Automatic Subtitling and Subtitle Compression: FBK at the IWSLT 2024 Subtitling track

Marco Gaido, Sara Papi, Mauro Cettolo, Roldano Cattoni, Andrea Piergentili, Matteo Negri, Luisa Bentivogli
Fondazione Bruno Kessler, Trento, Italy
{mgaido,spapi,cettolo,cattoni,apiergentili,negri,bentivo}@fbk.eu

Abstract

The paper describes the FBK submissions to the Subtitling track of the 2024 IWSLT Evaluation Campaign, which covers both the Automatic Subtitling and the Subtitle Compression task for two language pairs: English to German and English to Spanish. For the *Automatic Subtitling* task, we submitted two systems each covering one of the two proposed training conditions, namely constrained and unconstrained: *i)* a direct model, trained in constrained conditions, that produces the SRT files from the audio without intermediate outputs (e.g., transcripts), and *ii)* a cascade solution that integrates only free-to-use and freely trained components, either taken off-the-shelf or developed in-house. Results show that, on both language pairs, our direct model outperforms both cascade and direct systems trained in constrained conditions in last year's edition of the campaign, while our solution assemblying pre-trained models is competitive with the best 2023 systems, although they were fine-tuned on task specific training data. For the *Subtitle Compression* task, our primary submission involved prompting a Large Language Model in zero-shot mode to shorten subtitles that exceed the reading speed limit of 21 characters per second. Our results highlight the challenges inherent in shrinking out-of-context sentence fragments that are automatically generated and potentially error-prone, underscoring the need for future studies to develop targeted solutions.

1 Introduction

In response to the growing amount of audiovisual content produced every day, the task of automatically generating subtitles has seen increasing attention (Álvarez et al., 2015; Vitikainen and Koponen, 2021), with the goal of fostering the accessibility of the material by overcoming language barriers. In light of this, starting from the 2023 edition, the IWSLT Evaluation Campaign includes the Automatic Subtitling task, in which participants had to generate well-formed subtitles in German and Spanish starting from the corresponding English audio (Agarwal et al., 2023). In addition to requiring high-quality translations of the audio content, correct subtitles also need the translated text to be split into blocks (each of them possibly split into 2 lines) in a way that minimizes the users' cognitive effort (Bogucki, 2004; Khalaf, 2016; Cintas and Remael, 2021), and these blocks have to be presented on-screen with the correct timing, i.e. in sync with the original audio.

Although there is no absolute rule to determine the cognitive effort required to read a subtitle, typical constraints to keep it low include: *i)* not having more than 2 lines per block (LPB); *ii)* keeping the number of characters per line (CPL) below a given threshold, which was set to 42 in the IWSLT 2023 campaign; and *iii)* avoiding excessive reading speed expressed in the number of characters per second (CPS) to be read by the user, which was set to 21. Good subtitles should hence be displayed in text blocks that conform to these rules, and their adherence to the constraints can be measured as the percentage of blocks compliant with them. Since automatic subtitling systems can fail in fully matching all the above constraints, the IWSLT 2024 campaign introduced an additional Subtitle Compression sub-task,[1] which requires to reduce the number of characters in each block of pre-generated subtitles to an extent that satisfies the reading speed constraint, without compromising its semantic content.

This paper describes FBK's submissions to both tasks (Automatic Subtitling and Subtitle Compression) of the IWSLT 2024 Subtitling track. Our submitted systems cover both language directions under evaluation, namely English-German (en-de) and English-Spanish (en-es).

Regarding Automatic Subtitling, we explored

[1] https://iwslt.org/2024/subtitling

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 134–144
August 15-16, 2024

two approaches that led to two submissions, one for each training condition, constrained and unconstrained. On the one hand, following the promising results obtained by the first direct models for automatic subtitling (Papi et al., 2023a), we trained a direct subtitling model (§2.1) in constrained conditions, i.e. using only the data allowed by the organizers for this setting. We call this model *direct* as it generates the subtitles in the target languages (including block and line delimiters) as well as timestamps without any intermediate discrete content representation, such as textual transcripts of the audio. In this respect, it is different from the two *direct* models submitted in the 2023 edition as both required the generation of intermediate transcripts for the timestamps estimation, either by using an auxiliary automatic speech recognition (ASR) system (Bahar et al., 2023) or by using auxiliary modules of the direct speech translation (ST) system (Papi et al., 2023b). On the other hand, we created a pipeline system (§2.2) within the AI4Culture[2] EU project, which binds us to use only code and models released under licenses as permissive as possible. Lastly, our primary submission to the newly proposed Subtitle Compression task (§2.3) tackled the problem with an LLM-based approach. To this aim, we explored a first basic solution by prompting the model in zero-shot mode to shorten candidate hypotheses exceeding the 21 CPS limit, and compared it with simpler, word/character deletion strategies.

2 Systems Description

In this section, we first describe the direct (§2.1) and cascade (§2.2) Automatic Subtitling systems, and then our Subtitle Compression submissions (§2.3).

2.1 Direct Subtitling with SBAAM

Our direct subtitling system is based on an encoder-decoder architecture, made of a 12-layer Conformer[3] encoder (Gulati et al., 2020) and a 6-layer Transformer decoder (Vaswani et al., 2017). It is trained to predict the translation in the target language with end of line (`<eol>`) and end of block (`<eob>`) delimiters to learn both to translate and segment into subtitle units. Moreover, we add a Connectionist Temporal Classification (CTC) on target module (Yan et al., 2023) on top of the encoder that is trained with the same target as the autoregressive Transformer decoder. In addition, to reduce the computational cost of our model, we include a CTC compression module in the 8$^{\text{th}}$ encoder layer (Gaido et al., 2021). This module is trained to predict the transcription of the audio, but no transcript is generated at inference time and the module only averages similar vectors without producing any textual representation of the source.

The end-to-end training is realized with a composite loss ($\mathcal{L}$) that sums the label smoothing cross-entropy (CE) loss (Szegedy et al., 2016) on the decoder outputs with the CTC loss of the CTC on target module, and the CTC loss of the CTC compression module. By defining t as the transcript of an audio sample, and x and y as the target translation augmented with `<eob>` and `<eol>` delimiters, we can formalize the loss as:

$$\mathcal{L} = \lambda_1 \,\mathrm{CTC}(h_8, t) + \lambda_2 \,\mathrm{CTC}(h, y) + \lambda_3 \,\mathrm{CE}(\mathcal{D}(h, y), y)$$

where $\lambda_{1,2,3}$ control the relative weight of the losses, h_8 is the output of the 8$^{\text{th}}$ encoder layer, h is the encoder output, and D is the Transformer decoder. In our experiments, we follow the indication of (Yan et al., 2023) and set $(\lambda_1, \lambda_2, \lambda_3)$ to (1.0, 2.0, 5.0).

The inference phase, instead, combines only the probabilities predicted by the CTC on target module and by the decoder, following the joint CTC/attention framework with CTC rescoring (Watanabe et al., 2017; Yan et al., 2023). This method involves rescoring the next-token probabilities produced by the decoder using the probabilities of the candidate prefixes obtained from the CTC on target module (TgtCTC):

$$p = p_{\mathcal{D}}(y_i | h, y_{0,\dots,i-1}) + \alpha \, p_{\mathrm{TgtCTC}}(y_{0,\dots,i} | h)$$

where α is a hyperparameter that controls the weight of the CTC rescoring.

The output of this inference is the translated text with subtitle boundaries. As such, we still miss a key element for subtitles: the start and end timestamps of each block, which control how long and when they have to be displayed on the screen. To estimate them, we rely on the Speech Block Attention Area Maximization (SBAAM) method (Gaido et al., 2024). SBAAM leverages the encoder-decoder attention to create alignments

[2]https://pro.europeana.eu/project/ai4culture-an-ai-platform-for-the-cultural-heritage-data-space

[3]We use the padding-safe implementation tested with `pangolinn` by Papi et al. (2024).

between the generated subtitles and the source audio, as done in many works both in text-to-text scenarios (Tang et al., 2018; Zenkel et al., 2019; Garg et al., 2019; Chen et al., 2020) and, more recently, speech-to-text ones (Papi et al., 2023c; Alastruey et al., 2023). In fact, SBAAM first applies a mean-standard deviation normalization to the attention matrix on the text axis (clipping all negative values to a small $-\epsilon$ quantity to avoid penalizing in different ways unnecessary areas). Then, for each block boundary (<eob>) in the generated text, it iteratively determines the timing of the <eob> by selecting the splitting point that maximizes the area of the current block with the audio up to that point and the remaining blocks with the rest of the audio.

Once all the <eob>s in the output have been processed, all blocks will have start and end timings.

Experimental Details. The input of our models is represented by 80 Mel-filterbank features extracted every 10 ms with a window of 25 ms. The input features are then processed with two 1D convolutional layers with stride 2 that reduce the input length by a factor of 4. We use 512 for the encoder and the decoder embedding dimensions and 2048 hidden features in the feed-forward layers. The vocabularies are based on unigram SentencePiece (Kudo, 2018), with size 8,000 for the English source and 16,000 for the target (either German or Spanish). The total number of parameters of our models is 133M. The final models are obtained by averaging the last 7 checkpoints obtained from the trainings, which are performed on 4 NVIDIA Ampere GPU A100 (64GB VRAM). At inference time, when long unsegmented audios have to be subtitled, the audio is first segmented into smaller audio chunks with SHAS[4] (Tsiamas et al., 2022). The code used to create the models is available at: `https://github.com/hlt-mt/FBK-fairseq`.

Training Data. The models are trained on most of the datasets admitted for the "constrained" submission type. These include all the available ST corpora, namely MuST-Cinema (Karakanta et al., 2020), EuroParl-ST (Iranzo-Sánchez et al., 2020), and CoVoST v2 (Wang et al., 2020). Also, we leverage most of the available ASR datasets (CommonVoice (Ardila et al., 2020), LibriSpeech (Panayotov et al., 2015), TEDLIUM v3 (Hernandez et al., 2018), and VoxPopuli (Wang et al., 2021)), by automatically translating the transcripts into the two target languages using the NeMo MT models.[5] <eol> and <eob> tags are added to both transcripts and translations of all datasets, except for MuST-Cinema that already include them, using the multimodal segmenter by Papi et al. (2022).

2.2 Cascade Subtitling

As stated in the introduction, within the EU AI4Culture project, we developed a cascade subtitling system combining free-to-use components only. Most of them are taken off-the-shelf, while others were developed in-house. The entire system is publicly available at `https://github.com/hlt-mt/FBK-subtitler`.

The pipeline is shown in Figure 1 and concatenates the following modules:

Audio segmenter: Speech recognition and speech translation models are unable to process long audios, which then have to be split into shorter segments. As in the direct architecture, here too SHAS is used to carry out this task. It is worth noting that, in general, each audio segment contains multiple subtitles. SHAS code and models are released under the very permissive MIT license.

Speech recognition system: To transcribe the input speech, we opted for Whisper[6] (`large-v3`) to date one of the best ASR systems covering English, licensed under the MIT license. Whisper generates transcripts already split in subtitles, each supplied with start and end timestamps. However, two main issues can affect Whisper's outputs: hallucinations and lack of segmentation in lines, both handled by specific modules of the pipeline.

Hallucination removal filter: It removes hallucinations, a well-known concern of LLMs, which refers to the generation of text that is erroneous, nonsensical, or detached from reality. Here, only *shallow* hallucinations are considered, i.e. those involving the syntax of subtitles but not their semantics. We observed two types of shallow hallucinations, *within* and *across* subtitles. The first type refers to the repetition of single words or short n-grams many consecutive times within a subtitle. The second type refers to instances where the same transcript is repeated an anomalous number of times across consecutive subtitles. We implemented a script which heuristically detects and

[4]`https://github.com/mt-upc/SHAS`

[5]Publicly available at: `https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/nlp/machine_translation/machine_translation.html`

[6]`https://github.com/openai/whisper`

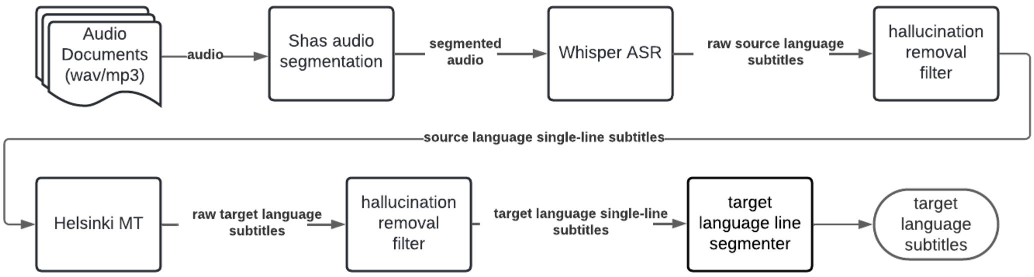

Figure 1: The cascade subtitling system based on pre-trained LLMs.

removes such phenomena from subtitles; in the pipeline, it is used downstream of both the ASR and the MT models.

Machine translation system: It performs the translation of a text (here: the text in each subtitle generated by Whisper, amended by hallucinations) from a source language into the target language. Various freely usable pre-trained LLMs have been tested in a preliminary investigation, namely NLLB,[7] mBART-50,[8] Helsinki Opus-MT.[9] The outcomes indicated the Helsinki Opus-MT as the best performer. Code and models are released under the MIT license.

Text segmenter: In general, its goal would be splitting the input text into fragments suitable, in terms of both quality and compliance to spatio-temporal constraints, to be displayed on the screen. However, since here the goal is solely to split too long, single line subtitles generated from the previous stages of the pipeline into two lines, we implemented a script that splits subtitles longer than 42 characters into two lines rewarding: the compliance of both lines with the 42-character limit, a similar length of the two lines, and the presence of a punctuation mark at the end of the first line.

2.3 Subtitle Compression

The newly introduced Subtitle Compression task required participants to rephrase subtitles provided by the task organizers that did not comply with the reading speed constraint of 21 CPS.

The material to be automatically processed was presented to participants as standard SRT (SubRip File Format) files that include: i) the text of sequentially numbered subtitles, which can be either one or two lines, and ii) timing information for each subtitle (i.e. timestamps in the format `hours:minutes:seconds,milliseconds`), indicating how long the subtitle should stay on the screen. As per the task guidelines, the goal was to exclusively work at the text level, compressing subtitles' text when necessary and without modifying the time boundaries. To achieve this, given the lack of indications on which automatic subtitles needed correction, we relied on the subtitle compliance script also provided by the task organizers. This allowed us to reliably identify the subtitle candidates requiring text compression and focus exclusively on rephrasing them.

The identified subtitles (39.8% and 30.0% of the total for en-de and en-es, respectively) underwent the compression phase, for which we devised two strategies. The first one, selected for our primary submission, is *user-oriented*: its goal is to target the CPS constraint with an LLM-based, fluency-driven approach aimed at preserving the readability of the compressed subtitles and, in turn, user experience. The second strategy, selected for our contrastive submissions, is more *metric-oriented*. Its goal is to shorten non-CPS-compliant subtitles by removing function words with varying levels of aggressiveness.

User-oriented approach (GPT – primary). Our LLM-based compression approach exploits GPT-4 (OpenAI, 2024) (model `gpt-4-0613`, with default parameters except for the temperature, which we set to 0), which was prompted in zero-shot mode with the instruction: "`Shorten this [LANGUAGE] text to a maximum of [TARGET_NUMCHARS] characters while preserving the original words as much as possible: [TEXT]`", where:

- `LANGUAGE` indicates the language of the subtitle, either "German" or "Spanish";
- `TARGET_NUMCHARS` specifies the maximum al-

[7] https://github.com/facebookresearch/fairseq/tree/nllb/

[8] https://huggingface.co/facebook/mbart-large-50

[9] https://huggingface.co/models?sort=trending&search=Helsinki-NLP

lowed length for the compressed subtitle, measured in characters including spaces. The target value is calculated based on the total on-screen time of the subtitle, which is determined by subtracting its start time from its end time and then multiplying this duration by 21 (e.g., with 3.2 seconds of on-screen time, `TARGET_NUMCHARS` is 67.2, truncated to 67);

- `TEXT` is the original subtitle that needs to be compressed.

The choice of the overall approach was driven by the aim to preserve the user experience by leveraging the generation capabilities of large language models. In fact, simpler and more aggressive methods, such as the metric-oriented ones presented in the next paragraph, can easily improve the rate of subtitles compliant with the CPS limit but at the cost of losing important information and detracting their readability. In an opposite direction, our LLM-based approach aims to strike a balance between improving CPS values and retaining the original information through targeted and meaning-preserving rephrasing.

Our zero-shot prompting strategy was primarily driven by fast-development reasons. In fact, we expect significant improvements by feeding the model with exemplars, i.e., via in-context learning (Brown et al., 2020). We opted for a simpler, cheaper, and more conservative approach to establish a starting point and a reference baseline for future in-depth comparative experiments. For similar reasons, we opted for a solution that concentrates on individual subtitles instead of operating on full sentences. Though likely more effective, letting the LLM reformulate *full* sentences in a shorter way would have introduced the additional burden of rearranging the resulting content into timed subtitles afterward. This is certainly a promising direction for future improvements.

Metric-oriented approach (`Del_*` – contrastive). For our contrastive submissions, we designed "metric-oriented" solutions that aim to improve CPS by aggressively reducing the length of subtitles through simple character or word deletions. The goal was to measure the extent to which this baseline approach affects the readability of subtitles. Along this direction, we explored a range of options which share the common trait of removing from the non-CPS-compliant subtitles specific categories of function words identified from pre-compiled lists downloaded from the web.[10] Word removal is carried out with varying levels of aggressiveness, ranging from *i)* the deletion of articles (`Del_articles`) to *ii)* the deletion of articles, prepositions, and adverbs (`Del_art/prep/adv`), and *iii)* the deletion of all function words (`Del_all-func-wrds`). On the one side, these strategies avoid the loss of important content in the original subtitles and the presence of incomplete words in the output, as it happens in the Baseline approach proposed by the task organizers. On the other side, they intervene in the syntactic structure of the subtitles, altering them in a way that improves CPS but penalizes both readability and automatic evaluation with reference-based metrics.

3 Results

As a recap, FBK submitted the following runs:

Automatic Subtitling task

- Primary run in Constrained condition: FBK_{24}^{drct} (§2.1)
- Primary run in Unconstrained condition: FBK_{24}^{cscd} (§2.2)

Subtitle Compression task

- Primary run: GPT (§2.3, paragraph "User-oriented approach")
- Contrastive1 run: `del all func wrds` (§2.3, "Metric-oriented approach")
- Contrastive2 run: `del art/prep/adv` (§2.3, "Metric-oriented approach")

3.1 Automatic Subtitling

Results on subtitling task are provided in Tables 1, 2, and 3. Table 1 compares the *SubER* (Wilken et al., 2022) scores,[11] the primary metric of the task, computed on the subtitles of the development set generated by our systems and by the best systems at IWSLT 2023 in constrained and unconstrained conditions. Table 2 shows global results, i.e., on subtitles of all domains, on test23 of our runs as provided to us by organizers, and of the best primary runs at IWSLT 2023, as published in (Agarwal et al., 2023). Table 3 gathers results, global and on each domain, on test24 of our runs

[10] `https://github.com/Yoast/javascript/tree/develop/packages/yoastseo/src/researches`

[11] When we do state otherwise, we compute SubER without casing and punctuation, as done in the previous evaluation campaign for the sake of fair comparison with previous scores.

en-de									
		TED		*ITV*		*PELOTON*		***AVG***	
system	**cnd**	SubER		SubER		SubER		SubER	
		cased	uncased	cased	uncased	cased	uncased	cased	uncased
$\mathtt{AppTek}_{23}^{\mathtt{cscd}}$	C	-	63.0	-	83.6	-	87.6	-	78.1
$\mathtt{FBK}_{23}^{\mathtt{drct}}$	C	69.4	-	83.7	-	79.1	-	77.4	-
$\mathtt{AppTek}_{23}^{\mathtt{cscd}}$	U	-	64.3	-	71.4	-	71.9	-	69.2
$\mathtt{FBK}_{24}^{\mathtt{drct}}$	C	61.6	62.1	80.0	80.7	75.6	78.2	72.4	73.7
$\mathtt{FBK}_{24}^{\mathtt{cscd}}$	U	69.0	69.0	79.3	78.9	73.4	76.1	73.9	74.7

en-es									
		TED		*ITV*		*PELOTON*		***AVG***	
system	**cnd**	SubER		SubER		SubER		SubER	
		cased	uncased	cased	uncased	cased	uncased	cased	uncased
$\mathtt{AppTek}_{23}^{\mathtt{cscd}}$	C	-	48.8	-	82.1	-	79.0	-	70.0
$\mathtt{FBK}_{23}^{\mathtt{drct}}$	C	52.5	-	82.2	-	80.3	-	71.7	-
$\mathtt{TLT}_{23}$	U	-	45.9	-	71.3	-	74.9	-	64.0
$\mathtt{FBK}_{24}^{\mathtt{drct}}$	C	49.5	47.5	79.1	79.5	79.3	80.8	70.3	70.3
$\mathtt{FBK}_{24}^{\mathtt{cscd}}$	U	49.2	48.0	72.2	73.5	73.9	76.9	65.1	66.1

Table 1: SubER (↓) comparison with the best cascade ($\mathrm{AppTek}_{23}^{\mathrm{cscd}}$ – Bahar et al. 2023 – and TLT_{23} – Perone 2023 – for en-es) and direct ($\mathrm{FBK}_{23}^{\mathrm{drct}}$) models trained on constrained/unconstrained (C/U of column cnd) conditions from the IWSLT 2023 Evaluation Campaign on automatic subtitling for en-de and en-es validation sets. The results of our systems are reported in bold.

			Subtitle quality	**Translation quality**			**Subtitle compliance**		
en-	**system**	**cnd**	SubER↓	BLEU↑	ChrF↑	BLEURT↑	CPS↑	CPL↑	LPB↑
-de	$\mathtt{FBK}_{24}^{\mathtt{drct}}$	C	74.26	13.08	34.77	.3742	72.75	89.35	99.96
	$\mathtt{AppTek}_{23}^{\mathtt{cscd}}$	C	77.14	12.40	33.17	.3300	93.01	100.00	100.00
	$\mathtt{FBK}_{24}^{\mathtt{cscd}}$	U	73.78	16.46	39.07	.4454	61.44	93.04	100.00
	$\mathtt{AppTek}_{23}^{\mathtt{cscd}}$	U	70.23	15.10	37.39	.4291	87.87	100.00	100.00
-es	$\mathtt{FBK}_{24}^{\mathtt{drct}}$	C	70.09	19.16	41.58	.3972	73.08	91.64	99.97
	$\mathtt{AppTek}_{23}^{\mathtt{cscd}}$	C	72.33	17.72	38.49	.3467	95.30	100.00	100.00
	$\mathtt{FBK}_{24}^{\mathtt{cscd}}$	U	66.02	23.87	46.53	.4811	67.56	94.25	100.00
	$\mathtt{TLT}_{23}$	U	67.29	22.54	46.40	.4993	85.51	99.53	100.00

Table 2: Global subtitling results (ALL) of 2024 FBK submissions and of 2023 best primary runs on test2023.

			Subtitle quality	**Translation quality**			**Subtitle compliance**		
en-	**system**	**dmn**	SubER↓	BLEU↑	ChrF↑	BLEURT↑	CPS↑	CPL↑	LPB↑
-de	$\mathtt{FBK}_{24}^{\mathtt{drct}}$	TED	57.50	25.79	54.78	.6114	83.10	83.69	100.00
		ITV	78.90	9.67	28.43,	.2911	70.45	90.04	99.97
		PLT	80.68	7.71	30.45	.3542	82.16	92.77	100.00
		ALL	73.99	13.48	36.12	.3775	76.19	88.86	99.99
	$\mathtt{FBK}_{24}^{\mathtt{cscd}}$	TED	63.26	22.94	53.70	.5872	79.99	89.52	100.00
		ITV	79.92	14.86	35.16	.4048	54.20	91.12	100.00
		PLT	78.34	11.30	34.13	.4202	76.52	96.99	100.00
		ALL	75.56	16.23	40.10	.4503	64.64	91.79	100.00
-es	$\mathtt{FBK}_{24}^{\mathtt{drct}}$	TED	39.86	45.63	69.63	.7441	82.43	86.59	100.00
		ITV	77.00	11.91	31.95	.2986	70.61	92.60	100.00
		PLT	79.70	11.88	40.05	.4329	82.26	89.58	100.00
		ALL	67.13	22.03	44.69	.4277	76.00	90.35	100.00
	$\mathtt{FBK}_{24}^{\mathtt{cscd}}$	TED	40.75	45.69	69.20	.7500	83.42	90.31	100.00
		ITV	70.82	18.92	40.17	.4262	60.85	93.46	100.00
		PLT	74.17	16.18	44.42	.5108	80.24	97.03	100.00
		ALL	63.01	26.60	49.64	.5174	69.97	93.28	100.00

Table 3: Detailed subtitling results of FBK submissions on test2024.

as provided to us by organizers. Besides SubER that measures overall subtitle quality, Table 2 and Table 3 include BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and TER (Snover et al., 2006) for translation quality and CPS, CPL and LPB conformity[12] for subtitling guideline compliance.

By looking at SubER scores of Table 1 and Table 2, we notice that our direct system outperforms not only the best direct system submitted last year but also the best cascade in constrained conditions. This superiority is consistent over all domains and language pairs. Also, focusing on Table 2, this is confirmed by all the translation quality metrics on test2023. In the unconstrained setting, instead, the results are less clear. Our cascade system achieves a lower (hence, better) SubER than the unconstrained submissions from last year on the en-es section of test2023 while, on the en-de section, it has a higher SubER than $AppTek_{23}^{cscd}$, in contrast with the definitely higher translation quality scores.

Back to the comparison between our direct constrained system and our cascade unconstrained solution, we notice consistent trends over all the evaluation sets (validation, test2023, test2024). The direct system achieves better scores on the TED domain, which is the only one covered by the training data allowed for the constrained setting, but falls behind by a large margin on the other two (ITV and PELOTON), especially on en-es. This result is not surprising as the unconstrained system has been trained on a wide range of domains and is therefore more robust to domain shifts. Regarding subtitle compliance, interesting trends emerge: the cascade system has higher CPL compliance (∼+3% across all settings), while the direct system outperforms it in terms of CPS compliance (+6-12%). The latter aspect may be motivated by the direct access to the source audio of the direct system (which is also guided by the CTC module that directly maps the audio sequence to the textual output).

3.2 Subtitle Compression

The results for the subtitle compression task are reported in Table 4 in terms of BLEURT and CPS (as a measure of reading speed compliance). BLEURT results are computed in two ways, either considering the provided subtitles as references or by using the actual subtitle references. The former results serve as a proxy of translation quality, as well as a way to measure the distance between the original subtitles to be modified and the resulting modified ones (i.e. as an indicator of how radical the applied changes are). The latter ones, instead, provide real translation quality measurements. For the sake of discussion, the table includes the results of the `Baseline` as provided by the task organizers and those of an unsubmitted metric-oriented solution (`Del_articles`), besides those of our official primary (GPT) and contrastive submissions (`Del_all-func-wrds` and `Del_art/prep/adv`).

Overall, the scores for the two languages indicate different levels of difficulty but exhibit similar trends. Specifically, en-es appears to be an easier direction, as indicated by higher translation quality (BLEURT) and reading speed compliance scores (CPS) compared to en-de. Unsurprisingly, the **BLEURT scores computed against the provided original subtitles (i.e., vs. [1])** are significantly higher than those computed against the actual references (vs. [0]). This indicates the tendency of the proposed methods to apply rather conservative changes. This holds particularly for the metric-oriented approaches (`Del_*`), which are actually designed to do so. Still, the relatively high BLEURT results of the user-oriented approach (`GPT`) are a symptom of local and rather moderate changes, which likely do not suffer from major issues related to hallucinations and/or under-generation into too short subtitles. Regarding the **BLEURT scores computed against the actual subtitle references (i.e., vs. [0])**, the results drop significantly, attesting that a large quality gap between all methods and human subtitles still exists. Interestingly, however, the gap between metric and user-oriented approaches shrinks on en-es and even disappears on en-de, where `GPT` achieves results that are substantially equivalent to those of `Del_art/prep/adv`.

For both languages and evaluation conditions the higher conservativeness of metric-oriented approaches is not sufficient to yield acceptable CPS results. First, the least aggressive one (the unsubmitted `Del_articles`), which consistently achieves the highest BLEURT computed on the provided references, is definitely the worst one in terms of CPS. Second, also the other ones (our contrastive submissions `Del_art/prep/adv` and `Del_all-func-wrds`) attain lower reading speed conformity compared to the LLM-based user-oriented approach. Aimed to strike a balance between translation quality and CPS conformity, our

[12]Computed with the script provided by Papi et al. (2023a): `https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/subtitle_compliance.py`

id	Subtitles		de BLEURT↑ vs. [0]	de BLEURT↑ vs. [1]	de CPS↑	es BLEURT↑ vs. [0]	es BLEURT↑ vs. [1]	es CPS↑
0	`Reference`		-	-	86.47	-	-	89.98
1	`Provided`		.1946	-	60.25	.2136	-	69.97
2	`Baseline`		.1720	.7871	100.00	.1892	.8766	100.00
	method	submission						
3	`Del_articles`	`not submitted`	-	.9230	65.92	-	.9700	73.80
4	`Del_art/prep/adv`	`FBK contrastive2`	.1890	.9071	67.94	.2113	.9572	75.74
5	`Del_all-func-wrds`	`FBK contrastive1`	.1811	.8365	83.36	.2033	.9123	87.48
6	`GPT`	`FBK primary`	.1895	.8370	84.81	.2063	.9062	90.66

Table 4: Subtitle Compression results. For both languages, BLEURT scores are computed both against the reference subtitles ([0]) and the provided original subtitles ([1]).

primary submission (`GPT`) consistently achieves the best CPS scores (84.81 for en-de, 90.66 for en-es). Paired with the above observations about translation quality, these results suggest that LLM-based approaches to subtitle compression are a promising direction for future explorations.

The trade-off between BLEURT and CPS is further highlighted by the plot in Figure 2 where, between the two extremes represented by `Provided` ([1]) and `Baseline` ([2]) subtitles, the subtitles generated through metric-oriented strategies ([4] and [5]) are placed according to a nearly linear relationship. The exception are `GPT`'results which slightly deviate from this linear trend, as a confirmation of our intuition: generative, user-oriented strategies are capable to perform pinpointed text reductions to pursue CPS compliance without a catastrophic loss of the original subtitles' meaning.

Overall, our results indicate that, even though it is a sub-task of a very complex problem such as automatic subtitling, subtitle compression has its own difficulties. On the one hand, the generative approach based on LLMs is intuitively promising because, unlike rough trimming strategies that are incompatible with the user experience, it targets a compression that is respectful of the subtitles' semantic content. On the other hand, however, this approach faces the challenge of reformulating text material that is potentially error-prone and often does not come in the form of well-formed sentences but rather as text spans representing sentence portions or words spanning contiguous phrases. At least in the zero-shot prompting modality, the combination of these two aspects makes the task extremely challenging for LLMs. As a matter of fact, upon preliminary analysis of the generated compressions, LLMs often reveal a tendency to generate sentence-like outputs, attempting to "complete" their generations with hallucinated content, a behavior that can only be exacerbated in the presence of errors in the subtitle to be compressed. The opposite potential issue, represented by "over-compressing" the subtitle beyond the allowed number of characters, is rarely observed.

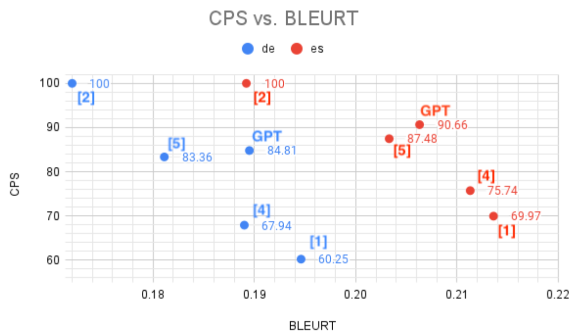

Figure 2: Scatter plot of compression results from Table 4 (BLEURT against the reference subtitles).

4 Conclusions

We presented the FBK's submissions to the Automatic Subtitling and Subtitle Compression tasks of the IWSLT 2024 Evaluation Campaign. For Automatic Subtitling, we proposed two systems: a direct model trained under constrained conditions and a cascade architecture integrating free-to-use components. Our direct model showcased superior performance compared to constrained direct and cascade submissions of the last year. The cascade solution proved competitive with top-performing unconstrained and fine-tuned 2023 runs. For Subtitle Compression, our primary submission exploits GPT in zero-shot prompting mode to shorten subtitles exceeding the reading speed limit of 21 CPS. While promising, this approach revealed the complexities of compressing out-of-context automatically generated sentence fragments, underscoring the necessity for further research in this area.

Acknowledgments

The work presented in this paper is co-funded by the European Union under the project *AI4Culture: An AI platform for the cultural heritage data space* (Action number 101100683). Marco Gaido is supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. We acknowledge the CINECA award (MAGIS) under the ISCRA initiative, for the availability of high-performance computing resources and support.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Belen Alastruey, Aleix Sant, Gerard I. Gállego, David Dale, and Marta R. Costa-jussà. 2023. Speechalign: a framework for speech translation alignment evaluation. *Preprint*, arXiv:2309.11585.

Aitor Álvarez, Carlos Mendes, Matteo Raffaelli, Tiago Luís, Sérgio Paulo, Nicola Piccinini, Haritz Arzelus, João Neto, Carlo Aliprandi, and Arantza Pozo. 2015. Automating live and batch subtitling of multimedia contents for several european languages. *Multimedia Tools and Applications*, 75:1–31.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France.

Parnia Bahar, Patrick Wilken, Javier Iranzo-Sánchez, Mattia Di Gangi, Evgeny Matusov, and Zoltán Tüske. 2023. Speech translation with style: AppTek's submissions to the IWSLT subtitling and formality tracks in 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 251–260, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Łukasz Bogucki. 2004. The constraint of relevance in subtitling. *The Journal of Specialised Translation*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online.

Jorge Díaz Cintas and Aline Remael. 2021. *Subtitling: Concepts and Practices*. Translation practices explained. Routledge.

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.

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)*, Bangkok, Thailand.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China.

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. In *Proc. Interspeech 2020*, pages 5036–5040.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer*, pages 198–208, Cham.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. MuST-cinema: a speech-to-subtitles corpus. In *Proc. of the 12th Language Resources and Evaluation Conference*, pages 3727–3734, Marseille, France.

Bilal Khalaf. 2016. An introduction to subtitling: Challenges and strategies. *International Journal of Comparative Literature and Translation Studies*, 3.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023a. Direct Speech Translation for Automatic Subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Sara Papi, Marco Gaido, and Matteo Negri. 2023b. Direct models for simultaneous translation and automatic subtitling: FBK@IWSLT2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 159–168, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Sara Papi, Marco Gaido, Andrea Pilzer, and Matteo Negri. 2024. When Good and Reproducible Results are a Giant with Feet of Clay: The Importance of Software Quality in NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand.

Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022. Dodging the data bottleneck: Automatic subtitling with automatically segmented ST corpora. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 480–487, Online only.

Sara Papi, Matteo Negri, and Marco Turchi. 2023c. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada. Association for Computational Linguistics.

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

Simone Perone. 2023. Matesub: The translated subtitling tool at the IWSLT2023 subtitling task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 461–464, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

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.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. *Preprint*, arXiv:2202.04774.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Kaisa Vitikainen and Maarit Koponen. 2021. Automation in the intralingual subtitling process: Exploring productivity and user experience. *Journal of Audiovisual Translation*, 4(3):44–65.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003, Online.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus. *Preprint*, arXiv:2007.10310.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

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).

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.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*.