Matesub: the Translated Subtitling Tool at the IWSLT2023 Subtitling task

Simone G. Perone
Translated srl
via Indonesia 23
00144 Rome - Italy
simone@translated.com

Abstract

This paper briefly describes Matesub, the subtitling tool Translated used to participate in the Subtitling shared task at IWSLT 2023. Matesub is a professional web-based tool that combines state-of-the-art AI with a WYSIWYG editor. The automatic generation of subtitles in Matesub is based on a cascade architecture, composed of ASR, text segmenter and MT neural models, which allows covering any pair from about 60 languages and their variants.

1 Matesub

Matesub[1] is a web-based tool released by Translated[2] that combines state-of-the-art AI with a WYSIWYG (What You See Is What You Get) editor for supporting professionals in the creation of subtitles for audio visual documents. Matesub generates subtitling suggestions through a processing pipeline which was used to participate in the Subtitling shared task at IWSLT 2023. This paper first describes the pipeline, and then presents and discusses the scores of the submission.

1.1 The subtitling pipeline

In Matesub, subtitles are automatically generated by a pipeline (Figure 1) which concatenates two main modules, based on neural models: an automatic speech recognition (ASR) system and a module providing the Captions & Subtitles Service. They are described in the following.

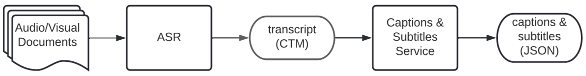

Figure 1: Architecture of the subtitling pipeline.

1.1.1 Automatic speech recognition

The ASR is in charge of the transcription of the speech content of an audio signal. In Matesub, this processing stage is provided either by an in-house ASR model or by a 3rd party commercial ASR service, according to the availability of the internal solution and its relative quality. In both cases, the word hypotheses are expected to be given in conversation time mark (CTM) format. This text file consists of records each having 5 fields, e.g.:

23.66	0.29	human	0.00998	False
23.96	0.40	beings.	0.01000	True
24.48	0.13	We	0.33000	False

whose meaning is given in Table 1.

field	meaning
1	start time (sec)
2	duration (sec)
3	token (i.e. word)
4	confidence
5	end of sentence (boolean)

Table 1: Fields in the CTM format.

Note that the transcription is punctuated and cased; moreover, the flag indicating the *end of sentence* is typically set *on* for acoustic reasons, like the presence of the pause between the tokens *begin.* and *We*, but - less frequently - also for "linguistic" evidence (learned by the ASR from training data).

1.1.2 Captioning and subtitling

The Captions and Subtitles Service is in charge of building, starting from a given CTM file, the SubRip Subtitle (SRT) files of the transcription contained in the CTM file and its translation; the two SRTs are finally merged in a single JSON file. As shown in Figure 2, this module consists of two main components, a text segmenter and a neural machine translation (NMT) system, in addition to a number of secondary sub-components.

The two main components are built using the same sequence-to-sequence neural modeling tech-

[1] https://matesub.com/
[2] https://translated.com/

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 461–464
July 13-14, 2023 ©2023 Association for Computational Linguistics

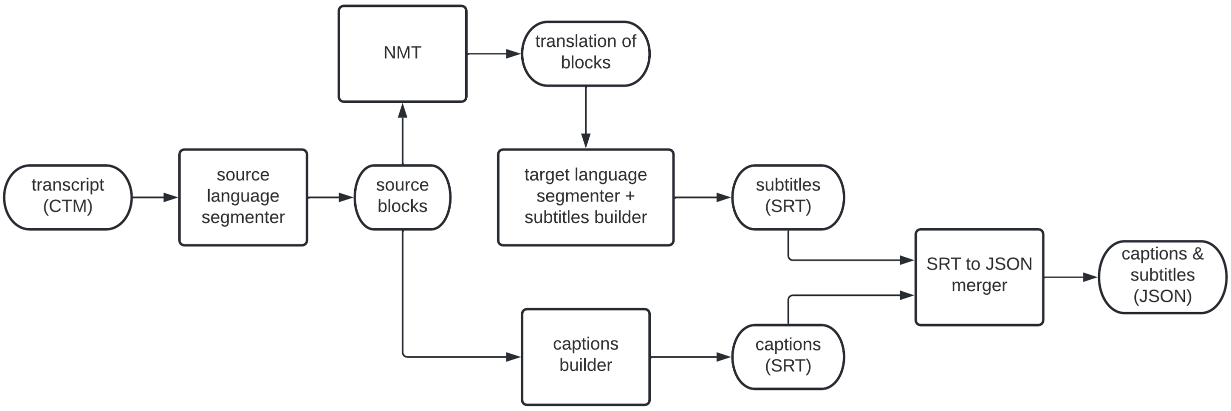

Figure 2: Captions and subtitles service.

nique. The segmenter, implemented as proposed in (Karakanta et al., 2020; Papi et al., 2022), inserts in an unsegmented input text - either in the source or in the target language - markers of segment boundaries. It is trained on pairs of unsegmented-segmented text, where segment boundaries are marked by means of two special symbols: <*eob*> to mark the end of block (caption or subtitle), and <*eol*> to mark the end of line. Figure 3 shows an example of a sentence after inserting the markers from the corresponding fragment of the SRT file.

```
164
00:08:57,020–>00:08:58,476
I wanted to challenge the idea

165
00:08:58,500–>00:09:02,060
that design is but a tool
to create function and beauty.
```

```
I wanted to challenge the idea <eob> that design is but a
tool <eol> to create function and beauty. <eob>
```

Figure 3: Subtitle file (top) and the full sentence annotated with the subtitle breaks (bottom). Figure taken from (Karakanta et al., 2020).

The neural machine translation engine performs the translation of the text from the source language (English, in the IWSLT 2023 context) into the corresponding text in the target language (here German and Spanish). Other processing modules are in charge of (i) generating captions/subtitles in SRT format (starting from transcripts, word timestamps, translations and segmentations), and (ii) merging the SRTs of captions and subtitles into a single JSON file. The main processing steps are:

1. Segmentation of the transcription on the basis of acoustic cues (*audio blocks*)
2. Segmentation of audio blocks into *caption blocks* (and lines) by means of the source language segmenter
3. Automatic translation of each caption block into the target language(s) (*subtitle blocks*)
4. Segmentation of subtitle blocks into lines by means of the target language segmenter
5. Timing projection from the CTM to the caption/subtitle blocks
6. Packaging of SRT and JSON files.

Note that the translation of each block in step 3 is done without looking at the context, i.e. at the surrounding blocks. On the one hand, this worsens the quality of the translation a little, but, on the other, it facilitates the satisfaction of the reading speed requirement through the n-best mechanism, sketched in the next section.

1.1.3 Machine translation

Neural machine translation is provided by ModernMT[3] (Bertoldi et al., 2021) through a REST API connection. ModernMT implements the Transformer (Vaswani et al., 2017) architecture; generic *big* models (about 200M parameters each), trained on both public and proprietary data, cover hundred of languages[4] in any direction, through a seamless integration of the pivot based approach, where the pivot language is English. Matesub requests ModernMT to provide the 16 best translations of

[3] https://www.modernmt.com/

[4] https://www.modernmt.com/api/#languages

each block (step 3 mentioned in the previous section); between them, the hypothesis with the highest probability and whose length permits to satisfy the reading speed constraint (given the duration of the block) is selected. If no such hypothesis exists, the shortest is chosen.

1.2 The editor

Matesub provides a WYSIWYG editor, which allows the user to review and correct the subtitles automatically generated and synced in the chosen target language by the back-end subtitling pipeline. Figure 4 shows a screenshot of the Matesub editor.

The editor permits the user to easily fix both translation and segmentation errors, thanks to the rich catalogue of functions and user-friendliness. Once the editing is over, subtitles can be embedded in the video or exported in production-ready SRT files or any other supported subtitles format.

2 Submission and Results

Translated participated in the Subtitling shared task at IWSLT 2023 with the back-end subtitling pipeline of Matesub. No adaptation of the general purpose pipeline was carried out, therefore the quality of subtitles generated for the audio-visual documents proposed in the shared task is that typically expected by the in-production system before the post-editing stage. Since neural models of Matesub (ASR, text segmenter and MT) were trained on more resources than those allowed for the constrained condition, we labelled our submission as *unconstrained*; it was also our unique submission, and as such it is the *primary* run.

Table 2 shows scores of our test set subtitles as computed by the organizers (Agarwal et al., 2023). They are in line with those we obtained on the dev sets.

Without knowing the results of the other submissions, it is hard to judge the results obtained. However, some considerations can be made:

- As expected, from the pure speech translation perspective, the TED domain is the easiest one by far
- Surprisingly, at least when German is the target language, the EPTV domain is as much challenging as ITV and PELOTON, which we expected to be the most difficult ones
- Assuming that BLEURT and ChrF are more reliable than BLEU and TER (according to (Kocmi et al., 2021), for example), it seems that the quality of TED and of Spanish EPTV subtitles is high, while subtitles of ITV, PELOTON and German EPTV documents would need major post-editing
- Since SubER is based on TER and Sigma on BLEU, their values match the scores of those metrics rather than BLEURT, ChrF and the subtitle compliance as measured by CPS/CPL/LPB, possibly affecting the final ranking of Matesub
- The compliance of subtitles is language independent
- Despite the fact that Matesub does not implement any hard rule, relying only on machine learning methods, CPL and CPL are (almost) perfect
- The reading speed (CPS) is under the max threshold of 21 characters per second in about 85% of subtitles; more in detail, the average is about 18.5 and only in 5% of cases it exceeds 30 characters per second, values that we consider satisfactory.

Acknowledgements

Matesub received funding from the European Institute of Innovation and Technology (EIT), a body of the European Union, through the MateDub (2020) and MateDub++ (2021) projects. Within them, the Matesub subtitling chain was developed in collaboration with the FBK's MT research unit. We especially thank Mauro Cettolo for his invaluable contribution to the success of this product and the support he gave us in the participation to the Subtitling track at IWSLT 2023.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Khalid Choukri, Alexandra Chronopoulou, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Benjamin Hsu, John Judge, Tom Ko, Rishu Kumar, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Matteo Negri, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Elijah Rippeth, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Mingxuan Wang, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In Proc. of IWSLT, Toronto, Canada.

		Subtitle quality		Translation quality				Subtitle compliance		
en-	domain	SubER↓	Sigma↑	BLEU↑	ChrF↑	TER↓	BLEURT↑	CPS↑	CPL↑	LPB↑
-de	EPTV	87.04	57.73	12.08	43.59	85.53	.4705	88.59	99.20	100.00
	TED	67.70	62.01	20.37	50.05	65.55	.5500	90.55	98.61	100.00
	ITV	73.11	67.04	14.92	37.13	71.27	.4501	80.21	99.47	100.00
	PELOTON	79.72	68.27	10.06	34.46	78.25	.4264	89.17	99.29	100.00
	ALL	75.41	65.22	14.81	39.50	73.60	.4591	84.97	99.25	100.00
-es	EPTV	74.47	59.59	21.06	54.11	72.08	.5728	90.15	99.44	100.00
	TED	45.94	66.85	40.36	65.72	43.81	.7047	92.62	99.48	100.00
	ITV	71.25	71.06	18.50	41.07	69.57	.4592	81.93	99.51	100.00
	PELOTON	74.87	70.99	15.96	41.86	73.88	.4666	88.27	99.60	100.00
	ALL	68.11	68.37	22.34	47.38	66.66	.5059	86.07	99.52	100.00

Table 2: Results of the Matesub submission.

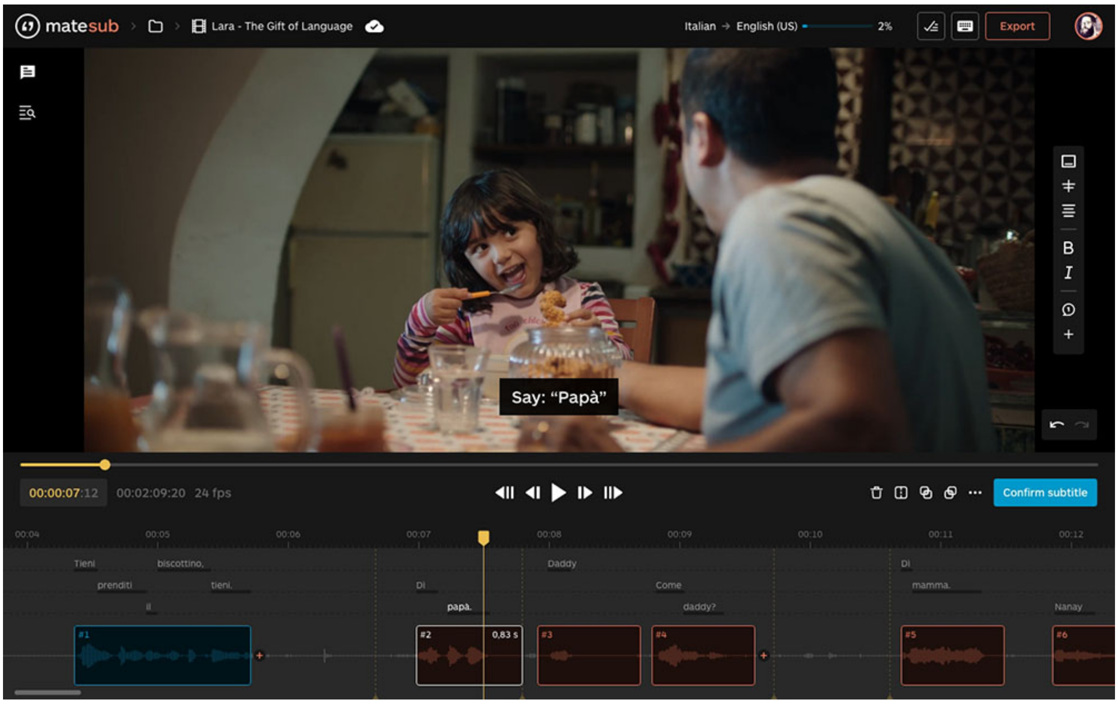

Figure 4: Screenshot of Matesub Editor

Nicola Bertoldi, Davide Caroselli, M. Amin Farajian, Marcello Federico, Matteo Negri, Marco Trombetti, and Marco Turchi. 2021. Translation system and method. US Patent 11036940.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. MuST-cinema: a speech-to-subtitles corpus. In Proc. of LREC, pages 3727–3734, Marseille, France.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In Proc. of WMT, pages 478–494.

Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022. Dodging the data bottleneck: Automatic subtitling with automatically segmented st corpora. In Proc. of AACL-IJCNLP, pages 480–487.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proc. of NIPS, pages 5998—-6008.