Are Subtitling Corpora really Subtitle-like?

Alina Karakanta[1,2], **Matteo Negri**[1], **Marco Turchi**[1]
[1] Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento - Italy
[2] University of Trento, Italy
{akarakanta,negri,turchi}@fbk.eu

Abstract

Growing needs in translating multimedia content have resulted in Neural Machine Translation (NMT) gradually becoming an established practice in the field of subtitling. Contrary to text translation, subtitling is subject to spatial and temporal constraints, which greatly increase the post-processing effort required to restore the NMT output to a proper subtitle format. In this work, we explore whether existing subtitling corpora conform to the constraints of: 1) length and reading speed; and 2) proper line breaks. We show that the process of creating parallel sentence alignments removes important time and line break information and propose practices for creating resources for subtitling-oriented NMT faithful to the subtitle format.

1 Introduction

Machine Translation (MT) of subtitles is a growing need for various applications, given the amounts of online multimedia content becoming available daily. Subtitling translation is a complex process consisting of several stages (transcription, translation, timing), and manual approaches to the task are laborious and costly. Subtitling has to conform to spatial constraints such as length, and temporal constraints such as reading speed. While length and reading speed can be modelled as a post-processing step in an MT workflow using simple rules, subtitle segmentation, i.e. where and if to insert a line break, depends on semantic and syntactic properties. Subtitle segmentation is particularly important, since it has been shown that a proper segmentation by phrase or sentence significantly reduces reading time and improves comprehension (Perego, 2008; Rajendran et al., 2013).

Hence, there is ample room for developing fully or at least partially automated solutions for subtitle-oriented NMT, which would contribute in reducing post-processing effort and speeding-up turn-around times. Automated approaches though, especially NMT, are data-hungry. Performance greatly depends on the availability of large amounts of high-quality data (up to tens of millions of parallel sentences), specifically tailored for the task. In the case of subtitle-oriented NMT, this implies having access to large subtitle training corpora. This leads to the following question: **What should data specifically tailored for subtitling-oriented NMT look like?**

There are large amounts of available parallel data extracted from subtitles (Lison and Tiedemann, 2016; Pryzant et al., 2018; Di Gangi et al., 2019). These corpora are usually obtained by collecting files in a subtitle specific format (*.srt*) in several languages and then parsing and aligning them at sentence level. MT training at sentence level generally increases performance as the system receives longer context (useful, for instance, to disambiguate words). As shown in Table 1, this process compromises the subtitle format by converting the subtitle blocks into full sentences. With this "merging", information about subtitle segmentation (line breaks) is often lost. Therefore, recovery of the MT output to a proper subtitle format has to be performed subsequently, either as a post-editing process or by using hand-crafted rules and boundary predictions. Integrating the subtitle constraints in the model can help reduce the post-processing effort, especially in cases where the input is a stream of data, such as in end-to-end Speech Neural Machine Translation. To date, there has been no study examining the consequences of obtaining parallel sentences from sub-

```
1
00:00:14,820 −− > 00:00:18,820
Grazie mille, Chris.
É un grande onore venire
2
00:00:18,820 −− > 00:00:22,820
su questo palco due volte.
Vi sono estremamente grato.
```

```
Grazie mille, Chris.
É un grande onore venire su questo palco due volte.
Vi sono estremamente grato.
```

Table 1: Subtitle blocks (top, 1-2) as they appear in an .srt file and the processed output for obtaining aligned sentences (bottom).

titles on preserving the subtitling constraints.

In this work, we explore whether the large, publicly available parallel data compiled from subtitles conform to the temporal and spatial constraints necessary for achieving quality subtitles. We compare the existing resources to an adaptation of MuST-C (Di Gangi et al., 2019), where the data is kept as subtitles. For evaluating length and reading speed, we employ character counts, while for proper line breaks we use the Chink-Chunk algorithm (Liberman and Church, 1992). Based on the analysis, we discuss limitations of the existing data and present a preliminary road-map towards creating resources for training subtitling-oriented NMT faithful to the subtitling format.

2 Related work

2.1 Subtitling corpora

Building an end-to-end subtitle-oriented translation system poses several challenges, mainly related to the fact that NMT training needs large amounts of high-quality data representative of the target application scenario (subtitling in our case). Human subtitlers translate either directly from the audio/video or they are provided with a template with the source text already in the format of subtitles containing time codes and line breaks, which they have to adhere to when translating.

Several projects have attempted to collect parallel subtitling corpora. The most well-known one is the OpenSubtitles[1] corpus (Lison and Tiedemann, 2016), extracted from 3.7 million subtitles across 60 languages. Since subtitle blocks do not always correspond to sentences (see Table 1), the blocks are merged and then segmented into sentences using heuristics based on time codes and punctuation. Then, the extracted sentences are aligned to create parallel corpora with the time-overlap algorithm (Tiedemann, 2008) and bilingual dictionaries. The 2018 version of OpenSubtitles has high-quality sentence alignments, however, it does not resemble the realistic subtitling scenario described above, since time and line break information are lost in the merging process. The same methodology was used for compiling MontenegrinSubs (Božović et al., 2018), an English – Montenegrin parallel corpus of subtitles, which contains only 68k sentences.

The Japanese-English Subtitle Corpus JESC (Pryzant et al., 2018) is a large parallel subtitling corpus consisting of 2.8 million sentences. It was created by crawling the internet for film and TV subtitles and aligning their captions with improved document and caption alignment algorithms. This corpus is aligned at caption level, therefore its format is closer to our scenario. On the other hand, non-matching alignments are discarded, which might hurt the integrity of the subtitling documents. As we will show, this is particularly important for learning proper line breaks between subtitle blocks.

A corpus preserving both subtitle segmentation and order of lines is SubCo (Martínez and Vela, 2016), a corpus of machine and human translated subtitles for English–German. However, it only consists of 2 source texts (∼150 captions each) with multiple student and machine translations. Therefore, it is not sufficient for training MT systems, although it could be useful for evaluation because of the multiple reference translations.

Slightly deviating from the domain of films and TV series, corpora for Spoken Language Translation (SLT) have been created based on TED talks. The Web Inventory of Transcribed and Translated Talks (Cettolo et al., 2012) is a multilingual collection of transcriptions and translations of TED talks. The talks are aligned at sentence level without audio information. Based on WIT, the IWSLT campaigns (Niehues et al., 2018) are annually releasing parallel data and their corresponding audio for the task of SLT, which are extracted based on time codes but again with merging operations to create segments. MuST-C (Di Gangi et al., 2019) is to date the largest multilingual corpus for end-to-end speech translation. It contains (audio-source language transcription-target

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

language translation) triplets, aligned at segment level. The process of creation is the opposite from IWSLT; the authors first align the written parts and then match the audio. This is a promising corpus for an end-to-end system which translates from audio directly into subtitles. However, the translations are merged to create sentences, therefore they are far from the suitable subtitle format. Given the challenges discussed above, there exists no systematic study of the suitability of the existing corpora for subtitling-oriented NMT.

2.2 Subtitle segmentation

Subtitle segmentation techniques have so far focused on monolingual subtitle data. Álvarez et al. (2014) trained Support Vector Machine and Logistic Regression classifiers on correctly/incorrectly segmented subtitles to predict line breaks. Extending this work, Álvarez et al. (2017) used a Conditional Random Field (CRF) classifier for the same task, also differentiating between line breaks (next subtitle line) and subtitle breaks (next subtitle block). Recently, Song et al. (2019) employed a Long-Short Term Memory Network (LSTM) to predict the position of the period in order to improve the readability of automatically generated Youtube captions. To our knowledge to date, there is no approach attempting to learn bilingual subtitle segmentation or incorporating subtitle segmentation in an end-to-end NMT system.

3 Criteria for assessing subtitle quality

3.1 Background

The quality of the translated subtitles is not evaluated only in terms of fluency and adequacy, but also based on their format. We assess whether the available subtitle corpora conform to the constraints of length, reading speed (for the corpora where time information is available) and proper line breaks on the basis of the criteria for subtitle segmentation mentioned in the literature of Audiovisual Translation (AVT) (Cintas and Remael, 2007) and the TED talk subtitling guidelines[2]:

1. **Characters per line**. The space available for a subtitle is limited. The length of a subtitle depends on different factors, such as size of screen, font, age of the audience and country. For our analysis, we consider max. 42 chars for Latin alphabets, 14 for Japanese (including spaces).

2. **Lines per subtitle**. Subtitles should not take up too much space on screen. The space allowed for a subtitle is about 20% of screen space. Therefore, a subtitle block should not exceed 2 lines.

3. **Reading speed**. The on-air time of a subtitle should be sufficient for the audience to read and process its content. The subtitle should match as much as possible the start and the end of an utterance. The duration of the utterance (measured either in seconds or in feet/frames) is directly equivalent to the space a subtitle should occupy. As a general rule, we consider max. 21 chars/second.

4. **Preserve 'linguistic wholes'**. This criterion is related to subtitle segmentation. Subtitle segmentation does not rely only on the allowed length, but should respect linguistic norms. To facilitate readability, subtitle splits should not "break" semantic and syntactic units. In an ideal case, every subtitle line (or at least subtitle block) should represent a coherent linguistic chunk (*i.e.* a sentence or a phrase). For example, a noun should not be separated from its article. Lastly, subtitles should respect natural pauses.

5. **Equal length of lines**. Another criterion for splitting subtitles relates to aesthetics. There is no consensus about whether the top line should be longer or shorter, however, it has been shown that subtitle lines of equal length are easier to read, because the viewer's eyes return to the same point on the screen when reading the second line.

While subtitle length and reading speed are factors that can be controlled directly by the subtitle software used by translators, subtitle segmentation is left to the decision of the translator. Translators often have to either compromise the aesthetics in favour of the linguistic wholes or resort to omissions and substitutions. Therefore, modelling the segmentation decisions based on the large available corpora is of great importance for a high-quality subtitle-oriented NMT system.

3.2 Quality criteria filters

In order to assess the conformity of the existing subtitle corpora to the constraints mentioned above, we implement the following filters.

Characters per line (CPL): As mentioned above, the information about line breaks inside

[2] https://www.ted.com/participate/translate/guidelines

subtitle blocks is discarded in the process of creating parallel data. Therefore, we can only assume that a subtitle fulfils the criteria 1 and 2 above by calculating the maximum possible length for a subtitle block; 2 * 42 = 84 characters for Latin scripts and 2 * 14 = 28 for Japanese. If $CPL > max\_length$ then the subtitle doesn't conform to the length constraints.

Characters per second (CPS): This metric relates to reading speed. For the corpora where time codes and duration are preserved, we calculate CPS as follows: $CPS = \frac{\#chars}{duration}$

Chink-Chunk: Chink-Chunk is a low-level parsing algorithm which can be used as a rule-based method to insert line breaks between subtitles. It is a simple but efficient way to detect syntactic boundaries. It relates to preserving linguistic wholes, since it uses POS information to split units only at punctuation marks (logical completion) or when an open-class or content word (chunk) is followed by a closed-class or function word (chink). Here, we use this algorithm to compute statistics about the type of subtitle block breaks in the data (punctuation break, content-function break or other). The algorithm is described in Algorithm 1.

Algorithm 1: Chink-Chunk algorithm

```
if POS_last in ['PUNCT', 'SYM', 'X'] then
    punc_break +=1;
else
    if POS_last in content_words and POS_next in function_words then
        cf_break +=1;
    else
        other_split +=1;
    end
end
return punc_break, cf_break, other_split
```

4 Experiments

For our experiments we consider the corpora which are large enough to train NMT systems; OpenSubtitles, JESC and MuST-C. We focus on 3 language pairs, Japanese, Italian and German, paired with English, as languages coming from different families and having a large portion of sentences in all corpora. We tokenise and then tag the data with Universal Dependencies[3] to obtain POS tags for the Chink-Chunk algorithm.

To observe the effect of merging processes on preserving the subtitling constraints, we create a version of MuST-C at a subtitle level. We obtain the same .srt files used to create MuST-C. We extract only the subtitles with matching timestamps from the common talks in the language pair without any merging operations. Table 2 shows the statistics of the extracted corpus. We randomly sample 1,000 sentence pairs and manually inspect their alignments. 94% were correctly aligned, 3% partially aligned and 3% misaligned.

LP	Total	Extracted	MuST-C
EN-IT	671K	452K / 3.4M	253K / 4.8M
EN-DE	575K	361K / 2.7M	229K / 4.2M
EN-JA	669K	399K / 3M	-

Table 2: Total number of subtitles vs. number of extracted subtitles (in lines) from TED talks .srt files vs. the original MuST-C corpus. The first number shows lines (or sentences respectively), while the second words on the English side.

We apply each of the criteria filters in Section 3.2 to the corpora both on the source and the target side independently. Then, we take the intersection of the outputs of all the filters to obtain the lines/sentences which conform to all the criteria.

5 Analysis

Table 3 shows the percentage of preserved lines/sentences after applying each criterion.

Length: The analysis of Characters per line filter shows that both OpenSubtitles and JESC conform to the quality criterion of length in at least 94% of the cases. Despite the merging operations to obtain sentence alignments, OpenSubtitles still preserves a short length of lines, possibly because of the nature of the text of subtitles. A manual inspection shows that the text is mainly short dialogues and the long sentences are parts of descriptions or monologues, which are more rare. On the other hand, the merging operations in MuST-C create long sentences that do not resemble the subtitling format. This could be attributed to the format of TED talks. TED talks mostly contain text written to be spoken, prepared talks usually delivered by one speaker with few dialogue turns. Among all corpora, MuST-C_subs shows the highest conformity to the criterion of length, since indeed no merging operations were performed.

Reading speed: Conformity to the criterion of reading speed is achieved to a lesser degree, as

[3] https://universaldependencies.org/

LP	Corpus	Format	Time	CPL (s/t) %	CPS (s/t) %	Chink-Chunk (s/t) %	Total%
EN-IT	MuST-C	segment	✓	49 / 48	78 / 72	99 / 99	45
	OpenSubtitles	segment	-	95 / 94	-	99 / 99	91
	MuST-C_subs	subtitle	✓	99 / 98	86 / 81	87 / 83	79
EN-DE	MuST-C	segment	✓	51 / 47	77 / 66	99 / 99	42
	OpenSubtitles	segment	-	95 / 95	-	99 / 99	92
	MuST-C_subs	subtitle	✓	99 / 98	84 / 75	87 /87	74
EN-JA	OpenSubtitles	segment	-	96 / 93	-	99 /98	91
	JESC	subtitle	-	97 / 94	-	88 / 87	85
	MuST-C_subs	subtitle	✓	99 / 94	85 / 99	92 / 91	83

Table 3: Percentage of data preserved after applying each of the quality criteria filters on the subtitling corpora independently. Percentages are given on source and target side (s/t), except for the *Total* where source and target are combined.

shown by the Characters per second filter. Except for Japanese, where the allowed number of characters per line is lower, all other languages range between 66%-86%. In general, MuST-C_subs, being in subtitling format, seems to conform better to reading speed. Unfortunately, time information is not present in corpora other than the two versions of MuST-C, therefore a full comparison is not possible.

Linguistic wholes: The Chink-Chunk algorithm shows interesting properties of the subtitle breaks for all the corpora. MuST-C and OpenSubtitles conform to the criterion of preserving linguistic wholes in 99% of the sentences, which does not occur in the corpora in subtitle format; JESC and MuST-C_subs. Since these two corpora are compiled by removing captions based on unmatched time codes, the integrity of the documents is possibly broken. Subtitles are removed arbitrarily, so consecutive subtitles are often not kept in order. This shows the importance of preserving the order of subtitles when creating subtitling corpora.

This observation might lead to the assumption that JESC and MuST-C_subs are less subtitle-like. However, a close inspection of the breaks shows that OpenSubtitles and MuST-C end in a punctuation mark in 99.9% of the cases. Even though they preserve logical completion, these corpora do not contain sufficient examples of line breaks preserving linguistic wholes. On the other hand, the subtitle-level corpora contain between 5%-11% subtitle breaks in the form of content-function word. In a realistic subtitling scenario, an NMT system at inference time will often receive unfinished sentences, either from an audio stream or a subtitling template. Therefore, line break information might be valuable for training NMT systems that learn to translate and segment.

The total retained material shows that OpenSubtitles is the most suitable corpus for producing quality subtitles in all investigated languages, as more than 90% of the sentences passed the filters. However, this is not a fair comparison, given that the data was filtered with only 2 out of the 3 filters. One serious limitation of OpenSubtitles is the lack of time information, which does not allow for modelling reading speed. We showed that corpora in subtitling format (JESC, MuST-C_subs) contain useful information about line breaks not ending in punctuation marks, which are mostly absent from OpenSubtitles. Since no information about subtitle line breaks (inside a subtitle block) is preserved in any of the corpora, the criterion of equal length of lines cannot be explored in this study.

6 Conclusions and discussion

We explored whether the existing parallel subtitling resources conform to the subtitling constraints. We found that subtitling corpora generally conform to length and proper line breaks, despite the merging operations for aligning parallel sentences. We isolated some missing elements: the lack of time information (duration of utterance) and the insufficient representation of line breaks other than at punctuation marks.

This raises several open issues for creating corpora for subtitling-oriented NMT; i) **subtitling constraints:** a subtitling corpus, in order to be representative of the task, should respect the subtitling constraints; ii) **duration of utterance:** since the translation of a subtitle depends on the duration of the utterance, time information is highly relevant; iii) **integrity of documents:** a subtitle often occupies several lines, therefore the order of subtitles should be preserved whenever possible; iv) **line break information:** while parallel sentence alignments are indispensable, they should not compromise line break and subtitle block information. Break information could be preserved by inserting special symbols.

We intend to use these observations for an adaptation of MuST-C, containing triplets (audio, source language subtitle, target language subtitle), preserving line break information and taking advantage of natural pauses in the audio. In the long run, we would like to train NMT systems which predict line breaks while translating, possibly extending the input context using methods from document level translation.

Acknowledgements

This work is part of a project financially supported by an Amazon AWS ML Grant.

References

Aitor Álvarez, Haritz Arzelus, and Thierry Etchegoyhen. 2014. Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 229–238, Cham. Springer International Publishing.

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. In *Speech Communication*, volume 88, pages 83–95. Elsevier BV.

Petar Božović, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić, and Vojko Gorjanc. 2018. Opus-Montenegrinsubs 1.0: First electronic corpus of the Montenegrin language. In *Conference on Language Technologies & Digital Humanities*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Jorge Diaz Cintas and Aline Remael. 2007. *Audiovisual Translation: Subtitling*. Translation practices explained. Routledge.

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June.

Mark Liberman and Kenneth Church. 1992. Text analysis and word pronunciation in text-to-speech synthesis. *Advances in Speech Signal Processing*, pages 791–831.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from Movie and TV subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*.

José Manuel Martínez Martínez and Mihaela Vela. 2016. SubCo: A learner translation corpus of human and machine subtitles. In *Language Resources and Evaluation Conference (LREC)*.

Jan Niehues, Roldano Cattoni, Mauro Cettool Sebastian Stuke an, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 evaluation campaign. In *Proceedings of IWSLT 2018*.

Elisa Perego. 2008. Subtitles and line-breaks: Towards improved readability. *Between Text and Image: Updating research in screen translation*, 78(1):211–223.

Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English Subtitle Corpus. *Language Resources and Evaluation Conference (LREC)*.

Dhevi J. Rajendran, Andrew T. Duchowski, Pilar Orero, Juan Martnez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21.

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). 9:1504.

Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*.