Is 42 the Answer to Everything in Subtitling-oriented Speech Translation?

Alina Karakanta
Fondazione Bruno Kessler
University of Trento
Trento - Italy
akarakanta@fbk.eu

Matteo Negri
Fondazione Bruno Kessler
Trento - Italy
negri@fbk.eu

Marco Turchi
Fondazione Bruno Kessler
Trento - Italy
turchi@fbk.eu

Abstract

Subtitling is becoming increasingly important for disseminating information, given the enormous amounts of audiovisual content becoming available daily. Although Neural Machine Translation (NMT) can speed up the process of translating audiovisual content, large manual effort is still required for transcribing the source language, and for spotting and segmenting the text into proper subtitles. Creating proper subtitles in terms of timing and segmentation highly depends on information present in the audio (utterance duration, natural pauses). In this work, we explore two methods for applying Speech Translation (ST) to subtitling: a) a direct end-to-end and b) a classical cascade approach. We discuss the benefit of having access to the source language speech for improving the conformity of the generated subtitles to the spatial and temporal subtitling constraints and show that length[1] is not the answer to everything in the case of subtitling-oriented ST.

1 Introduction

Vast amounts of audiovisual content are becoming available every minute. From films and TV series, informative and marketing video material, to home-made videos, audiovisual content is reaching viewers with various needs and expectations, speaking different languages, all across the globe. This unprecedented access to information through audiovisual content is made possible mainly thanks to subtitling. Subtitles, despite being the fastest and most wide-spread way of translating audiovisual content, still rely heavily on human effort. In a typical multilingual subtitling workflow, a subtitler first creates a subtitle template (Georgakopoulou, 2019) by transcribing the source language audio, timing and adapting the text to create proper subtitles in the source language. These source language subtitles (also called captions) are already compressed and segmented to respect the subtitling constraints of length, reading speed and proper segmentation (Cintas and Remael, 2007; Karakanta et al., 2019). In this way, the work of an NMT system is already simplified, since it only needs to translate matching the length of the source text (Matusov et al., 2019; Lakew et al., 2019). However, the essence of a good subtitle goes beyond matching a predetermined length (as, for instance, 42 characters per line in the case of TED talks). Apart from this spatial dimension, subtitling relies heavily on the temporal dimension, which is incorporated in the subtitle templates in the form of timestamps. However, templates are expensive and slow to create and as so, not a viable solution for short turn-around times and individual content creators. Therefore, skipping the template creation process would greatly extend the application of NMT in the subtitling process, leading to massive reductions in costs and time and making multilingual subtitling more accessible to all types of content creators.

In this work, we propose Speech Translation as an alternative to the template creation process. We experiment with cascade systems, i.e. pipelined ASR+MT architectures, and direct, end-to-end ST systems. While the MT system in the pipelined approach receives a raw textual transcription as input, the direct speech translation receives temporal and prosodic information from the source language input signal. Given that several decisions about the form of subtitles depend on the audio (e.g. subtitle segmentation at natural pauses, length based on utterance duration), in this work we ask the question

[1] Speaking of subtitles and their optimum length of 42 characters per line, we could not help but alluding to the book and series *The Hitchhiker's Guide to the Galaxy* by Douglas Adams, where the number 42 is the "Answer to the Ultimate Question of Life, the Universe, and Everything", calculated by a massive supercomputer named Deep Thought for over 7.5 million years.

Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT), pages 209–219
July 9-10, 2020.
https://doi.org/10.18653/v1/P17

whether end-to-end ST systems can take advantage of this information to better model the subtitling constraints and subtitle segmentation. In other words, we investigate whether, in contrast to text translation (where one mainly focuses on returning subtitles of a maximum length of 42 characters), in speech translation additional information can help to develop a more advanced approach that goes beyond merely matching the source text length.

Our contributions can be summarised as follows:

- We present the first end-to-end solution for subtitling, completely eliminating the need of a source language transcription and any extra segmenting component;
- We conduct a thorough analysis of cascade vs. end-to-end systems both in terms of translation quality and conformity to the subtitling contraints, showing that end-to-end systems have large potential for subtitling.

2 Background

2.1 Subtitling

Subtitling involves translating the audio (speech) in a video into text in another language (in the case of interlingual subtitling). Subtitling is therefore a multi-modal phenomenon, incorporating visual images, gestures,[2] sound and language (Taylor, 2016), but also an intersemiotic process, in the sense that it involves a change of channel, medium and code from speech to writing, from spoken verbal language to written verbal language (Assis, 2001).

The temporal dimension of subtitles and the relation between audio and text has been stressed also in professional subtitling guidelines. In their subtitling guidelines, Carroll and Ivarsson (1998) mention that: *"The in and out times of subtitles must follow the speech rhythm of the dialogue, taking cuts and sound bridges into consideration"* and that: *"There must be a close correlation between film dialogue and subtitle content; source language and target language should be synchronized as far as possible"*. Therefore, a subtitler's decisions are guided not only by attempting to transfer the source language content, but also by achieving a high correlation between the source language speech and target language text.

[2]While we recognise the importance of the visual dimension in the process of subtitling, incorporating visual cues in the NMT system is beyond the scope of this work.

In addition, subtitles have specific properties in the sense that they have to conform to spatial and temporal constraints in order to ensure comprehension and a pleasant user experience. For example, due to limited space on screen, a subtitle cannot be longer than a fixed number of characters per line, ranging between 35-43 (Cintas and Remael, 2007). When it comes to the temporal constraints, a comfortable reading speed (about 21 chars/second) is key to a positive user experience. It should be ensured that viewers have enough time to read and assimilate the content, while at the same time their attention is not monopolised by the subtitles. Lastly, the segmentation of subtitles should be performed in a way that facilitates comprehension, by keeping linguistic units (e.g. phrases) in the same line. For the reasons mentioned above, subtitlers should ideally always have access to the source video when translating. However, working directly from the video can have several drawbacks. The subtitler needs to make sense of what is said on the screen, deal with regional accents, noise, unintelligible speech etc.

One way to automatise this labour-intensive process, especially in settings where several target languages are involved, is creating a subtitle template of the source language (Georgakopoulou, 2019). A subtitle template is an enriched transcript of the source language speech where the text is already compressed, timed and segmented into proper subtitles. This template can serve as a basis for translation into other languages, whether the translation is performed by a human or an MT system.

In the case of templates, optimal duration, length and proper segmentation are ensured, since the change of code between oral and written in the source language has already been curated by an expert. Due to the high costs and time required, creating a subtitle template is not a feasible solution both for small content creators and in the case of high volumes and fast turn-around times.

In the absence of a subtitle template, an automatic transcription of the source language audio could seem an efficient alternative. However, Automatic Speech Recognition (ASR) systems produce word-for-word transcriptions of the source speech, not adapted to the subtitling constraints, and where all information coming from the speech is discarded. This purely textual transcription is then translated by the MT system. Therefore, it is highly probable that a higher post-editing effort is

required not only in terms of translation errors, but chiefly for repairing the form of the subtitles.

Direct, end-to-end speech translation receives audio as input. Therefore, the model receives two types of information from the spectrogram: 1) information about the temporal dimension of the speech, e.g. duration, and 2) information related to the frequency, such as pitch, power and other prosodic elements. While intonation and stress are mostly related to semantic properties, speech tempo directly affects the compression rate of subtitles, and pauses often correspond to prosodic chunks which can determine the subtitle segmentation. Given that, it is worth asking the question whether access to this information can lead to better modelling of the subtitling constraints and subtitle segmentation.

2.2 Speech translation

Traditionally, the task of Speech-to-Text Translation has been addressed with cascade systems consisting of two components: an ASR system, which transcribes the speech into text, and an MT system, which translates the transcribed text into the target language (Eck and Hori, 2005). This approach has the benefit that it can take advantage of state-of-the-art technology for both components and leverage the large amount of data available for both tasks. On the other hand, it suffers from error propagation from the ASR to the MT, since transcription errors are impossible to recover because the MT component typically does not have access to the audio. Several works have attempted to make MT robust to ASR errors (Di Gangi et al., 2019b; Sperber et al., 2017) by working on noisy transcripts.

One further drawback of the cascaded approach, particularly relevant for the task of subtitling, is that any transcript, no matter how accurate, is subject to information loss in the semiotic shift from the richer audio representation to the poorer text representation. This limitation has been addressed in the past in speech-to-speech translation cascades chiefly for improving the naturalness of the synthesised speech and for resolving ambiguities. This has been performed through acoustic feature vectors related to different prosodic elements, such as duration and power (Kano et al., 2013), emphasis (Do et al., 2015, 2016) and intonation (Aguero et al., 2006; Anumanchipalli et al., 2012).

By avoiding intermediate textual representations, end-to-end speech translation (Bérard et al., 2016) can cope with the above limitations. However, its performance and suitability for reliable applications has been impeded by the limited amount of training data available. In spite of this data scarcity problem, it has been recently shown that the gap between the two approaches is closing (Niehues et al., 2018, 2019), especially with specially-tailored architectures (Di Gangi et al., 2019c; Dong et al., 2018) and via effective data augmentation strategies (Jia et al., 2019). Despite an increasing amount of works attempting to improve the performance of ST for general translation, there has been almost no work on comparing the two technologies on specific problems and applications, which is among the focus points of this work.

2.3 Machine Translation for subtitling

Despite the relevance of developing automatic solutions for subtitling both for the industry and academia, there have been very limited attempts to customise MT for subtitling. Previous works based on Statistical MT (SMT) used mostly proprietary data and led to completely opposite outcomes. Volk et al. (2010) developed SMT systems for the Scandinavian TV industry and reported very good results in this practical application. Aziz et al. (2012) reported significant reductions in post-editing effort compared to translating from scratch for DVD subtitles for English-Portuguese. On the other hand, the SUMAT project (Bywood et al., 2013, 2017), involving seven European language pairs, concluded that subtitling poses particular challenges for MT and therefore a lot of work is still required before MT can lead to real improvements in audiovisual translation workflows (Burchardt et al., 2016).

Recently, after the advent of the neural machine translation paradigm, Matusov et al. (2019) presented an NMT system customised to subtitling. The main contribution of the paper is a segmenter module trained on human segmentation decisions, which splits the resulting translation into subtitles. The authors reported reductions in post-editing effort, especially regarding subtitle segmentation. On a different strand of research, Lakew et al. (2019) proposed two methods for controlling the output length in NMT. The first one is based on adding a token, as in Multilingual NMT (Johnson et al., 2017; Ha et al., 2016), which in this setting represents the length ratio between source and target, and the second inserts length information in the positional encoding of the Transformer.

The application of MT (either SMT or NMT) in the works described above is possible only because of the presence of a "perfect" source language transcript, either for the translation itself or for computing the length ratio. To our knowledge today, no work so far has experimented with direct end-to-end ST in the domain of subtitling.

3 Experimental Setup

3.1 Data

For the experiments we use the MuST-Cinema corpus (Karakanta et al., 2020),[3] which contains (*audio*, *transcription*, *translation*) triplets where the breaks between subtitles have been annotated with special symbols. The symbol <eol> corresponds to a line break inside a subtitle block, while the symbol <eob> to a subtitle block break (the next subtitle comes on a different screen), as seen in the following example from the MuST-Cinema test set:

This kind of harassment keeps women <eol>
from accessing the internet – <eob>
essentially, knowledge. <eob>

We experiment with 2 language pairs, English→French and English→German, as languages with different syntax and word order. The training data consist of 229K and 275K sentences (408 and 492 hours) for German and French respectively, while the development sets contain 1088/1079 sentences and the test sets 542/544 sentences.

3.2 MT and ST systems

The **Cascade** system consists of an ASR and an MT component. The ASR component is based on the KALDI toolkit (Povey et al., 2011), featuring a time-delay neural network and lattice-free maximum mutual information discriminative sequence-training (Povey et al., 2016). The audio data for acoustic modelling include the clean portion of LibriSpeech (Panayotov et al., 2015) (∼460h) and a variable subset of the MuST-Cinema training set (∼450h), from which 40 MFCCs per time frame were extracted. A MaxEnt language model (Alumäe and Kurimo, 2010) is estimated from the corresponding transcripts (∼7M words). The MT component is based on the Transformer architecture (big) (Vaswani et al., 2017) with similar settings to the original paper. The system is first trained on the OPUS data, with 120M sentences for EN→FR and 50M for EN→DE and then fine-tuned on MuST-Cinema. Considering that the ASR output is lower-cased and without punctuation, we lowercase and remove the punctuation from the source side of the parallel data used in pre-training the MT system. To mitigate the error propagation between the ASR and the MT, for fine-tuning, we use a version of MuST-Cinema where the source audio has been transcribed by the tuned ASR.

For the **End-to-End** system, we experiment with two data conditions, one where we only use the MuST-Cinema training data (**E2E-small**) and a second one where we pre-train on a larger amount of data and fine-tune on MuST-Cinema (**E2E**). This will allow us to detect whether there is any trade-off between translation quality and conformity to constraints when increasing the amount of training data that are not representative of the target application (subtitling). The architecture used is S-Transformer, (Di Gangi et al., 2019c), an ST-oriented adaptation of Transformer, which has been shown to achieve high performance on different speech translation benchmarks. We remove the 2D self-attention layers and increase the size of the encoder to 11 layers, while for the decoder we use 4 layers. This choice was motivated by preliminary experiments, where we noted that replacing the 2D self-attention layers with normal self-attention layers and adding more layers in the encoder increased the final score, while removing a few decoder layers did not negatively affect the performance. As distance penalty, we choose the logarithmic distance penalty. We use the encoder of the ASR model to initialise the weights of the ST encoder and achieve faster convergence (Bansal et al., 2019).

Since the E2E-small system, trained only on MuST-Cinema, is disadvantaged in terms of the amount of training data compared to the cascade, we utilise synthetic data to boost the performance of the ST system (E2E). To this aim, we automatically translate into German and into French the English transcriptions of the data available for the IWSLT2020 offline speech translation task[4] (whenever the translation is not available in the respective target language). To this aim, we use an MT Transformer model achieving 43.2 BLEU points on the WMT'14 test set (Ott et al., 2018) for EN→FR.

[3] Must-Cinema has been derived from the MuST-C corpus (Di Gangi et al., 2019a), which currently represents the largest multilingual corpus for ST.

[4] http://iwslt.org/doku.php?id=offline_speech_translation

Our EN→DE Transformer model using similar settings achieves 25.3 BLEU points on the WMT'14 test set. The resulting training data (both real and synthetic) amount to 1.5M sentences on the target side. We use a different tag to separate the real from the synthetic data. We further use SpecAugment (Park et al., 2019), a technique for online data augmentation, with augment rate of 0.5. Finally, we fine-tune on MuST-Cinema.

For comparison, we also report MT results when starting from a "subtitle template" (**Template**). In this setting, we use the textual source side of MuST-Cinema, which contains the human transcriptions of the source language audio and its segmentation into subtitles. In this way, the input to the MT system is already split in subtitles using the special symbols, respecting the subtitling constraints and with proper segmentation. This will allow us to have an upper-bound of the performance that NMT can achieve when provided with input already in the form of subtitles. We pre-train large models with the OPUS data used in the cascade but without lowercasing or removing punctuation and then fine-tune them on the full training set of MuST-Cinema. It should be noted that only the MuST-Cinema data contain break symbols. We use the same Transformer architecture as in the cascade system. For all the experiments we use the fairseq toolkit (Gehring et al., 2017). Models are trained until convergence. Byte-Pair Encoding (BPE) (Sennrich et al., 2016) is set to 8K operations for the E2E-small and E2E systems while to 50K joint operations for the Cascade and the Template systems.

3.3 Evaluation

To evaluate translation quality we use BLEU (Papineni et al., 2002) against the MuST-Cinema test set, both with the break symbols and after removing them (BLEU-nob). For BLEU, an incorrect break symbol would account for an extra n-gram error in the score computation, while BLEU-nob allows us to evaluate only the translation quality without taking into account the subtitle segmentation.

For evaluating the conformity to the constraint of length, we calculate the percentage of subtitles with a maximum length of 42 characters per line (CPL), while for reading speed the percentage of sentences with maximum 21 characters per second (CPS). Since the MuST-Cinema data come from TED talks, these values were chosen according to the TED subtitling guidelines[5].

Finally, for judging the goodness of the segmentation, i.e. the position of the breaks in the translation, we mask all words except for the break symbols and compute Translation Edit Rate (Snover et al., 2006) only for the breaks against the reference translation (TER-br). This will allow us to determine the effort required by a human subtitler to manually correct the segmentation.

4 Results

4.1 Translation quality

The results are shown Table 1. As far as translation quality is concerned, the best performance is reached, as expected, in the Template setting, where the MT system is provided with "perfect" source language transcriptions. On the MuST-Cinema test set, this leads to BLEU scores of 30.62 and 22.08 respectively for French and German. The Cascade setting follows with a BLEU score reduction of 8 points for French and 4 points for German, which can be attributed to error propagation from the ASR component to the MT.

The E2E-small model (trained solely on MuST-Cinema) achieves 18.76 and 11.92 BLEU points for French and German respectively, which is a relatively low performance compared to the rest of the systems. This "low-resource" setting is a didactic experiment aimed at exploring how far the data-hungry neural approach can go with the limited amount of data available in the domain of ST for subtitling (280K and 234K sentences). It should be also noted that MuST-Cinema is the only Speech Translation corpus of the subtitle genre, both respecting the subtitling constraints and containing break symbols. Consequently, the interference of other data may hurt the conformity to the subtitling constraints, despite improving the translation performance. On the other hand, pre-training has evolved in a standard procedure for coping with the data-demanding nature of NMT. Therefore, pre-training also in the case of end-to-end speech translation offers a comparable setting with the template and the cascade experiments. Indeed, after fine-tuning the pre-trained model on MuST-Cinema, E2E reaches 22.22 and 17.28 BLEU points for French and German. The difference in translation quality is not statistically significant between the Cascade and the E2E, with the

[5] https://translations.ted.com/TED_Translator_Resources:_Main_guide

		BLEU↑	BLEU-nob↑	CPL↑	CPS↑	TER-br↓
FR	Template	30.62	28.86	91%	68%	18
	Cascade	22.41†	22.06	93%	72%	22
	E2E-small	18.76	18.03	95%	70%	23
	E2E	22.22†	21.9	95%	70%	20
DE	Template	22.08	21.10	90%	56%	17
	Cascade	17.81†	17.82	90%	56%	21
	E2E-small	11.92	11.38	93%	55%	24
	E2E	17.28†	16.90	92%	56%	20

Table 1: Results for translation quality (BLEU, BLEU-nob), for conformity to the subtitling constraints (CPL, CPS) and for subtitle segmentation (TER-br) for the four systems. Results marked with † are not statistically significant.

Cascade scoring higher with 0.2/0.6 BLEU points for French/German respectively. This shows that when increasing the size of the training data for the E2E the gap between the cascade and the end-to-end approach is closed and that end-to-end approaches may have finally found a steady ground for flourishing in different applications.

4.2 Conformity to the subtitling constraints and subtitle segmentation

When it comes to the conformity to the subtitling constraints, the results show a different picture (see CPS and CPL of Table 1). E2E exceeds all models at achieving proper length of subtitles, with 95% and 93% of the subtitles having length of maximum 42 characters. E2E achieves higher conformity with length even compared to the Template, for which the segmentation is already provided to the system in the form of break symbols, while the cascade is behind by 2%. The same tendency is observed in the TER-br results computed to measure the proper placement of the break symbols. While the Template benefits from the source language segmentation and therefore requires less edits to properly segment the subtitles, the Cascade is disadvantaged in guessing the correct position of the break symbols, as shown by a 22 and 21 TER score. For E2E-small TER-br is higher, possibly due to the low translation quality. However, E2E outperforms the Cascade by 1 TER point in this respect, showing that less effort would be required to segment the sentences into subtitles. This suggests that the E2E system receives information compared to the Cascade, which allows for better guessing the positions of the break symbols in the translation. This is another indication that subtitle segmentation decisions are not solely determined by reaching a maximum length of 42 characters, but a combination of multiple (possibly intersemiotic) factors can offer a better answer to automatic subtitling.

5 Analysis

The higher scores for CPL and TER-br in Section 4 suggest that the E2E system is better at modelling the subtitling constraints of length and proper segmentation. In this section we shed more light into this aspect by analysing factors which might be determining the system's behaviour in relation to the insertion of the break symbols <eol> and <eob>.

One question quickly arising is how the system can determine whether to insert a subtitle break symbol <eob> (which means that the next subtitle will follow on a new screen) or a line break symbol <eol> (which means that the next line of the subtitle will appear on the same screen). Since the maximum number of lines allowed per subtitle block is 2, a simple answer would be to alternate between <eob> and <eol> such that all subtitles would consist of two lines (two-liners). However, anyone having watched a film with subtitles is aware that subtitles can be two-liners or one-liners. Coming back to the example in Section 3.1, depending on the choice of break symbols (except for the last symbol which should always be an <eob>), there are two possible renderings of the subtitle:

```
10
00:00:31,066 --> 00:00:34,390
This kind of harassment keeps women
from accessing the internet --
11
00:00:34,414 --> 00:00:36,191
essentially, knowledge.
```

and

```
10
00:00:31,066 --> 00:00:34,390
This kind of harassment keeps women
11
00:00:34,414 --> 00:00:36,191
from accessing the internet --
essentially, knowledge.
```

Only the first rendering is acceptable because it satisfies the reading speed constraint but also

corresponds to the speech rhythm, since the speaker makes a pause after uttering the word "internet". In this case, how can the MT system determine which type of break symbol to insert?

We have mentioned in Section 2.1 that the *in* and *out* times of a subtitle should follow the rhythm of speech. Therefore, we expect that the end of a subtitle block, which in our setting is signalled by the break symbol <eob>, should correspond to the end of a speech act, a pause or a terminal juncture. On the other hand, line breaks inside the same subtitle block, which in our work correspond to the break symbol <eol>, have a different role. While line breaks can still overlap with pauses or signal the change of speaker, their function is to split a long subtitle into two smaller parts in order to fit the screen. The decision of where to insert a line break inside a subtitle block is determined by two factors: achieving a more or less equal length of the upper and the lower subtitle line and inserting the break in a position such that syntactic units are kept together. Consequently, the insertion of <eol> is determined more by the length and the syntactic properties of the subtitle and less by the natural rhythm of the speech.

If the hypothesis above holds, the choice of whether to insert an <eob> or an <eol> symbol is defined by prosodic properties and not solely by reaching the maximum length of 42 characters. As a consequence, it is not a simple alternating procedure.

To test this hypothesis, we compute the duration of the pause coming after each word in the source side of the MuST-Cinema test set. To achieve this, we perform forced alignment of the transcript against the audio and subtract the end time of each word from the start time of the next word:

$$pause_{w1w2} = start\_time_{w2} - end\_time_{w1} \quad (1)$$

Then we separate the pauses in 3 groups: *i)* pauses corresponding to positions where <eob> is present, *ii)* pauses corresponding to positions where <eol> is present and *iii)* pauses after which there is no break symbol (*None*). In Table 2 we report average and standard deviation of the pause duration for each category.

Pauses corresponding to the positions where <eob> symbols are present are more than x10 longer than the pauses in positions without any break symbols (None). Even if we take the most extreme cases (based on standard deviation), any pause above 0,37 seconds requires the insertion of <eob>. Pauses corresponding to <eol> symbols are on average x2 longer, but there is an overlap between the possible durations of the *None* and the <eol> category. This confirms our hypothesis about the different roles of the two subtitle breaks. Therefore, prosodic information is an important factor which can help the ST system determine the subtitling segmentation according to the speech rhythm. This finding provides strong evidence towards a clear limitation of the cascade setting, where the raw textual transcription from the ASR does not provide any prosodic information to the MT system. The MT system in the cascade setting is disadvantaged by the inability to: *i)* recover from possible ASR errors, and *ii)* make decisions determined by factors other than text.

Pause type	Avg	Stdev
None	0.039	0.022
<eob>	0.551	0.181
<eol>	0.074	0.027

Table 2: Average pause duration and standard deviation (in seconds) for the category without breaks (None), and for the categories with the two types of break symbols <eob> and <eol>.

With this knowledge, we analyse the breaks in the results of the two systems. In order to control for differences in translation, we select all sentences with at least 2 breaks and with the same number of break symbols (regardless of whether <eob> or <eol>) between the reference, the output of the Cascade and of the E2E. The resulting sentences are 137 for French and 158 for German. We calculate the accuracy of the type of break symbols for the two systems. For French the accuracy is 89% for the Cascade and 93% for the E2E. For German the accuracy is 85% for the Cascade and 88% for the E2E. This difference in accuracy suggests that the E2E is aided by the acoustic information and specifically by the pause duration in determining the correct break symbol.

Table 3 presents some examples, evaluated also against the video. In the first example, the decision of which type of break to insert between the two sentences can only be determined by the duration of the pause that comes between them. Indeed, the speaker in the video asks the question and then leaves some time to the audience before giving the answer. The pause between the two sentences is about 2 seconds. In this case, the first sentence

EN	“Who do you report to?” <eob> “It depends”.
CS	Wen melden Sie an? <eol> Es hängt davon ab.
E2E	Wen berichten Sie? <eob> Es kommt darauf an.
REF	“An wen schickst du deine Berichte?” <eob> “Das kommt darauf an”.
One executive at another company <eob> likes to explain how he used to be <eol> a master of milestone-tracking.	
Un cadre d’une autre entreprise aime expliquer <eob> comment il était jadis un maître <eol> du trek capital.	
Un dirigeant d’une autre entreprise <eob> aime expliquer comment il était <eol> un maître de traçage en pierre.	
Un cadre d’une autre entreprise <eob> aime raconter comme il était passé maître <eol> dans la surveillance des étapes.	
But you know how they say <eol> that information is a source of power?	
Vous savez comment dire que l’information <eol> est une source de pouvoir.	
Saviez-vous comment dire <eol> que l’information est une source de pouvoir ?	
Mais vous savez qu’on dit que <eol> l’information est source de pouvoir ?	

Table 3: Examples of translations by the cascade (CS) and the end-to-end model (E2E) compared to the source sentence (EN) and the reference (REF). The sentence-final <eob> has been removed.

should be in one subtitle block, then disappear, and the second sentence should come in the next subtitle block in order not to reveal the answer before it was spoken by the speaker. This information is only available to the E2E system.

In the second example, although both systems have chosen the right type of break, there are differences in the actual positions of the breaks. The E2E inserts the first break symbol in the same position as in the source and reference (*entreprise*), while the Cascade inserts it at a later position (*expliquer*), resulting in a subtitle of 46 characters, which is above the 42-character length limit. The Cascade correctly inserts the second break (*maître*), as in the reference. Here, the E2E copies the break position from the source sentence, which is in a different position compared to the reference (after the word *be* instead of the word *master* as in the reference). The E2E is faithful to the segmentation of the source language when it corresponds to the pauses of the speaker. The positions chosen by the Cascade to insert the break symbols are before a conjunction (*comment*) and a preposition (*du*). Contrary to the E2E, the Cascade’s decisions are based more on syntactic patterns, learned from the existing human segmentation decisions in the training data.

The third example shows that prosody is important also for other factors related to the translation. The Cascade, not receiving any punctuation, was not able to reproduce the question in the translation, while the intonation might have helped the E2E to render the sentence as a question despite using the wrong tense (*saviez* instead of *savez*).

All in all, these examples confirm our analysis and once again indicate the importance of considering the intersemiotic nature of subtitling when developing MT systems for this task.

6 Conclusion

We have presented the first Speech Translation systems specifically tailored for subtitling. The first system is an ASR-MT cascade, while the second a direct, end-to-end ST system. These systems allow, for the first time, to create satisfactory subtitles both in terms of translation quality and conformity to the subtitling constraints in the absence of a human transcription of the source language speech (template). We have shown that while the two systems have similar translation quality performance, the E2E seems to be modelling the subtitle constraints better. We show that this could be attributed to acoustic features, such as natural pauses, becoming available to the E2E system through the audio input. This leads to a segmentation closer to the speech rhythm, which is key to a pleasant user experience. Our work takes into account the intersemiotic nature of subtitling by avoiding conditioning the translation on the textual source language length, as in previous approaches to NMT for subtitling, arriving to the conclusion that 42 is not the answer to everything in the case of NMT for subtitling. Rather, key elements for good automatic subtitling are prosodic elements such as intonation, speech tempo and natural pauses. We hope that this work will pave the way for developing more comprehensive approaches to NMT for subtitling.

Acknowledgments

This work is part of the “End-to-end Spoken Language Translation in Rich Data Conditions” project,[6] which is financially supported by an Amazon AWS ML Grant.

[6] https://ict.fbk.eu/units-hlt-mt-e2eslt/

References

Pablo Daniel Aguero, Jordi Adell, and Antonio Bonafonte. 2006. Prosody generation in the speech-to-speech translation framework. In *Acoustics, Speech and Signal Processing, ICASSP 2006*, volume 1. IEEE International Conference.

Tanel Alumäe and Mikko Kurimo. 2010. Efficient estimation of maximum entropy language models with n-gram features: an SRILM extension. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1820–1823.

G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black. 2012. Intent transfer in speech-to-speech machine translation. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 153–158.

Rosa Alexandra Assis. 2001. Features of oral and written communication in subtitling. *(Multi)Media Translation. Concepts, Practices and Research*, pages 213–221.

Wilker Aziz, Sheila Castilho Monteiro de Sousa, and Lucia Specia. 2012. Cross-lingual sentence compression for subtitles. In *16th Annual Conference of the European Association for Machine Translation*, EAMT, pages 103–110, Trento, Italy.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

Aljoscha Burchardt, Arle Lommel, Lindsay Bywood, Kim Harris, and Maja Popović. 2016. Machine translation quality in an audiovisual context. *Target*, 28(2):206–221.

Lindsay Bywood, Panayota Georgakopoulou, and Thierry Etchegoyhen. 2017. Embracing the threat: machine translation as a solution for subtitling. *Perspectives*, 25(3):492–508.

Lindsay Bywood, Martin Volk, Mark Fishel, and Panayota Georgakopoulou. 2013. Parallel subtitle corpora and their applications in machine translation and translatology. *Perspectives*, 21(4):595–610.

Mary Carroll and Jan Ivarsson. 1998. *Code of Good Subtitling Practice*. Simrishamn: TransEdit.

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. 2019a. 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 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota.

Mattia Antonino Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019b. Robust neural machine translation for clean and noisy speech transcripts. In *16th International Workshop on Spoken Language Translation (IWSLT)*.

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2019c. Adapting transformer to end-to-end spoken language translation. In *INTERSPEECH*, pages 1133–1137.

Quoc Truong Do, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. 2016. Transferring emphasis in speech translation using hard-attentional neural network models. In *17th Annual Conference of the International Speech Communication Association (InterSpeech 2016)*, San Francisco, California, USA.

Quoc Truong Do, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2015. Improving translation of emphasis with pause prediction in speech-to-speech translation systems. In *12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 Evaluation Campaign. In *International Workshop on Spoken Language Translation (IWSLT)*.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Panayota Georgakopoulou. 2019. Template files: The holy grail of subtitling. *Journal of Audiovisual Translation*, 2(2):137–160.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *International Workshop on Spoken Language Translation (IWSLT*.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Takatomo Kano, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2013. Generalizing continuous-space translation of paralinguistic information. In *14th Annual Conference of the International Speech Communication Association (InterSpeech 2013)*, Lyon, France.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2019. Are Subtitling Corpora really Subtitle-like? In *Sixth Italian Conference on Computational Linguistics, CLiC-It*.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. MuST-Cinema: a Speech-to-Subtitles Corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the Output Length of Neural Machine Translation. In *Proceedings of the 16th International Workshop on Spoken Language Translation, (IWSLT)*.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Matteo Negri, Marcho Turchi, and Marcello Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 15th International Workshop on Spoken Language Translation, (IWSLT)*, Bruges, Belgium.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Matteo Negri, Marcho Turchi, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico. 2019. The IWSLT 2019 Evaluation Campaign. In *Proceedings of the 16th International Workshop on Spoken Language Translation, (IWSLT)*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Belgium, Brussels. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proc. of Interspeech*, pages 2751–2755.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of Translation Edit Rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *International Workshop on Spoken Language Translation (IWSLT)*.

Christopher Taylor. 2016. The multimodal approach in audiovisual translation. *Target*, 2(28).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine Translation of TV Subtitles for Large Scale Production. In *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC'10)*, pages 53–62, Denver, CO.