# Unsupervised Subtitle Segmentation with Masked Language Models

**David Ponce**[*1,2] and **Thierry Etchegoyhen**[*1] and **Victor Ruiz**[1]
[1] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2] University of the Basque Country UPV/EHU
{adponce,tetchegoyhen,vruiz}@vicomtech.org

## Abstract

We describe a novel unsupervised approach to subtitle segmentation, based on pretrained masked language models, where line endings and subtitle breaks are predicted according to the likelihood of punctuation to occur at candidate segmentation points. Our approach obtained competitive results in terms of segmentation accuracy across metrics, while also fully preserving the original text and complying with length constraints. Although supervised models trained on in-domain data and with access to source audio information can provide better segmentation accuracy, our approach is highly portable across languages and domains and may constitute a robust off-the-shelf solution for subtitle segmentation.

## 1 Introduction

Subtitling is one of the principal means of providing accessible audiovisual content. With the ever increasing production of audiovisual content in multiple domains and languages, in the current digital era, subtitle provision can benefit from automation support, via Automatic Speech Recognition and/or Machine Translation (Volk et al., 2010; Aliprandi et al., 2014; Etchegoyhen et al., 2014; Tardel, 2020; Bojar et al., 2021).

Subtitles are subject to specific constraints in order to achieve adequate readability, including layout, on-screen duration and text editing. Among these constraints, segmentation addresses the maximum number of characters per line, the number of lines per subtitle, and breaks at natural linguistic frontiers. Segmentation has been shown to be an important readability factor (Perego et al., 2010; Rajendran et al., 2013), with improperly segmented subtitles resulting in increased cognitive effort and reading times for users. Thus, automated subtitling systems need to generate properly segmented subtitles to achieve readability.

A typical baseline for subtitle segmentation, still used in some production systems, is simple character counting, whereby line breaks are inserted before reaching the maximum allowed number of characters per line. Although simple and fast, this approach does not address the need for linguistically correct segments and, therefore, falls short in terms of readability. Several approaches have been proposed to improve segmentation by automated means. Álvarez et al. (2014) proposed a machine learning method where subtitle breaks are predicted by Support Vector Machine and Linear Regression models trained on professionally-created subtitles. A similar method based on Conditional Random Fields was then shown to improve over these results (Alvarez et al., 2017). Approaches that directly generate subtitle breaks within Neural Machine Translation have also been proposed in recent years (Matusov et al., 2019; Karakanta et al., 2020a). Recently, Papi et al. (2022) developed a multilingual segmenter which generates both text and breaks and may be trained on textual input only, or on joint text and audio data.

Although quality subtitle segmentation may be achieved with the aforementioned approaches, they require supervised training on segmented subtitle corpora. At present, the largest subtitle corpus is Open Subtitles (Lison et al., 2018), which mainly covers entertainment material, contains subtitles mostly created by non-professionals or automatically translated, and does not include line breaks. The MuST-Cinema corpus (Karakanta et al., 2020b), on the other hand, is a multilingual speech translation corpus that includes subtitles breaks, but is only available for 8 languages at the moment. Considering the vast amount of languages and domains in audiovisual content, the lack of segmented training data hinders the development of robust automated subtitling systems.

In this work, we describe a novel unsupervised method based on pretrained masked language mod-

[*]These authors contributed equally to this work.



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 2: Short Papers*, pages 771–781

els (MLM), where line and subtitle breaks are inserted according to the likelihood of a segment acting as an isolated unit, as approximated by the probability of a punctuation mark occurring at a given segmentation point. In our experiments, this novel approach obtained competitive results on most metrics, while also fully preserving the original text and complying with length constraints. Our system may thus be used as a simple yet efficient subtitle segmenter with any pretrained masked language model, for any language covered by the model.

## 2 Approach

Our approach is based on the standard view that the more appropriate subtitle segments are those that may function as isolated grammatical chunks. We further hypothesise that a relevant approximation for the identification of this type of unit is the likelihood of a punctuation mark being inserted at the end of a candidate segment, as punctuation may mark the closure of a syntactic unit and is often associated with discursive pauses. To test this hypothesis, we compute the likelihood of punctuation marks at different segmentation points, as predicted by a pretrained MLM, and select the insertion point with the highest likelihood.[1]

The segmentation candidates are determined under a sliding-window approach over the entire input text. We first generate the list of all pairs <$\alpha$, $\beta$> over the unprocessed portion of the text, where $\alpha$ is a segmentation candidate of length under a specified limit $K$, corresponding to the maximum number of characters per line, and $\beta$ is the remaining portion of the text to be segmented.

We then score all segmentation candidates $\alpha$ with one of the LM scoring variants described below. A segmentation marker, either end-of-line (<eol>), or end-of-block indicating the end of a subtitle (<eob>), is then appended to the best scoring candidate, and $\beta$ becomes the input text to be segmented in a recursive iteration of the process.

Since our method does not rely on any additional information, such as an audio source, to determine the segmentation type, an <eob> tag is inserted every even segment or when $\beta$ is empty; otherwise, an <eol> tag is inserted. We thus generate subtitles with a maximum of two lines, following a standard recommendation in subtitling. We also define a minimal number of characters ($min$) in $\alpha$ for the segmentation process to apply, and do not segment lines that are under the specified character limit.

We evaluated three approaches to compute segmentation scores over each candidate pair <$\alpha$, $\beta$>:

- **Substitution:** The last token of $\alpha$ is masked and the score is the highest MLM probability among punctuation marks on this mask.
- **Insertion:** A mask is appended to $\alpha$ and the score is the highest MLM probability among punctuation marks on this mask.
- **LM-Score:** The score is the average of the perplexity of $\alpha$ and $\beta$, as derived from the MLM probabilities for each token in the corresponding sequence.

The first two methods are variants of our core approach. The third method, while also based on the same pretrained MLM, relies instead on the pseudo-perplexity of the sequences according to the MLM, computed following Salazar et al. (2020). We included this latter variant to measure the potential of using LM scoring directly, without resorting to the likelihood of punctuation marks.

## 3 Experimental Setup

**Corpora.** For all experiments, we used the MustST-Cinema corpus (Karakanta et al., 2020b), which is derived from TED talks and contains both line and subtitle break markers. In addition to being publicly available, it also allows for a direct comparison with the supervised models of Papi et al. (2022). We report results of our approach on the 6 MuST-Cinema datasets for which comparative results were available, directly predicting segmentation on the test sets without any training.[2]

**Methods.** For our approach, we tested the three variants described in Section 2. We used BERT (Devlin et al., 2019) as our MLM for all languages.[3]. Additionally, we included a variant called *overt clueing* (*OC*), where an overt punctuation mark at the end of a candidate segment increments the mask score by 1. We then compared the results of the best LM-based variant with those obtained by alternative approaches. In all cases, our results were computed with $min = 15$, as this value obtained the best results overall over the development

[1] Throughout our experiments, we used the following punctuation marks: '.', ',', '?', '!', ':' and ';'.

[2] Our results on all remaining languages of the MuST-Cinema datasets are presented in Appendix B.

[3] Specifically *bert-base-uncased* as available on HuggingFace (https://huggingface.co/), accessed on November 2022.

| | English | | | Spanish | | | German | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Sigma | EOL | EOB | Sigma | EOL | EOB | Sigma | EOL | EOB |
| Substitution | 71.65 | +19.86 | -10.96 | 69.34 | **+12.36** | -5.74 | 69.31 | +19.05 | -7.05 |
| Insertion | **76.77** | **+19.18** | -9.91 | **73.47** | +12.98 | -4.91 | **70.85** | +18.53 | -7.96 |
| LM-Score | 69.97 | +21.40 | **-8.66** | 67.70 | +13.29 | **-5.37** | 64.07 | **+16.45** | **-6.51** |

Table 1: Sigma and break coverage test set results for LM-based segmentation variants

sets, although the differences were minor with the other values we tested (1, 10 and 20).[4]

We used the simple character counting approach (hereafter, *CountChars*) as baseline, and, as representative supervised methods on the selected datasets, the models described by (Papi et al., 2022). Their core supervised approach is based on a Transformer (Vaswani et al., 2017) architecture with 3 encoder layers and 3 decoder layers, trained on textual MuST-Cinema input only (*MC.Text*), or on complementary audio data as well via an additional speech encoder with 12 encoder layers (*MC.Multi*). They trained each variant on either monolingual data alone (*mono*), or in a multilingual setting (*multi*). Finally, they also report results for a variant (*OS.Text*) trained on the Open Subtitles corpus (Lison et al., 2018) for their zero-shot experiments.

**Evaluation.** We use the subtitle-oriented metric Sigma (Karakanta et al., 2022), which computes the ratio of achieved BLEU (Papineni et al., 2002) over an approximated upper-bound BLEU score, on text that includes line and subtitle breaks. Sigma is meant to support the evaluation of imperfect texts, i.e. text that differs from the reference when breaks are omitted. Although our approach does not produce imperfect text, achieving perfect BLEU scores when breaks are ignored, we used this metric for comparison purposes. We also report break coverage results (Papi et al., 2022), defined as the ratio of predicted breaks over reference breaks, which we computed separately for the EOL and EOB breaks. Finally, we include length conformity results (CPL), measured as the percentage of subtitle lines whose length is under the maximum number of characters defined by the subtitle guidelines (42 in the TED guidelines[5]).

[4]See Appendix C for results with different values of the *min* parameter.

[5]https://www.ted.com/participate/translate/subtitling-tips

## 4 LM-based Segmentation Variants

We first compared the three methods described in Section 2 on the English, Spanish and German datasets, with the results described in Table 1. In terms of Sigma, the Insertion method obtained the best results in all cases. It also obtained the best scores in terms of coverage for the EOL marker, except in Spanish, although all three variants tend to overgenerate end-of-line markers to similar extents. The LM-Score variant obtained the worst results in terms of Sigma, but outperformed the alternatives in terms of EOB coverage, a metric on which the three variants performed markedly better than on EOL coverage. Considering the overall results, we selected the Insertion variant as the most balanced one for all remaining experiments reported below.

## 5 Comparative Results

In Table 2, we present the results obtained by the selected approaches on the languages for which results were available with supervised models trained on in-domain data. Overall, our approach outperformed the *CountChars* baseline across the board, and was in turn outperformed by the supervised variants in terms of Sigma scores. Although it is clear from these results that training segmentation models on in-domain data, with or without audio data, provides clear advantages in terms of subtitle segmentation, it is worth noting that Sigma does not, by design, reflect the actual BLEU score without breaks, i.e. the generation of imperfect text, which is a by-product of the above supervised approaches and non-existent in ours.[6] In terms of CPL, all supervised models generate subtitle lines that overflow the limit, to a significant degree, whereas the selected unsupervised models trivially respect the length constraint.

[6]The results indicated in Table 3 on unseen data seem to indicate that their *MC.Multi* model can reach BLEU scores close to 100, thereby limiting the negative impact of imperfect text generation in these cases.

| | | English | | French | | German | | Italian | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Training | Sigma | CPL | Sigma | CPL | Sigma | CPL | Sigma | CPL |
| CountChars | N/A | 63.71 | **100%** | 62.87 | **100%** | 62.34 | **100%** | 61.49 | **100%** |
| MC.Text | mono | 84.87 | 96.6% | 83.68 | 96.7% | 83.62 | 90.9% | 82.22 | 90.0% |
| | multi | 85.98 | 88.5% | 84.56 | 94.3% | 84.02 | 90.9% | 83.04 | 91.2% |
| MC.Multi | mono | 85.76 | 94.8% | 84.25 | 93.9% | 84.22 | 91.4% | 82.62 | 89.9% |
| | multi | **87.44** | 95.0% | **86.49** | 94.1% | **86.40** | 89.9% | **85.33** | 90.0% |
| MLM | N/A | 76.77 | **100%** | 73.78 | **100%** | 70.85 | **100%** | 71.38 | **100%** |
| MLM+OC | N/A | 77.89 | **100%** | 76.07 | **100%** | 75.63 | **100%** | 74.20 | **100%** |

Table 2: Comparative results between unsupervised methods and supervised approaches trained on in-domain data

| Dutch | | | | | |
|---|---|---|---|---|---|
| Method | BLEU | Sigma | CPL | EOL | EOB |
| CountChars | **100** | 63.2 | **100%** | -21.2 | -7.1 |
| OS.Text | 89.5 | 64.4 | 71.2% | -31.4 | -51.3 |
| MC.Text | 61.3 | 74.4 | 77.8% | -23.4 | -9.9 |
| MC.Multi | 99.9 | **80.3** | 91.4% | -27.2 | **0.4** |
| MLM | **100** | 68.7 | **100%** | **+20.4** | -10.0 |
| MLM+OC | **100** | 73.9 | **100%** | +21.2 | -10.0 |
| **Spanish** | | | | | |
| Method | BLEU | Sigma | CPL | EOL | EOB |
| CountChars | **100** | 63.2 | **100%** | -24.6 | **-4.4** |
| OS.Text | 92.6 | 64.1 | 71.2% | -32.3 | -45.4 |
| MC.Text | 69.6 | 75.8 | 70.1% | -47.6 | -19.3 |
| MC.Multi | 99.6 | **78.7** | 91.8% | -22.4 | 4.7 |
| MLM | **100** | 73.5 | **100%** | **+13.0** | -4.9 |
| MLM+OC | **100** | 75.6 | **100%** | +13.4 | -4.6 |

Table 3: Comparative results between unsupervised methods and zero-short supervised approaches

In Table 3, we show the comparative results between the selected unsupervised methods and the supervised variants, in languages where zero-shot results were available for the latter approaches. In this scenario, in terms of Sigma our approach obtained results on a par with the supervised *MC.Text* models trained on in-domain MuST-Cinema data, outperformed the *OS.Text* models trained on Open Subtitles data, and was surpassed by the *MC.Multi* model, which exploits additional audio information, by 3.1 and 6.4 points. In terms of break coverage, in most cases our unsupervised method outperformed the supervised variants, to a significant degree compared to the text-based *OS.Text* and *MC.Text* models. Regarding BLEU scores without breaks, only the *MC.Multi* model reaches a score close to the perfect one achieved by the unsupervised models, whereas the *MC.Text* model is outperformed by 38.7 and 31.4 points in Dutch and Spanish, respectively. In all cases, the CPL scores indicate that none of the supervised approaches fully meet the length constraint, leading to overflowing lines in 8.2% of the cases at best and 29.9% at worst. In this scenario as well, the unsupervised approaches fully meet the length constraint, by design.

Overall, overt clueing improved over our core method by an average of 3.12 Sigma points, indicating that some likely punctuation configurations were not properly captured by our MLM approximation. In general, our approach tends to overgenerate EOL markers, whereas the opposite is true for the selected supervised models. Determining which of these tendencies leads to better subtitle readability would require a specific human evaluation which we leave for future research.

Although the zero-shot Sigma results obtained by the supervised *MC.Multi* method show the potential of this approach to provide pretrained models applicable to other languages, two important aspects are worth considering. First, the available zero-shot results were obtained on datasets in the same domain as the data seen to train the supervised models. A more complete assessment of the capabilities of these models in zero-shot settings, which would be the most frequent scenario consid-

ering the lack of training data across domains and languages, would require specific evaluations in other domains. Secondly, although segmentation is a key aspect for subtitle readability, length conformity is an equally important constraint, if not more so considering that subtitles with lines over the CPL limit are considered invalid in subtitling. Our proposed unsupervised method can thus be seen as a pragmatic approach which guarantees valid subtitles while also providing quality segmentation across the board.[7]

## 6 Conclusions

We described an unsupervised approach to subtitle segmentation, based on pretrained masked language models, where line or subtitle breaks are inserted according to the likelihood of punctuation occurring at candidate segmentation points.

Although supervised models, trained on in-domain data with audio support, were shown to perform better that this simple textual approach in terms of the Sigma metric, they tend to generate imperfect text to varying degrees, while also failing to fully meet length constraints that are essential for subtitling.

In contrast, our LM-based textual approach outperformed supervised models in most cases in terms of break generation coverage, while also fully preserving the original text, complying with length constraints, and obtaining competitive results in terms of Sigma. This simple approach may thus provide a highly portable complementary solution for subtitle segmentation across languages and domains.

## 7 Limitations

The first clear limitation of our approach is its text-based nature. This prevents important audio information, typically silences in speech patterns, from being exploited to generate subtitle breaks. A more complete system could be devised though, for instance by associating our text-based approach with the information provided by a forced alignment toolkit, whenever audio information is available. A simple method along these lines could be the following: 1. Apply our MLM-based segmentation but only generating a unique segmentation tag SEG; 2. Insert EOB markers wherever the silence between two aligned words is above a specified threshold; 3. Traverse the text sequentially and replace SEG with EOL if there exists a previous marker of type EOB, otherwise replace with EOB. We left this use of our method in combination with audio information for future research, as audio alignment for subtitles typically involves additional factors such as non-literal transcriptions.

Additionally, our method is limited in its adaptability to specific segmentation guidelines, which may be company-specific. The main adaptable parameters of our methods are the minimum and maximum parameters of the segmentation window, and the set of predefined punctuation marks over which masking is computed, neither of which could fully model idiosyncratic segmentation guidelines. However, in our experience at least, segmentation in real professional data tends to display varying degrees of consistency with respect to guidelines, and natural linguistic breaks seem to be the dominant factor for subtitle segmentation. A specific evaluation would be needed on data from varied professional datasets to determine the extent to which our method might deviate from specific guidelines.

Finally, other aspects of subtitling, such as the recommendation in some guidelines for subtitles to appear in a pyramidal view, i.e. with the first line shorter than the second line, have not been taken into consideration in this work. Our aim was to evaluate our core LM-based approach without additional variables that can vary across guidelines and may also have led to results that are more difficult to interpret overall. Our approach could nonetheless be easily augmented with constraints on relative line lengths within subtitles, by incrementing the scores of segmentation candidates that respect this surface-level constraint.

## 8 Ethical Considerations

Our approach involves the use of large pretrained language models, whose computational performance is typically higher when deployed in more powerful environments with GPUs. Under such usage, electric consumption and associated carbon footprint are likely to increase and users of our method under these conditions should be aware of this type of impact. However, subtitle segmentation is often performed offline, where efficient processing is less of a concern, and lower-cost CPU deployments are an entirely viable option. All our results were obtained with a single large LM de-

[7] Examples of segmented subtitles can be found in Appendix A.

ployed on CPU, with the aim of reducing energy consumption at inference time.

Additionally, our method requires no training for the task at hand and thus removes the cost of model training associated with the supervised methods with which we compare our results. For instance, Papi et al. (2022) indicate that they use four K80 GPUs to train their models, which we took as comparison points, with 1 day of training for their text-only models and 1 week for their multimodal segmenters. Therefore, given the large number of potential language pairs and domains in need of segmented subtitle content, our approach can provide competitive results with a comparatively lesser impact on energy resource consumption.

## Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work was partially supported by the Department of Economic Development and Competitiveness of the Basque Government (Spri Group) through funding for the StreAmS project (ZL-2021/00700).

## References

Carlo Aliprandi, Cristina Scudellari, Isabella Gallucci, Nicola Piccinini, Matteo Raffaelli, Arantza del Pozo, Aitor Álvarez, Haritz Arzelus, Renato Cassaca, Tiago Luis, et al. 2014. Automatic live subtitling: state of the art, expectations and current trends. In *Proceedings of NAB Broadcast Engineering Conference: Papers on Advanced Media Technologies, Las Vegas*, volume 13.

Aitor Álvarez, Haritz Arzelus, and Thierry Etchegoyhen. 2014. Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 229–238. Springer.

Aitor Alvarez, Carlos-D Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. 2017. Improving the automatic segmentation of subtitles through conditional random field. *Speech Communication*, 88:83–95.

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Peter Polák, Ebrahim Ansari, Mohammad Mahmoudi, Rishu Kumar, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stüker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. 2021. ELITR multilingual live subtitling: Demo and strategy. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 271–277, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 46–53, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alina Karakanta, François Buet, Mauro Cettolo, and François Yvon. 2022. Evaluating subtitle segmentation for end-to-end generation systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3069–3078, Marseille, France. European Language Resources Association.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020a. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online. Association for Computational Linguistics.

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

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022. Dodging the data bottleneck: Automatic subtitling with automatically segmented ST corpora. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint*

*Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 480–487, Online only. Association for Computational Linguistics.

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

Elisa Perego, Fabio Del Missier, Marco Porta, and Mauro Mosconi. 2010. The cognitive effectiveness of subtitle processing. *Media psychology*, 13(3):243–272.

Dhevi J Rajendran, Andrew T Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Anke Tardel. 2020. Effort in semi-automatized subtitling processes: speech recognition and experience during transcription. *Journal of Audiovisual Translation*, 3(2):79–102.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine translation of TV subtitles for large scale production. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 53–62, Denver, Colorado, USA. Association for Machine Translation in the Americas.

## A Segmentation Examples

Table 4 provides examples of subtitles in the MuST-Cinema test sets segmented with either the character counting baseline or our LM-based approach, in its insertion variant without resorting to overt punctuation clueing.

In these examples, the MLM approach generates end-of-line and end-of-subtitle breaks that are overall in line with natural linguistic breaks, contrary to the character counting baseline. As such, on either short, medium or longer input, the readability of the generated subtitles is significantly enhanced with our approach.

## B Extended Results

The results presented in Section 5 were limited to the subset of languages and metrics for which published comparative results were available on the MuST-Cinema datasets. In Table 5, we present the complete list of results obtained with our method, for all languages and metrics. The selected variant of our method is the insertion masking approach, which was selected for the main results in our paper, with a segmentation window starting at 15 characters and ending at 42. We do not include BLEU scores computed over text that includes segmentation breaks, as the results are identical to those obtained with the Sigma metric for our approach, which does not generate imperfect text.

Across languages, the results are relatively uniform, with the best Sigma scores obtained in English and the lowest in Dutch, for a difference of 4.1 points between the two languages. In terms of break coverage, the best results were obtained for Spanish and the worst for Romanian, although results were also relatively uniform across languages. In all cases, overt clueing, where overt punctuation marks raised the LM score by 1, improved Sigma scores, although it had less of an impact on break coverage results, where both variants performed similarly overall.

## C Results With Different *min* Parameters

As noted in Section 3, considering preliminary results over the development set we selected a default value of 15 for the *min* parameter, which indicates the number of characters after which the segmentation process applies. In Table 6, we present comparative results on the test sets with different *min* values. In terms of Sigma, values of 15 and 20 led to rather similar results; values of 1 and 10 resulted in slightly lower results, with the lowest results achieved with the former.

In terms of <eol> and <eob> coverage, the former increases with larger *min* values, which is expected given the more restricted space to insert these end-of-line markers as the value increases; for <eob>, the restricted insertion space results in increased under-generation, which in turn results in better scores for lower values of the *min* parameter.

| CountChars | MLM |
|---|---|
| They're things you access through your <eol><br>computer. <eob> | They're things you access <eol><br>through your computer. <eob> |
| Every row of data is a life whose story <eol><br>deserves to be told with dignity. <eob> | Every row of data is a life <eol><br>whose story deserves to be told <eob><br>with dignity. <eob> |
| During the winter, struggling to get <eol><br>warm, my neighbors would have no choice <eob><br>but to bypass the meter after their heat <eol><br>was shut off, just to keep their family <eob><br>comfortable for one more day. <eob> | During the winter, struggling to get warm, <eol><br>my neighbors would have no choice <eob><br>but to bypass the meter <eol><br>after their heat was shut off, <eob><br>just to keep their family comfortable <eol><br>for one more day. <eob> |

Table 4: Examples of subtitles segmented via character counting and MLM-based mask insertion

| Language | Method | BLEU | Sigma | EOL | EOB | CPL |
|---|---|---|---|---|---|---|
| DE | MLM | 100 | 70.85 | 18.53 | -7.96 | 100% |
| | MLM+OC | 100 | 75.63 | 19.81 | -7.78 | 100% |
| EN | MLM | 100 | 76.77 | 19.18 | -9.91 | 100% |
| | MLM+OC | 100 | 77.89 | 19.86 | -9.73 | 100% |
| ES | MLM | 100 | 73.47 | 12.98 | -4.91 | 100% |
| | MLM+OC | 100 | 75.59 | 13.45 | -4.63 | 100% |
| FR | MLM | 100 | 73.78 | 16.51 | -6.58 | 100% |
| | MLM+OC | 100 | 76.07 | 17.47 | -6.12 | 100% |
| IT | MLM | 100 | 71.38 | 18.49 | -9.55 | 100% |
| | MLM+OC | 100 | 74.20 | 20.34 | -8.57 | 100% |
| NL | MLM | 100 | 68.71 | 20.37 | -9.96 | 100% |
| | MLM+OC | 100 | 73.88 | 21.22 | -9.96 | 100% |
| PT | MLM | 100 | 71.59 | 20.03 | -10.81 | 100% |
| | MLM+OC | 100 | 75.50 | 19.87 | -10.02 | 100% |
| RO | MLM | 100 | 69.45 | 23.37 | -10.44 | 100% |
| | MLM+OC | 100 | 74.13 | 23.37 | -10.09 | 100% |

Table 5: Complete results with MLM mask insertion on the MuST-Cinema test sets ($min$=15)

| Language | *min* | BLEU | Sigma | EOL | EOB |
|---|---|---|---|---|---|
| DE | 1 | 100 | 72.31 | 28.75 | **-0.18** |
| | 10 | 100 | 73.96 | 22.68 | -4.43 |
| | 15 | 100 | **75.63** | 19.81 | -7.78 |
| | 20 | 100 | 75.28 | **14.54** | -11.21 |
| EN | 1 | 100 | 74.30 | 37.33 | **-0.98** |
| | 10 | 100 | 77.14 | 24.49 | -7.77 |
| | 15 | 100 | **77.89** | 19.86 | -9.73 |
| | 20 | 100 | 77.16 | **15.24** | -12.68 |
| ES | 1 | 100 | 73.00 | 20.87 | **0.28** |
| | 10 | 100 | 74.32 | 18.24 | -2.04 |
| | 15 | 100 | 75.59 | 13.45 | -4.63 |
| | 20 | 100 | **75.83** | **8.66** | -7.87 |
| FR | 1 | 100 | 73.89 | 24.68 | **-0.73** |
| | 10 | 100 | 75.26 | 20.83 | -3.93 |
| | 15 | 100 | 76.07 | 17.47 | -6.12 |
| | 20 | 100 | **76.75** | **12.5** | -10.05 |
| IT | 1 | 100 | 72.01 | 29.75 | **-3.66** |
| | 10 | 100 | 73.75 | 24.71 | -6.61 |
| | 15 | 100 | **74.20** | 20.34 | -8.57 |
| | 20 | 100 | 73.66 | **14.62** | -11.61 |
| NL | 1 | 100 | 72.16 | 26.83 | **-5.47** |
| | 10 | 100 | 73.56 | 23.26 | -8.47 |
| | 15 | 100 | 73.88 | 21.22 | -9.96 |
| | 20 | 100 | **74.40** | **16.81** | -12.43 |
| PT | 1 | 100 | 72.87 | 26.38 | **-6.24** |
| | 10 | 100 | 74.53 | 22.15 | -8.08 |
| | 15 | 100 | **75.50** | 19.87 | -10.02 |
| | 20 | 100 | 74.98 | **14.17** | -13.36 |
| RO | 1 | 100 | 72.05 | 32.3 | **-4.51** |
| | 10 | 100 | 73.76 | 26.98 | -7.52 |
| | 15 | 100 | 74.13 | 23.37 | -10.09 |
| | 20 | 100 | **74.89** | **17.53** | -12.83 |

Table 6: Test set results with the MLM+OC method and different values of the *min* parameter

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*7*

☒ A2. Did you discuss any potential risks of your work?
*8*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☐ Did you use or create scientific artifacts?

*Not applicable. Left blank.*

☐ B1. Did you cite the creators of artifacts you used?
*Not applicable. Left blank.*

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Not applicable. Left blank.*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Not applicable. Left blank.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Not applicable. Left blank.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Not applicable. Left blank.*

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Not applicable. Left blank.*

### C ☑ Did you run computational experiments?

*3*

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*We didn't trained any models for this paper, and inference was performed on CPU.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*3*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*5*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*3*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Not applicable. Left blank.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Not applicable. Left blank.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Not applicable. Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. Left blank.*