A-TASC: Asian TED-Based Automatic Subtitling Corpus

Anonymous ACL submission

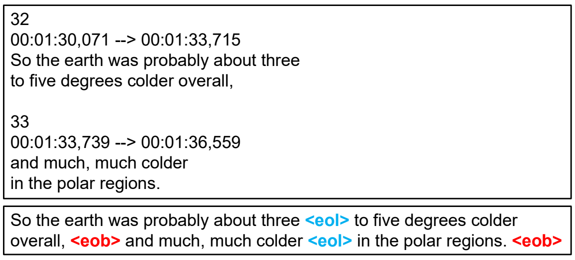
Abstract

Subtitles play a crucial role in improving the accessibility of the vast amount of audiovisual content available on the Internet, allowing audiences worldwide to comprehend and engage with those contexts in various languages. Automatic subtitling (AS) systems are essential in alleviating the substantial workload of human transcribers and translators. However, existing AS corpora and the primary metric SubER focus on European languages. This paper introduces A-TASC, an Asian TED-based automatic subtitling corpus derived from English TED Talks, comprising nearly 800 hours of audio segments, aligned English transcripts, and subtitles in Chinese and Japanese. Meanwhile, we present SacreSubER, a modification of SubER, to enable the reliable evaluation of the subtitle quality. Experimental results of an end-to-end AS system and pipeline approaches based on strong ASR and LLMs on our corpora confirm the quality of the proposed corpus and reveal differences in AS performance between European and Asian languages.

1 Introduction

The immense amount of audiovisual content has become a primary medium for information sharing, education, and entertainment. Subtitles play a vital role in allowing non-native speakers to access such content in their languages. However, the subtitling workflow is complex; 1) transcribing the audio content, 2) annotating the start and end timestamps of the transcriptions, and 3) translating the transcriptions into the target language. Thus, there is a growing demand for automatic subtitling (AS) systems to reduce the heavy workload of subtitling.

The growing demand for automatic subtitling urged researchers to generate subtitles automatically by leveraging automatic speech recognition (ASR) and machine translation (MT) (pipeline approaches) or spoken language translation (SLT) (end-to-end approaches). The major obstacle to the development of AS systems is the lack of language resources containing subtitle segmentation and timing information for training and evaluation. Such information is absent in the existing corpora for MT (Lison et al., 2018) and SLT (Di Gangi et al., 2019). Although the MuST-Cinema corpus (Karakanta et al., 2020) has been developed for automatic subtitling from an SLT corpus MuST-C (Di Gangi et al., 2019), the target languages are limited to European languages (German, Spanish, French, Italian, Dutch, Portuguese, and Romanian), which are close languages to the source language, English, and challenges in automatic subtitling to distant languages remain to be clarified. Furthermore, the primary metric for automatic subtitling, SubER (Wilken et al., 2022), leverages spaces to tokenize text and cannot be directly applied to *scriptio continua* languages such as Chinese and Japanese. These limitations obstruct the development and evaluation of multilingual AS systems supporting more languages.



```
32
00:01:30,071 --> 00:01:33,715
So the earth was probably about three
to five degrees colder overall,

33
00:01:33,739 --> 00:01:36,559
and much, much colder
in the polar regions.
```

```
So the earth was probably about three <eol> to five degrees colder
overall, <eob> and much, much colder <eol> in the polar regions. <eob>
```

Figure 1: Example of a sentence in the .srt subtitle file (top) and that annotated by subtitle breaks (bottom).

In this study, aiming to address the lack of resources for automatic subtitling, we present A-TASC, an Asian TED-based automatic subtitling corpus, and SacreSubER, the SubER metric integrated with SacreBLEU (Post, 2018)'s tokenizer for TER metric (Snover et al., 2006). A-TASC is composed of (audio, transcription, translation) triplets, where the translation contains special tokens marking the subtitle breaks (Figure 1). A-

TASC can be thereby used for AS as well as ASR, MT, and SLT. The data acquisition and processing scripts are released under CC BY-4.0 license.[1]

To confirm the quality and utility of the proposed corpus, we evaluate the latest AS model SBAAM (Gaido et al., 2024) on our corpus with different training set size and audio-text alignment approaches. We next compare the AS performance for different languages and analyze the gap between the latest end-to-end AS system and a pipeline approach that adopts Whisper (Radford et al., 2023) as the ASR model and DeepSeek-V3 (Liu et al., 2024) as the LLM for the MT model.

Our contributions are summarized as follows:

- We propose A-TASC, a large-scale TED-based AS corpus from English to two Asian languages, Chinese and Japanese.
- We present SacreSubER, which modifies SubER (Wilken et al., 2022) metric to support the evaluation of subtitles in Asian languages.
- We empirically confirm the utility and quality of the proposed corpus via the evaluation of end-to-end and pipeline AS approaches.
- We mention the limitation of SubER to evaluate automatic subtitling into distant target languages such as Japanese for English audio.

2 Related Work

In this section, we first introduce the subtitle-based corpora for tasks other than automatic subtitling. Next, we introduce the only existing corpus for automatic subtitling task and point out its limitations. Finally, we explain the task setting of AS and the recent development of AS systems.

2.1 Subtitle-based Corpora for Non-AS Tasks

The subtitles of audiovisual content have been exploited to create language resources for MT and SLT. The OpenSubtitles corpus (Lison et al., 2018) contains millions of parallel sentences extracted from movie and TV show subtitles, making it one of the largest publicly available parallel corpora across 60 languages. However, since it is aimed to be a corpus for MT, the audiovisual content is not involved in the corpus and is generally protected by copyright. Besides, the information of subtitle breaks is removed to obtain the aligned parallel text. Thus, it is hard to make use of it for AS task.

MuST-C (Di Gangi et al., 2019) is to date the largest multilingual corpus for SLT, aiming to provide sizeable resources for training and evaluating SLT systems. It is built from TED talks published between 2007 and April 2019, and contains (audio, transcription, translation) triplets aligned at sentence level. However, the subtitles were merged to create full sentences and the information about the subtitle breaks was removed. Thus, it cannot be used for the training of end-to-end AS systems.

2.2 Automatic Subtitling Corpora

To address the unique challenge of automatic subtitling (Ahmad et al., 2024) in segmenting the translated text into subtitles compliant with constraints that ensure high-quality user experience, MuST-Cinema (Karakanta et al., 2020) is developed and has been the only corpus for training and evaluating end-to-end AS systems. It is built on top of MuST-C, by annotating the transcription and the translation with two special tokens, `<eob>` and `<eol>`, to represent the two types of subtitle breaks: 1) block breaks, i.e., breaks denoting the end of the subtitle on the current screen, and 2) line breaks, i.e., breaks between two consecutive lines (wherever two lines are present) inside the same subtitle block. However, the target languages in MuST-Cinema are limited to seven European languages, and the subtitle breaks are inserted automatically, instead of actual subtitle breaks.

In this work, following the corpus creation method of MuST-Cinema, we create an automatic subtitling corpus for Asian languages while overcoming the above limitations. Moreover, unlike MuST-Cinema, we release the script to create the corpus from TED talk data, enabling easier data extension from the newly released TED talks.

2.3 Automatic Subtitling Systems and Metrics

Given an audio file, the goal of AS systems is to generate a subtitle file composed of subtitle blocks, each of which include a piece of translated text and the corresponding start and end timestamps. In what follows, we introduce existing AS approaches and metrics to evaluate AS systems.

AS systems can be categorized into pipeline and end-to-end approaches. The pipeline approach usually adopts ASR to generate transcriptions, and use a segmentation model trained on data with subtitle break annotations to segment the transcriptions into subtitle blocks. With the timed word list provided by the ASR system and the segmented tran-

[1] https://anonymous.4open.science/r/atasc-E087/

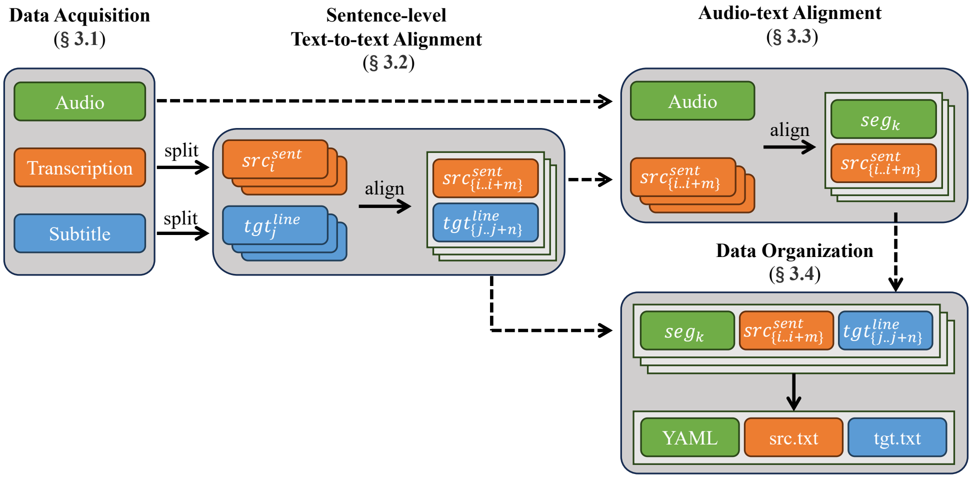


Figure 2: Overview of the corpus creation workflow of A-TASC.

scriptions, the timestamps of each block can be calculated. To generate the output subtitles, the text in each block are translated by an MT system and the timestamps are kept as the same. On the other hand, end-to-end approach generates translations with special symbols marking the subtitle line/block breaks directly from the audio. The special symbols are then managed to be aligned with audio frames to calculate the timestamps. According to a recent study (Gaido et al., 2024), the latest end-to-end system outperforms the best pipeline system, demonstrating the effectiveness of learning the translation and segmentation at the same time.

SubER (Wilken et al., 2022) has been the primary metric to evaluate the overall subtitle quality (Ahmad et al., 2024). Specifically, inspired by TER (Snover et al., 2006) metric, SubER computes the number of word edits and block/line edits required to match the reference, where hypothesis and reference words are allowed to match only within subtitle blocks that overlap in time. Therefore, it can provide a holistic evaluation of subtitles, encompassing translation quality, block/line segmentation accuracy, and timing quality. Because SubER tokenizes the subtitle text by spaces, it is not applicable for *scriptio continua* languages such as Chinese and Japanese.

3 Corpus Creation

In this section, we introduce the corpus creation method of A-TASC, the Asian TED-based Automatic Subtitling Corpus, which is composed of sentence-level (audio, transcription, translation) triplets. For fair comparison with MuST-Cinema (Karakanta et al., 2020), we mostly follow their method except for necessary adaptations to Chinese and Japanese subtitles.

Figure 2 overviews the corpus creation workflow. We first obtain the data of TED Talks including audio and transcription files in English and subtitle files in each target language (§ 3.1). We next split text in the transcription and subtitle files, and align them in sentence level (§ 3.2). We then align the audio file with the transcription sentences generated in the previous step to obtain audio segments (§ 3.3). We finally organize the aligned audio segments, transcription sentences, and subtitle lines into a YAML file and two text files, respectively (§ 3.4). In what follows, we detail each step.

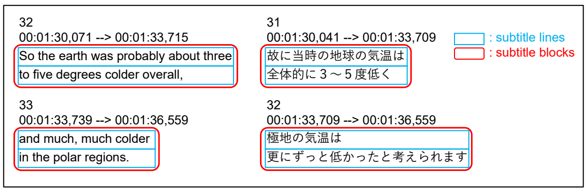
3.1 Data Acquisition

Like MuST-C and MuST-Cinema, the data of A-TASC is derived from TED Talks, where all subtitles go through transcription, translation, and review steps by qualified volunteers before publishing. Besides, dozens of hours of TED Talks are subtitled into multiple languages each year, which contributes to around a total of 800 hours of talks containing Chinese and Japanese subtitles. In addition, these talks are presented by presenters from all over the world, spanning over 300 different topics, *e.g.*, science, education, and society, as shown in Table 1. This contributes to large-scale corpora that have high-quality subtitles and high topic coverage, which are meant for creating a large-scale high-quality corpus for automatic subtitling.

We obtained the source data from all the English TED Talks with both Chinese and Japanese subtitles uploaded before November 2024. These audio files, transcription files, and subtitle files are all obtained from the official website. For later pro-

topic	science	technology	animation	education	social change	culture	history	society	health	business
#talks	1444	1301	1076	1029	862	844	689	674	664	657
hours	254	264	86	119	183	168	100	146	112	142

Table 1: Top-10 topics in A-TASC. Each talk has multiple topic tags, spanning over 300 different topics.



```
32
00:01:30,071 --> 00:01:33,715
So the earth was probably about three
to five degrees colder overall,

33
00:01:33,739 --> 00:01:36,559
and much, much colder
in the polar regions.
```

```
31
00:01:30,041 --> 00:01:33,709
故に当時の地球の気温は
全体的に３～５度低く

32
00:01:33,709 --> 00:01:36,559
極地の気温は
更にずっと低かったと考えられます
```

Figure 3: Example of a sentence in a subtitle file containing two subtitle blocks and four subtitle lines.

cessing, the audio files are transformed from .m4a into .wav format with a sample rate of 16,000 Hz. In addition to these key information, we also provide metadata, including title, presenter, duration, uploaded year, and topics, for possible future use.

3.2 Sentence-level Text-to-text Alignment

Having obtained the source data, the first step is to align the transcription text with the subtitle text at the sentence level. The purpose of this step is to prevent incomplete sentences in the subtitle blocks from hindering the training of AS systems. While the English transcriptions can be easily split by sentences based on punctuation-based heuristics, it is challenging to do sentence segmentation for the Chinese or Japanese subtitles resulting from the possible absence of strong punctuation marks.

As demonstrated in Figure 2, we thus split the translations into subtitle lines[2] instead of sentences. As illustrated in Figure 3, "subtitle blocks" are the subtitles presented on the screen for a specified period of time, and "subtitle lines" are the lines contained in each subtitle block. Unlike MuST-C, which uses an aligner supporting European languages only, we align the subtitle **lines** with the transcription **sentences** using Bertalign (Liu and Zhu, 2023), a sentence aligner based on the LaBSE (Feng et al., 2022) model, which supports 109 languages. Specially, the alignment is performed in a sequential order, including one-to-one, one-to-many, many-to-one, and many-to-many relations. We removed all parenthesized contents before the alignment, because most of them were absent from the speech to be translated. Finally, we obtain the aligned pairs of transcription sentences and the corresponding subtitle lines. This method eliminates the dependency on punctuations, hence can be applied to all languages supported by LaBSE, which is suitable for future extension.

3.3 Audio-text Alignment

The second step is to locate the audio segments from the audio file that aligned with the transcription sentences obtained in the previous section. A straightforward approach is to identify the minimum set of subtitle blocks that fully encompass the aligned transcription sentences and then locate the audio segments from the start time of the first block to the end time of the last block. However, there are incorrectly annotated timestamps for unknown reasons. To mitigate the impact of this, we follow the MuST-C's approach and employ a forced-aligner, Gentle,[3] to locate the audio segments aligned with the transcription sentences.

Specifically, Gentle generates the start and end timestamps for each word in the transcription text, and some of the words may not be recognized successfully. To discard the possibly noisy talks, we filter out entire talks when the proportion of unrecognized words is equal to or greater than 15% of the total. Then, we attempt to set the start time of the first word as the start time of the transcription sentence and the end time of the last word as the end time of the transcription sentence. If the first word is unrecognized, we assign the end time of the last word from the previous sentence as the start time. Similarly, if the last word is unrecognized, we assign the start time of the first word from the following sentence as the end time. If the start time or end time cannot be successfully assigned after these processes, we filter out that sentence. In this process, about 1.8% of the sentences are discarded.

3.4 Data Organization and Statistics

Finally, we organize our corpus in the same format as MuST-Cinema. Specifically, for each target language, we list the aligned transcription and translation sentences in two text files; for each sentence,

[2] Although we have applied some sentence segmentation models to the translations, none of them meet our expectations, probably because their training data are fully punctuated.

[3] https://github.com/lowerquality/gentle

lang	train	dev	test
MuST-Cinema (∼400h per language)			
De	229K	1,088	542
Es	265K	1,095	536
Fr	275K	1,079	544
It	253K	1,049	545
Nl	248K	1,023	548
Pt	206K	975	542
Ro	236K	494	543
A-TASC (∼800h per language)			
Zh	406k	1,392	738
Ja	370k	1,285	687

Table 2: Numbers of examples of MuST-Cinema and A-TASC for training, development, and test sets.[4]

the start time, duration, and the source .wav file of the corresponding audio segment are included in a YAML file. Then, we randomly split the talks into training, development, and test sets, where development and test sets contain 20 and 10 talks, respectively. Note that the two-step alignment is not necessary for test set, because AS systems are desired to be able to generate the subtitle files solely based on the audio files. To ensure the quality of the test set, we manually check and modify both the translation and timing quality of the subtitles.

Table 2 lists the statistics of our corpus and MuST-Cinema. For both languages, the training set is composed of more than 4K talks, containing around 400K examples and 800 hours of speech, which is about twice as large as MuST-Cinema.

4 Experiments

In this section, we present three sets of experiments, which are respectively aimed to i) empirically validate the quality of the A-TASC corpus and demonstrate the baseline results for future comparison (§ 4.2), ii) compare the AS performance across languages and analyze the causes of the performance gap (§ 4.3), and iii) compare the latest end-to-end AS model and a strong pipeline system (§ 4.4).

4.1 Settings

4.1.1 Automatic Subtitling Models

We evaluate the following end-to-end and pipeline AS systems on our A-TASC corpus.

SBAAM (Gaido et al., 2024) is the first end-to-end AS model which entirely eliminates any dependence on intermediate transcriptions for the whole subtitle generation process. It is a direct autoregressive encoder-decoder model, where the encoder is composed of three blocks: i) an acoustic encoder made of two 1D CNNs and eight Conformer (Gulati et al., 2020) layers, ii) a length adaptor leveraging the CTC Compression (Gaido et al., 2021) module, and iii) a semantic encoder made of four Conformer layers. The encoder output is then fed to an autoregressive decoder and a CTC on Target (TgtCTC) module (Yan et al., 2023). During the generation, it translates the audio segments into translations with `<eob>` and `<eol>` tokens. Each token is then aligned with the audio frames, so that the timestamps of generated subtitles can be computed according to the audio frames corresponding to `<eob>` tokens. Since it computes the timestamps relying solely on translations, the timing quality of the generated subtitles is proved to be better than in the existing pipeline approaches. The training settings of SBAAM are described in Appendix A.

Whisper(X)+DS are pipeline systems we evaluate in the third experiment (§ 4.4). We use vanilla Whisper (Radford et al., 2023) and WhisperX (Bain et al., 2023)[5] (both based on large-v2) as the ASR model[6] and DeepSeek-V3 (Liu et al., 2024) as LLM model for MT. DeepSeek-V3 is claimed to be comparable to GPT-4o (Hurst et al., 2024) while having a higher price–performance ratio and possibly higher MT performance for Asian languages. Finally, we segment the translated text by the same LLM as a postprocess. This is the first time LLM is incorporated and evaluated in the AS pipeline. The prompts to the LLM are shown in Appendix B.

4.1.2 Data Processing

For the training and development sets, we follow the instruction of SBAAM (Gaido et al., 2024) to preprocess our data, where the log Mel 80-dimensional filter-bank features are extracted as the input features, and the unigram tokenizer is applied to the aligned transcription and subtitle text for each language and 8,000 vocabulary size.

For the test set, following existing work (Papi et al., 2023), we use SHAS (Tsiamas et al., 2022) to segment the original audio files into segments

[4] It is hard to compute the accurate duration of audio due to the data filtering, thus we report the approximate value.

[5] WhisperX enhances the timing ability by a phoneme-based ASR model based on wav2vec 2.0 (Baevski et al., 2020).

[6] Whisper has the translation mode, but it can only translate speech in other languages into English text.

lang	CPL	CPS
EN	33.0	17.2
Zh	12.7	5.4
Ja	14.3	7.1

Table 3: CPL (character per line) and CPS (character per sec.) for English, Chinese, and Japanese in A-TASC.

lang	CPL	CPS
Zh	≤ 16	≤ 9
Ja	≤ 13	$\leq 4 \rightarrow 6$

Table 4: Chinese and Japanese subtitling constraints derived from Netflix.

TGT	200h	400h	600h	800h
Zh	48.88	46.08	44.62	43.27
Ja	63.41	61.12	60.79	59.67

Table 5: SacreSubER (↓) scores of SBAAM trained on training sets with different sizes.

TGT	Forced-Aligner	Timestamps
Zh	46.08	46.80
Ja	61.12	61.90

Table 6: SacreSubER (↓) scores of SBAAM trained on corpus aligned by forced-aligner and timestamps.

less than 16 seconds to prevent the input audio segments from being too long to be processed by the AS models. Note that the segmentation in this step is different from that performed by the AS models, which is aimed at generating subtitles with appropriate length. It also differs from the segmentation in the training set, where the audio segments are aligned with the aligned transcription sentences.

4.1.3 Metrics

To evaluate the quality of subtitles, we have to consider the translation quality, timing quality, and the compliance with subtitling constraints at the same time. In addition to evaluating the overall subtitle quality, we thereby evaluate the translation quality and the compliance with subtitling constraints.

Overall Quality. To address the language dependency problem mentioned in § 2.3, we introduce **SacreSubER**, which integrates SacreBLEU (Post, 2018)'s TER tokenizer with SubER (Wilken et al., 2022) before computing the number of edits.

Translation Quality. We adopt **AS-BLEU** and **AS-ChrF** (Matusov et al., 2005), which realign the system and reference subtitles based on the Levenshtein distance before computing the BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) scores.

Compliance with subtitling constraints. We use **CPL** (character per line) and **CPS** (character per second), following the existing studies (Ahmad et al., 2024). Unlike these studies, we set the standards based on Netflix[7] instead of TED Talks,[8] because TED Talks applies the English standard (CPL ≤ 42, CPS ≤ 21) to subtitles in all languages, which we believe is not appropriate. As listed in Table 3, the CPL and CPS computed in our corpus are significantly different between English and the two Asian languages, Chinese and Japanese. The subtitling constraints derived from Netflix are illustrated in Table 4. Here, we adjust the CPS constraint of Japanese from four to six, because four seems to be too strict according to our experiment results. Besides, an empirical study (Sasaki, 2017) proves that most participants preferred 6 CPS versions of subtitled films, indicating that the traditional 4 CPS rule may be a bit outdated for today's audience.

4.2 Experiment 1: Corpus Quality and Utility

The quality plays the most important role of the usefulness of a corpus. In this study, we verify the usefulness of A-TASC by observing the enhancement of baseline's performance with the increment of training set size, and with the effort of mitigating the impact of the incorrectly annotated timestamps.

4.2.1 Influence of the Training Set Size

In this experiment, we randomly select talks in the training set until the total duration of the audio segments reaches 200, 400, and 600 hours. Then, SBAAM is trained on these subsets and the full training set (800h), respectively. The results shown in Table 5 indicate that the performance of the baseline model continues to improve as the training data size grows. This verifies the quality of A-TASC. In addition, the large performance gain from 400h (size of MuST-Cinema) to 800h highlights the necessity of a larger corpus for automatic subtitling.

4.2.2 Forced-Aligner vs. Timestamps

To investigate whether realigning the audio segments by the forced-aligner can mitigate the noises in the raw timestamps and improve the AS performance, we align the audio and text of the same set

[7] https://partnerhelp.netflixstudios.com/hc/en-us/sections/22463232153235-Timed-Text-Style-Guides

[8] https://www.ted.com/participate/translate/subtitling-tips

TGT	Overall	Translation		Readability	
	(Sacre)SubER (↓)	**AS-BLEU** (↑)	**AS-ChrF** (↑)	**CPL** (↑)	**CPS** (↑)
Evaluated on MuST-Cinema (Gaido et al., 2024)					
De	59.8	-	-	90.1	75.7
Es	47.5	-	-	94.6	79.7
Fr	53.4	-	-	91.0	72.5
It	51.6	-	-	89.3	78.5
Nl	48.7	-	-	85.1	81.7
Pt	45.5	-	-	89.4	82.1
Ro	49.3	-	-	93.7	84.0
Evaluated on A-TASC (our corpus)					
Zh	46.1	22.4	20.0	97.0	95.7
Ja	61.1	19.8	18.2	85.1	59.6

Table 7: Results of SBAAM for European languages in MuST-Cinema and Asian languages in A-TASC, where the results on MuST-Cinema are directly derived from the original paper (Karakanta et al., 2020).

of talks as the 400h training subset on the basis of timestamps, as mentioned in § 3.3. The results in Table 6 verifies the effectiveness of the forced-aligner for mitigating the negative impact caused by the noises in the original subtitle files.

4.3 Experiment 2: Performance on Different Languages

In this experiment, we compare the performance of SBAAM trained on the proposed Asian corpus A-TASC and that trained on the European corpus MuST-Cinema. For fairness, we trained on the 400h subset of A-TASC, which is comparable to the size of MuST-Cinema.

Table 7 lists the results. We observe the SacreSubER results of SBAAM on the two Asian languages are roughly within the range of the reported SubER results (Gaido et al., 2024) on European languages, where the overall quality of Japanese is worse than that of Chinese. More specifically, comparing to the difference in terms of the translation-only metrics, the difference of the overall metric between Chinese and Japanese is much larger. This result indicates bad segmentation or timing quality may contribute more to the worse overall quality of Japanese, which will be further explained in § 4.4.

For CPL and CPS conformity, the results for Japanese are lower as well, which may attribute to the relatively stricter constraints. Still, it makes no sense to apply the English constraints to the Asian languages, which leads to CPL and CPS conformity that close to 100% for both Chinese and Japanese.

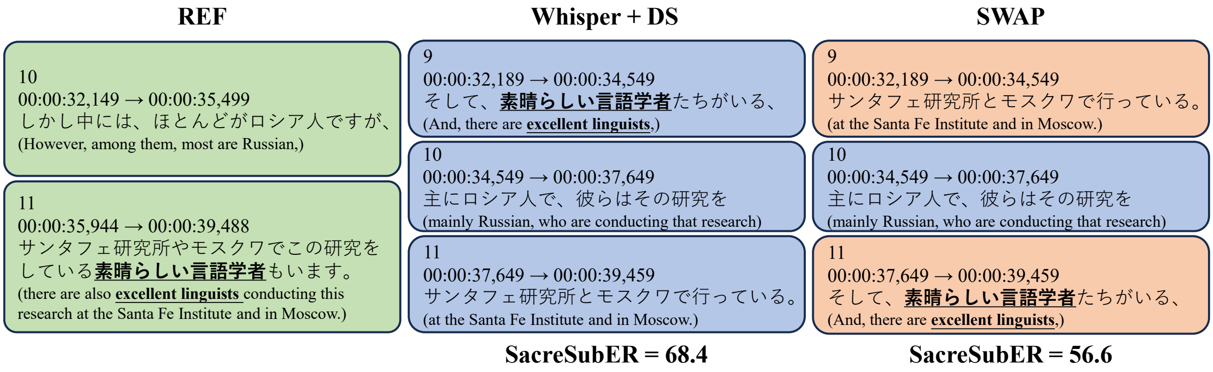
4.4 Experiment 3: End-to-end vs. Pipeline

Table 8 lists the results of the two pipeline approaches together with the results of SBAAM. Firstly, we observe that the pipeline approaches achieve much better translation quality thanks to the strong LLM. However, the SacreSubER result becomes even worse. This result indicates that the latest AS model trained on A-TASC can achieve better timing and segmentation quality than the pipeline approaches. Secondly, we notice that for Japanese, WhisperX achieves a better overall score than Whisper with a slightly lower translation quality, showing the plenty room for improvement regarding the timing quality for Whipser-based tools. Thirdly, considering the significantly worse results of pipeline approaches for Japanese, we assume this may not only attribute to the unsatisfactory timing and segmentation quality of the generated subtitles, but also the incapability of SubER to evaluate the SOV languages that frequently have word order swaps between subtitles. Figure 4 shows this fundamental problem of SubER. In this example, the pipeline system generates a more literal translation, the word order of which is similar to the English speech while different from the nature Japanese word order in the reference subtitles. Specifically, we observe although the last block of the reference column and the first block of the system column both contain the boldfaced phrase, they do not overlap in time. Therefore, this phrase would be considered as "not translated" when evaluated by SubER, which is based on the time-constrained TER metric. To further confirm this problem, we swap the text of the first and the last block of the system subtitles and compare the SacreSubER score[9] with the score computed before the swap. We observe that when the blocks

[9] The scores are computed on the subtitle files containing these blocks only.

TGT	model	Overall	Translation		Readability		
		SacreSubER	AS-BLEU	AS-ChrF	CPL	CPS	LPB
Zh	**SBAAM**	**43.3**	25.2	22.1	96.6	**96.4**	**99.9**
	Whisper + DS	44.9	**30.2**	**26.1**	**98.7**	93.9	97.0
	WhisperX + DS	45.1	26.2	22.8	95.7	90.1	88.7
Ja	**SBAAM**	**59.7**	21.3	19.4	**84.3**	**55.7**	**99.9**
	Whisper + DS	63.3	**28.0**	**25.9**	79.1	23.6	85.5
	WhisperX + DS	62.8	27.7	24.6	68.0	24.3	72.6

Table 8: Results of SBAAM and pipeline approaches on A-TASC.


REF

10
00:00:32,149 → 00:00:35,499
しかし中には、ほとんどがロシア人ですが、
(However, among them, most are Russian,)

11
00:00:35,944 → 00:00:39,488
サンタフェ研究所やモスクワでこの研究を
している**素晴らしい言語学者**もいます。
(there are also **excellent linguists** conducting this research at the Santa Fe Institute and in Moscow.)

Whisper + DS

9
00:00:32,189 → 00:00:34,549
そして、**素晴らしい言語学者**たちがいる、
(And, there are **excellent linguists**,)

10
00:00:34,549 → 00:00:37,649
主にロシア人で、彼らはその研究を
(mainly Russian, who are conducting that research)

11
00:00:37,649 → 00:00:39,459
サンタフェ研究所とモスクワで行っている。
(at the Santa Fe Institute and in Moscow.)

SacreSubER = 68.4

SWAP

9
00:00:32,189 → 00:00:34,549
サンタフェ研究所とモスクワで行っている。
(at the Santa Fe Institute and in Moscow.)

10
00:00:34,549 → 00:00:37,649
主にロシア人で、彼らはその研究を
(mainly Russian, who are conducting that research)

11
00:00:37,649 → 00:00:39,459
そして、**素晴らしい言語学者**たちがいる、
(And, there are **excellent linguists**,)

SacreSubER = 56.6

Figure 4: Example of SubER failing to properly evaluate the overall subtitle quality for SOV languages like Japanese. The height of the blocks represents the time overlapping among subtitles.

containing this phrase have time overlapping, the SacreSubER improves substantially, even though the translation is incomprehensible.

For readability metrics, we additionally report LPB (lines per block) besides CPL and CPS. The LPB constraint is set to two for both languages according to Netflix's guidelines. We observe SBAAM performs better than the pipeline approaches, which may result from multiple reasons. First, we use the audio segments split by SHAS as input for the sake of fairness, which may hinder Whisper to use the context information, resulting in suboptimal audio segmentation. Second, unlike Whisper, WhisperX segments the audio based solely on voice action detection (VAD) to enable the batched inference, which leads to longer subtitles. Third, although the subtitle segmentation post-processed by LLM can contribute to better CPL, it is not helpful for CPS, and may lead to worse LPB if the translations in subtitle blocks are too long.

In all, the end-to-end AS model, SBAAM, trained on A-TASC achieves better overall results and a better compliance with subtitling constraints than the LLM-based zero-shot pipeline approaches, regardless of the worse translation quality.

5 Conclusions

We present A-TASC, an Asian TED-Based Automatic Subtitling Corpus, including about 800 hours of audio segments and the aligned transcriptions and subtitles in Chinese and Japanese. A-TASC is the first corpus for automatic subtitling that includes Asian languages, and has the largest scale per language among the existing corpora. We propose SacreSubER, which supports the overall evaluation of subtitles in Asian languages. We empirically validate the quality of A-TASC, compare the AS performance between European and Asian languages, and discuss the possible reason of the worse SubER results for Japanese.

In the future, we would like to include more Asian languages in our corpus, such as Korean, Vietnamese, Thai, etc. Considering the low language dependency of our corpus creation method, little adaptation would be needed. In addition, we plan to deeply investigate the AS performance for Asian languages, and develop a AS metric that is more suitable to evaluate the subtitle quality in Asian languages.

Limitations

While we address the problems of the only existing AS corpus by developing A-TASC, an Asian TED-based automatic subtitling corpus, there are still some limitations of this work.

Firstly, we only involve Chinese and Japanese as target languages, ignoring other Asian languages, such as Korean, Vietnamese, Thai, etc. Nevertheless, considering the low language dependency of our corpus creation method, little adaptation would be needed for the incorporation of other languages. Secondly, our corpus creation workflow is lack of manual validation, e.g., to sample the aligned pairs from the two-step alignments and check the quality. Finally, although we modify the SubER metric to make it applicable for Asian languages, it still has fundamental problems when evaluating SOV languages like Japanese. Therefore, proposing a new metric for Asian languages is a possible direction for future work.

References

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 4489–4493. ISCA.

Mattia A 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 1 (Long and Short Papers)*, pages 2012–2017. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. CTC-based compression for direct speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 690–696, Online. Association for Computational Linguistics.

Marco Gaido, Sara Papi, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. 2024. SBAAM! eliminating transcript dependency in automatic subtitling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3673–3691, Bangkok, Thailand. Association for Computational Linguistics.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *CoRR*, abs/2005.08100.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. MuST-cinema: a speech-to-subtitles corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lei Liu and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023. Direct speech translation for automatic subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Runa Sasaki. 2017. Is the four-character-per-second word limitation outdated? an empirical study of japanese film subtitling. *Interpreting and Translation Studies: The Journal of the Japan Association for Interpreting and Translation Studies*, 17:149–165.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv preprint arXiv:2202.04774*.

Patrick Wilken, Panayota Georgakopoulou, Athena Consultancy, and Evgeny Matusov. 2022. Suber: A metric for automatic evaluation of subtitle quality. *IWSLT 2022*, page 1.

Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, and Shinji Watanabe. 2023. CTC alignments improve autoregressive translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1623–1639, Dubrovnik, Croatia. Association for Computational Linguistics.

Optimizer	AdamW
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.98$
Source CTC weight	1.0
Target CTC weight	2.0
CE weight	5.0
CE label smoothing	0.1
Learning Rate scheduler	Noam
Learning Rate	2e-3
Warmup steps	10,000
Weight Decay	0.001
Dropout	0.1
Clip Normalization	10.0
Training steps	100,000
Maximum tokens	40,000
Update frequency	2

Table 9: Training settings for SBAAM.

A Training Settings

For the training of SBAAM (Gaido et al., 2024), we follow the instruction described in FBK's repository[10]. Specifically, the training pipeline includes three phases: 1) an ASR training, 2) an ST training (with the encoder weights initialized from the ASR), 3) Subtitling fine-tuning from the ST model with the inclusion of the CTC on target module. For ASR training, since A-TASC and MuST-Cinema corpora share the same source language, we directly adopt the available checkpoint. The training settings for ST training and subtitling finetuning are demonstrated in Table 9. The model is validated for every 1,000 steps, and the early stop patience is set to 10. After the training of both phases, the last 7 checkpoints are averaged as the final model checkpoints. All trainings are executed on one NVIDIA RTX A6000 GPU (48GB VRAM).

B LLM Prompts

Figure 5 demonstrates the prompt for the LLM used in Experiment 3 (§ 4.4) to translate and segment the subtitles.

[10] `https://github.com/hlt-mt/FBK-fairseq/blob/master/fbk_works/SBAAM.md`

Translation Prompt:

You are a professional subtitle translation assistant, skilled in translating English subtitles line by line into **{language}.**
Your tasks are:

1. Carefully read the English subtitle text provided by the user, fully understanding the context.
2. Since the subtitle text is generated by an ASR model, there may be recognition errors. You need to infer the correct content based on the context and translate it accordingly.
3. Ensure the translation is accurate and natural, conforming to the expression habits of **{language}**.
4. Maintain logical coherence in the translation with the context, avoiding taking sentences out of context.
5. Output the translation results line by line, without including any information other than the translated text.
6. As a subtitle translation assistant, you need to reference the original text to break sentences appropriately, conforming to the normal word order of **{language}**.
7. Strictly maintain the same number of lines in the output translation as in the input subtitles by appropriately breaking sentences, and do not use blank lines to fill.

Segmentation Prompt:

You are a professional subtitle proofreader, skilled in segmentation for **{language}** subtitles. Your tasks are:

1. Split the given sentence at appropriate points, ensuring that each line does not exceed **{CPL}** characters, and the total number of lines does not exceed 2.
2. If the original sentence already meets the requirements in 1 without modification, do not alter it and output the original sentence directly.
3. Only output the final result after segmentation, without including any additional information.

Figure 5: Prompt for the LLM to translate and segment the subtitles. **{language}** is replaced by the target languages, and **{CPL}** is replaced by the CPL constraint of the target language.