INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Subtitle Translation as Markup Translation

Colin Cherry, Naveen Arivazhagan, Dirk Padfield, Maxim Krikun

Google Research, USA
{colincherry, navari, padfield, krikun}@google.com

Abstract

Automatic subtitle translation is an important technology to make video content available across language barriers. Subtitle translation complicates the normal translation problem by adding the challenge of how to format the system output into subtitles. We propose a simple technique that treats subtitle translation as standard sentence translation plus alignment driven markup transfer, which enables us to reliably maintain timing and formatting information from the source subtitles. We also introduce two metrics to measure the quality of subtitle boundaries: a Timed BLEU that penalizes mistimed tokens with respect to a reference subtitle sequence, and a measure of how much Timed BLEU is lost due to suboptimal subtitle boundary placement. In experiments on TED and YouTube subtitles, we show that we are able to achieve much better translation quality than a baseline that translates each subtitle independently, while coming very close to optimal subtitle boundary placement.
Index Terms: machine translation, speech translation, subtitles

1. Introduction

Video content is becoming increasingly important as a means of distributing and consuming information and entertainment, including how-to videos, lectures, journalism and many other formats. As video content becomes more prominent, the most efficient way to make it available outside its recorded language is to provide foreign-language subtitles. One can view automatic foreign-language subtitling either as a speech translation task with audio input and presentation constraints on the output [1], or as a subtitle translation task, where we assume the input is in the form of a subtitle template consisting of carefully formatted source text [2]. Both views are valuable: the former enables translation of any content, while the latter enables high-quality translation of content that has already been subtitled. Figure 1 shows a pair of translated subtitles. Its source is an example of template input, and its target shows the desired output of either approach.

In this work, we focus on the case of translating source-language subtitles. A lot of human intelligence goes into placing subtitle boundaries, which account not only for the syntactic and semantic structure of the text to improve readability, but also audio and visual cues. Audio cues include prosody and pauses, while visual cues include the presence and position of a speaker on the screen, and the amount of other textual content in the video, which might lead the subtitler to compress their transcription [3]. Furthermore, the source subtitles may contain important display meta-data beyond timestamps, indicating the position, color or style of the text. Throughout this work, we maintain a one-to-one mapping between source and target subtitles, preserving all subtitle timing and meta-data, along with whatever intelligence went into these annotations.

Our primary technical contribution is to build upon a baseline that has been forgotten in recent work [1, 4], namely that a subtitle template typically takes the form of marked-up source text, and therefore subtitle translation can be treated as markup translation [5]. This allows us to easily maintain meta-data and to benefit from work put into markup translation for other tasks, such as HTML translation. We present a concise, alignment-driven approach to markup translation that enables the incorporation of additional constraints. Furthermore, we present a novel evaluation metric for the quality of subtitle boundary placement given reference subtitles. We define a version of BLEU that is sensitive to timing decisions (Timed BLEU), and then isolate the contribution of boundary placement by reporting how much Timed BLEU could be recovered with optimal boundaries. Finally, we present an evaluation of our markup translation approach on TED and YouTube data.

Source	00:00:08.180 --> 00:00:11.579 I went to school, I hung out with my friends, 00:00:11.580 --> 00:00:13.965 I fought with my younger sisters.
Target	00:00:08.180 --> 00:00:11.579 Ich ging zur Schule, hing mit Freunden ab, 00:00:11.580 --> 00:00:13.965 stritt mich mit meinen jüngeren Schwestern.

Figure 1: *English to German subtitle translation.*

2. Related work

Recent work in automatic foreign language subtitling has assumed access to training data with reference subtitles. Matusov et al. [4] propose a neural model to predict subtitle boundaries in target text. The model only has access to the target sentence, but the inference procedure combines the model's score with other heuristics that account for the target length constraints and the length of the corresponding source subtitles. Karakanta et al. [1] propose an end-to-end speech translation approach, where the translation system learns to go from audio to a target string that contains subtitle boundaries as special tokens. This enables them to account for audio cues when placing subtitle boundaries. They do not assume access to source subtitles, and therefore cannot benefit from the information contained therein. Interestingly, the end-to-end approach often has better subtitle placement than their oracle that has access to the source subtitles, which indicates that there is room to refine both the metrics used to measure boundary placement and the algorithms used to transfer boundaries to translated subtitles. Our method differs from both of these approaches in that we use a simple, untrained tag transfer mechanism to project source subtitle boundaries onto a target translation. This guarantees a one-to-one mapping between source and target subtitles, enabling meta-data transfer. Since the method is untrained, it requires target subtitles only for evaluation.

An orthogonal line of work looks at length controls for neu-

http://dx.doi.org/10.21437/Interspeech.2021-744

ral machine translation [6, 7]. These works devise methods to constrain the translation to some relative length of the source sentence, often using an encoding that tells the decoder at each step how many tokens are left until reaching the desired length. These works use HTML translation and subtitle translation as motivating cases, and we see them as enabling technologies for our tag transfer approach, enabling us to focus on correct semantic boundary placement. We assume that the translation can be made to fit any characters-per-second or characters-per-line length constraints imposed by the subtitling guides.

Another enabling technology for our approach is automatic subtitling from audio. Here, the task is to use automatic speech recognition to transcribe the spoken content of the video, with the added challenge of adjusting transcription length and formatting to satisfy display constraints. Desired lengths can be achieved with length-controlled speech recognition [8], or by fine-tuning on subtitles rather than full transcriptions [3].

3. Subtitle Boundary Projection

Our proposed method treats subtitle boundary placement as a problem of markup tag translation. In particular, we adopt the detag + project method, which has been shown to be roughly as good as methods that incorporate tag translation directly into neural machine translation (NMT) [5]. This approach can be decomposed into word-alignment followed by projection.

We opt to treat alignment as a post-translation process that aligns the source to the output translation, rather than attempt to derive an alignment from the NMT attention mechanism. Our aligner is an HMM aligner [9] trained on the output of the NMT model.

Given a word alignment A, represented as a set of links between source and target tokens, we project the boundaries onto the output with the following algorithm:

- For each target token t, find all linked source tokens $s : (s, t) \in A$, and take the index of the corresponding source caption segment; this gives for each target token t a set $X[t]$ (possibly empty) of integers.
- Find a non-decreasing sequence $z[t]$ such that the size of the set $\{t : z[t] \notin X[t]\}$ is minimized
- Define target segment n to include all target tokens for which $z[t] = n$.

We find that such an algorithm has two nice properties – it is able to ignore some word alignment errors, and the optimization problem can be easily modified to include additional constraints, such as desired segment length.

4. Evaluation Metrics

When translating subtitles, there are two main characteristics of interest: translation quality and subtitle boundary quality. We report case-sensitive, tokenized BLEU [10] as our primary quality metric. Since some of the methods under test are not guaranteed to generate exactly one target sentence per input source sentence, we first align the unsegmented translation with the provided reference sentences by minimizing WER [11].

Our test sets provide human subtitle boundaries not only in the source but also in the reference. Therefore, we can measure the quality of our subtitle boundaries by comparing them to the reference boundaries. We assume that the human subtitler who produced the reference was sensitive to all the competing syntactic, semantic, audio and visual pressures on subtitle boundaries. Karakanta et al. [1] propose TER-br to measure boundary errors, where subtitle boundaries are rendered as special tokens in the text, and TER is calculated with all other tokens masked. We found this metric to lack sufficient resolution in preliminary experiments. Instead, we prefer the "S mode" BLEU from Matusov et al. [4] (S-BLEU), which calculates BLEU on subtitles instead of sentences, so that any target words that appear in the wrong subtitle count as errors.

S-BLEU assumes that the subtitles in the target and the reference match, both in their number and their timings. While this would hold if both were derived from the source template, this is not always true for our reference subtitles. Therefore, we generalize S-BLEU to *Timed BLEU (T-BLEU)*, which assigns timestamps to each token in the target by linearly interpolating the subtitle timings. The target is then aligned to the reference's subtitle segments based on temporal overlap. Let y and y^* be target and reference subtitle segments, with timings given by functions $\mathrm{start}(\cdot)$ and $\mathrm{end}(\cdot)$, and sequence length in tokens given by $|\cdot|$. We then assign start times to tokens y_t for $0 \leq t < |y|$:

$$\mathrm{start}(y_t) \coloneqq \mathrm{start}(y) + t\frac{\mathrm{end}(y) - \mathrm{start}(y)}{|y|} \quad (1)$$

Token y_t is aligned with the reference segment y^* iff $\mathrm{start}(y^*) \leq \mathrm{start}(y_t) < \mathrm{end}(y^*)$. This temporal alignment creates target-reference segment pairs, over which BLEU can be calculated normally. In the scenario where target and reference subtitlings are temporally identical, T-BLEU reduces to S-BLEU.

T-BLEU combines translation quality and boundary quality in a manner that can be difficult to interpret. We therefore propose a novel metric that isolates the boundary component: T-BLEU Headroom (TBHR). We first calculate an upper-bound on T-BLEU: we ignore the system's predicted subtitle boundaries, and instead align the unsegmented translation with the reference *subtitles* by minimizing WER. This gives us subtitle boundaries that approximately maximize T-BLEU. We then report the difference between this upper bound and T-BLEU. This provides a lower-is-better boundary error rate, interpretable as the amount of T-BLEU that could be recovered by improving only boundary placement.

5. Experimental setup

5.1. Data

5.1.1. TED talks

We create our first test set by directly scraping raw transcripts and translations from ted.com.[1] We scrape the 12 talks from the IWSLT 2015 test set [12]. These transcripts are annotated with subtitle boundaries and timing, which we use when evaluating T-BLEU and TBHR. TED subtitles are translated and reviewed by volunteers, so scraping directly from TED also enables us to get all available target languages: we chose Arabic (ar), German (de), French (fr), Japanese (ja), Korean (ko), and Russian (ru) to get a wide variety of character sets, morphological properties and word orders. Since we do not need audio for our subtitle translation experiments, we can report results both in the natural En→XX direction, and in the inverse XX→En.

5.1.2. YouTube videos

TED talks have relatively high quality captions and translations, but they are a constrained domain: well-practiced, educational,

[1] `ted.com/talks/subtitles/id/$TALK_ID/lang/$LANG`

single-speaker talks. They are also predominantly English-original. Therefore, we also evaluate on a test set drawn from YouTube. We collect videos that have user-contributed source- and target-language subtitle tracks in addition to an automatically derived (through speech recognition) source subtitle track. The lack of review in user subtitle contribution means the quality of foreign-language subtitles can vary substantially. We collect a sample of 100 videos for each of our directional language-pairs. Furthermore, since YouTube covers much more content, we can collect videos with either an English speaker and another language in the translation, or a non-English speaker and an English translation. Following a long-standing [13], but recently abandoned [14], tradition of WMT shared tasks, for each non-English language XX, we combine the En-audio and XX-audio raw data into a single test set of 200 videos, which can be used for either En→XX or XX→En evaluation. This enables us to study whether it affects evaluation to use the original text as the source or the reference, as this can be a source of bias in standard MT research [15].

Finally, for English-original videos, we experiment with a test set where the source subtitle track is automatically derived from speech recognition, formatted into subtitles using a combination of pause-based cues and a 42-character line length limit. This enables us to test how well our system's automatically-derived foreign-language subtitle track compared to one provided by a human uploader.

5.2. Baselines

Our baseline system translates each subtitle independently, making the alignment to the source subtitles trivial. This may seem linguistically offensive by omitting crucial in-sentence context, but it is a reasonable baseline for the same reasons that subtitle boundary placement is important and hard: subtitles are often coherent stand-alone semantic units. However, since semantic coherence is just one of the properties optimized by human subtitle boundaries, this baseline tends to favor subtitle boundary quality at the expense of translation quality.

5.3. Implementation

In the spirit of maximal reuse of existing components, we carry out subtitle translation using an internal MT system similar to the one available from the Google Cloud Translation API. This MT engine has been trained on large amounts of web-mined parallel text and is considered to be a general-domain system. We send input to the translation engine either one sentence at a time using an internal sentence-breaker, or one subtitle at a time for the baseline. In the case of sentence-level translation, we re-establish subtitle boundaries using the alignment-based projection described in Section 3.

6. TED Results

Our experimental results on the TED En→XX test sets are shown in Table 1. This table demonstrates that translating whole sentences and projecting boundaries offers substantially better BLEU (+3.1) on average. This comes at the cost of reduced boundary accuracy: our TBHR is 1.2 points higher than the baseline, but we believe that this is a good trade, since absolute TBHR of our system is relatively low, with an average of 2.0.[2] This indicates that attempts to improve boundary placement further have limited headroom. As one might expect, the languages with the most substantial word-order differences due to their subject-object-verb syntax (Japanese, Korean) have the highest absolute TBHR scores, likely due to a combination of reduced word alignment quality, and increased difficulty of finding a good projection.

Table 1: *TED En→XX results.*

Target	System	BLEU ↑	T-BLEU ↑	TBHR ↓
ar	base	18.9	19.3	0.7
	project	20.9	19.8	1.7
de	base	27.5	28.3	0.6
	project	33.3	31.1	1.8
fr	base	39.5	39.3	1.0
	project	42.6	40.9	1.5
ja	base	14.4	10.1	1.0
	project	16.2	10.9	2.5
ko	base	13.8	12.6	1.3
	project	16.0	12.7	3.4
ru	base	21.1	21.6	0.3
	project	25.0	23.7	1.0

Table 2: *TED XX→En results.*

Source	System	BLEU ↑	T-BLEU ↑	TBHR ↓
ar	base	27.3	29.4	1.6
	project	36.9	33.6	2.9
de	base	31.4	31.3	2.1
	project	36.1	33.3	3.3
fr	base	39.0	38.3	1.4
	project	42.0	40.0	2.1
ja	base	10.3	9.5	0.8
	project	12.8	9.3	2.8
ko	base	16.4	14.6	2.3
	project	22.2	17.3	4.8
ru	base	24.6	25.5	0.3
	project	28.9	27.6	1.2

For the TED XX→En results in Table 2, the broad trends established in En→XX hold, although the absolute numbers have changed. With respect to the baseline, there continue to be large improvements from sentence-level translation and similar degradations in boundary quality. Absolute TBHR values for our projection have an average of 2.9, so there is more room for improved boundary placement in this direction but still not much overall.

The example in Figure 2 contrasts the two systems. The baseline's translation for subtitle 2 is defensible without the context of what was being critiqued, but it is nonsensical in context. The projection uses sentential context to produce a better translation, but it also makes a boundary error that would be impossible for the baseline, moving a comma from the end of subtitle 1 to the beginning of 2.

7. YouTube Results

Our YouTube results for En→XX are shown in Table 3; XX→En is omitted for space, but is similar. The trends here are broadly similar to those observed in TED, which is encourag-

[2] We assume T-BLEU and TBHR behave similarly to BLEU, allowing us to transfer intuitions on meaningful score differences; for example, we expect a difference of 1.0 to be visible to attentive users.

	Reference	Baseline	Projection
1:	Now, Edsger Dijkstra, when he wrote this,	Edsger Dijkstra, in writing it,	Edsger Dijkstra, in writing it
2:	intended it as a criticism	ran it like a review	, directed it as a critique
3:	of the early pioneers of computer science,	pioneers in computer science,	of pioneers in computer science,

Figure 2: *Example of baseline versus projected translation of Fr→En.*

Table 3: *YouTube En→XX results.*

Target	System	BLEU ↑	T-BLEU ↑	TBHR ↓
ar	base	19.2	19.5	0.4
	project	19.3	19.4	0.7
de	base	32.8	33.7	0.2
	project	34.2	34.0	0.6
fr	base	35.5	34.5	1.4
	project	36.5	35.1	1.5
ja	base	13.1	10.4	0.2
	project	14.2	10.7	1.1
ko	base	12.2	11.5	0.5
	project	12.8	11.4	1.2
ru	base	19.2	18.4	0.1
	project	21.1	19.5	0.3

Table 4: *YouTube De→En by audio language of evaluation set.*

Audio	System	BLEU ↑	T-BLEU ↑	TBHR ↓
de	base	44.1	44.4	0.8
	project	43.8	43.6	1.1
en	base	36.1	36.2	1.2
	project	39.3	37.8	1.8
Combined	base	40.7	41.0	0.9
	project	41.9	41.2	1.4

ing since YouTube has more domain diversity and half of each test set was recorded from non-English speakers. For Arabic and Korean, we now see Timed BLEU regressions, indicating that for the translation-versus-boundary-quality trade-off captured by this combined metric, we are not recovering enough translation quality with respect to the baseline to make up for reduced boundary quality. Also, for translation out of English, absolute THBR numbers for both the baseline and the proposed projection are extremely low (averaging less than 1 T-BLEU point lost), indicating that boundary placement is very close to human placement.

7.1. Biases in translated references

We observed a marked difference on YouTube in the performance of our projection depending on whether captions in the audio language are used as the source or the reference for evaluation. Table 4 shows De→En results as an example. Our projection fails to beat the baseline on any metric for German-original videos, while it shows a 3-point quality improvement on English-original. Looking into the references manually, we found that this discrepancy could be explained by the tendency of YouTube's user-contributed translations to translate each subtitle independently, not necessarily picking incorrect translations but awkward ones influenced by source subtitle boundaries. When translating into the original language, there is no opportunity for boundaries to influence word choices, so the picture is more in keeping with observations on TED data. This highlights a subtitle-specific problem when evaluating against translated references: boundary bias can be added to concerns regarding translationese and naturalness [15].

Table 5: *YouTube En_ASR →XX results.*

Target	System	BLEU ↑	T-BLEU ↑	TBHR ↓
ar	base	10.7	9.8	1.6
	project	13.9	12.7	1.9
de	base	14.6	12.8	1.7
	project	20.2	17.1	2.6
fr	base	19.8	17.9	2.4
	project	25.3	22.4	3.0
ja	base	11.9	7.4	0.9
	project	14.4	8.0	2.6
ko	base	7.7	5.9	1.4
	project	11.2	7.5	3.0
ru	base	10.6	9.0	1.5
	project	15.4	12.7	2.2

7.2. Results on Automatic Subtitles

Results on Youtube ASR data can be found in Table 5. Comparing BLEU scores against Table 3, we see that ASR errors have a drastic impact on translation quality. As already mentioned, subtitle boundaries in ASR data are heuristically inserted based on time-stamps and subtitle length constraints. They are unlikely to match semantic boundaries like in human captions. For this reason, our projection approach yields a much larger improvement to BLEU than the baseline, which no longer benefits from carefully inserted human boundaries. In a similar vein, both approaches yield larger TBHRs because the source ASR boundaries are unlikely to match human subtitle boundaries on the target.

8. Conclusion

We have presented an alignment-based markup transfer baseline for subtitle translation. Our method guarantees a one-to-one mapping between source subtitles and target subtitles, enabling well-defined transfer of all meta-data. We have also proposed two metrics for subtitle boundary quality: Timed BLEU, which combines translation and boundary quality, and Timed BLEU Headroom, which measures the amount of Timed BLEU that is lost due to suboptimal boundary placement. Experiments on TED and YouTube videos show that markup transfer results in near-optimal boundary placement, with less than 2 Timed BLEU points that can be recovered going out of English, and less than 3 going into English. In the future, we plan to integrate length-controlled translation into this approach to ensure fully length-compliant, properly formatted, high-quality subtitle translations.

9. References

[1] A. Karakanta, M. Negri, and M. Turchi, "Is 42 the answer to everything in subtitling-oriented speech translation?" in *Proceedings of the 17th International Conference on Spoken Language Translation*. Online: Association for Computational Linguistics, Jul. 2020, pp. 209–219. [Online]. Available: https://www.aclweb.org/anthology/2020.iwslt-1.26

[2] P. Georgakopoulou, "Template files: The holy grail of subtitling," *Journal of Audiovisual Translation*, 2019.

[3] D. Liu, J. Niehues, and G. Spanakis, "Adapting end-to-end speech recognition for readable subtitles," in *Proceedings of the 17th International Conference on Spoken Language Translation*. Online: Association for Computational Linguistics, Jul. 2020, pp. 247–256. [Online]. Available: https://www.aclweb.org/anthology/2020.iwslt-1.30

[4] E. Matusov, P. Wilken, and Y. Georgakopoulou, "Customizing neural machine translation for subtitling," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 82–93. [Online]. Available: https://www.aclweb.org/anthology/W19-5209

[5] G. Hanneman and G. Dinu, "How should markup tags be translated?" in *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1160–1173. [Online]. Available: https://www.aclweb.org/anthology/2020.wmt-1.138

[6] S. M. Lakew, M. Di Gangi, and M. Federico, "Controlling the Output Length of Neural Machine Translation," in *Proceedings of the 16th International Conference on Spoken Language Translation*, Nov. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3524957

[7] J. Niehues, "Machine translation with unsupervised length-constraints," in *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Virtual: Association for Machine Translation in the Americas, Oct. 2020, pp. 21–35. [Online]. Available: https://www.aclweb.org/anthology/2020.amta-research.3

[8] K. Angerbauer, H. Adel, and N. T. Vu, "Automatic Compression of Subtitles with Neural Networks and its Effect on User Experience," in *Proc. Interspeech 2019*, 2019, pp. 594–598. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1750

[9] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996. [Online]. Available: https://www.aclweb.org/anthology/C96-2141

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040

[11] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating machine translation output with automatic sentence segmentation," in *International Workshop on Spoken Language Translation (IWSLT)*, 2005.

[12] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, "The IWSLT 2015 evaluation campaign," in *International Workshop on Spoken Language Translation (IWSLT)*, 2015.

[13] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, P. Koehn, and C. Monz, "Findings of the 2018 conference on machine translation (WMT18)," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 272–303. [Online]. Available: https://www.aclweb.org/anthology/W18-6401

[14] L. Barrault, M. Biesialska, O. Bojar, M. R. Costa-jussà, C. Federmann, Y. Graham, R. Grundkiewicz, B. Haddow, M. Huck, E. Joanis, T. Kocmi, P. Koehn, C.-k. Lo, N. Ljubešić, C. Monz, M. Morishita, M. Nagata, T. Nakazawa, S. Pal, M. Post, and M. Zampieri, "Findings of the 2020 conference on machine translation (WMT20)," in *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1–55. [Online]. Available: https://www.aclweb.org/anthology/2020.wmt-1.1

[15] M. Freitag, I. Caswell, and S. Roy, "APE at scale and its implications on MT evaluation biases," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 34–44. [Online]. Available: https://www.aclweb.org/anthology/W19-5204