HW-TSC's submission to the IWSLT 2024 Subtitling track

Yuhao Xie, Yuanchang Luo, Zongyao Li, Zhanglin Wu, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Hengchao Shang, Jiaxin Guo, Daimeng Wei, Hao Yang
Huawei Translation Service Center, Beijing, China
{xieyuhao2, luoyuanchang1, lizongyao, wuzhanglin2, chenxiaoyu35, raozhiqiang, lishaojun18, shanghengchao, guojiaxin1, weidaimeng, yanghao30}@huawei.com

Abstract

This paper introduces HW-TSC's submission to the IWSLT 2024 Subtitling track. For the automatic subtitling track, we use an unconstrained cascaded strategy, with the main steps being: ASR with word-level timestamps, sentence segmentation based on punctuation restoration, further alignment using CTC or using machine translation with length penalty. For the subtitle compression track, we employ a subtitle compression strategy that integrates machine translation models and extensive rewriting models. We acquire the subtitle text requiring revision through the CPS index, then utilize a translation model to obtain the English version of this text. Following this, we extract the compressed-length subtitle text through controlled decoding. If this method fails to compress the text successfully, we resort to the Llama2 few-shot model for further compression.

1 Introduction

In recent years, the demand for subtitles across various media platforms has surged, driving the need for efficient and high-quality subtitling solutions. Two main approaches have emerged for automatic subtitle generation: cascaded strategies and end-to-end models.

Cascaded Strategies Traditional cascaded strategies involve a multi-step pipeline (Bentivogli et al., 2021), where each component handles a specific subtask. This typically begins with an Automatic Speech Recognition (ASR) system that transcribes the audio into text. The transcribed text is then segmented into subtitles, accounting for timing constraints and reading speeds. Finally, the segmented subtitles may undergo text compression to ensure they fit within spatial limitations while retaining critical information.

End-to-End Strategies In contrast, end-to-end models (Berard et al., 2016) aim to directly generate subtitles from audio or audio-visual inputs using a single unified framework, typically leveraging recent advances in deep learning and sequence-to-sequence modeling. Such models can jointly learn and optimize all subtitling tasks, mitigating error propagation issues.

In this paper, we employ a cascaded strategy. Due to Whisper (Radford et al., 2023)'s remarkable achievements across multiple domains, the cascaded strategy is expected to perform well. At the same time, it allows us to leverage our existing text-to-text machine translation capabilities.

In the process of automatic subtitle generation discussed above, regardless of the method employed, subtitle compression emerges as a pivotal element. This is due to the restricted display space for subtitles, and the necessity of adapting subtitles to the playback speed of the video, as well as the reading speed of the audience. Consequently, once the automatic generation of the subtitle file is finalized, it becomes essential to compress content for overly long subtitles. By retaining the basic information and meaning, this compression significantly enhances the quality of the subtitles.

Traditional text compression strategies encompass Deletion-oriented approach (Moran, 2009) and Substitution-oriented approach (Yang et al., 2010). In addition to the aforementioned methods, training sequence-to-sequence models with parallel data of both the original and compressed text can enhance efficiency in text compression while more effectively preserving semantic integrity (Angerbauer et al., 2019).

In this paper, we leverage a model generation approach to accomplish the task of subtitle compression. Uniquely, in the absence of extensive parallel data of original and compressed text for model training, we deviate from traditional model compression methods. Instead, we employ a machine translation model to execute the task. This requires the compression and reformation of text, and the deployment of large language models to

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 334–338

manage the compression task on certain texts that pose challenges for rewriting.

2 Automatic Subtitling

We propose a Whisper-based cascaded automatic subtitling strategy, with the details as follows:

2.1 Automatic Speech Recognition (ASR)

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification. We use the large-v3 version of Whisper for ASR, and output word-level timestamps, which will help with re-segmentation after punctuation restoration.

2.2 Punctuation-Restoration-Based Segmentation

Bert-restore-punctuation[1] model is a punctuation restoration model for the general English language. Through punctuation restoration, we can obtain sentence segmentation information that is more semantically consistent, thereby obtaining better segments. Aided by the word-level timestamps from the previous step, we perform sentence segmentation at the predicted punctuation marks (commas, periods, exclamation marks, question marks), and generate corresponding timestamps.

2.3 CTC-Alignment

We use wav2vec2-large-960h-lv60[2] for forced alignment, which is pretrained and fine-tuned on 960 hours of Libri-Light and Librispeech on 16kHz sampled speech audio.

2.4 Machine Translation

Since the timestamps generated by the ASR system are good enough, when generating subtitles, we only translate the English into the target language, keeping the timestamps unchanged.

This track contains two language directions: English to German and English to Spanish, with the details as follows:

2.4.1 Data

The training data includes domains such as travel, subtitles, applications, and technology. The data size is shown in Table 1.

[1]https://huggingface.co/felflare/bert-restore-punctuation

[2]https://huggingface.co/facebook/wav2vec2-large-960h-lv60

	en2de	en2es
Baseline Data	5.8M	8.4M
Subtitle Data	1.3M	1.1M

Table 1: Data sizes of MT corpus.

2.4.2 Baseline models

We directly employ the en2de model we trained for the IWSLT 2024 Offline track and we employ our online-server en2es model. The training strategies include the following steps:

Regularized Dropout Regularized Dropout (R-Drop) (Wu et al., 2021) improves performance over standard dropout, especially for recurrent neural networks on tasks with long input sequences. It ensures more consistent regularization while maintaining model uncertainty estimates. The consistent masking also improves training efficiency compared to standard dropout. Overall, Regularized Dropout is an enhanced dropout technique that often outperforms standard dropout.

Back Translation Augmenting parallel training data with back-translation (BT) (Sennrich et al., 2016; Wei et al., 2023) has been shown effective for improving NMT using target monolingual data. Numerous works have expanded the understanding of BT and investigated various approaches to generate synthetic source sentences. Edunov et al. found that back-translations obtained via sampling or noised beam outputs tend to be more effective than those via beam or greedy search in most scenarios. For optimal joint use with FT, we employ sampling back-translation (ST)

Forward Translation Forward translation (FT) (Abdulmumin, 2021) uses source-side monolingual data to improve model performance. The general procedure of FT involves three steps: (1) randomly sampling a subset from large-scale source monolingual data; (2) using a "teacher" NMT model to translate the subset into the target language, thereby constructing synthetic parallel data; and (3) combining the synthetic and authentic parallel data to train a "student" NMT model.

2.4.3 Domain Adaptation Models

We used domain data to fine-tune the baseline model to achieve domain adaptation. The domain data came from three sources: 1. Directly crawled from the internet. 2. Obtained domain data from general domain data through curriculum learning.

Curriculum Learning A practical curriculum

learning (CL) (Zhang et al., 2019) approach for NMT should address two key issues: ranking training examples by difficulty, and modifying the sampling procedure based on ranking. For ranking, we estimate example difficulty using domain features (Wang et al., 2020). The domain feature is calculated as:

$$q(x,y) = \frac{\log P(y|x;\theta_{in}) - \log P(y|x;\theta_{out})}{|y|} \quad (1)$$

Where θ_{in} is an in-domain NMT model, while θ_{out} is an out-of-domain model. The subtitle domain is treated as in-domain.

We fine-tune the model on the validation set to get the teacher model and select top 40% of the highest scoring data for fine-tuning.

2.4.4 Settings

In the training, each model undergoes training utilizing 8 NPUs. The encoder-decoder layers is 25-6. The batch size remains fixed at 6144, the update frequency is 2, the dropout is 0.1, and the learning rate is maintained at 5e-4. A total of 4000 warmup steps are executed, and the model is saved every 2000 steps. Additionally, λ is set to 5 for R-Drop. During inference, the beam size is set to 5 for both models.

2.5 Experiment

We conduct our experiments on the IWSLT 2023 development data (including itv, peloton and TED), and calculate the SubER (Wilken et al., 2022), shown in Table 2 and Table 3. Here are the systems we submitted:

Pipeline We used the strategies mentioned in sections 2.1, 2.2, and 2.4.

Length-Penalty In addition to the pipeline system, we incorporated a length penalty when performing machine translation. We set the length normalization parameter to 10 and the word penalty parameter to 15.

CTC-alignment In addition to the pipeline system, we performed CTC-alignment on the transcription results.

3 Subtitle Compression

In the task of subtitle compression, our explicit objective is to rewrite the original subtitle text, leveraging its content to fulfill the parameters of characters per second (CPS (Papi et al., 2023)) and BLEURT (Sellam et al., 2020) indicators to the highest degree possible.

SubER-en2de	itv	peloton	TED	avg
Matesub	73.11	79.72	67.70	73.51
AppTek	71.40	71.90	64.30	69.20
FBK	83.70	79.10	69.40	77.40
Pipeline	74.41	78.92	72.03	75.10
+Length-Penalty	74.32	78.77	65.52	**72.86**
+CTC-alignment	74.21	79.30	71.24	74.91

Table 2: SubER in en2de

SubER-en2es	itv	peloton	TED	avg
Matesub	71.25	74.87	45.94	64.02
AppTek	82.10	79.00	48.80	69.97
FBK	82.20	80.30	52.50	71.67
Pipeline	71.87	79.98	52.49	68.11
+Length-Penalty	69.18	78.31	49.03	**65.50**
+CTC-alignment	71.41	80.27	51.27	67.62

Table 3: SubER in en2es

3.1 Strategy

Given the constraints that only the original subtitle file can be utilized and its timestamp information remains unalterable, our compression strategy is confined to sentence-level rewriting tasks. It implies that compression needs to retain the original semantics, but sentence-level fusion compression is unfeasible.

In the absence of large volumes of parallel data comprising original and compressed text, and the presence of substantial bilingual data, we suggest a subtitle compression approach that blends machine translation model rewriting and large model rewriting. Our subtitle compression framework is delineated in Figure 1.

We employ the same training data and strategies used for automatic subtitles to train the bidirectional translation model between English and German, and between English and Spanish. For large language models, we utilize Llama2 to accomplish the subtitle text rewriting task.

3.2 Experiment

We performed exploratory studies on the IWSLT 2023 development data and computed CPS and BLEURT, utilizing the compressed subtitle text as the benchmark reference. The computation details are presented in Table 4. We have listed below the systems submitted for consideration:

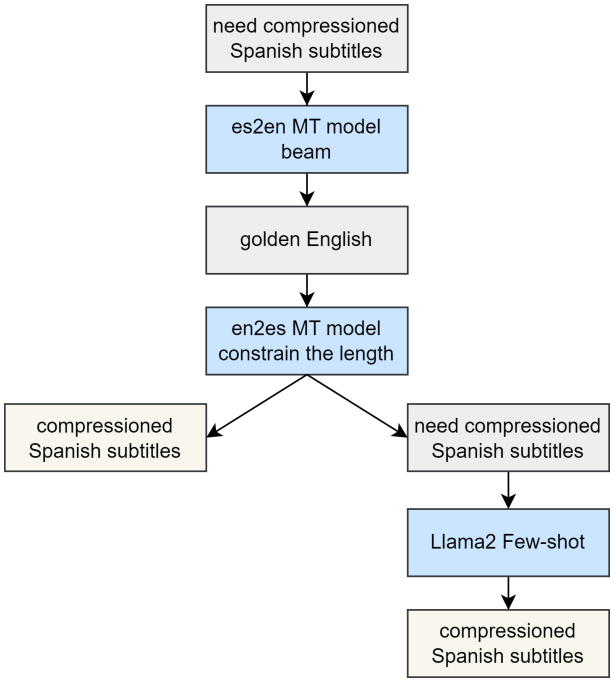

Figure 1: The subtitle compression framework

Reformulation using the machine translation model(system A) Within the inference architecture of the machine translation model, two parameters exist that can potentially constrain the length of the generated text:

1. the length normalization parameter: Divide translation score by pow(translation length, arg)

2. the word penalty parameter: Subtract (arg * translation length) from translation score

We also carry out the task of subtitle compression based on this feature. Using Spanish subtitle files as an example, we first utilize the CPS index to identify subtitle texts that do not conform to the required length specifications and, therefore, need compression. These texts are then processed through the Spanish-to-English translation model to generate golden English. These English versions are then subjected to re-translation back into Spanish, but we apply a length penalty during this translation to yield a compressed subtitle text. In this study, the length normalization parameter is set to 10 while the word penalty parameter is set to 15.

Revision based on the Llama2(systems B) The large-scale model exhibits robust reasoning capabilities, which can also be harnessed to accomplish the task of rephrasing subtitle text. Although Llama2 may not have been specifically trained for text condensation tasks, we adopt a few-shot methodology during inference. More precisely, a number of subtitle texts are chosen at random, and the condensed text is achieved through the aforementioned approach based on machine translation model rephrasing. During each inference, the large-scale model is initially presented with these instances, and then permitted to carry out the condensation and rephrasing assignment. The specific guidelines are as follows:

Tienes una gran capacidad de reescritura. Ahora necesitas reescribir el español en oraciones más concisas y cortas manteniendo la mayor cantidad posible de semántica del texto original.

1. Texto original: - ¿Cómo ayudará este impuesto a Europa a salir de la crisis económica? Texto después de reescribir: - ¿Cómo ayudará este impuesto a Europa a salir de la crisis?

2. Texto original: - Al fin y al cabo es un gesto político, nada más. Texto después de reescribir: - Al final es un gesto político, nada más.

3. Texto original: - Creo que la realidad es que, con sólo 11 países en el mundo, han adoptado este impuesto de manera efectiva Texto después de reescribir: - Creo que la realidad es que, con sólo 11 países efectivos en el mundo, adoptan este impuesto

Revision strategy utilizing machine translation models and large model amalgamation(systems A and B) Given that the machine translation model's output is derived from the golden English text, it holds a higher BLEURT score juxtaposed with Spanish, implying lesser semantic loss. Therefore, the initial consideration is leveraging a machine translation model for rewriting. However, for texts that pose higher rewriting complexities, a rewriting approach based on the Llama2 model is explored. Despite the potential for some semantic loss, this strategy ensures compliance with the prescribed length requirements for subtitle text.

System	CPS	CPS_mean	BLEURT
System A	75.3	19.9	0.78
System B	71.8	19	0.71
Systems A and B	81.2	18.6	0.62

Table 4: CPS and BLEURT in Spanish dev set

4 Conclusion

Although our performance in the experiment did not achieve best results, the comparison between our own systems can also illustrate some issues:

1. In automatic subtitling track, the results of machine translation with length penalty performed the best, indicating that compared to real subtitles, machine translation results tend to be longer.

2. In the subtitle compression track, the machine translation model produces rewritten text with a low BLEURT loss. However, the mean CPS value is 19.9, higher than the mean value of 19 from Llama2-based rewrites. This suggests that the machine translation model prioritizes translation quality and struggles to compress long sentences significantly. On the other hand, rewrites from Llama2 show lower CPS but higher BLEURT loss, indicating that the larger model possesses stronger reasoning abilities and can tackle challenging compression tasks effectively with prompts, albeit at the cost of potentially losing some sentence semantics.

References

Idris Abdulmumin. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers*, volume 1350, page 355. Springer Nature.

Katrin Angerbauer, Heike Adel, and Ngoc Thang Vu. 2019. Automatic compression of subtitles with neural networks and its effect on user experience. In *Interspeech*, pages 594–598.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *ArXiv*, abs/1612.01744.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 489. Association for Computational Linguistics.

Siobhan Moran. 2009. *The effect of linguistic variation on subtitle reception*. York University Toronto.

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.

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.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7944–7959, Toronto, Canada.

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. SubER - a metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Jie Chi Yang, Chia Ling Chang, Yi Lung Lin, and M Shih. 2010. A study of the pos keyword caption effect on listening comprehension. *SL Wong et al*.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1903–1915.