translation Edit Rate (HTER) and Subtitle Edit Rate (SER) on the post-edited parts 2 and 3 of the *Lucy: The Bean Queen* documentary. *The comparison of the adapted NMT on part 2 is against the human post-editing of the baseline NMT output.

in Section 4.3, and the improvement could only be explained by the additional context.

6.4 Evaluation of human post-editing effort

We computed the HTER scores (TER against the post-edited MT output) for the parts 2 and 3 of both files. We also computed the *subtitle error rate* (SER), that we defined as the percentage of subtitles which were changed by the post-editor (not counting possible corrections of the line breaks within a subtitle). Table 3 shows the HTER and SER results for the *Home* documentary. We see that the HTER is consistently better for both post-editors when the adapted MT output is used. PE1 especially finds the adapted MT output acceptable and keeps approximately $1/3$ of the subtitles completely unchanged. The second post-editor makes more corrections in general, but also for him the number of corrections made on the adapted MT output is significantly lower. The numbers above have to be taken with a grain of salt, since there was no other way but to compare the post-editing effort on different parts of the file. However, even when we compare the adapted MT output on part 2 against the post-edited baseline MT output, we obtain lower HTER and SER scores than for the baseline MT output itself. This again underlines the high quality of the adapted MT output with proper subtitle segmentation.

Similar conclusions can be made from the HTER and SER results in Table 4 for the *Lucy* sitcom. Here, the number of corrections is generally higher, but the reduction of the post-editing effort when post-editing the adapted vs. baseline MT is very significant, e.g. from 73.8 to 44.0% HTER for PE1 (as measured on different parts of the file with similar translation difficulty).

File	Post-editor	PE speed subs/min	Gain (%)	
			Product.	Time
Home baseline	PE1	1.36	15.98	13.78
	PE2	2.00	-3.64	-3.77
Lucy baseline	PE1	1.43	-0.29	-0.30
	PE2	2.28	13.79	12.12
Home adapted	PE1	1.87	58.70	36.99
	PE2	2.15	3.75	3.62
Lucy adapted	PE1	1.86	28.91	22.43
	PE2	3.12	56.10	35.94

Table 5: Productivity and time gain by using baseline/adapted MT output as compared to translating "from scratch".

6.5 Post-editing efficiency

We also performed an analysis of productivity gain and time efficiency by comparing translator speeds when post-editing the baseline and adapted MT outputs against their benchmark speed (Section 6.2). The results are presented in Table 5.

Productivity gain is the estimated percentage of additional work a translator would be able to complete when performing an MTPE task versus translating the same text from scratch. Time efficiency is the estimated percentage of time a translator would save when performing an MTPE task versus translating the same material from scratch.

As we can see, both productivity gain and time efficiency were achieved for both post-editors overall. The average productivity gain was 6.46% on the baseline MT output and 36.87% on the adapted MT output, the time efficiency increased by 5.46% and 24.74%, respectively. There were borderline productivity losses in one file per translator when working on the baseline output, but more than significant productivity increases on the adapted output. The same trend is observed with time efficiency as well, verifying our initial hy-

pothesis regarding the usability of the adapted MT output.

Though no conclusion may be drawn from the results of two post-editors only, but given that their overall profiles are quite similar with a marked difference in their translation speed, it is interesting to note that the slower of the two benefits more overall in an MTPE workflow. It should be pointed out, however, that PE1 did have some experience with MT in subtitles, whereas PE2 did not, which might indicate that PE2 had to go through a learning curve and, hence, explain his slow speed on *Home* and the large increase in his post-editing speed on *Lucy*.

The % of subtitles changed (SER) is analyzed in Section 6.4. 100-SER is the percentage of subtitles that were left unchanged after post-editing, including both punctuation and capitalization aspects. This metric does not take into account the pertinence or complexity of the changes made by a post-editor to the rest of the subtitle file. As a result, from a time efficiency perspective, it does not necessarily indicate the effort a post-editor needs to invest when post-editing the entire file. It is still expected though that with lower SER, a translator's time efficiency is likely to increase. The results above corroborate this assumption, and we note that where translators saved more time when performing an MTPE task, they were also using more of the MT output without making any changes to it. A marked increase in the usage of MT output with zero edits was noted in the adapted MT output with the average overall subtitles unchanged at 25% across all files and both post-editors versus 8% for the baseline MT output.

6.6 User survey

A qualitative evaluation with the two translators that were involved in the MTPE task was also performed, in the form of a survey. The MQM[6] framework was used to define the dimensions of MT output quality the translators were asked about, and the following definitions were provided to them:

- Accuracy: Meaning, e.g. mistranslations, omissions, additions, untranslated words
- Fluency: Well-formedness of text, e.g. spelling, grammar, word order, consistency, typography, style
- Design: Physical presentation of text, e.g. line length, readability, line and subtitle breaks

The translators were asked to rank the MT outputs they worked on with respect to each of the three quality dimensions above, as well as on the basis of the overall MT quality and regarding the post-editing experience itself. A ranking scale of 1-5 was used in this survey (5 being best). All results were consistent, with translators ranking both quality and post-editing experience for the baseline MT output as a 2 on average, i.e. poor, and for the adapted MT output as 3, i.e. fair.

The translators confirmed the improvement in quality in the adapted MT output, which corroborates previous findings and our initial hypothesis for this case study. When asked additional questions regarding the perceived MT impact on their productivity and on the quality of the final product, PE2 confirmed he "felt" the increase in productivity he witnessed on *Lucy* and explained that his experience with *Home* would have been similar had it not been for a particularly difficult section in the source text in the last part of *Home* that slowed him down substantially. Yet PE2 felt it was only the easier parts of *Lucy*, the simpler sentences, on which the MT was perfect, while it still translated most of the slang and puns (i.e. the creative part) wrongly.

PE1 noted the difficulty in finding his own writing style when post-editing, but also explained that he became much faster once he understood what to expect from the MT and found a rhythm. He was impressed by the correct terminology in the MT output of *Home*, one of the main reasons why both translators reported they would consider using output such as that of the adapted MT on documentary genres like *Home* in their daily work.

Both subtitlers raised concerns about the influence the MT has on their productivity (PE1) and on the quality of the final product (PE2) as uncommon but correct translations in the target language would not be corrected in a post-editing workflow, potentially affecting the overall result[7]. Finally, both translators said that there were only few cases in the baseline MT output where expressions from European Spanish were used and had to be corrected; they reported only one such case in the adapted MT output (see Section 4.2 on why).

[6] http://www.qt21.eu/mqm-definition/definition-2014-06-06.html

[7] Cf. also the findings of Farrell (2018) on this matter.

6.7 Discussion

Both the automatic measures as well as the productivity/time gain evaluation with independent subtitlers indicate that the adapted MT output significantly outperformed the baseline MT in terms of quality. All of the metrics, whether on Part 1 against a gold reference file, or on Parts 2 and 3, against the post-edited files correlate and verify the conclusion above. A ranking scale qualitative evaluation by the translators also confirmed the above findings, and translators provided further insights as to the post-editing process itself.

7 Conclusions

In this paper, we described how a state-of-the-art NMT system can be effectively customized for subtitling. We proposed a simple way to integrate inter-sentence context for translation of short utterances and dialog turns, adapted the NMT system to language variation (Latin American Spanish) and subtitling style and domain. We introduced a novel algorithm for subtitle segmentation that combines a recurrent neural network model with hard and soft subtitle length and duration constraints in a beam search. We performed an extensive automatic and human evaluation, which showed notable improvements in quality of the adapted MT output segmented into subtitles with our proposed algorithm as compared to the baseline MT system output with heuristics-based line breaks. This quality improvement led to significant productivity and time gains when the adapted MT output was post-edited by independent professional translators, compared both to translation from scratch and post-editing the translations of the baseline MT system. Finally, we received positive qualitative feedback on the adapted MT output from the post-editors involved in our study.

In the future, we plan to use more sophisticated document-level features for better consistency of the translations. We also started to expand the language coverage and trained similar adapted systems with learned segmentation for the language pairs Spanish-to-English and English-to-Russian. Examples of automatic subtitles created by these systems when using or ignoring inter-sentence context are shown in Figure 1 and examples of heuristics-based vs. model-based segmentation for En→Es, Es→En, and En→Ru NMT output are shown in Figure 2 in the Appendix.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Aitor Álvarez, Carlos-D Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. 2017. Improving the automatic segmentation of subtitles through conditional random field. *Speech Communication*, 88:83–95.

Aitor Álvarez Muniain, Marina Balenciaga, Arantza del Pozo Echezarreta, Haritz Arzelus Irazusta, Anna Matamala, and Carlos D Martínez Hinarejos. 2016. Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3049–3053.

Parnia Bahar, Jan Rosendahl, Nick Rossenbach, and Hermann Ney. 2017. The RWTH Aachen machine translation systems for IWSLT 2017. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 29–34.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard Van Loenhout, Arantza Del Pozo, Mirjam Sepesy Maucec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In *LREC*, pages 46–53.

Michael Farrell. 2018. Machine translation markers in post-edited machine translation output. In *The 40th Conference Translating and the Computer, London, United Kingdom*, pages 15–16.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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

Catherine Kobus, Josep Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Pierre Lison and Raveesh Meena. 2016. Automatic turn segmentation for movie & TV subtitles. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 245–252. IEEE.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller and Martin Volk. 2013. Statistical machine translation of subtitles: From OpenSubtitles to TED. In *Language processing and knowledge in the Web*, pages 132–138. Springer.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Hye-Jeong Song, Hong-Ki Kim, Jong-Dae Kim, Chan-Young Park, and Yu-Seop Kim. 2019. Inter-sentence segmentation of YouTube subtitles using long-short term memory (LSTM). *Applied Sciences*, 9(7):1504.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. *arXiv preprint arXiv:1708.05943*.

Ottokar Tilk and Tanel Alumäe. 2015. LSTM for punctuation restoration in speech transcripts. In *Sixteenth annual conference of the international speech communication association*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Coverage-based neural machine translation. *CoRR*, abs/1601.04811.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine translation of TV subtitles for large scale production. In *JEC 2010; November 4th, 2010; Denver, CO, USA*, pages 53–62. Association for Machine Translation in the Americas.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. *arXiv preprint arXiv:1704.04347*.

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

Albert Zeyer, Tamer Alkhouli, and Hermann Ney. 2018. RETURNN as a generic flexible neural toolkit with application to translation and speech recognition. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 128–133.

A Supplementary Material

Source text:	MT without context:	Document-level MT:
18 00:05:43,751 --> 00:05:45,083 ¿Y por qué lo aceptaste? 19 00:05:45,125 --> 00:05:47,792 Porque hablé con él. Creo que es inocente. 20 00:05:47,876 --> 00:05:49,250 No, yo no estoy tan segura. 21 00:05:49,292 --> 00:05:50,959 Todas las pruebas están en su contra. 22 00:05:51,083 --> 00:05:53,417 Pasó la noche con ella. Fue el último que la vio con vida. 23 00:05:53,542 --> 00:05:54,501 De acuerdo. 27 00:06:02,501 --> 00:06:07,626 Sin embargo, Carlos conoció a esa mujer esa misma noche. 28 00:06:08,584 --> 00:06:10,626 Al día siguiente, se iba a casar con Alejandra. 29 00:06:10,667 --> 00:06:12,542 ¿Qué motivos tendría para matarla? 30 00:06:12,667 --> 00:06:14,417 -Yo no lo sé. -Ninguno.	18 00:05:43,751 --> 00:05:45,083 **And why** did you accept it? 19 00:05:45,125 --> 00:05:47,792 Because I talked to him. I think he's innocent. 20 00:05:47,876 --> 00:05:49,250 No, I'm not so sure. 21 00:05:49,292 --> 00:05:50,959 All the evidence is against **you**. 22 00:05:51,083 --> 00:05:53,417 **She** spent the night with her. **He** was the last one to see her alive. 23 00:05:53,542 --> 00:05:54,501 Okay. 27 00:06:02,501 --> 00:06:07,626 However, Carlos met that woman that same night. 28 00:06:08,584 --> 00:06:10,626 The next day, **he** was going to marry Alejandra. 29 00:06:10,667 --> 00:06:12,542 What's the point of killing her? 30 00:06:12,667 --> 00:06:14,417 - I don't know. - None.	18 00:05:43,751 --> 00:05:45,083 **Then why** did you accept it? 19 00:05:45,125 --> 00:05:47,792 Because I talked to him. I think he's innocent. 20 00:05:47,876 --> 00:05:49,250 No, I'm not so sure. 21 00:05:49,292 --> 00:05:50,959 All the evidence is against **him**. 22 00:05:51,083 --> 00:05:53,417 **He** spent the night with her. **He** was the last one to see her alive. 23 00:05:53,542 --> 00:05:54,501 Agreed. 27 00:06:02,501 --> 00:06:07,626 However, Carlos met that woman that same night. 28 00:06:08,584 --> 00:06:10,626 The next day, **she** was marrying Alejandra. 29 00:06:10,667 --> 00:06:12,542 What motive would **she** have to kill her? 30 00:06:12,667 --> 00:06:14,417 -I don't know. -None.

Figure 1: Examples of Spanish-to-English subtitle translation with and without inter-sentence context available to the NMT system.

Source text:	MT without segmentation algorithm:	MT with segmentation algorithm:
17 00:01:39,160 --> 00:01:42,400 You can skip the eulogy, I'm not gone yet.	17 00:01:39,160 --> 00:01:42,400 Можешь пропустить надгробную речь, я еще не ушел.	17 00:01:39,160 --> 00:01:42,400 Можешь пропустить надгробную речь, я еще не ушел.
35 00:02:48,400 --> 00:02:51,480 That proves that you have confidence in my work.	35 00:02:48,400 --> 00:02:51,480 Это доказывает, что ты уверен в моей работе.	35 00:02:48,400 --> 00:02:51,480 Это доказывает, что ты уверен в моей работе.
42 00:03:11,440 --> 00:03:16,480 A contract with the Royal Furniture Company for $1,500?	42 00:03:11,440 --> 00:03:16,480 Контракт с Королевской мебельной компанией за $1500?	42 00:03:11,440 --> 00:03:16,480 Контракт с Королевской мебельной компанией за $1500?
45 00:07:29,040 --> 00:07:32,800 Thanks to them, the carbon drained from the atmosphere 46 00:07:32,920 --> 00:07:35,400 and other life forms could develop.	45 00:07:29,040 --> 00:07:32,800 Gracias a ellos, el carbono drenado de la atmósfera y otras formas de vida 46 00:07:32,920 --> 00:07:35,400 podrían desarrollarse.	45 00:07:29,040 --> 00:07:32,800 Gracias a ellos, el carbono drenado de la atmósfera 46 00:07:32,920 --> 00:07:35,400 y otras formas de vida podrían desarrollarse.
49 00:07:47,240 --> 00:07:50,200 which enabled it to break apart the water molecule 50 00:07:50,320 --> 00:07:52,240 and take the oxygen.	49 00:07:47,240 --> 00:07:50,200 lo que le permitió romper 50 00:07:50,320 --> 00:07:52,240 la molécula de agua y tomar el oxígeno.	49 00:07:47,240 --> 00:07:50,200 lo que le permitió romper la molécula de agua 50 00:07:50,320 --> 00:07:52,240 y tomar el oxígeno.
10 00:05:24,000 --> 00:05:27,125 No, bueno, es que todos fueron a comer 11 00:05:27,375 --> 00:05:29,918 y te quería decir que, si quieres, podemos salir a comer juntos.	10 00:05:24,000 --> 00:05:27,125 No, well, they all went to eat and I wanted to tell you, if you want, 11 00:05:27,375 --> 00:05:29,918 we can go out and eat together.	10 00:05:24,000 --> 00:05:27,125 No, well, they all went to eat 11 00:05:27,375 --> 00:05:29,918 and I wanted to tell you, if you want, we can go out and eat together.
13 00:05:32,000 --> 00:05:34,083 Tengo mucho trabajo. Y con esto del caso de Ibarra, 14 00:05:34,209 --> 00:05:35,876 estoy saturado, Olivia.	13 00:05:32,000 --> 00:05:34,083 I have a lot of work. 14 00:05:32,000 --> 00:05:34,083 And with this Ibarra case, 15 00:05:34,209 --> 00:05:35,876 I'm saturated, Olivia.	13 00:05:32,000 --> 00:05:34,083 I have a lot of work. And with this Ibarra case, 14 00:05:34,209 --> 00:05:35,876 I'm saturated, Olivia.

Figure 2: Examples of subtitle segmentation using model-based approach vs. heuristics-based approach (English-to-Russian, English-to-Spanish, and Spanish-to-English translation).