AppTek's Automatic Speech Translation: Generating Accurate and Well-Readable Subtitles

Frithjof Petrick, Patrick Wilken, Evgeny Matusov, Nahuel Roselló, Sarah Beranek
AppTek, https://www.apptek.ai
Aachen, Germany
{fpetrick,pwilken,ematusov,nbeneitez,sberanek}@apptek.com

Abstract

We describe AppTek's submission to the subtitling track of the IWSLT 2025 evaluation. We enhance our cascaded speech translation approach by adapting the ASR and the MT models on in-domain data. All components, including intermediate steps such as subtitle source language template creation and line segmentation, are optimized to ensure that the resulting target language subtitles respect the subtitling constraints not only on the number of characters per line and the number of lines in each subtitle block, but also with respect to the desired reading speed. AppTek's machine translation with length control plays the key role in this process, effectively condensing subtitles to these constraints. Our experiments show that this condensation results in high-quality translations that convey the most important information, as measured by metrics such as BLEU or BLEURT, as well as the primary metric subtitle edit rate (SubER).

1 Introduction

Subtitle translation is a complex task that includes much more than recognizing what was uttered in an audio/video recording and translating it into the target language. Accurate timing of the subtitles, segmentation into syntactically and/or semantically coherent units, and comfortable reading speed are important aspects affecting the viewing experience, in addition to mere translation quality (Gerber-Morón et al., 2018; Liao et al., 2021).

In this paper, we describe how AppTek approaches the task with our in-house automatic speech recognition (ASR) and neural machine translation (NMT) systems, which we couple with our intelligent subtitle line segmentation algorithm (Matusov et al., 2019). In addition to algorithmic and modeling improvements, our unconstrained submissions benefit from in-domain data that was either available to us or was automatically extracted from public (parallel) data.

The paper is structured as follows. In the following Section, we describe the three main components of our subtitle translation approach: speech recognition in Section 2.1, machine translation in 2.2 and subtitle segmentation in 2.3. Section 3 gives details of our domain adaptation strategy for adapting to the entertainment and financial news domains which includes the usage of domain tags and fine-tuning on in-domain data. The effectiveness of domain adaptation is supported with experimental results at the end of the section. Next, we focus on the newest enhancements of our system with regard to subtitling constraints - the restrictions on the maximum number of characters per line (CPL), lines per subtitle block (LPB), and the desired maximum reading speed measured in characters per second (CPS). Section 4 explains in detail our approach of space-constrained MT, which includes elaborate MT length control combined with targeted re-translation and line segmentation optimizations. The trade-off between translation quality and compliance with the constraints such as the reading speed is explained at the end of the section, with experimental results showing how AppTek's NMT produces condensed translations that fulfill subtitling constraints just like subtitles from a professional subtitle translator, yet without a significant drop of the core MT quality. Finally, we summarize our findings in Section 5.

2 Subtitle Translation

We follow a cascaded speech translation approach - first, the speech signal is automatically transcribed into an English subtitle file, which is then automatically translated into the target language. Additional challenges related to subtitles are handled between these two components and after one or more translations of an utterance are obtained.

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 222–231
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2.1 Speech Recognition

AppTek's automatic speech recognition component is implemented as a hybrid conformer/HMM system. The acoustic model is trained on approximately 30K hours of transcribed, mixed-bandwidth English speech data including broadcast news, telephony, and publicly available open-source datasets. The training corpus includes a broad distribution of English dialects and accents. The acoustic model operates on 80-dimensional log Mel filterbank features and estimates posterior probabilities over 9K tied triphone states. The model architecture is based on a deep Conformer network using approximately 1 billion parameters. This conformer model was trained for 20 epochs using an OCLR-inspired learning rate schedule (Smith and Topin, 2018). Frame-level alignments and state tying were obtained from our previous best conformer-based acoustic model with 350M parameters. This model serves as the general-purpose English ASR system.

For adaptation to the entertainment domain, the general-purpose English ASR model is fine-tuned for approximately 1.5 epochs on 100K hours of in-domain audio, supplemented with 40 hours of music-only data. In both general and domain-specific systems, the language model (LM) is based on the LSTM architecture with over 300M parameters, combined with a count-based n-gram LM used for look-ahead pruning within the hybrid ASR framework. The vocabulary used across both LMs consists of approximately 250K words.

Punctuation marks and word casing are predicted on the raw ASR output with a separate LSTM sequence labeling model. The predicted sentence-final punctuation (period, question mark, exclamation mark) is then used to define sentence-like units for translation.

Optionally, we also apply inverse text normalization (ITN) to convert spoken numbers, dates, monetary amounts, and other entities involving numbers to their well-formatted text form that uses digits. This is done with an attention-based RNN sequence-to-sequence model trained to recover the original English written text from a synthetic spoken form, which we create by applying hand-crafted text normalization rules. For the submission, we make use of this ITN system for the Asharq-Bloomberg task, as the number of numeric entities in financial news is high and it is beneficial to have them correctly represented already in the source language. For ITV, we skip this step and instead rely on the MT system's ability to do the conversion to written form as part of the translation process (see Section 2.2).

2.2 Text Translation

AppTek's NMT system is a variant of the Transformer Big architecture (Vaswani et al., 2017) that uses a factored embedding representation on the source and target side for encoding word case, subword segmentation and glossary transfer information (Wilken and Matusov, 2019; Dinu et al., 2019). The system is trained to support additional input signals, represented with special tokens on the source or target side (Ha et al., 2016).

Part of the training data for which document labels are available is processed to include the context of the previous sentence with a separator, following the approach of Tiedemann and Scherrer (2017), with the difference that also sentences without context are used in training, so that the ability to benefit from the extra context can be turned on or off during inference.

As mentioned in the previous section, the MT system supports "spoken" input, i.e. without punctuation and casing, and with numbers and other numeric entities represented with words. For this, a part of the MT parallel training data is duplicated, and then the source language side of the copy is processed using our rule-based text normalization to create this spoken form.

Our MT systems support genre tags for approximately 20 genres, including a genre "news" for news-like content and "dialogs" for movie-like dialog or subtitle content. We use tags for these two genres in the experiments below. The training data is partitioned into genres using a sentence-level classifier trained on monolingual English training data, for which genre information is known.

All AppTek's MT systems support a length control mechanism which we applied in previous IWSLT submissions (Bahar et al., 2023). It is based on prefix tokens added to the target-side training data which represent length classes according to the target-to-source character count ratio. The boundaries of the length classes are chosen so that an equal number of training examples falls into each class (most extreme classes are chosen to be half the size). We use five length bins for the English to German and seven for the English to Arabic system. For training data that originates from subtitles in particular, our assumption is that it naturally consists of a mix of verbatim as well as differently condensed translation examples, and via the length token one can select between these condensation levels at inference time.

For certain language pairs, AppTek's MT systems also support style and speaker gender tags (Matusov et al., 2020). For English-to-German, the style tag controls the formality level of the output. The default tag value is "undefined" - this means that the system decides what formality level to use (e.g. formal second-person German pronoun "Sie" or informal "du") based solely on the context of the sentence. When the formality control is set, the system chooses the desired formality level. In the experiments below, we set the formality level to "formal" for the Asharq-Bloomberg financial news translation, where a formal translation style is expected.

It is worth mentioning that all of the above controls are implemented in AppTek's MT API as parameters[1] which a user can set based on prior knowledge or information derived from the upstream components (such as speaker gender).

2.3 Subtitle Segmentation

To combine ASR, MT and additional components into a subtitling pipeline, we follow the source template approach described in a previous edition's submission (Bahar et al., 2023). It consists of two steps: creation of captions in the original language of the video (here, English) and translation of these captions while keeping the subtitle blocks including their timings fixed. In both steps, a neural segmentation model is used to place line and block boundaries at semantically meaningful positions in the text, while additional hard constraints make sure the predicted segmentation adheres to the subtitling constraints (42 characters per line, 2 lines per block; while creating source language blocks also minimum and maximum block duration of 0.83 and 7 seconds, respectively). Automatically predicted punctuation and pauses of 3 seconds or longer are used to separate sentences, which are processed independently by the line segmentation algorithm (Matusov et al., 2019). Block timings are created from ASR word timings of the first and last word in a block - these are extremely accurate in a hybrid ASR system. For translation, the sentences as defined above - which may span several subtitle blocks - are sent to the MT component and are re-inserted into the source template using an additional hard constraint that enforces translations to be segmented into the existing blocks in terms of number and approximate relative sizes. More details on the subtitling pipeline can be found in Bahar et al. (2023).

In this year's submission we improve this subtitling pipeline in particular by focusing on reading speed compliance. This is achieved by tightly integrating MT length control with the space constraints of subtitling. The method will be described in Section 4.1.

3 Domain Adaptation

In the following we describe how we adapt our system to the specific domains of the subtitling tasks of this year.

3.1 Domain and Style Tags

As described in Section 2.2, AppTek's MT system supports domain and style control at inference time. This allows it to adapt to a specific domain without retraining the model.

For the ITV data we set the domain tag 'dialogs', but enforce no style control as it varies throughout each movie. For Bloomberg, we set the domain to 'news' and the style to 'formal'.

3.2 Fine-tuning with Parallel Data

While domain and style parameters optimize the controllability of a single model, stronger domain adaptation can be achieved by creating specialized models via fine-tuning on in-domain data.

For the ITV domain, we fine-tune both the ASR and the MT systems on movie subtitling data provided by one of AppTek's major media and entertainment customers. The ASR system adapted for the task is described in Section 2.1. To adapt the MT system, we extract sentences from English and German subtitles and obtain their sentence alignment using Vecalign (Thompson and Koehn, 2019). After filtering, a total of 12.6M sentence pairs with 130M running words on the English side is obtained. Using this parallel corpus, we fine-tune our general domain English-to-German MT system for approximately one epoch with a reduced learning rate.

For Bloomberg, we only adapt the MT system. In case of the Bloomberg English-to-Arabic task, we have access to a parallel corpus provided by Asharq Business with Bloomberg as part of AppTek's partnership with this company. It con-

[1] https://docs.apptek.com/reference/machine-translation

tains human-curated English captions and their human translations into Arabic, with a total of 240K sentence pairs with 7.3M running words counted on the English side. The source of the data are Bloomberg news programs similar to the ones used as development and test data at the IWSLT evaluation. We made sure that only historical data excluding the IWSLT dev/test data was used for training. The segmentation of this in-domain training data is mostly based on speaker turns and pauses and in most cases includes full sentences or speaker turns, so that an additional sentence alignment step is not necessary. Fine-tuning of our general-domain English-to-Arabic MT model is then performed for approximately one epoch.

3.3 Data Filtering Based on Development Set

For some domains and languages, it is challenging to find parallel in-domain datasets. This is, for example, the case for translating Bloomberg content into German. We use a simple filtering approach to collect parallel sentences that are similar to specific seed data, here, the German references of the Bloomberg development set.

Our filtering approach maps the sentences of the development set and the target-side of the general-domain parallel data into a shared embedding space. Sentence embeddings are obtained by averaging GloVe embeddings (Pennington et al., 2014) of each word in the sentence, as described in Arora et al. (2017). We embed the entire seed data into a single vector v_{seed} by averaging the embeddings of its target sentences. To filter a given corpus $\mathcal{C}$ down to size n, we choose the n sentences $e \in \mathcal{C}$ with the highest dot product $v(e)^T v_{\text{seed}}$ against the sentence embedding $v(e)$.

In our submission, we create German Bloomberg adaptation data from these corpora: a) ECB, Europarl, JRC-Acquis, NewsCommentary and DGT corpora from OPUS (Tiedemann, 2009) which all are closely related to the financial news domain, b) CCMatrix (Schwenk et al., 2021) as a large crawled corpus, as well as c) OpenSubtitles (Lison and Tiedemann, 2016) and TED2020 from OPUS to better match the translation style to the subtitling domain. We use pre-trained German word embeddings provided by Ferreira et al. (2016) to calculate the sentence embeddings. From each corpus, we select $n = 30$K sentences using the method above, ending up with a total of 157K deduplicated sentence pairs for fine-tuning.

The data obtained from our filtering is sentence-level. During fine-tuning, we mix it with TED 2020 samples with document-level context in a 1:1 ratio and otherwise follow the same recipe as described above for fine-tuning on parallel data.

ASR	MT	SubER	Bleu
ITV German			
General	General	73.4	18.8
General	+ domain tag	72.5	19.3
General	Fine-tuning	69.6	20.8
Adapted	General	71.4	19.5
Adapted	+ domain tag	71.3	20.1
Adapted	Fine-tuning	67.2	21.7

Table 1: Domain-adaptation of ASR and MT models on the ITV task, metrics measured on the 2025 development set. No subtitle condensation is applied (Section 4).

MT System	SubER	Bleu
Bloomberg German		
General	61.8	26.6
+ domain & formality tag	61.4	26.6
Finetuned (filter via dev set)	59.6	27.1
Bloomberg Arabic		
General	62.5	20.6
+ domain tag	61.8	20.9
Finetuned (in-domain data)	61.3	21.3

Table 2: Domain-adaptation for the German and Arabic Bloomberg tasks, on the 2025 development set. No subtitle condensation is applied (Section 4).

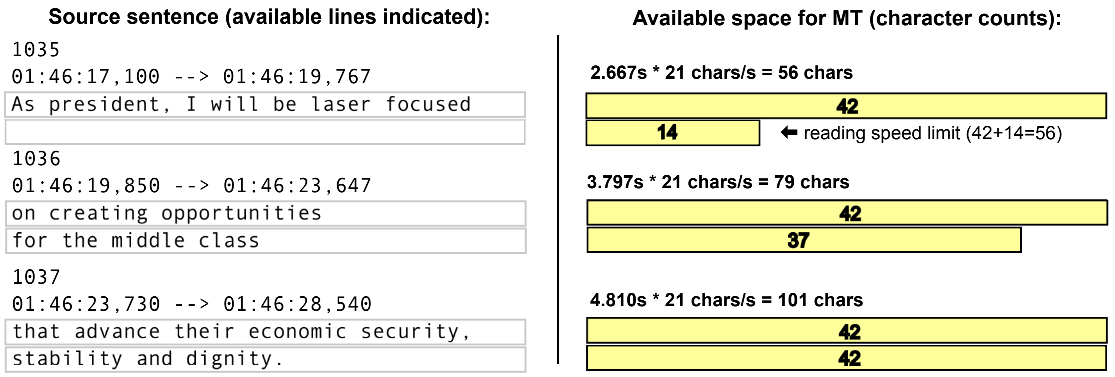
3.4 Results

The results for adapting the ASR and MT models for the ITV German task are shown in Table 1. For Bloomberg, we only adapt the MT model and show the results in Table 2.

We report case- and punctuation-sensitive subtitle edit rate SubER (Wilken et al., 2022) and BLEU (Papineni et al., 2002; Post, 2018) scores. Both are calculated with the SubER tool[2] using the final MT hypothesis in subtitle format. For BLEU calculation, the reference subtitles are converted into plain text sentences based on sentence final-punctuation. The hypothesis is aligned to the reference with an edit distance based algorithm, similar to the one implemented in mwerSegmenter (Matusov et al., 2005).

Comparing the upper and lower half of Table 1, one can see a clear positive effect of ASR domain

[2] https://github.com/apptek/SubER



Source sentence (available lines indicated):

```
1035
01:46:17,100 --> 01:46:19,767
As president, I will be laser focused

1036
01:46:19,850 --> 01:46:23,647
on creating opportunities
for the middle class

1037
01:46:23,730 --> 01:46:28,540
that advance their economic security,
stability and dignity.
```

Available space for MT (character counts):
2.667s * 21 chars/s = 56 chars
42
14
reading speed limit (42+14=56)
3.797s * 21 chars/s = 79 chars
42
37
4.810s * 21 chars/s = 101 chars
42
42

Figure 1: Calculation of MT space constraint. Translation of a sentence from the source subtitle template (left) has to fit into the same blocks it is extracted from. Blocks are limited to 2 lines of 42 characters. In addition, the reading speed limit defines a maximum character count per block (block duration multiplied by CPS value) to which we truncate the available lines. The list of character counts (here: `42,14,42,37,42,42`) is passed to the MT component to request a translation that fits into this space. Notes: source blocks with only one line allow for an additional line to be added during translation; last block in the example has long enough duration so that reading speed limit is fulfilled as a consequence of line and character limit; the calculated space does *not* limit the length of individual lines in the final subtitle file, e.g. the second line can be longer than 14 characters if the first line is shortened accordingly.

adaptation on the end-to-end ITV task performance. Presumably, improved robustness to difficult audio conditions such as background sounds and music, as well as adaptation to a wide range of forms of speech (mumbling, shouting, laughter, etc.) are some of the key factors. We refrain from calculating word error rates (WER) using the English development set subtitles as reference, as they are not exact verbatim transcriptions. However, for a manually annotated in-house test set of similar entertainment content, we measure a WER improvement from 13.1% to 12.0% as the result of fine-tuning.

Regarding machine translation, both adaptation methods – setting domain tags and fine-tuning – improve SubER and BLEU over the unadapted system; yet fine-tuning is consistently more effective. The biggest gains are observed on the ITV task, where the fine-tuning corpus is the largest and already in subtitle form. In particular, the MT system learns to produce shorter translations that better match the space constraints of subtitling, even without the explicit length control that will be introduced in the next section. On the Bloomberg task, fine-tuning is less effective. For English-to-Arabic, this may be partially explained by the fact that a portion of the fine-tuning data was already included in the training data of the general domain MT system. When comparing the effect of fine-tuning for English-to-German vs. English-to-Arabic, the automatically constructed English-German in-domain data set seems to achieve a similar positive effect to the real in-domain customer data used for Arabic as the target language.

While combining fine-tuning with the domain and formality tags is possible, we do not observe any significant improvements over the fine-tuned system, as it already learned domain and formality from the training data.

4 Subtitle Condensation

Subtitles do not only need to provide accurate translations, they also need to follow readability constraints to not hinder an immersive viewing experience. In last year's edition of the subtitling track, only one out of eight submitted systems for the English-to-German task achieved a compliance with the desired maximum reading speed of 21 characters per second (CPS) for more than 80% of the subtitles. We therefore focus on the reading speed constraint this year while at the same time aim to keep translation quality at a high level.

4.1 Space-constrained MT

We have made use of MT length control, as described in Section 2.2, for automatic subtitling in past IWSLT editions (Bahar et al., 2023; Ahmad et al., 2024; Wilken and Matusov, 2022). Here, we improve our approach by making the length constraints more exact and by basing them on the reading speed limit, not only the line limit. To do this, we make the following changes to the subtitle translation pipeline described in Section 2.3:

1. When extracting sentences from the source subtitle template, we calculate the available space the translation has to fit into, see Figure 1. Because line breaks cannot occur at arbitrary character positions, only at word boundaries, calculating a single character count value as the length limit for the translation would be imprecise. This problem is illustrated in Figure 2. Instead, we express the space constraint in terms of a list of character count values per line. Usually we have 2 lines of 42 characters available for each block, but this gets truncated by the character limit per individual block, which is block duration multiplied by the reading speed limit.

2. We implement an iterative process in the MT component. In the first iteration, translation is done without a length constraint by letting the model predict the length token itself. In the second iteration, the optimal length token is guessed based on the length ratio between total available space and the source character count and is then forced in the first decoding step. In each subsequent iteration the next shorter length class is selected. This process stops as soon as all words of the translation can be put into the space calculated in step 1 without overflow, or if the shortest length class is reached.[3]

3. Additional logic is added to the line segmentation algorithm (Section 2.3) to guarantee that a translation which *can* fulfill all space constraints indeed *does* fulfill all constraints after segmentation. This involves look-ahead pruning of partial hypotheses for which the remaining words do not fit into the remaining available blocks/lines.

4. Before translation, we decrease the source-side reading speed by shifting block end times beyond the duration of the actual speech onto the start of the next block or until a targeted CPS value is met. This way, space constraints for MT are relaxed and more content can be preserved, leading to improved translation quality scores. Here, we even use a CPS value of 17 instead of 21 to increase the effect. We repeat this block duration extension after translation (using 21 CPS). However, there it affects less than 1% of the blocks.

4.2 Results

To put the following evaluation of subtitle condensation into context, we first analyze how well the human-created reference subtitles adhere to the subtitling constraints. Table 3 shows that while the lines-per-block and characters-per-line limits are strictly followed, the reading speed limit is not. In fact, in practice it is often viewed as a soft limit that is expected to be met only for the majority of subtitles of a given film/show, but not necessarily for all of them. In addition, it can be seen that the reading speed limit is violated more often for German. German sentences are longer on average than their English equivalents, making it harder to fulfill a certain character limit.

Task	Ref. Compliance [%]		
	LPB	CPL	CPS
ITV German	100.0	100.0	88.6
Bloomberg German	100.0	100.0	78.4
Bloomberg Arabic	99.9	100.0	97.4

Table 3: Fraction of subtitles in the 2025 development set reference compliant with the 2 lines-per-block (LPB), 42 characters-per-line (CPL) and 21 characters-per-second (CPS) limits.

To show the effect of subtitle condensation, in Figure 3 we plot the translation quality as measured in BLEU against CPS compliance (always based on 21 characters per second) while using different CPS values to constrain the translation lengths. We see a trade-off between the two, which is expected as more content can be preserved in longer translations, leading to more n-gram matches with the reference. Especially if the CPS compliance surpasses the one of the human reference we see a clear drop in BLEU score.

Even when calculating length constraints using the targeted value of 21 characters per second, the compliancy does not reach 100%. Manual inspection reveals that the remaining violations are indeed cases where the speaking rate is so high that even the shortest MT length class does not lead to a compliant translation. This in particular happens for very short sentences containing no superfluous words. Notably, already the generated English source templates, which – apart from some ASR errors – contain verbatim transcripts, have a CPS compliancy of only around 80%, indicating that there are many fast-paced dialogues in the development set videos which would require heavy condensation. For extreme cases, we see a limitation of our approach of sentence-by-sentence translation, because whole sentences might have to be left out in the subtitles to keep up with the video, or mul-

[3] This iterative approach is an efficient alternative to the "Length ROVER" (Wilken and Matusov, 2022) with similar output but a significantly reduced number of translation passes, therefore suitable for commercial application.

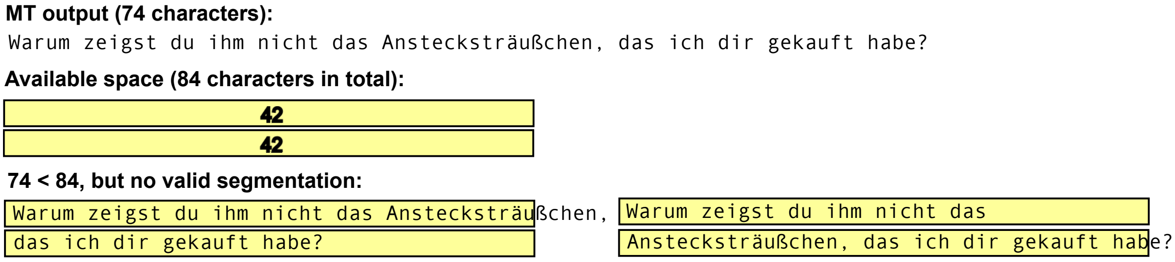

MT output (74 characters):
Warum zeigst du ihm nicht das Ansteckstäußchen, das ich dir gekauft habe?
Available space (84 characters in total):
42
42
74 < 84, but no valid segmentation:
Warum zeigst du ihm nicht das Ansteckstäußchen,
das ich dir gekauft habe?
Warum zeigst du ihm nicht das
Ansteckstäußchen, das ich dir gekauft habe?

Figure 2: Example illustrating that a simple translation length limit in terms of total character count is too imprecise for subtitle template translation. Assuming a given source sentence is contained within a single block, it is not enough to limit translation length to 84 characters, which one would naively derive from a block size of 2 lines with 42 characters. In fact, as shown, even a translation of 74 characters – depending on the specific word lengths – may not fit into one block as line breaks may only occur at word boundaries[4]. This is the reason we compute an exact line-wise space constraint according to Figure 1 and use it as compliancy check while selecting MT length variants.

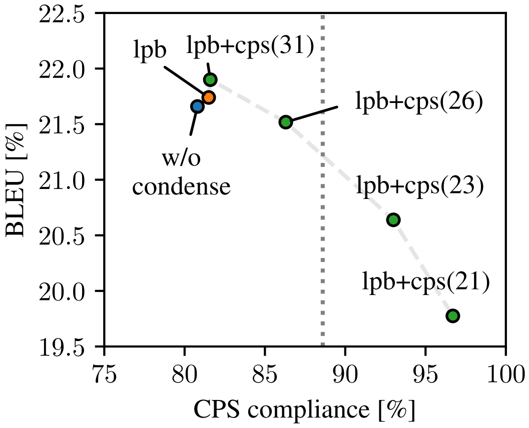


Figure 3: Different subtitle condensation levels on the English-to-German ITV development set: we start by condensing translations to fit into 2 lines per block (lpb), and then introduce different reading speed values (cps) as additional constraint. The more constrained settings lead to worse translation quality (BLEU), but better 21-characters-per-second compliance (CPS). The reference set has a compliance of 88.6 % (dotted vertical line).

tiple sentences would have to be condensed into a single one.

We determine which condensation setting to use for each task by re-translating with increasingly strict length control when necessary (see Section 4.1): we always try to condense subtitles to fit into two lines (LPB compliance), and then compare condensation with different target reading speeds (CPS). The MT reading speed limit is then chosen so that the subtitles have a reasonably high CPS compliance, while the translation quality does not deteriorate too much. For the ITV task, we set the condensation parameter to 23 CPS, while for both Bloomberg tasks we set 21 CPS. These settings are also optimal in terms of SubER.

Table 4 shows the main results of the subtitle condensation for all three tasks. We report SubER and BLEU as described previously in Section 3.4. In addition to the BLEU score, we also compute the neural metric BLEURT (Sellam et al., 2020) on the MT plain text hypotheses aligned to plain text reference translations by the SubER tool. The compliance metrics are calculated with the script provided by Papi et al. (2023).

Subtitle condensation leads to more compliant and thus overall shorter subtitles for all three tasks. In fact, the subtitles generated by our systems are more CPS compliant than the human-created reference translations of the development set (Table 3). Our implementation further guarantees a CPL compliance of 100%, as we chose to violate the LPB limit instead by adding additional lines in cases where the translation does not fit into the available space.

Although condensation does have a negative impact on the translation quality when measured in BLEU or BLEURT, it does lead to an improvement in SubER. One reason for this is that SubER does not penalize word omissions as harshly. In case of BLEU, short hypotheses are explicitly penalized with the brevity penalty (all three of the condensed systems have a hypothesis/reference length ratio of less than 1).

The IWSLT findings report (Abdulmumin et al., 2025) verifies that our system also produced highly space compliant subtitles on the evaluation data: We achieve 100.0% CPL and more than 99% LPB compliancy on all three tasks, and have 93.8%, 92.4% and 99.8% LPB compliancy on the ITV,

[4] Hyphenation to split words across lines is not common in subtitling.

Task	Condense	SubER (↓)	Bleu (↑)	Bleurt (↑)	Compliance [%] (↑) LPB	 CPL	 CPS
ITV German	✗	67.2	21.7	0.454	98.0	100.0	80.8
	✓	64.9	20.6	0.448	100.0	100.0	93.0
Bloomberg German	✗	59.6	27.1	0.548	95.3	100.0	76.9
	✓	59.2	25.6	0.536	99.3	100.0	92.1
Bloomberg Arabic	✗	61.3	21.3	0.568	99.6	100.0	96.4
	✓	61.2	20.8	0.563	100.0	100.0	99.8

Table 4: Subtitle condensation results on all three IWSLT tasks, reported on the 2025 development set. All models are domain-adapted (via fine-tuning). The three models with condensation enabled correspond to our submitted primary systems.

Bloomberg German and Arabic test sets respectively, while maintaining high translation quality.

Looking back to the IWSLT 2023, we report a significant improvement over the last two years. While we obtained the best automated SubER scores on the ITV test set among all submissions (Agarwal et al., 2023), the 2023 system's SubER score on the development set (the same one used in this year's evaluation) was 71.4 (Bahar et al., 2023). In comparison, our system from this year scores an substantial 6.5 points better in this metric.

5 Conclusions

This paper presented AppTek's submission for the unconstrained subtitling track of the IWSLT 2025. Our focus this year was to a) boost translation quality by domain- and style-specific adaptation, and b) deliver readable subtitles that adhere to given space and reading-speed constraints.

Our key findings are as follows:

- **Domain and style adaptation matter.** Fine-tuning the system on in-domain data yields a large quality gain as measured in BLEU and BLEURT. The simpler, tag-based adaption approach does not require additional training but is less effective.

- **Length-aware condensation works.** The proposed condensation algorithm generates subtitles with characters-per-line (CPL), lines-per-block (LPB) and character-per-second (CPS) compliance scores similar to or better than human references, while only marginally decreasing BLEU and BLEURT scores. The trade-off between space and reading speed constraints and the general translation quality can further be controlled gradually.

Together, these methods result a substantial boost in SubER – the track's primary evaluation metric.

Acknowledgments

This work was (partially) supported by the project RESCALE within the program *AI Lighthouse Projects for the Environment, Climate, Nature and Resources* funded by the Federal Ministry of Germany for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV), funding ID: 6KI32006B.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 30 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austia (in-person and online). Association for Computational Linguistics. To appear.

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, and 43 others. 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.

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, and 25 others. 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.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Olivia Gerber-Morón, Agnieszka Szarkowska, and Bencie Woll. 2018. The impact of text segmentation on subtitle reading. *Journal of Eye Movement Research*, 11(4):10–16910.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Sixin Liao, Lili Yu, Erik D Reichle, and Jan-Louis Kruger. 2021. Using eye movements to study the reading of subtitles in video. *Scientific Studies of Reading*, 25(5):417–435.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC‘16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

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

Evgeny Matusov, Patrick Wilken, and Christian Herold. 2020. Flexible customization of a single neural machine translation system with multi-dimensional metadata inputs. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 204–216, Virtual. Association for Machine Translation in the Americas.

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023. Direct speech translation for automatic subtitling. *Trans. Assoc. Comput. 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

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.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6490–6500. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Leslie N. Smith and Nicholay Topin. 2018. Super-convergence: Very fast training of neural networks using large learning rates. *Preprint*, arXiv:1708.07120.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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@ACL 2022, Dublin, Ireland (in-person and online), May 26-27, 2022*, pages 1–10. Association for Computational Linguistics.

Patrick Wilken and Evgeny Matusov. 2019. Novel applications of factored neural machine translation. *Preprint*, arXiv:1910.03912.

Patrick Wilken and Evgeny Matusov. 2022. AppTek‘s submission to the IWSLT 2022 isometric spoken language translation task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 369–378, Dublin, Ireland (in-person and online). Association for Computational Linguistics.