Cross-Lingual Sentence Compression for Length-Constrained Subtitles in Low-Resource Settings

Tollef Emil Jørgensen and Ole Jakob Mengshoel
Norwegian University of Science and Technology,
Data and Artificial Intelligence Group
{tollef.jorgensen, ole.j.mengshoel}@ntnu.no

Abstract

This paper explores the joint task of machine translation and sentence compression, emphasizing its application in subtitle generation for broadcast and live media for low-resource languages and hardware. We develop CLSC (Cross-Lingual Sentence Compression), a system trained on openly available parallel corpora organized by compression ratios, where the target length is constrained to a fraction of the source sentence length. We present two training methods: 1) Multiple Models (MM), where individual models are trained separately for each compression ratio, and 2) a Controllable Model (CM), a single model per language using a compression token to encode length constraints. We evaluate both subtitle data and transcriptions from the EuroParl corpus. To accommodate low-resource settings, we constrain data sampling for training and show results for transcriptions in French, Hungarian, Lithuanian, and Polish and subtitles in Albanian, Basque, Malay, and Norwegian. Our models preserve high semantic meaning and metric evaluations for compressed contexts.

1 Introduction

Subtitles and captions are becoming increasingly important due to the abundance of audiovisual content we produce daily. Furthermore, regulations for universal design are continuously developing, often requiring captions for published media, which is especially vital for the hard of hearing (Burgstahler, 2009).

Sentence compression is a natural part of the subtitling process, as there are strict requirements concerning features like the on-screen visual constraints of the text itself (Cintas and Anderman, 2008; Corston-Oliver, 2001). These constraints are commonly categorized as follows. First, in terms of *space*, we aim for brevity, with subtitles being one or two lines. Second, concerning *time*, the goals are accurate timing and on-screen reading time. Third, regarding *presentation*, subtitles must be positioned well and not obstruct critical visual elements. Subtitle compression is closely related to sentence compression, the task of reducing or summarizing the contents of a single sentence. For cross-lingual subtitles, however, introducing translation increases complexity. Furthermore, we may perform *compression* to reduce or rephrase spoken language into subtitles. A subtitle may be as short as an "Ok!" instead of a longer spoken utterance such as "I agree, let's go for it!". Therefore, automated subtitling from a source to a target language has requirements beyond typical translation applications, where the complete meaning behind the text should be kept.

Transcription (English):
Okay, Mr. Conroy, can you tell me your location?
I don't know. I'm in a coffin, I don't know where.
Please help me, I'm scared.
↓
Subtitle (Norwegian Bokmål):
Hvor befinner du deg? (Where are you located?)
Inni en kiste. Jeg vet ikke hvor. (Inside a coffin. I don't know where)

Figure 1: Example of a compressed subtitle from the movie *Buried* (2010). Underlined text is used for the translation and subtitling.

This paper is motivated by observations of the compression of text and terms from English transcriptions to the translated subtitles in broadcast TV. The specific types of compression are evaluated in the thesis by Sandvold (2019) for Norwegian, although the same constraints are found across different languages (Karakanta et al., 2020).

Challenges

Figure 1 shows a compressed subtitle, used as an example in *Guidelines for good subtitling in Norway* (Språkrådet, 2017). Here, the reference to "Please help me, I'm scared" is entirely omitted in the subtitle (arguably because the scene and multimodal context infer it), and the other sentences

Proceedings of the 31st International Conference on Computational Linguistics, pages 6447–6458
January 19–24, 2025. ©2025 Association for Computational Linguistics

Compression rate	MM (single model per compression)	CM (controllable model)
0.5	*Can you say where you are?* *I am in a coffin, I don't know where*	*Can you say where you are?* *I am in a coffin.*
0.7	*Can you say where you are, Mr. Conroy?* *I am in a coffin, I don't know where*	*Can you say where you are?* *I am in a coffin, I don't know where*
1.0	*Can you say where you are, Mr. Conroy?* *I am in a coffin, I don't know where I am*	*Can you tell me where you are?* *I am in a coffin, I don't know where*
Original pre-trained	*- Ok, Mr. Conroy, can you tell me where you are?* *I am in a coffin, I don't know where*	

Table 1: Output examples for our proposed MM and CM models for varying compression rates for the original transcription in Figure 1. All results are back-translated from Norwegian to English.

are *compressed*. The same excerpt is used for Table 1, where our proposed CLSC models CM and MM are configured to specific compression ratios (0.5 denotes models trained for 50% of the target length). Compression ratios and our models are further explained in Section 3. Automating reduction and compression, justified due to redundancy and upholding the constraints of subtitles, is one of the many challenges to resolve in cross-lingual subtitle translation. Limitations are discussed at the end.

Source: English	**Target: Norwegian (English)**
Oh father in-law-son in-law-bonding	Vi knytter bånd. (We're bonding.)
Hey guys before Anu gets here can I talk about the seating situation?	Kan vi snakke om sitteplassene? (Can we talk about the seating?)
Hey have you checked the dates on these? They're all expired	Disse har gått ut på dato. (These have expired.)

Table 2: Examples from The Big Bang Theory data by (Sandvold, 2019). Underlined text is used for the translation and subtitling.

We use data from the thesis *Audiovisual Translation: A Comparative Text Analysis of English Speech and Norwegian Subtitles in The Big Bang Theory* (Sandvold, 2019) to inspect challenges with compressed translations. This data contains manually transcribed spoken English utterances and official Norwegian subtitle pairs from season 12 of the sitcom *The Big Bang Theory* (BBT). Excerpts are found in Table 2, showcasing different forms of *compression*. One of the many challenges in the dataset is texts that rely heavily on external world knowledge, such as cultural phrases, idioms, and localized terms (e.g., movie titles will have their translated versions across languages). Consider an idiom like “It's raining cats and dogs,” which famously translates poorly to other languages. A Norwegian translation is “det bøtter ned” (it's bucketing down). In these cases, literal translations may result in misunderstandings.

Contributions

We show that small and efficient models, when trained on length-constrained data, are viable for highly accessible cross-lingual subtitle compression. Such subtitles are suitable for, e.g., assistive tools for subtitlers and applications for the hard of hearing and language learners. Empirical results are provided for OpenSubtitles data (Lison and Tiedemann, 2016) and transcribed speeches from the EuroParl corpus (Koehn, 2005). Finally, our work includes many tools to access, manipulate, and visualize parallel corpora, easily adaptable to new sources and problems.[1]

Paper structure

Section 2 discusses related work on translation and compression with a focus on subtitling. In Section 3, we introduce a system that performs both translation and compression to create subtitles, in addition to datasets, evaluation, and modeling setup. Section 4 presents experimental results and the degree of compression. In Section 5 we discuss our results and point to potential future research. Limitations of the work are discussed at the end.

2 Related work

Challenges in audiovisual subtitling and translation are discussed in detail in the books by Cintas et al. (Cintas, 2013; Cintas and Anderman, 2008; Cintas and Remael, 2020). Within natural language generation, Gupta et al. (2019) describes *problem categories* for automated translation of subtitles and TV shows, some specifically for textual translation.

[1] Code available at `https://github.com/tollefj/CLSC`

In their findings, some of the most prominent errors when translating from English to German, French, and Spanish are *paraphrased translations, word structure errors, word ordering* and *language nuances*. Interestingly, 20% to 30% of the errors are contributed by paraphrasing and are thus related to the output length. Examples show that the machine-translated results for German were almost double the length of the human translations. Similar conclusions about the importance of word ordering and paraphrasing were found for Dutch sentence compression (Marsi et al., 2009).

Early work on compression and simplification of subtitles includes the use of tagging, chunking, and shallow parsing, in addition to systems for avoiding ungrammatical sentences by techniques such as keeping determiners and pronouns related to the heads of noun phrases and alignment of syntactic trees (Vandeghinste and Pan, 2004; Daelemans et al., 2004). Continuing these developments, *sentence compression* became the suggested method to achieve shorter subtitles (Bouayad-Agha et al., 2006; Melero et al., 2006). Marsi et al. (2009) consult approaches for data-driven sentence compression, along with details on the subsequences found in sentence compression tasks within the same language. However, we cannot rely on subsequences in a cross-lingual setting.

While these earlier methods for sentence compression often included removing, splitting, and merging text, we no longer need to rely on manual edit operations to achieve a grammatical and *compressed* sentence, as pre-trained transformer models now excel at these tasks (Park et al., 2021).

2.1 Subtitle translation

An alternative to using subsequences or phrases within the same language would be to employ back-translation in parallel corpora. However, in a review on multi-domain adaptation for machine translation (MT) tasks, Saunders (2022) found that using back-translated data entirely may result in 'translationese' domains.

Aziz et al. (2012) studied translation and compression on a selection of subtitles in TV series from a previous edition of the OpenSubtitle corpus (Tiedemann, 2012). A phrase-based statistical MT system was developed using the Moses toolkit (Koehn et al., 2007) for English to Portuguese to have the translations comply with the time and space constraints of subtitles. Their system, unfortunately, is not available for reproduction.

Research on MT for subtitling is well described by Volk et al. (2010) and Bywood et al. (2014). Volk et al. (2010) developed a statistical machine translation (SMT) system for Swedish to Danish and Norwegian (used in production from 2008), three closely related languages with similar grammars but differing orthography. Similar to Aziz et al. (2012), the system was developed using several older software packages, and no reproducible code is available. Bywood et al. (2014) provided a comprehensive evaluation of statistical methods from a collected multilingual corpus (Petukhova et al., 2012), stating the need for quality-controlled subtitle pairs. However, the data is no longer openly available.

2.2 From Statistical to Neural

Recent advancements in natural language generation, including MT, have seen a shift from statistical methods to neural and transformer-based architectures (Gehring et al., 2017; Kalchbrenner et al., 2016; Zhao et al., 2019; Gao et al., 2022; Jauregi Unanue et al., 2021). Niehues (2020) explores the ability of encoder-decoder architectures to constrain the generation length of translation models. They manipulate the generation procedure of models directly by 1) restricting search space by altering the probability of the end-of-sentence token and 2) including a length-aware modeling scheme throughout the decoding step, requiring the target length during training. The last step poses a problem for new data, as the target length is unknown. Moreover, the system modifies the transformer architecture directly instead of adding flexible layers on top of pre-trained models.

Svensson and Troksch (2022) suggest methods to control the length of translations for subtitles based on models with the MarianMT architecture (Junczys-Dowmunt et al., 2018). They find that the *length ratio* of the data produces the best results for transfer learning of sequence-to-sequence models. Their work introduces new tokens to represent token lengths of different compression categories (short, normal, and long) and ratios (e.g., 0.5). Our CM approach builds upon this compression token to increase the amount of training data (focusing on the low-resource aspect) and normalizes the compression ratio based on the relative length ratios for different languages, described in detail in Section 3.

Perković et al. (2023) examine the reduced output of translation models by extending English source texts from OPUS-100 (Zhang et al., 2020a) with data from the Paraphrase Database (Ganitkevitch et al., 2013), keeping the target text as-is. Results show marginal improvements while introducing uncertainties around the quality of back-translations.

3 Methods and Data

With our Cross-Lingual Sentence Compression (CLSC) system, illustrated in Figure 2, we suggest a simplistic and efficient approach to subtitle compression. CLSC has two main components: preprocessing (discussed in Section 3.1) and modeling (Section 3.2). Through these components, sentence compression becomes easily accessible for target languages with limited access to data.

3.1 CLSC Preprocessing

Training data is sourced from the OpenSubtitles corpus (Lison and Tiedemann, 2016; Tiedemann, 2016), a dataset of sentence-aligned subtitles spanning 62 languages. Unlike methods relying on data augmentation techniques such as back-translation (Lu et al., 2021; Svensson and Troksch, 2022; Perković et al., 2023), we preserve the original data to ensure consistency and straightforward transferability across languages, including low-resource ones.

Compression is expressed using a compression ratio, c, representing the proportion of the target text length relative to the source length. To account for inherent length differences between languages, we introduce a language-specific normalization factor, α_{lang}, which adjusts c based on the average character length per sentence of the target language relative to the source. The normalized compression ratio is computed as $c \cdot \alpha_{\text{lang}}$. We evaluate a range of compression ratios, $c = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$, allowing us to control compression to suit different applications and analyze its effects in our experiments.

The availability of parallel sentence pairs varies significantly across languages and compression ratios, with some ratios (e.g., $c = 0.5$ or $c = 0.6$) having just a few thousand annotated pairs (see Table 3). To limit the effects of data imbalance between the training and evaluation of different languages, we cap the data to 250,000 pairs per compression ratio, a value chosen to balance length reduction, translation quality, and consistency across languages. For ratios with fewer than 250,000 pairs (highlighted in Table 3), we use all available data. After sampling, the data is split into training, validation, and test sets in an 80:10:10 ratio. For evaluation on EuroParl (Koehn, 2005), data is sampled as-is, without length-based filtering or splitting.

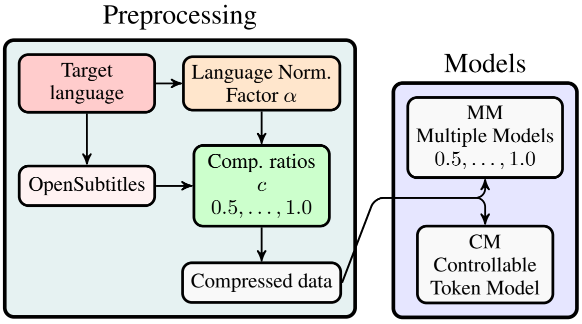
Preprocessing
Target language
Language Norm. Factor α
OpenSubtitles
Comp. ratios c $0.5, \ldots, 1.0$
Compressed data
Models
MM Multiple Models $0.5, \ldots, 1.0$
CM Controllable Token Model

Figure 2: The CLSC system. Data from a selected target language is processed and split into subsets of varying compression ratios. The processed data is passed to one of the two models supported: MM or CM.

3.2 CLSC Models

CLSC is based on the MarianMT architecture (Junczys-Dowmunt et al., 2018), using a collection of pre-trained encoder-decoder models from OPUS-MT (Tiedemann and Thottingal, 2020). We implement two training methods:

Multiple Models (MM) For each compression ratio $c \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, we train a separate model. These models generate compressed translations by conditioning the target sequence (T) length to be lower than the length of the source sentence (S): $T_{\text{length}} \leq \lfloor c \cdot S_{\text{length}} \rfloor$.

Controllable Model (CM) A single model is trained across all compression ratios, with a compression token t_c prepended to the input (source) sequence S. CM uses the same constraints as MM, but is trained with data for all values of c. Thus, it allows for adjustment of compression ratio during inference: $S = [t_c; S]$, where t_c is set to the string "@c" (e.g. @0.5).

All models are based on multilingual models for specific language families where available, denoted by the `en-*family*` format in Table 4. In cases where no models exist for narrower branches of a language family, which is the case for Albanian and Basque, we opt for monolingual models instead of larger language families. For multilingual models, we append language-specific prefixes to source sequences (e.g., `»fra«` for

Language	Length ratio	Samples (K) per compression ratio (c)					
	α	0.5	0.6	0.7	0.8	0.9	1.0
Hungarian	1.01	1,560	3,175	5,938	10,178	15,290	21,523
French	1.11	1,807	3,548	6,301	10,142	14,265	20,503
Polish	1.02	2,106	3,811	6,276	9,399	13,174	17,491
Norwegian	0.96	416	750	1,224	1,813	2,514	3,261
Albanian	1.05	**103**	**196**	331	528	769	1,076
Lithuanian	0.99	**69**	**134**	**241**	373	533	665
Malay	1.12	**37**	**74**	**129**	**211**	364	460
Basque	1.09	**8**	**16**	**28**	**49**	**73**	**106**

Table 3: Number of parallel sentences for each compression ratio (length $\leq c$). The *length ratio* is the average character length per sentence of the target language relative to the English source ($\geq 1 \Rightarrow$ longer sentences). The length ratio equals the α parameter described in Section 3.1. Sample sizes below 250,000 are highlighted.

Language	ISO 639	Language Family	Model Name
Albanian	sq (sqi)	Albanoid	en-sq
Basque	eu (eus)	Isolated	en-eus
French	fr (fra)	Romance	en-roa
Hungarian	hu (hun)	Uralic	en-urj
Lithuanian	lt (lit)	Baltic	en-bat
Malay	ms (msa)	Austronesian	en-map
Norwegian	no (nob)	North Germanic	en-gmq
Polish	pl (pol)	West Slavic	en-zlw

Table 4: Included languages, language codes, families, and Opus-MT model names.

French). Furthermore, the models have fewer than 75M parameters, trainable with less than 2GB of VRAM in full precision.[2] We use the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 5×10^{-6} and a linear decay with a warmup ratio of 10%, and a batch size of 16. Based on the findings by Svensson and Troksch (2022), where the training beyond the initial epoch showed marginal improvements, we train for a single epoch.

3.3 Evaluation

Evaluation is a challenging aspect of any natural language generation problem. We include well-established metrics for machine translation: BLEU, ROUGE METEOR, ChrF, and BERTScore. BLEU (Papineni et al., 2002) calculates n-gram overlap with a brevity penalty for a range of n (the default $n = 4$ is used) between the prediction and reference texts. ROUGE-1 and -2 (Lin, 2004) are commonly used for summarization tasks, calculating the uni- and bi-gram overlaps between texts. METEOR (Banerjee and Lavie, 2005) uses the harmonized score of word matches, including their stemmed forms and meanings, with added penalties for differences in word order. ChrF (Popović, 2015) is a character-level n-gram matching metric, and finally, BERTScore (Zhang et al., 2020b) uses the contextualized embeddings to calculate cosine similarities at a token level. OpenSubtitles evaluations for Albanian, Basque, French, Hungarian, Lithuanian, Malay, Norwegian, and Polish are made on a compression ratio $c = 0.5$ to study how the models perform on highly compressed target sequences. The same models are evaluated on EuroParl for French, Hungarian, Lithuanian, and Polish. As illustrated in Figure 4, the length distributions differ from the compressed OpenSubtitles. The CLSC system is easily expanded to include new languages supported by the evaluation data.

4 Experiments and Results

Before training on multiple languages, we studied the effects of training on smaller data samples to set a threshold sample size for low-resource languages. We train our CM model on English to Norwegian text using 250k, 100k, 25k, and 5k downsampled datasets. Results are evaluated entirely by observations on the BBT data. Table 5 shows the results (translated to English for readability). Results are calculated from five bootstrap resamples of 1000 sentences. Due to space limitations, we show the results from $c \in \{0.5, 0.7, 1.0\}$.[3] For the in-domain evaluation, the reported standard deviation is near zero for all metrics, and we thus only show the mean for simplicity. Mean and std are reported for the out-of-domain evaluation. English equivalents are provided, translated with Claude-3 and GPT-4, and may contain errors. We conclude that

[2] Tested with Accelerate (Gugger et al., 2022). In a minimal setting, the models require peak VRAM of 1.1GB with a batch size of 1.

[3] Full results are available in the CLSC repository `https://github.com/tollefj/CLSC`

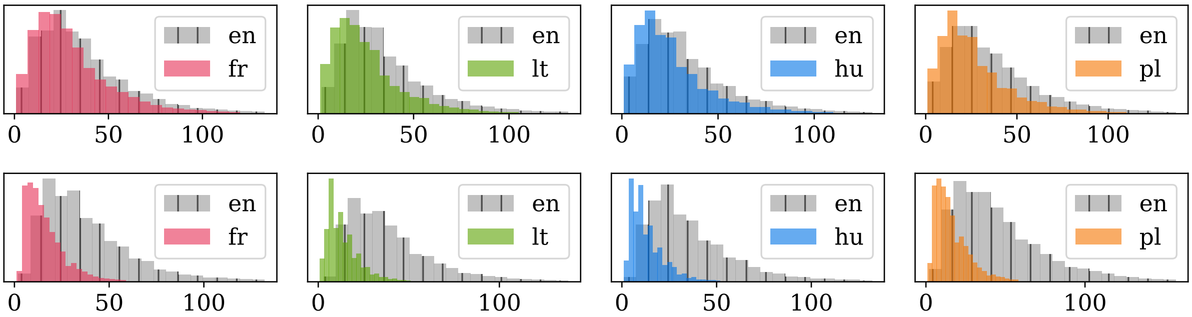

Figure 3: OpenSubtitles sentence length distributions (character count) of English to French, Lithuanian, Hungarian, and Polish. Compression levels: $c = 1.0$ (top row) and $c = 0.5$ (bottom row).

Samples	$c = 0.5$	$c = 0.7$	$c = 1.0$
5000	I want to show Howard that I can play this game.	I want to show Howard that I can play this game.	I want to show Howard that I can play this game
25000	I can play this game.	I want to show Howard that I can play.	I soon want to show Howard that I can play this game.
100000	I can play.	I want to show Howard this game.	I want to show Howard right away that I can play.
250000	I want to show Howard.	I want to show Howard this game.	I want to show Howard that I can play.

Table 5: Results for various sample sizes of compression levels 0.5, 0.7, and 1.0. Source: “In a minute I wanna show Howard I can play this game” with translations backtranslated from Norwegian for readability. Subtitle (as on TV): “I want to show that I can do it”. Note that while some outputs are purely extractive, the model is capable of abstractive compressions .

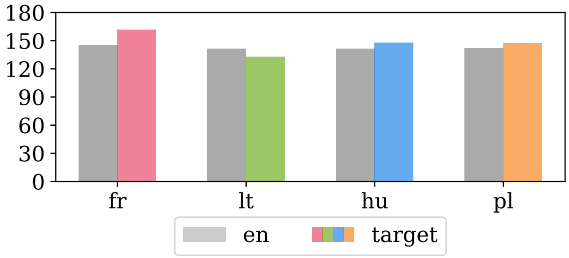

Figure 4: EuroParl mean sentence lengths (character count) of French, Lithuanian, Hungarian, and Polish.

250k samples provided a well-balanced threshold of high-quality language and better extractions for compression.

4.1 OpenSubtitles

The hold-out $c = 0.5$ test set from the compressed OpenSubtitles dataset, as described in the preprocessing steps of CLSC in Section 3.1, is evaluated to determine whether the two model types, CM and MM, can adapt to heavily compressed conditions. Results for all languages are shown in Table 6. OpenSubtitles presents challenges due to sporadic misalignments in parallel sentences, likely contributing to the lower scores. Despite this, both CLSC models outperform the baselines. Moreover, the more flexible CM model consistently surpasses MM across most c-values while maintaining better control of generation length.

4.2 EuroParl

Evaluation of EuroParl data, with its closely aligned length ratios between English and target languages (as seen in Figure 4), enables studying results under different conditions, where the target is *not* compressed. In other words, the aim is for CLSC models near $c = 1.0$ to perform on par with the baseline. Only Hungarian, French, Polish, and Lithuanian are available in the EuroParl corpus from the earlier selected languages. CLSC results are in Table 8. The baseline models, which do not apply compression, consistently achieve the highest scores across all metrics but with significantly longer output lengths. Compressed models, e.g., at $c = 0.5$, maintain reasonable scores in several metrics, especially BERTScore. Thus, despite reductions in sentence length, they preserve the se-

Lang/Model	BLEU	R-1	R-2	ChrF	MET	BERT	α
Basque							
baseline	2.77	0.19	0.03	24.35	0.24	0.72	3.18
0.5_{MM}	4.12	0.26	0.05	23.64	0.29	**0.76**	**2.06**
0.5_{CM}	**6.44**	**0.27**	**0.06**	26.79	**0.30**	0.75	2.46
0.7_{MM}	5.14	**0.27**	0.05	26.63	**0.30**	**0.76**	2.44
0.7_{CM}	5.42	0.26	0.05	**26.97**	**0.30**	0.75	2.60
1.0_{MM}	3.08	0.25	0.05	25.88	0.29	0.74	3.03
1.0_{CM}	3.80	0.24	0.04	26.31	0.28	0.74	2.90
French							
baseline	3.97	0.21	0.08	25.16	0.29	0.71	3.28
0.5_{MM}	9.05	0.31	**0.12**	28.49	0.34	**0.77**	1.67
0.5_{CM}	**9.39**	**0.32**	**0.12**	**28.71**	**0.35**	**0.77**	**1.65**
0.7_{MM}	6.30	0.28	0.11	27.51	0.32	0.74	2.25
0.7_{CM}	6.66	0.29	0.11	27.91	0.33	0.75	2.17
1.0_{MM}	4.65	0.25	0.09	26.14	0.31	0.73	2.83
1.0_{CM}	4.96	0.25	0.10	26.89	0.32	0.73	2.72
Hungarian							
baseline	3.07	0.22	0.09	23.42	0.26	0.69	3.51
0.5_{MM}	9.32	**0.30**	**0.12**	25.79	0.31	**0.76**	1.65
0.5_{CM}	**10.07**	**0.30**	0.13	**25.97**	**0.32**	**0.76**	**1.60**
0.7_{MM}	5.48	0.25	0.11	24.11	0.28	0.74	2.29
0.7_{CM}	6.19	0.27	0.11	24.93	0.29	0.75	2.15
1.0_{MM}	4.08	0.24	0.10	23.72	0.27	0.72	2.88
1.0_{CM}	4.52	0.24	0.10	24.11	0.28	0.73	2.73
Lithuanian							
baseline	2.42	0.17	0.05	20.83	0.25	0.69	3.76
0.5_{MM}	5.73	0.23	0.07	21.94	0.28	0.75	1.89
0.5_{CM}	**7.59**	**0.26**	**0.08**	**24.33**	**0.31**	**0.77**	**1.76**
0.7_{MM}	5.19	0.23	0.07	22.96	0.28	0.74	2.39
0.7_{CM}	5.33	0.24	**0.08**	23.71	0.29	0.75	2.29
1.0_{MM}	3.63	0.21	0.06	22.12	0.27	0.72	2.88
1.0_{CM}	3.91	0.22	0.06	22.56	0.28	0.73	2.77
Malay							
baseline	1.18	0.14	0.02	17.10	0.22	0.71	3.69
0.5_{MM}	**5.60**	**0.27**	**0.06**	25.13	**0.32**	**0.77**	**2.04**
0.5_{CM}	5.29	**0.27**	**0.06**	26.29	**0.32**	**0.77**	2.21
0.7_{MM}	4.92	0.26	**0.06**	26.31	0.31	0.76	2.41
0.7_{CM}	4.89	0.26	**0.06**	**26.40**	0.31	0.76	2.48
1.0_{MM}	3.38	0.23	0.04	24.83	0.30	0.74	3.09
1.0_{CM}	3.98	0.25	0.05	25.85	0.30	0.75	2.87
Norwegian							
baseline	3.71	0.21	0.07	23.82	0.27	0.71	3.46
0.5_{MM}	6.45	0.26	0.09	25.96	0.30	0.75	2.46
0.5_{CM}	**10.47**	**0.31**	**0.11**	**27.86**	**0.34**	**0.78**	**1.76**
0.7_{MM}	5.18	0.24	0.08	25.55	0.29	0.73	2.89
0.7_{CM}	7.36	0.28	0.10	26.95	0.31	0.76	2.33
1.0_{MM}	4.70	0.23	0.08	25.14	0.29	0.73	3.11
1.0_{CM}	5.65	0.25	0.09	26.11	0.30	0.74	2.86
Polish							
baseline	3.59	0.19	0.07	22.01	0.25	0.71	3.06
0.5_{MM}	**7.35**	0.24	**0.10**	23.16	0.28	**0.75**	**1.88**
0.5_{CM}	7.25	**0.25**	**0.10**	**23.38**	**0.29**	**0.75**	**1.88**
0.7_{MM}	4.87	0.22	0.09	22.76	0.27	0.73	2.41
0.7_{CM}	5.09	0.23	0.09	22.75	0.27	0.74	2.35
1.0_{MM}	4.07	0.20	0.08	22.61	0.26	0.72	2.80
1.0_{CM}	4.33	0.21	0.08	22.70	0.26	0.73	2.69
Albanian							
baseline	3.39	0.21	0.07	21.81	0.27	0.71	3.50
0.5_{MM}	6.85	0.30	0.10	24.48	**0.33**	**0.77**	**2.11**
0.5_{CM}	**6.96**	**0.31**	**0.11**	**25.35**	**0.33**	**0.77**	2.20
0.7_{MM}	6.01	0.29	**0.11**	24.90	0.32	0.75	2.51
0.7_{CM}	5.62	0.29	0.10	24.39	0.32	0.75	2.58
1.0_{MM}	4.37	0.25	0.09	23.25	0.29	0.73	2.98
1.0_{CM}	4.70	0.26	0.09	23.89	0.30	0.74	2.87

Table 6: CLSC results for OpenSubtitles testing data. Subscripts indicate model types. The highest scores for each language/model setup are highlighted. Observe the length ratio α for the varying compression ratios c compared to the baseline.

	English–French (OpenSubtitles)
source	Now, look here. I never was one to spoil a good time... but enough's enough. That's what I say...
target	**Je suis pas rabat-joie**... mais ça suffit ! (I'm not a **killjoy**... but that's enough!)
baseline	Je n'ai jamais été l'un pour gâcher un bon moment... mais c'est assez, c'est ce que je dis... (I've never been one to spoil a good time... but that's enough, that's what I say...)
0.5_{CM}	Je n'ai jamais gâché un bon moment, mais ça suffit. (I've never spoiled a good time, but that's enough.)
	English–Lithuanian (OpenSubtitles)
source	Like why do I have to be in camouflage? So the big bad quail doesn't see me?
target	**Ateinu**! **(I'm coming!)**
baseline	Kodėl aš turiu būti kamuflaže? (Why do I have to be in camouflage?)
0.5_{CM}	Tai didysis putpelis manęs nemato? (So the big quail doesn't see me?)
	English–French (EuroParl)
source	We have to get information.
target	Nous sommes tenus de recueillir des informations. (We are required to gather information.)
baseline	Nous devons obtenir de l'information. (We must obtain information.)
0.5_{CM}	On doit en savoir. (We need to know about it.)
	English–Lithuanian (EuroParl)
source	However, naturally, other measures are required now.
target	Tačiau suprantama, kad šiuo metu reikalingos kitos priemonės. (However, it is understood that other measures are needed now.)
baseline	Tačiau savaime suprantama, kad dabar reikia imtis kitų priemonių. (However, it goes without saying that other measures must now be taken.)
0.5_{CM}	Tačiau dabar reikia imtis kitų priemonių. (However, other measures must now be taken.)

Table 7: Excerpts from OpenSubtitles and EuroParl for French and Lithuanian with baseline and CLSC CM translations with $c = 0.5$. Several samples in the OpenSubtitles datasets have misalignment problems. See the highlighted English–Lithuanian example – "Ateinu!". The French example shows a requirement to infer phrases such as "killjoy" based on the longer description of "one to spoil a good time".

mantic information verified upon inspection (see Table 7, with full results in the provided code).

5 Discussion and Future Work

Subtitle quality relies on numerous factors such as on-screen context, speaker personality, and language (Cintas and Anderman, 2008). However, our results demonstrate that smaller models, trained on minimal data, can deliver acceptable translations and compressions across languages in a mapping between source and target. From examples in Table 7, we highlight text that is difficult for an automated system of this kind, where the term "killjoy" is being used to describe one that "does not want to spoil a good time" in the shorter sentence. In cases where the target language does not have an appropriate equivalent, these kinds of inferences become difficult. Moreover, the Lithuanian target "Ateinu!" is entirely misaligned with its source text "Like why do I have to be in camouflage? So the big bad quail doesn't see me?". See more examples in Table 9 from the validation set of OpenSubtitles data with $c = 0.5$.

Original English	**Norwegian**
I should never have stayed in a client's house.	Det blir alltid... (It always becomes...)
Our opportunity? How's that?	Hva? (What?)
That is totally unexceptable!	Kom ut! (Come out!)
Oh, isn't that something?	Ja, visst. (Yes, indeed.)
Keep a safe distance from him! Didn't I tell you that?	Forsvinn, Jack. (Go away, Jack.)
You don't have to do that, kid. - Right.	Ja. (Yes.)

Table 9: Examples of misalignment in English–Norwegian subtitle pairs (including back-translations) in the OpenSubtitles validation set with $c = 0.5$.

Future experiments should consider post-alignment, such as applying sentence similarity or relatedness. However, the immediate problem is that shorter sentences, by definition, omit parts of the source and will typically receive lower similarity scores, thus resulting in considerably reduced data sizes. In the EuroParl evaluation (Table 8), French maintained high quality at low compression levels, while Hungarian and Polish showed degradation. Polish's baseline model generated longer texts ($\alpha \approx 1.5$), but the $c = 0.5$-models (both MM and CM) halved this length effectively. Lithuanian proved more challenging to compress. These observations are interesting for future implementations to study specific languages in-depth.

We argue that BERTScore best aligns with subtitle quality and suggest that future work should evaluate metrics as a function of generation length to promote fair comparisons in compression and paraphrasing tasks, as we can observe for ROUGE-scores in the work by Schumann et al. (2020).

Concluding Remarks. This study explores sentence translation and compression with minimal resources, finding that controllable models are adequate for adjustable compression, avoiding multiple trained models. Though we observe high quality upon manual inspection, current metrics cannot fully capture this. Future work will involve more extensive experiments, going in-depth into specific languages to study the potential for improved metrics better aligned with human judgment. Code and all evaluation results are available at `https://github.com/tollefj/CLSC`.

Limitations

CLSC does not directly support low-resource languages outside the OpenSubtitles corpus, requiring minor modifications for new data sources. The OpenSubtitles dataset contains several cases of misalignment between the target and source language, making it challenging to map source sentences directly to shorter target sentences without additional context. Finally, the metrics BLEU, ROUGE, ChrF, and METEOR fail to correctly evaluate compressed and translated outputs where interchanging words and phrases may be desired. These metrics focus on surface-level n-gram or token similarity, not capturing how well the outputs preserve meaning, as metrics like BERTScore aim to resolve.

Lang/Model	BLEU	ROUGE-1	ROUGE-2	ChrF	METEOR	BERTScore	Length ratio α
French							
baseline	39.75 (0.45)	**0.67** (0.00)	**0.50** (0.00)	**66.67** (0.38)	**0.64** (0.00)	**0.88** (0.00)	1.20 (0.06)
0.5_{MM}	30.40 (0.67)	0.59 (0.01)	0.42 (0.01)	54.11 (0.53)	0.49 (0.00)	0.85 (0.00)	**0.76** (0.00)
0.5_{CM}	33.28 (0.51)	0.60 (0.00)	0.43 (0.00)	57.65 (0.29)	0.52 (0.00)	0.86 (0.00)	0.80 (0.01)
0.7_{MM}	37.76 (0.52)	0.65 (0.00)	0.47 (0.00)	62.32 (0.41)	0.59 (0.01)	**0.88** (0.00)	0.93 (0.00)
0.7_{CM}	37.91 (0.51)	0.65 (0.01)	0.47 (0.00)	62.55 (0.33)	0.59 (0.00)	0.87 (0.00)	0.93 (0.00)
1.0_{MM}	**39.82** (0.67)	**0.67** (0.00)	0.49 (0.00)	65.58 (0.57)	0.63 (0.00)	**0.88** (0.00)	1.04 (0.00)
1.0_{CM}	39.75 (0.63)	**0.67** (0.00)	0.49 (0.00)	64.55 (0.48)	0.62 (0.00)	**0.88** (0.00)	1.01 (0.00)
Hungarian							
baseline	**34.12** (0.98)	**0.69** (0.01)	**0.50** (0.00)	**64.60** (0.54)	**0.61** (0.01)	**0.88** (0.00)	1.05 (0.01)
0.5_{MM}	20.79 (0.65)	0.55 (0.00)	0.39 (0.00)	46.19 (0.55)	0.42 (0.01)	0.83 (0.00)	0.64 (0.01)
0.5_{CM}	18.66 (0.58)	0.52 (0.00)	0.36 (0.00)	43.95 (0.32)	0.39 (0.01)	0.82 (0.00)	**0.60** (0.01)
0.7_{MM}	30.02 (0.73)	0.65 (0.00)	0.47 (0.01)	58.83 (0.46)	0.55 (0.01)	0.87 (0.00)	0.87 (0.00)
0.7_{CM}	28.25 (0.77)	0.64 (0.00)	0.45 (0.00)	57.15 (0.43)	0.53 (0.01)	0.86 (0.00)	0.84 (0.01)
1.0_{MM}	33.49 (0.87)	0.68 (0.01)	0.49 (0.01)	63.12 (0.49)	0.60 (0.01)	**0.88** (0.00)	0.99 (0.00)
1.0_{CM}	32.42 (0.92)	0.68 (0.00)	0.49 (0.00)	62.04 (0.45)	0.59 (0.01)	**0.88** (0.00)	0.96 (0.00)
Lithuanian							
baseline	**33.67** (0.91)	**0.62** (0.01)	**0.42** (0.01)	**65.13** (0.56)	**0.60** (0.01)	**0.87** (0.00)	1.06 (0.02)
0.5_{MM}	30.78 (0.81)	0.58 (0.01)	0.38 (0.01)	59.56 (0.61)	0.53 (0.01)	0.86 (0.00)	0.90 (0.01)
0.5_{CM}	26.54 (0.87)	0.53 (0.01)	0.33 (0.01)	55.62 (0.61)	0.47 (0.01)	0.84 (0.00)	**0.81** (0.01)
0.7_{MM}	30.01 (0.94)	0.58 (0.01)	0.38 (0.01)	59.32 (0.70)	0.54 (0.01)	0.86 (0.00)	0.90 (0.01)
0.7_{CM}	29.83 (0.87)	0.58 (0.01)	0.37 (0.01)	59.82 (0.59)	0.54 (0.01)	0.86 (0.00)	0.93 (0.01)
1.0_{MM}	31.90 (0.55)	0.60 (0.01)	0.39 (0.01)	62.43 (0.45)	0.57 (0.01)	**0.87** (0.00)	1.00 (0.02)
1.0_{CM}	31.26 (0.82)	0.60 (0.01)	0.39 (0.01)	61.61 (0.60)	0.56 (0.01)	0.86 (0.00)	0.99 (0.01)
Polish							
baseline	12.28 (0.66)	0.44 (0.00)	0.24 (0.00)	48.84 (0.71)	0.44 (0.00)	0.80 (0.00)	1.51 (0.09)
0.5_{MM}	18.46 (0.58)	0.48 (0.00)	0.27 (0.00)	44.59 (0.41)	0.41 (0.00)	0.83 (0.00)	**0.75** (0.00)
0.5_{CM}	21.37 (0.52)	0.50 (0.00)	0.29 (0.01)	48.79 (0.55)	0.43 (0.00)	0.83 (0.00)	0.77 (0.01)
0.7_{MM}	22.31 (0.52)	0.53 (0.00)	0.31 (0.01)	51.11 (0.40)	0.47 (0.00)	0.85 (0.00)	0.88 (0.00)
0.7_{CM}	23.93 (0.58)	0.54 (0.00)	0.33 (0.00)	52.52 (0.42)	0.49 (0.00)	0.85 (0.00)	0.88 (0.00)
1.0_{MM}	**26.17** (0.46)	**0.57** (0.00)	**0.34** (0.00)	**55.80** (0.33)	**0.52** (0.00)	**0.86** (0.00)	0.97 (0.00)
1.0_{CM}	26.05 (0.38)	0.56 (0.00)	**0.34** (0.00)	55.11 (0.35)	0.52 (0.01)	**0.86** (0.00)	0.95 (0.00)

Table 8: CLSC results on EuroParl. Subscripts indicate model types. Reported mean and standard deviation (in parentheses) over five bootstrap resamples of 1000 sentences. We remind the reader that the sampled sentences are not filtered by length, and thus, we expect the baseline models to score higher. Observe the *Length* specifically; compared to metric evaluations, the lower compression models can uphold high scores compared to the output size.

References

Wilker Aziz, Sheila C. M. de Sousa, and Lucia Specia. 2012. Cross-lingual sentence compression for subtitles. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 103–110, Trento, Italy. European Association for Machine Translation.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Nadjet Bouayad-Agha, Angel Gil, Oriol Valentin, and Victor Pascual. 2006. A sentence compression module for machine-assisted subtitling. In *Computational Linguistics and Intelligent Text Processing: 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006. Proceedings 7*, pages 490–501. Springer.

Sheryl Burgstahler. 2009. Universal design of instruction (udi): Definition, principles, guidelines, and examples. *Do-It*.

L Bywood, T Etchegoyhen, Panayota Georgakopoulou, M Fishel, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, A Turner, M Volk, and M Maucec. 2014. Machine translation for subtitling: A large-scale evaluation. In *LREC 2014, Ninth International Conference on Language Resources and Evaluation*, pages 46–53.

Jorge Díaz Cintas. 2013. Subtitling: Theory, practice and research. In *The Routledge handbook of translation studies*, pages 273–287. Routledge.

Jorge Díaz Cintas and Gunilla Anderman. 2008. *Audiovisual translation: Language transfer on screen*. Springer.

Jorge Díaz Cintas and Aline Remael. 2020. *Subtitling: Concepts and practices*, 1st edition. Routledge.

Simon Corston-Oliver. 2001. Text compaction for display on very small screens. In *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 89–98. Association for Computational Linguistics.

Walter Daelemans, Anja Höthker, and Erik F Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *LREC 2004, Fourth International Conference on Language Resources and Evaluation*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Bi-simcut: A simple strategy for boosting neural machine translation. *arXiv preprint arXiv:2206.02368*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`.

Prabhakar Gupta, Mayank Sharma, Kartik Pitale, and Keshav Kumar. 2019. Problems with automating translation of movie/tv show subtitles. *arXiv preprint arXiv:1909.05362*.

Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. 2021. BERTTune: Fine-tuning neural machine translation with BERTScore. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 915–924, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. Must-cinema: a speech-to-subtitles corpus. *Preprint*, arXiv:2002.10829.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion*

Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erwin Marsi, Emiel Krahmer, Iris Hendrickx, and Walter Daelemans. 2009. On the limits of sentence compression by deletion. In *Conference of the European Association for Computational Linguistics*, pages 45–66. Springer.

Maite Melero, Antoni Oliver, and Toni Badia. 2006. Automatic multilingual subtitling in the etitle project. *Proceedings of Translating and the Computer*, 28:1–18.

Jan Niehues. 2020. Machine translation with unsupervised length-constraints. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 21–35, Virtual. Association for Machine Translation in the Americas.

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.

Yo-Han Park, Gyong-Ho Lee, Yong-Seok Choi, and Kong-Joo Lee. 2021. Sentence compression using bert and graph convolutional networks. *Applied Sciences*, 11(21):9910.

Andrej Perković, Jernej Vičič, Dávid Javorskỳ, and Ondřej Bojar. 2023. Shortening of the results of machine translation using paraphrasing dataset.

Volha Petukhova, Rodrigo Agerri, Mark Fishel, Sergio Penkale, Arantza Del Pozo, Mirjam Sepesy Maucec, Andy Way, Panayota Georgakopoulou, and Martin Volk. 2012. Sumat: Data collection and parallel corpus compilation for machine translation of subtitles. In *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, pages 21–28.

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.

Amanda Langeland Sandvold. 2019. Audiovisuell oversettelse: en komparativ tekstanalyse av engelsk tale og norsk teksting i the big bang theory. Master's thesis, University of Stavanger, Norway.

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:351–424.

Raphael Schumann, Lili Mou, Yao Lu, Olga Vechtomova, and Katja Markert. 2020. Discrete optimization for unsupervised sentence summarization with word-level extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5032–5042, Online. Association for Computational Linguistics.

Språkrådet. 2017. Retningslinjer for god teksting.

Joakim Svensson and Victor Troksch. 2022. Generating subtitles with controllable length using natural language processing.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, volume 2012, pages 2214–2218. Citeseer.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Vincent Vandeghinste and Yi Pan. 2004. Sentence compression for automated subtitling: A hybrid approach. In *Text Summarization Branches Out*, pages 89–95, Barcelona, Spain. Association for Computational Linguistics.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine translation of tv subtitles for large scale production. In *JEC 2010; November 4th, 2010; Denver, CO, USA*, pages 53–62. Association for Machine Translation in the Americas.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Guangxiang Zhao, Xu Sun, Jingjing Xu, Zhiyuan Zhang, and Liangchen Luo. 2019. Muse: Parallel multi-scale attention for sequence to sequence learning. *arXiv preprint arXiv:1911.09483*.