Toward Genre Adapted Closed Captioning

François Buet, François Yvon

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Toward Genre Adapted Closed Captioning

François Buet, François Yvon

Université Paris-Saclay, CNRS, LISN, France
{francois.buet,francois.yvon}@limsi.fr

Abstract

This paper studies the generation of intralingual closed captions from automatic speech transcripts, with the aim to assess techniques for multi-genre captioning. Captions and subtitles greatly vary in form and content depending on the programs genres and subtitling styles, resulting for instance in significantly different compression rates and lexical content. Borrowing ideas from the multi-domain machine translation literature, we implement and contrast several adaptation methods on a diverse set of programs broadcast on the French public TV. Our results show that such multi-domain adaption techniques are effective and help to improve our automatic subtitling system.
Index Terms: Speech Transcription, Automatic Captioning, Text Simplification

1. Introduction

Accessibility of video contents requires monolingual closed captioning for the deaf and hard-of-hearing audience, which has become a legal obligation for major TV channels in France (as of 2011),[1] as in many other countries. Closed captioning can also be useful for online talks and educational video contents and may also facilitate the comprehension of speech by language learners. This general context has stimulated work aimed at automating the generation of subtitles [1, 2, 3], a task that is nowadays performed with large neural architectures trained in a end-to-end fashion [4, 5]. Complete automation remains difficult since closed captioning is subject to multiple prescriptions related to the position on screen, the length of the text, its size, color, and display duration, synchronization with speech, etc. This means that TV captions significantly depart from automatic transcripts. A fully automated solution would thus not only necessitate automatic speech recognition (ASR), but also other processing modules such as text simplification, speaker diarization and sound event detection.

We focus here solely on the text generation component for same-language or *intralingual captioning*,[2] and study ways to better control and adapt captioning to the TV genres. For some programs, closed captioning is performed in an online fashion; while for pre-recorded contents, captioning is prepared offline, which can induce different compression strategies. Using a baseline architecture made of three main steps: ASR, caption generation, and caption segmentation, we consider several ways to make the second step dependent upon the program genre and style. Using a Transformer model [6], our main contribution is to experimentally assess the benefits of using of control tags that are interleaved within the text and the use of adapted compression ratios.

Our experiments with a diverse set of programs show that tag-based genre adaptation strategies are effective in delivering improved captions, and that further improvements can be obtained through self-learning.

[1] It is now required for all public TV channels.

[2] Even though we borrow much from the Machine Translation domain, interlingual captioning is out of the scope of this study.

Table 1: *Average Speech rate (SpR), Compression rate (CpR), BLEU score and Flesch Reading Ease (FRE) for various TV genres. Magazines contain a mixture of* `online` *and* `stock`*, while the other genres only contain* `stock` *[s] or* `online` *[o] types. SpR is the number of spoken words per hour; CpR is the ratio between the number of words in the transcripts and in the captions; BLEU compares the transcripts and the caption texts (higher means more similar); FRE is a measure of textual complexity aggregating sentence and word lengths in the captions (higher means simpler).*

Genre	SpR	CpR	BLEU	FRE
Cartoon [s]	6325	0.95	38.3	79.8
Documentary [s]	7770	0.86	52.7	79.1
Education [s]	10164	0.83	51.6	80.4
Quizzes [s]	8218	0.69	28.9	80.1
Series [s]	7198	0.86	33.3	81.5
Magazine [s/o]	11566	0.70	36.1	76.6
News [o]	10496	0.86	59.0	67.0
Politics [o]	12157	0.69	39.5	71.4

2. Motivations

2.1. Genres in TV shows

The automation of closed captioning is an early application of automatic speech transcription and machine translation technologies [1, 2, 3] and progress in this area has been steady, owing to the improvement of the underlying technologies. While initially developed as sophisticated pipelines, neural sequence-to-sequence models have opened the prospect of integrated, end-to-end training for these systems [4, 5]. Recent efforts have mainly focused on subtitling internet content, such as talks and classes. We explore here a much broader range of genres, considering the full spectrum of TV programs, from news bulletins to fictions, documentary and games.

Owing to production constraints, professional captions are either produced online or prepared offline. For subtitling live content such as News or talk shows, with real-time constraints, it is custom to use respeaking, speech-to-text technologies and manual post-editing. For prepared content, the redaction of subtitles is less constrained, yielding a different captioning style. Within each main style, other important differences subsist between genres (eg. games and magazines) and domains (eg. food vs. health-oriented TV magazines).

These differences are reflected in Table 1 where we report basic statistics regarding our training material, a representative set of programs from the French public TV (more details in Section 3.1). A first difference exist between `online` and `stock`

http://dx.doi.org/10.21437/Interspeech.2021-1762

(prepared) captioning styles: the former are more verbose, with about 11.3K words/hour, while the average speech rate for latter is only about 9.6K. `stock` captions however yield a larger number of sentences, that are also shorter (the average length is 7.7 words vs. 12.7 words for `online` sentences), reflecting again the well-prepared nature of these texts.

Table 1 also shows that differences in genres yield significant differences in speech and compression rates. Variations in the BLEU score reflect both the fact that closed captions for News are much closer to the transcripts than for Quizzes or Series and that the former are better recognized by the ASR.[3] Unsurprinsingly, the complexity metrics FRE clearly distinguishes Quizzes and Series from Politics and Educational contents.

In this paper, we study ways to adapt caption generation models and take these differences into account. Since the problem is analogous to building sequence-to-sequence models robust to variability in domains or styles, we propose to borrow techniques from these related tasks.

2.2. Toward multi-genre captioning

Such issues have been notably addressed for Machine Translation (MT) applications, where a significant body of recent work studies explore ways to simultaneously integrate process texts from multiple domains [7, 8, 9] or languages (eg. [10, 11, 10, 12]). As we view simplification (and also segmentation) as a form of monolingual translation process [13], aimed to convert the (noisy) verbatim speech transcript into a simplified (and segmented) caption text, we are in a position to reuse techniques originally proposed for multilingual or multi-domain MT. In a nutshell, these techniques rely on three main ingredients: (a) the use of domain / language tags that are inserted on the source side to generate adapted representations; (b) the use of adversarial techniques aimed to neutralize differences between genres; (c) specializations of subparts of the neural network whose parameters are adapted to one domain. In our experiments below, we use the former method, which is much simpler to implement and also deliver strong results across the board.

3. Methods

3.1. Training and test data

Our main material comprises of a variety TV shows of all genres and styles (news bulletin, documentaries, games, entertainment, magazines, fiction etc.) that was played on French public TV between 2018 and 2020, accompanied with their closed captions. Alltogether, they represent 1.6K hours of videos, which have been automatically transcribed by in house ASR system, then aligned with the reference subtitle. Alignments are performed at the level of speech segments recognized by the ASR, which often correspond to several captions - more than four on average in our data. This reflects best the test condition where caption content and time alignment have to be automatically derived from the transcripts. Basic statistics for this corpus are in Table 2. This "parallel" corpus is the main training data for the neural captioning systems described below; by analogy to machine translation, we refer to the transcripts as the *source side*, and to the captions as the *target side* of this corpus.

A diverse subset of $\approx$ 10h has also been randomly selected with no overlap with the training data and is used as our main test set. This test set has also been transcribed manually so as to evaluate the impact of transcription errors on the final output. We have finally used an additional set of 1.3K hours of subtitles to generate a *pseudo-parallel corpus* pairing actual captions with pseudo transcripts. These were automatically computed in a process mimicking *back-translation* in MT [14]. To this end, we have trained a system "translating" written captions into fake transcripts, that where then resegmented by merging consecutive captions, thus simulating the regular training data. Apart from its input and outputs, this system is identical to the baseline Transformer detailed below.

[3] For our test set, we observed WERs ranging between less than 10 and more than 40, depending on the program type.

Table 2: *Statistics for the train and test corpora*

Size	**Training**		**Test**
	aligned	pseudo	
Hours (h)	1,620	1,276	9.6
Captions	1,625,105	1,186,250	8,840
Segments	410,545	286,080	2,189
Words (speech)	17,043,840	–	103,487
Words (caption)	12,199,060	8,843,879	68,195

3.2. Baseline captioning systems

Our baseline systems rely on our own reimplementation of the encoder/decoder Transformer architecture [6] with the following parameters: all representations have dimension $d_{\text{model}} = 256$, with the feedforward sublayers having dimension $d_{\text{ff}} = 1024$; encoder and decoder contain 6 layers each, with $h = 8$ heads in each layer. Optimization is performed with Adam [15] using ($\beta_1 = 0,9, \beta_2 = 0,98, \epsilon = 10^{-9}$). Following a warm-up stage of 4 000 steps, we then trained all models until validation loss did not increase for 5 epochs. During preprocessing, the punctuated and capitalized ASR output is slightly adjusted by removing obvious dysfluencies and filled pauses; both the transcripts and associated reference captions are then tokenized and further decomposed using a subword vocabulary of 16K units computed with the *Sentencepiece* Toolkit [16].

In our baseline, a further post-processing step takes care of segmenting the compressed text into a valid closed caption, based on a handful of rules implementing the following constraints: a caption is made of one or two lines; each line contains at most 36 characters,[4] the duration of each caption is adapted to ensure it stays on screen for a sufficient period of time.

3.3. Adapted Transformers

We explore two main strategies to better control and adapt the level of compression and the segmentation of the automatic captioning system. The first one relies on length-controlled versions of the encoder/decoder architecture, where an additional length constrainst is input to the decoder. Such extensions were initially proposed for RNN architectures [17], then transposed to the Transformer model in [18] and [4], which is our main source of inspiration. It mainly consists in manipulating the positional encoding of the Transformer model by inputting either information regarding the distance (in words) till the end of the line, or a compression rate. More formally, recall that the baseline Transformer encodes absolute token positions pos into a vector of d_{model} dimensions, where component i is:

[4] Based on the recommandation of the French regulation authorities.

Automatic transcript and tag:
<stock><mag>1er invité écrire est-ce trahir, hé bien c'est la question que se pose également, Jean- Luc Coatalem lorsqu'il cherche à briser le silence qui entoure la mort de son grand- père, un grand-père qui n'a pas connu, arrêté en 1943 puis déporté en Allemagne et dont on a toujours refusé de parler dans une famille qui considère comme une trahison toute tentative d'explication Jean- Luc. 1rst invitee is writing a treason, well this is the question also asked by, Jean-Luc Coatalem when he tries to break the silence regarding is grand-father, a grand father whom he never knew, arrested in 1943 then deported to Germany and whom his family has always refused to discuss, considering that any attempt for an explaination would be a treason Jean-Luc.
Reference caption:
Ecrire, est-ce trahir ?<p>C'est la question que se pose
J.-L. Coatalem.<p>Il cherche à briser le silence
qui entoure la mort<p> de son grand-père, qu' il n'a pas
 connu, arrêté en 1943, puis déporté<p> en Allemagne.<p> On a toujours refusé d' en parler
 dans sa famille.<p> On considère comme une trahison
 toute tentative d' explication.<p>Writing, is it a treason?<p>This is the question asked by
J.-L. Coatalem.<p> He tries to break the silence
regarding the death<p>of his grand-father, whom he did not
know, arrested in 1943, then deported<p>to Germany.<p>One has always refused to discuss this
in his family.<p>One considers as a treason
any attempted explaination.<p>

Figure 1: *A complete training instance, made of several captions, and associated input and output tags. <stock> for prepared subtitle, <mag> for the 'magazine' genre; <p> denotes end of a caption,
 end of a line. Our own translation into English is in grey.*

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}),$$
$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}).$$

The *LRPE* and *LDPE* variants [18] instead respectively encode information regarding the total expected length l, or the difference (in words) from the current word as:

$$\mathrm{LRPE}_{(pos,l,2i)} = \sin(pos/l^{2i/d_{model}}),$$
$$\mathrm{LRPE}_{(pos,l,2i+1)} = \cos(pos/l^{2i/d_{model}}),$$
$$\mathrm{LDPE}_{(pos,l,2i)} = \sin((l-pos)/10000^{2i/d_{model}}),$$
$$\mathrm{LDPE}_{(pos,l,2i+1)} = \cos((l-pos)/10000^{2i/d_{model}}).$$

In our experiments, we have adjusted the length requirements in LRPE and LDPE so as to generate sentences that would either match a fixed or program-specific compression rate r (in which case the expected length caption length l is simply r times the input length), or to match a constant display frequency f (in which case the expected length l is equal to the total caption duration times f). For these models, compression rates apply at the speech segment level.

In addition, we explore a tag-based approach to generate adapted closed captions: on the "source" (speech side), we include tags to denote the caption (`online` or `stock`) and/or the program types, using the following categories: cartoon, documentary, education, magazine, news, politics, quiz and series. Tag-based approach have been effectively used to better control neural text generation in multiple studies: eg. to control domain, politeness, style or output language [19, 20, 21]; even closer to our work, [4] uses the same mechanism to control the output length of subtitles. Following also [22], on the "target" side, we insert caption and line boundaries in the training material, and ask the captioning system to simultaneously generate the textual content and the segmentation marks. A complete training instance with its tags is in Figure 1.

3.4. Using artificial transcripts

We also attempt to better exploit the wealth of available (untranscribed) reference captions. Following a common practice in the Machine Translation community [14, 23] we generate artificial training examples through "back-translation" (BT). To this end, we use our regular training data to learn a system turning reference captions (including segmentation marks) into artificial transcripts, thereby generating a training set of 286K lines. Note that prior to back-translation, we randomly merge consecutive captions into larger chunks, to better match our training material. Using BT, we expect to (a) improve the quality of lexical embeddings; (b) learn better distinctions between program and caption styles with supplementary instances; (c) improve the segmentation generation process, which will be given a much larger set of reference, well segmented, captions.

3.5. Evaluation metrics

In our evaluation, we aim to measure multiple factors, each of which accounts for a facet of subtitle quality. Regarding the quality of the text, we use the BLEU score [24] with respect to reference captions, as well as the SARI [13] metric, which not only scores the similarity to the reference but also rewards divergences (likely simplifications) from the input text.[5] Another measure of simplification is the Flesch Reading Ease (FRE) index (larger is simpler), adapted to French in [26].

To measure the satisfaction of length requirements, we compute the percentage of deviant captions, ie. segments that are either too long (*CPL>36*), or do not remain long enough on screen[6] (*CPS>15*). Finally, to evaluate the quality of segmentation, we follow [27, 22] and use BLEU-br (a variant of the BLEU metrics that also takes tags into account), as well as an adapted version of the Translation Edit Rate (TER) [28], which scores segmentation tags, but ignores words which are replaced by a placeholder before running the evaluation.

4. Experimental results

4.1. Baseline systems

We first evaluate the baseline systems, reporting detailed scores for all metrics in Table 3. First note that these averages mask the variance of scores across programs types: for the BLEU-Br metrics, they range from 22.8 (Talk show) to more than 44 (for news and cooking magazine). Using clean transcripts yields only a small gains for all metrics, suggesting that our system somewhat learns to fix transcription errors better that it learns to compress the input. Evaluating the system with reference captions and rule-based segmentations helps to realize the impact

[5] For both metrics, we use the implementation of [25].
[6] The reference is based on a frequency of 15 $char/s$.

Table 3: *Evaluation scores for averaged over programs. We also report scores obtained with clean transcripts and automatic segmentations (*), with reference captions and automatic segmentation (**) and with reference captions and segmentations (***).*

Systems	BLEU-br	BLEU	SARI	FRE	TER-br	CPL	CPL>36	CPS>15	CpR
Baselines									
Transformer + rules	38.1	43.3	52.2	87.3	0.39	23.8	0.0	63.5	0.82
Transformer + rules*	40.7	45.8	50.2	87.8	0.37	23.5	0.0	48.3	0.72
Reference + rules**	73.1	100	100	88.0	0.13	22.9	0.0	31.0	0.69
Reference***	100	100	100	88.0	0.0	25.4	0.0	46.2	0.69
Length control, genre agnostic systems									
+ tags	41.5	43.8	52.9	88.4	0.38	26.0	6.1	63.8	0.82
+ tags + BT	41.6	44.0	53.0	88.9	0.38	25.1	2.9	64.1	0.82
+ tags + LRPE (r = 0.75)	35.6	35.9	49.8	89.7	0.35	25.6	5.5	16.8	0.61
+ tags + LRPE (f = 14.5)	38.7	39.5	51.2	89.4	0.31	25.6	5.3	1.2	0.65
+ tags + LDPE (r = 0.75)	34.5	35.2	49.5	89.3	0.35	25.9	7.0	16.3	0.61
+ tags + LDPE (f = 14.5)	37.3	38.7	50.6	89.1	0.32	26.2	7.6	0.8	0.64
Length control, genre adapted systems									
+ tags + genre	42.8	44.5	53.2	88.2	0.34	26.2	7.0	58.0	0.79
+ tags + BT + genre	42.4	44.6	53.4	88.6	0.37	25.7	3.3	62.1	0.81
+ tags + LRPE (r adapted)	34.9	35.6	49.7	89.7	0.34	25.5	5.5	12.2	0.59
+ tags + LRPE (f adapted)	39.9	40.9	51.5	89.3	0.31	25.6	5.3	10.8	0.67
+ tags + LRPE (f) + genre	41.2	42.2	52.2	89.2	0.31	25.4	5.6	13.9	0.68

of segmentation errors on the associated scores (BLEU-BR and TER-Br). They also show that our segmentation rules are even stricter than the manual segmentation, as they get better scores on all related metrics (average segment length, average character rate, % of segments that exceed the recommended character rate). Finally, comparing reference FRE scores with the baseline shows that our outputs texts have right level of complexity.

4.2. Length control

We now study the effect of length control for the two strategies considered (LRDE, LRPE). Results are in Table 3 (middle part). Inserting and predicting segmentation tags in the output text is extremely beneficial on almost all accounts, with a notable increase of the BLEU-Br and SARI scores. The only downfall is an increase of the number of segments that exceed the 36 char limit. Throwing in additional BT data provides a tiny additional boost, and almost solves length related issues. Note that the effect of BT is very variable accross programs, with significant gains and losses: this is because the distributions of genres in the BT data is profitable for some (eg. News) and detrimental for others. Using an explicit length control mechanism in the decoder yields results that are worse than the baseline in terms of content, with no clear winner between LRPE and LRDE. It seems however that using the frame rate to compute the expected length is more effective than using a fixed compression rate of 0.75. The former strategy has the merit to yield segment that comply with the character rate constraints, while the latter both generates segments that are on average too short, but still often violates the 36 char limit.

4.3. Generating genre adapted subtitles

We finally study the effect of specializing the captioning system with genre tags. Our results are in Table 3 (bottom). Including caption and program type tags improves most of the scores, with a significant variability between the programs in our test set, confirming that the task is not as easy for all genres. As before, additional BT data does not help much in general, except for a notable gain with respect to the CPL constraint. The adaptation of length control brings mixed results: giving an unique compression rate for each genre does not improve over using a fixed rate of 0.75. Yet, we see a relative gain when adapting the compression level to the time constraint (LRPE f adapted) , which is probably closer to what human subtitlers do. However, except for the respect of the CPS limit and the TER-br, length control models underperform the simpler corresponding Transformer. Combining type tags with adapted length control helps to narrow this gap, showing that the two sources of information are not entirely redundant.

5. Conclusion

In this paper, we have studied the generation of closed captions that depend on the program type and genre. Using techniques borrowed from the multi-domain machine translation literature, we have proposed and evaluated several approaches to condition the subtitle compression process on the program. Using these, as well as other standard techniques (BPEs, back-translation), we were able to improve our baseline systems for almost all metrics, with clear improvements in terms of segmentation. In our future work, we intend to continue exploring multi-genre adaptation techniques, notably the use of adapter layers [29, 30]; another direction we wish to consider is to make a better use of back-translated data. For this, we will need to be more cautious in the design of the back-translated program mix with respect to the test distribution of programs and genres.

6. Acknowledgments

This study has been funded by the BPI-France investment programme "Grands défis du numérique", as part of the ROSETTA-2 project (Subtitling RObot and Adapted Translation). We warmly thank France-TV access for granting access to their recordings and reference subtitles, and J.-L. Gauvain for the automatic transcription tool. This study has been made possible thanks to the Saclay-IA computing platform.

7. References

[1] S. Piperidis, I. Demiros, P. Prokopidis, P. Vanroose, A. Hoethker, W. Daelemans, E. Sklavounou, M. Konstantinou, and Y. Karavidas, "Multimodal, multilingual resources in the subtitling process." in *Proceedings of LREC*, 2004.

[2] M. Melero, A. Oliver, and T. Badia, "Automatic multilingual subtitling in the eTITLE project," *Proceedings of Translating and the Computer*, vol. 28, pp. 1–18, 2006.

[3] M. Volk, R. Sennrich, C. Hardmeier, and F. Tidström, "Machine translation of tv subtitles for large scale production," in *Second Joint EM+/CNGL Workshop*, November 2010, pp. 53–62.

[4] S. M. Lakew, M. D. Gangi, and M. Federico, "Controlling the output length of neural machine translation," in *Proceedings of IWSLT'2019*, 2019.

[5] D. Liu, J. Niehues, and G. Spanakis, "Adapting end-to-end speech recognition for readable subtitles," in *Proceedings of the 17th International Conference on Spoken Language Translation*, Online, Jul. 2020, pp. 247–256.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6000–6010.

[7] M. A. Farajian, M. Turchi, M. Negri, and M. Federico, "Multi-domain neural machine translation through unsupervised adaptation," in *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, 2017, pp. 127–137.

[8] J. Su, J. Zeng, J. Xie, H. Wen, Y. Yin, and Y. Liu, "Exploring discriminative word-level domain contexts for multi-domain neural machine translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pp. 1–1, 2019.

[9] M. Q. Pham, J. Crego, and F. Yvon, "Revisiting multi-domain machine translation," *Transactions of the Association for Computational Linguistics*, vol. 9, no. 0, pp. 17–35, 2021.

[10] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 866–875.

[11] T.-H. Ha, J. Niehues, and A. Waibel, "Toward multilingual neural machine translation with universal encoder and decoder," in *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Vancouver, Canada, 2016.

[12] C. Chu and R. Dabre, "Multilingual and multi-domain adaptation for neural machine translation," in *Proceedings of the 24st Annual Meeting of the Association for Natural Language Processing*, ser. NLP 2018, Okayama, Japan, 2018, pp. 909—912.

[13] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016.

[14] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 86–96.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[16] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71.

[17] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, and M. Okumura, "Controlling output length in neural encoder-decoders," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 1328–1338.

[18] S. Takase and N. Okazaki, "Positional encoding to control output sequence length," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3999–4004.

[19] R. Sennrich, B. Haddow, and A. Birch, "Controlling politeness in neural machine translation via side constraints," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 35–40.

[20] C. Kobus, J. Crego, and J. Senellart, "Domain control for neural machine translation," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Varna, Bulgaria, Sep. 2017, pp. 372–378.

[21] L. Martin, É. de la Clergerie, B. Sagot, and A. Bordes, "Controllable sentence simplification," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4689–4698.

[22] A. Karakanta, M. Negri, and M. Turchi, "Is 42 the answer to everything in subtitling-oriented speech translation?" in *Proceedings of the 17th International Conference on Spoken Language Translation*. Online: Association for Computational Linguistics, Jul. 2020, pp. 209–219.

[23] F. Burlot and F. Yvon, "Using monolingual data in neural machine translation: a systematic study," in *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels, October 2018, pp. 144–155.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[25] F. Alva-Manchego, L. Martin, C. Scarton, and L. Specia, "EASSE: easier automatic sentence simplification evaluation," *CoRR*, vol. abs/1908.04567, 2019.

[26] L. Kandel and A. Moles, "Application de l'indice de Flesch à la langue française," *Cahiers Etudes de Radio-Télévision*, vol. 19, no. 1958, pp. 253–274, 1958.

[27] E. Matusov, P. Wilken, and Y. Georgakopoulou, "Customizing neural machine translation for subtitling," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Florence, Italy, Aug. 2019, pp. 82–93.

[28] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, vol. 200, no. 6. Cambridge, MA, 2006.

[29] A. Bapna and O. Firat, "Simple, scalable adaptation for neural machine translation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019, pp. 1538–1548.

[30] M. Q. Pham, J. M. Crego, F. Yvon, and J. Senellart, "A study of residual adapters for multi-domain neural machine translation," in *Proceedings of the Fifth Conference on Machine Translation*, Online, 2020, pp. 615–626.