



Global-SEG: Text Semantic Segmentation Based on Global Semantic Pair Relations

Wenjun Sun, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Mickaël Coustaty, Antoine Doucet

► To cite this version:

Wenjun Sun, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Mickaël Coustaty, Antoine Doucet. Global-SEG: Text Semantic Segmentation Based on Global Semantic Pair Relations. ICDAR 2024, Aug 2024, Athens, Greece. pp.253-269, 10.1007/978-3-031-70546-5_15 . hal-04695730

HAL Id: hal-04695730

<https://hal.science/hal-04695730v1>

Submitted on 26 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Global-SEG: Text Semantic Segmentation Based on Global Semantic Pair Relations

Wenjun Sun¹[0009-0002-7857-8737], Hanh Thi Hong
Tran^{1,2,3}[0000-0002-5993-1630], Carlos-Emiliano
González-Gallardo¹[0000-0002-0787-2990], Mickaël
Coustaty¹[0000-0002-0123-439X], and Antoine Doucet¹[0000-0001-6160-3356]

¹ University of La Rochelle, L3i, La Rochelle, France

² Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

³ Jožef Stefan Institute, Ljubljana, Slovenia

{firstname.lastname,thi.tran,carlos.gonzalez_gallardo}@univ-lr.fr

Abstract. Text semantic segmentation is a crucial task in language understanding, as subsequent natural language processing tasks often require cohesive semantic blocks. This paper introduces a new perspective on this task by utilizing global semantic pair relations from both token- and sentence-level language models. This approach addresses the limitations of prior work, which concentrated solely on individual semantic units like sentences. Our model processes both local and global levels of sentence semantics via encoders and then combines the semantics obtained at each stage into a semantic embedding matrix. This matrix is then fed through a convolutional neural network and finally used as input through another encoder. This process enables the identification of semantic segmentation boundaries by describing the relationships of global semantic pairs. Furthermore, we utilize semantic embeddings from large language models and consider the positional information of text within the document to assess their efficacy in augmenting semantics. We test our model with both contemporary and historical corpora, and the results demonstrate that our approach outperforms benchmarks on each dataset.

Keywords: Text semantic segmentation · Semantic pair relation · Historical documents

1 Introduction

Text semantic segmentation consists of analyzing the semantic relationships between sentences or paragraphs based on the input text and next dividing them into coherent semantic blocks [15,28]. An example is shown in Figure 1. By identifying semantics between different blocks, this task can assist in numerous downstream natural language processing (NLP) tasks, such as dialogue analysis [31] and automatic summarization [20].

Researchers have proposed several models to address this task [1,19,18], all of which perform single-sentence semantic analysis on text based on the output

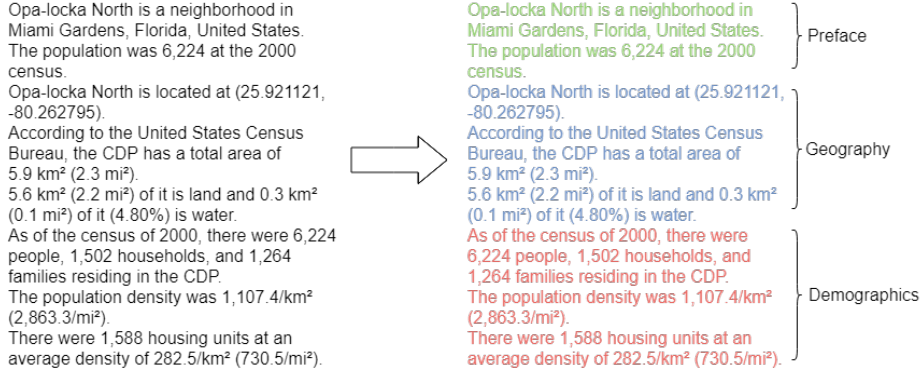


Fig. 1. Example of text segmentation

of a language model (LM) and ultimately output segmentation labels. However, these models suffer from one drawback: they neglect to analyze the relationships between groups of sentences. In addition, these approaches intend to obtain the final sentence semantic vectors either by combining the token-level embeddings or by directly using some sentence-level embeddings. But none of them have attempted to combine these two levels of information for a more accurate embedding. These two types of LMs are pre-trained on extensive data and offer two different perspectives: the relationship between tokens and the relationship between sentence meanings. Therefore, we aim to combine them and leverage the advantages of both types of LMs simultaneously, integrating the local and global viewpoints they provide.

We introduce **Global-SEG**, a new semantic segmentation model for text that leverages the global relationship between semantic pairs that utilizes both token-level and sentence-level LMs. A first encoder with transformer [29] architecture is used to obtain new sentence vectors from token embeddings. The outputs of the first encoder are then input into another encoder to get the relations between single sentences. Subsequently, the original sentence embedding from the sentence-level LM is combined with the output of these two encoders, forming a semantic embedding matrix. This matrix undergoes convolution by a convolutional neural network (CNN) along the semantic dimension to generate the embeddings for the sentence pairs. Finally, a last encoder aggregates the relationship between the sentence pairs and a linear layer predicts the semantic division. This general architecture is shown in Figure 2.

We experimented on both contemporary and historical corpora. Concerning the contemporary corpora, we used the Diseases and City subsets of the Wiki-section semantic segmentation dataset [1]. Regarding the historical corpora, we utilized the NewsEye [11] dataset, where we evaluated the semantic segmentation

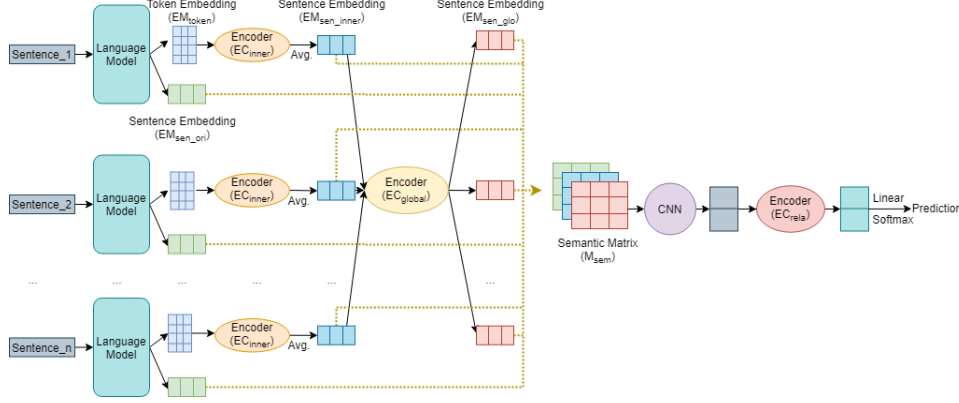


Fig. 2. General architecture of Global-SEG

on Finnish⁴ and French⁵ historical texts. In addition, we tested the effect of text position on semantic segmentation since the Newseye dataset is a multimodal corpus containing both text and images. This idea was inspired by the Layoutlm model [32], which feeds both word embeddings and position embeddings of text into BERT [8] for document analysis and achieves excellent performance on related tasks like table comprehension and classification. Overall, Global-SEG outperforms benchmarks in each corpus and verifies the importance of analyzing global semantic relationships. Furthermore, we discuss the effects of including text location information and the use of large language models (LLMs) into Global-SEG⁶.

The paper is organized as follows: In Section 2, we review existing research related to text semantic segmentation. Section 3 delineates the architecture of the model, accompanied by a comprehensive description of the data. The experimental setup and evaluation metrics are expounded upon in Section 4. Section 5 encompasses the presentation of results and related discussions. Section 6 offers the conclusion, highlights potential avenues for future research and provides the analysis of the limitations of our work.

2 Related Work

Text semantic segmentation typically consists of computing the text semantics (converting text into a vector) and then segmenting the whole text into blocks based on the semantic similarities. First researches implemented this task using a statistical approach [3], which constructs incremental exponential models to

⁴ <https://zenodo.org/record/5654858>

⁵ <https://zenodo.org/record/5654841>

⁶ The code of this paper is available at https://github.com/WenjunSUN1997/text_seg

detect boundaries by analyzing the semantic changes of topic and cue words. There are also methods based on Bayesian models [6,24], which calculate the text topic by analyzing the word list and the word distribution of the text. Others have tried to build relationship graphs to describe semantic relevance to segment text by unsupervised methods [12].

With the development of neural networks (NNs), the common approach today is to use LMs to project the text into a vector space and then detect semantic transformations. The transformer architecture proposed by [29] has significantly improved the performance of the LMs in various domains. Based on this architecture, several pre-trained models and LLMs such as BERT [8], GPT [22], and LLAMA [26] have emerged, followed by reinforcement learning (RL) models such as GPT-4 [21] and Llama 2-Chat [27].

Based on the token semantics, different models for meaning understanding at the sentence level have been proposed, including InferSent [7], Sentence-BERT [23], and SimCSE [10] as top three solutions. For the detection of the segmentation boundary, the long short-term memory (LSTM) architecture and its variants are heavily employed due to the sensitivity to sequential data and the ability to remember long-term information [1,2].

In the context of detecting semantic changes, the attention architecture of the transformer architecture has been also explored. In a similar manner, the cross-segment method [19] employs the token embedding of multiple sentences as input to the transformer for semantic processing. Furthermore, researchers have attempted to directly input sentence vectors to accomplish this task [13,18]. However, these approaches primarily focus on the relationships between individual sentences while neglecting the semantic relationships among pairs or groups of sentences. Additionally, they do not leverage both token-level and sentence-level LMs to compensate for the limitations of each approach. To overcome these challenges, we propose a novel architecture that takes advantage of the global semantic relations information for text semantic segmentation.

3 Model

Global-SEG is composed of an intra-sentence encoder (EC_{inner}) that is used to process the token-level semantics, and a global semantic encoder (EC_{global}) set to process the relation between individual sentences of the whole document. Both sentence semantics conform a semantic embedding matrix (M_{sem}) for integrating and representing the semantics of each level more efficiently. Since semantic segmentation occurs between two semantically different sentences, the segmentation point is determined between sentence pairs. A CNN (CNN) is used for sentence pair information processing. Finally, an encoder (EC_{rela}) is set to model sentence pair relations and amplify the difference between the sentence pairs where the segmentation point is located and other semantically harmonized sentence pairs.

Directly capturing the semantics of a sentence using a pre-trained LM has shown to be a challenge [16]. Furthermore, contemporary LMs, trained on present-

day corpora, lack the robustness required to handle the inherent noise in historical text [5,25]. The inclusion of the EC_{inner} encoder serves to address these issues, and its efficacy has been empirically validated [5].

EC_{global} is used to aggregate the various levels of embeddings and process the semantics of individual text units from a global perspective. Then, these semantic embeddings obtained from different perspectives form M_{sem} and the convolution results of semantic unit pairs are obtained by CNN . The reason for using both token- and sentence-level LMs is that the former can be combined with an encoder to reduce the noise and semantic bias of historical text and digitization errors but lacks the ability to describe the semantic similarity at a sentence-level [23]. Conversely, the latter can provide semantic assistance. Building a global semantic embedding matrix for analysis has already yielded positive results on the entity-level relation task [33]; we extend this same concept to global semantic segmentation and formulate M_{sem} .

The complete pipeline of Global-SEG is the following. Initially, a sentence undergoes processing by the LM to acquire both token-level (EM_{token}) and sentence-level (EM_{sen_ori}) embedding representations. Subsequently, the first encoder, denoted as EC_{inner} , operates on the token embeddings, generating the second sentence-level vector representation EM_{sen_inner} as the average of the encoder’s output, i.e., $EM_{sen_inner} = Avg(EC_{inner}(EM_{token}))$.

All the sentence embeddings from the previous encoder EM_{sen_inner} are then fed into the global syntactic encoder EC_{global} to obtain the embedding EM_{sen_glo} that represents the relationship between sentences. After that, the embeddings EM_{sen_ori} , EM_{sen_inner} and EM_{sen_glo} are concatenated to form a semantic embedding matrix

$$M_{sem} = [EM_{sen_ori}, EM_{sen_inner}, EM_{sen_glo}], \quad (1)$$

where $M_{sem} \in \mathbb{R}^{3 \times Sen_num \times Dim}$. Sen_num corresponds to the number of sentences and Dim to the dimension of the semantic representation. For the multimodal data, we also introduced bounding box information EM_{pos} of the text area to enhance the semantics. Since the bounding box information contains two-dimensional coordinates on the top left and bottom right, i.e., two x-axis coordinates and two y-axis coordinates, we set up four embedding layers to embed these four values and use them as position embeddings. Hence, the dimension of the semantic embedding matrix becomes $M_{sem} \in \mathbb{R}^{7 \times Sen_num \times Dim}$.

The CNN is responsible for convolving the semantic embedding matrix M_{sem} to obtain a representation of the sentence pairs, after which the EC_{rela} encoder processes the output of the CNN to express the relationships among the individual sentence pairs.

Finally, linear and softmax layers are used to obtain the labels for the semantic segmentation as follows:

$$Seg = Softmax(Linear(EC_{rela}(CNN(M_{sem}))))). \quad (2)$$



Fig. 3. An example from the NewsEye dataset. Different articles are marked with different colors. As shown in the red area, in addition to the text block’s text annotation, there is also its bounding box annotation.

4 Experiments

In this section, we first introduce the datasets used for the experiments and then describe the evaluation metrics used to measure Global-SEG’s performance as well as other experimental details.

4.1 Datasets

Our experiments were performed over two dataset types: contemporary and historical corpora.

Contemporary Corpus We utilized the Diseases and City subsets of Wikisection [1] sourced from Wikidata’s disease category in English. This dataset encompasses abstracts and complete texts of pertinent articles and was used on the Sector [1] and Transformer² [18] methods.

Historical Corpus We selected the Finnish (fi) and French (fr) sections of the Newseye dataset [11]. This dataset, whose texts are derived from media sources (i.e., newspapers) between 1848 and 1918, was compiled by the national

libraries of France, Finland, and Austria. It contains newspapers in German, French, and Finnish; however, only the sections in French and Finnish are publicly available. Evaluation campaigns like HIPE-2022 [9] have used this dataset to test the capacity of state-of-the-art LM-based systems to process historical texts given that the semantic information of numerous tokens has undergone changes over time. Since most LMs are trained on contemporary corpora, they are likely to encounter challenges in accurately embedding the desired semantics of certain historical content. Furthermore, this dataset contains substantial noise, including OCR errors, which introduces additional hurdles for individual text semantic segmentation models. Leveraging the multimodal nature of this corpus, which incorporates both text and image data with specific textual locations labeled within the images, we explore the integration of positional embeddings into Global-SEG as an additional aid to enhance text semantic comprehension. An example can be seen in Figure 3.

For the division of the contemporary dataset into train, development, and test, we used the official 70%, 20%, and 10% partitions. Meanwhile, we divided the different languages of the historical corpus into train (60%), development (20%), and test (20%). Table 1 provides the number of articles (Docs), sentences, paragraphs, and average number of sentences per paragraph (Avg. sents) of each dataset.

Table 1. Statistics of each dataset

Dataset	Docs	Sentences	Paragraphs	Avg. sents
Diseases	3,590	237,671	24,248	9.80
City	19,539	1,238,133	114,103	10.85
Newseye (fr)	182	50,698	6,792	7.46
Newseye (fi)	200	22,042	6,348	3.47

4.2 Metrics

P_k [3] is a metric specifically designed to measure the result of segmentation. It uses a sliding window of size k to determine whether the nodes on the two edges of the window belong to the same topic. To perform the evaluations we used the default value of k which corresponds to half of the average size of each block of the standard segmentation.

Precision, recall, and F1 are the common evaluation metrics to measure the model’s ability to correctly identify targets in the test set (precision) and the quality of the model’s own prediction (recall). F1 corresponds to the harmonic mean of precision and recall and measures the model’s performance in a comprehensive manner. These mentioned metrics are also widely used in text semantic segmentation [30,17,1,19,14,31], thus our results are directly comparable to the benchmarks.

4.3 Language Models

LMs are responsible for embedding the input textual information into a continuous vector space. Moreover, token-level LMs vectorize individual tokens lacking a comprehensive understanding of the text as a whole. To address this limitation, sentence-level LMs capable of mapping entire sentences into vectors have been introduced. While still based on tokens, they offer superior performance in tasks like sentence meaning comprehension.

Considering the importance of both token-level and sentence-level understanding, our proposed model combines these two approaches. For token-level modeling, we selected BERT, while for sentence-level modeling, we opted for Sentence-BERT. Besides, we used the LLaMA-7B version to analyze the impact of LLMs. Given the dimensionality of the semantic vector generated by LLaMA is 4,096 and by Sentence-BERT is 768, we took the average value of individual token embeddings as the initial semantic vector for each sentence when conducting experiments with LLaMA.

4.4 Baselines and Benchmarks

LMs We take into account four models as baselines: (1) BERT, (2) Sentence-BERT, (3) hmBERT [25] and (4) LLaMA. While BERT and LLaMA consider the mean value as the sentence vector, Sentence-BERT takes advantage of the model’s output directly. HmBERT is a language model trained using media sources that includes the historical corpus from the experiment in its training data. We use the cosine similarity to measure the semantic difference and set the threshold to 0.0. This means that if the cosine similarity between 2 sentences is less than 0.0, segmentation is needed. Otherwise, the two sentences belong to the same text block.

Sector [1] is a model that relies on Bi-LSTM and sentence vectors. Firstly, it obtains the sentence semantics with a LM which is then fed to the Bi-LSTM network to derive a topic embedding. Then, a classification layer is employed to analyze these embeddings and identify topic shifts within the text. In our experiments, Sentence-BERT is used to vectorize the sentences.

Transformer² [18] is an optimization of the two-level transformer model [13]. In the original model, token embeddings are obtained using fastText [4] and input along with token positions into a transformer to generate sentence vectors. A second transformer is then employed to further process the sentence vector and perform the final prediction. In contrast, Transformer² improves the process of obtaining sentence vectors. Using BERT or its variants, the [CLS] vector of the sentence is obtained, after which it is spliced with the [CLS] of the sentence pair to describe the final sentence semantics. Finally, the semantic segmentation is performed by a similar method as the two-level transformer. Since the Newseye dataset has no topic annotation, the topic prediction layer of Transformer² is not involved during the training of the historical corpus.

4.5 Detailed Setup and Hyperparameters

In our experiments, transformers are used as encoders and the number of attention heads for both EC_{inner} and EC_{global} encoders are set to 8, while 4 for EC_{rela} . All the encoders' layers are set to 2. When using BERT and LLaMA, the model's encoder dimensionality is 768 and 4,096 respectively. The backbone LM is not fine-tuned and is frozen during training. The CNN network is of dimensions (2 ; 192) and (2 ; 1,024) for BERT and LLaMA respectively. Batch size is set to 16 and the random seed is set to 3,407 with a window length of 8 and a step size of 7 to load the data for training and validation. We set the learning rate of the AdamW optimizer to 10^{-5} with a dropout of 0.5. In our experiments, we used an NVIDIA RTX A6000 GPU with 48Gb.

5 Result and Discussion

In this section, we first summarize the experimental results according to contemporary and historical corpora. Secondly, we analyze the reasons for these results. In tables 2 to 4, the symbol $\downarrow$ means that lower values are better, while $\uparrow$ is the opposite. The *-Inner*, *-Global*, and *-Rela* mean that the corresponding encoder has been removed for the model. This is used to perform the ablation study. The results of Sector[1] and Transformer²[18] in the contemporary corpus were taken from its original paper, while in the historical corpus, we reproduced the works and used the results from our experiments.

5.1 Results

Contemporary Corpus Global-SEG outperformed the baselines and benchmarks in precision, recall, and F1 for the Diseases and City datasets as reported in tables 2 and 3. It is important to notice that in some cases, when the model achieves a high F1 value it also obtains the worst P_k . This phenomenon is due to the calculation method of the P_k . The window size used in the calculation process is set to the default, which is half the average size of each semantic block. Consequently, during the sliding window process, multiple segmentation boundaries can be assigned to the same window, causing several erroneous predictions of the model to be counted only once. Additionally, the P_k calculation method considers the text at both ends of the window to be correctly assigned to the corresponding semantic block, without considering the accuracy of the segmentation position. Regarding LMs, employing solely BERT yielded the least favorable results across all three datasets. BERT seems to miss a lot of segment boundaries which results in a low coverage compared to the other methods. It is worth noting that Sentence-BERT outperformed LLaMA in F1. It is possible that the performance of BERT and LLaMA models could be improved by optimizing the computational methods and setting new thresholds. However, the experimental results demonstrate that the models trained specifically for the sentence comprehension task can better respond to semantic changes under the current conditional settings.

It is crucial to take into consideration that all models, except BERT, have higher recall than precision in prediction. Meanwhile, Global-SEG achieves the smallest difference between precision and recall and obtains the highest F1, which means that it is more reliable compared to the benchmarks.

Table 2. Results on Diseases dataset. *-Inner*, *-Global*, and *-Rela* are the results of removing the corresponding encoder.

Models	$P_k(\%) \downarrow$	Precision($\% \uparrow$)	Recall($\% \uparrow$)	F1($\% \uparrow$)
BERT	42.06	27.99	4.17	7.26
Sentence-Bert	44.60	21.10	72.03	32.65
LLaMA	55.85	11.75	99.39	21.02
Sector	26.80	—	—	56.70
Transformer ²	18.80	—	—	—
Global-SEG _{BERT}	1.37	97.23	98.89	98.05
Global-SEG _{-Inner}	3.72	87.73	99.76	93.46
Global-SEG _{-Global}	3.67	86.94	99.96	92.99
Global-SEG _{-Rela}	4.08	87.49	99.57	93.14
Global-SEG _{LLaMA}	1.44	95.37	100.00	97.63

Table 3. Results on City dataset. *-Inner*, *-Global*, and *-Rela* are the results of removing the corresponding encoder.

Models	$P_k(\%) \downarrow$	Precision($\% \uparrow$)	Recall($\% \uparrow$)	F1($\% \uparrow$)
BERT	43.58	6.19	5.88	6.03
Sentence-Bert	45.79	18.82	89.07	31.08
LLaMA	43.44	18.00	81.81	29.50
Sector	14.40	—	—	71.60
Transformer ²	9.10	—	—	—
Global-SEG _{BERT}	8.58	62.50	96.63	75.90
Global-SEG _{-Inner}	11.96	49.58	99.35	66.29
Global-SEG _{-Global}	12.59	53.12	100.00	69.38
Global-SEG _{-Rela}	8.34	61.02	98.03	75.21
Global-SEG _{LLaMA}	9.54	60.24	88.23	71.59

Historical Corpus As seen in tables 4 and 5, Global-SEG continues to outperform the benchmarks in terms of F1 and P_k for the historical corpus, while BERT consistently exhibits the least favorable performance. For the French dataset, BERT cannot determine the correct segmentation position at all. This could be related to the threshold that has been set, but overall, BERT has trouble adapt-

ing historical text. On the other hand, LLaMA outperforms Sentence-BERT in the Finnish texts, with a higher F1 of 4.34%.

Introducing positional information into the model (Global-SEG_{Bbox}) resulted in a 3.27% improvement in P_k and a 4.58% increase in F1 for the Finnish dataset. And for the French dataset, compared to Global-SEG_{BERT}, Global-SEG_{Bbox} improved 0.77% in P_k , but was 0.98% less in F1.

Table 4. Results on Newseye French dataset. *-Inner*, *-Global*, and *-Rela* are the results of removing the corresponding encoder.

Models	$P_k(\%) \downarrow$	Precision($\% \uparrow$)	Recall($\% \uparrow$)	F1($\% \uparrow$)
BERT	39.32	0	0	0
Sentence-Bert	45.47	20.77	54.56	30.08
hmBert	57.58	14.58	22.88	17.81
LLaMA	60.67	13.60	98.83	23.91
Sector	29.70	38.98	51.16	44.25
Transformer ²	29.38	31.14	65.44	42.20
Global-SEG _{BERT}	27.97	37.91	62.29	47.14
Global-SEG _{-Inner}	28.72	42.78	43.35	43.06
Global-SEG _{-Global}	27.18	47.23	39.70	43.14
Global-SEG _{-Rela}	27.16	52.78	36.21	42.95
Global-SEG _{LLaMA}	30.08	41.46	42.35	41.90
Global-SEG _{Bbox}	27.20	43.25	49.50	46.16

Table 5. Results on Newseye Finnish dataset. *-Inner*, *-Global*, and *-Rela* are the results of removing the corresponding encoder.

Models	$P_k(\%) \downarrow$	Precision($\% \uparrow$)	Recall($\% \uparrow$)	F1($\% \uparrow$)
BERT	44.77	42.85	0.46	0.91
Sentence-Bert	36.06	40.86	35.02	37.71
hmBert	47.73	31.91	26.42	28.90
LLaMA	55.54	28.07	83.71	42.05
Sector	27.73	50.91	77.50	61.45
Transformer ²	23.59	59.16	71.88	64.90
Global-SEG _{BERT}	18.60	64.81	84.28	73.27
Global-SEG _{-Inner}	32.17	46.60	78.27	58.42
Global-SEG _{-Global}	22.38	58.80	83.35	68.96
Global-SEG _{-Rela}	23.50	57.42	85.82	68.80
Global-SEG _{LLaMA}	21.93	63.70	76.27	69.42
Global-SEG _{Bbox}	14.33	79.67	76.11	77.85

5.2 Discussion

Ablation Study In tables 2 to 5, Global-SEG_{-[Inner|Global|Rela]} refers to the performance of the model when the corresponding encoder is removed. A comparison of the results on the contemporary and historical corpus shows that the segmentation results are all negatively affected when one of the encoders in the model is removed. This suggests that each encoder is playing a role in the semantic segmentation task. In particular, on the historical corpus, the model’s performance shows a significant drop when the encoder EC_{inner} is removed, which further illustrates that the models pre-trained with the modern corpus require additional components for semantic enhancement. However, on the French dataset, the model performance gap is not as large as in Finnish after removing the different encoders, which suggests that the removal of the encoder has had a greater impact on the Finnish than on French.

Impact of LLMs By comparing the results on each dataset it can be seen that the introduction of a LLM does not improve Global-SEG’s performance. The reason for this phenomenon could be explained by the procedure we followed to extract the sentence embedding from LLaMA using the average value of token embedding. LLaMA is designed to target language generation rather than language understanding, so using a more appropriate LLaMA sentence embedding extraction method or switching to a prompt engineering approach may improve Global-SEG performance with LLaMA.

Impact of Historical Corpus By comparing the performance of a model on both the contemporary and historical corpus, the scores are consistently better on the former than on the latter. Part of the reason for this behavior is that the semantic changes in the historical text and the noise produced by the OCR challenge the semantic embedding ability of LMs trained on contemporary data. Also, since the historical data is obtained from newspapers, there are many texts that cannot be accurately located, such as publication date, article author, etc., which may be marked as a separate paragraph, but which are not clearly semantically related to the news text. In addition, the granularity of text annotation varies greatly in the historical corpus, with some texts having only a few words while others having whole paragraphs, which also poses a high challenge to the embedding ability of the LMs and the segmentation model. It can be seen that hmBERT performs much better than the LMs that have not been trained on historical corpora, however it does not outperform Sentence-Bert given that it has not been trained for semantic similarity.

Impact of Text Position and Layout Regarding the historical corpus, we can see that all models perform better in Finnish than in French. Moreover, the introduction of positional embedding improves Global-SEG’s performance of both P_k and F1 in Finnish, while the performance of F1 in French is slightly reduced. This indicates that depending on the complexity of the text layout, simply using text location information does not reliably add semantic information. We have

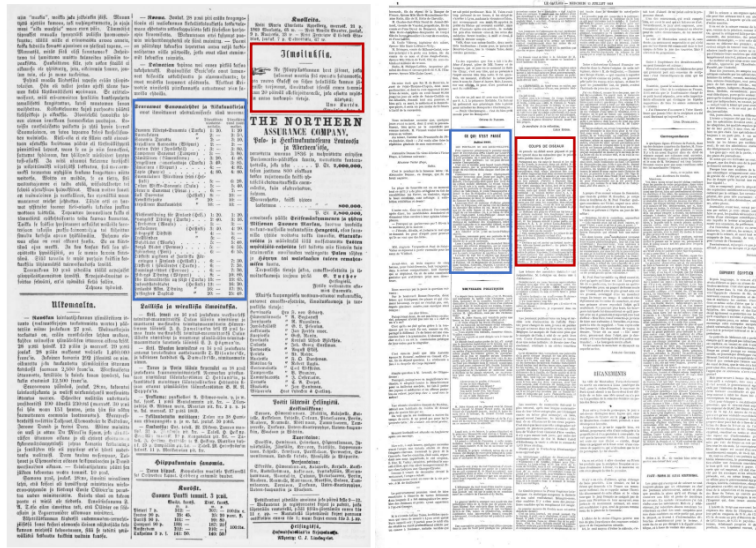


Fig. 4. Example of comparison of two dataset layouts, Finnish newspaper on the left and French on the right

only used a simple embedding layer to vectorize the bounding box coordinates of the text, which could potentially be changed if a more optimal approach is introduced.

As shown in Figure 4, the text layout of Finnish newspapers is less complex than the corresponding layout of French. This means that for Finnish, a simple positional embedding can reflect the text’s positional and semantic relations. However, this approach does not work for the French dataset due to its complex layout. If two adjacent text blocks are in the same horizontal line but not in the same column, they have adjacent coordinates but do not belong to the same semantic segment. In this case, the similar position embedding will instead interfere with the simple semantic vector. Figure 4 also shows the effect of the position information of adjacent paragraphs on the semantic segmentation, where two paragraphs have similar y-axis coordinates, but belong to different semantic blocks. This is why the aforementioned Layoutlm model can enhance the semantics with the positional embedding. As Layoutlm uses fine-grained positions of tokens, adjacent tokens tend to have similar semantics where they are likely to be in the same semantic unit (e.g., the same phrase, sentence, or paragraph). However, unlike Layoutlm, in our experiments with Global-SEG, the location of paragraphs is used.

6 Conclusion

In this paper, we introduced Global-SEG to address the text semantic segmentation task with global semantic relations by combining token- and sentence-level LMs and LLMs. The results over several corpora reveal that Global-SEG outperforms the benchmarks and that including LLaMA does not improve the semantic segmentation ability of the model. We observed that when applied to the historical corpus, the performance of all models (baselines, benchmarks, and Global-SEG) decreases compared to the contemporary corpus. This indicates that historical texts present certain particularities related to spelling, casing, grammar, and semantics that neither LMs nor LLMs manage to capture.

Furthermore, since the historical corpus contains information about the location of the text, we included it to improve the semantics. However, while this addition had a positive impact on Finnish texts, the opposite was true for French texts. We hypothesize that using the bounding box embedding method alone to represent positional information would result in a weakening of the semantic differences between juxtaposed text segments. This aggregation improved Global-SEG’s performance due to the simpler layout of Finnish compared to French, but the presence of more text columns in French data exacerbates the above-mentioned problem.

To the best of our knowledge, current text-semantic segmentation models must follow the reading order, i.e., the input text is sorted in advance according to human reading logic. If the input of these models is the OCR output text of a complex structured text such as a newspaper (no reading order), they will fail the task. This is a problem that needs to be solved for text semantic segmentation. In addition, when dealing with multimodal data, introducing location information alone may require more training data to mitigate the problem of juxtaposed text location confusion, or directly introducing visual embeddings to assist the model.

The biggest shortcoming of Global-SEG is that the inputs have to follow a reading logic and it cannot automatically sort and analyze unordered inputs. This shortcoming means that the model requires higher-quality annotation when facing multimodal text analysis scenarios (e.g., layout+text or image+text). Also, since the embedding of the layout uses only the most direct embedding network, this makes the positional embedding of text blocks with adjacent positions to each other to have a negative effect on the model instead. This is due to the fact that they usually have the same x or y coordinates. In addition, since the texts in the experiments for French and Finnish are taken from historical newspapers and the majority of LMs are now trained using modern corpora, the embedding of historical texts is insufficient. To handle this complication, we added an encoder to further process the text after obtaining the word embedding in order to compensate for the errors in the backbone LM. As a result, the complexity of the model is increased.

Acknowledgements

This work has been supported by the ANNA (2019-1R40226), TERMITRAD (AAPR2020-2019-8510010), Pypa (AAPR2021-2021-12263410), and Actuadata (AAPR2022-2021-17014610) projects funded by the Nouvelle-Aquitaine Region, France.

References

1. Arnold, S., Schneider, R., Cudré-Mauroux, P., Gers, F.A., Löser, A.: Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics* **7**, 169–184 (2019)
2. Barrow, J., Jain, R., Morariu, V., Manjunatha, V., Oard, D.W., Resnik, P.: A joint model for document segmentation and segment labeling. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 313–322 (2020)
3. Beeferman, D., Berger, A., Lafferty, J.: Statistical models for text segmentation. *Machine learning* **34**, 177–210 (1999)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Transactions of the association for computational linguistics* **5**, 135–146 (2017)
5. Boroş, E., Hamdi, A., Pontes, E.L., Cabrera-Diego, L.A., Moreno, J.G., Sidere, N., Doucet, A.: Alleviating digitization errors in named entity recognition for historical documents. In: *Proceedings of the 24th conference on computational natural language learning*. pp. 431–441 (2020)
6. Chen, H., Branavan, S., Barzilay, R., Karger, D.R.: Global models of document structure using latent permutations. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. p. 371–379. Association for Computational Linguistics (2009)
7. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pp. 670–680. Association for Computational Linguistics (2017)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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)*. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). <https://doi.org/10.18653/v1/N19-1423>, <https://aclanthology.org/N19-1423>
9. Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., Clematide, S., Faggioli, G., Ferro, N., Hanbury, A., Potthast, M.: Extended overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In: *CEUR Workshop Proceedings*. pp. 1038–1063. No. 3180, CEUR-WS (2022)
10. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. In: *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. pp. 6894–6910. Association for Computational Linguistics (ACL) (2021)

11. Girdhar, N., Coustaty, M., Doucet, A.: Benchmarking nas for article separation in historical newspapers. In: International Conference on Asian Digital Libraries. pp. 76–88. Springer (2023)
12. Glavaš, G., Nanni, F., Ponzetto, S.P.: Unsupervised text segmentation using semantic relatedness graphs. In: Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics. pp. 125–130. Association for Computational Linguistics (2016)
13. Glavaš, G., Somasundaran, S.: Two-level transformer and auxiliary coherence modeling for improved text segmentation. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 7797–7804 (Apr 2020). <https://doi.org/10.1609/aaai.v34i05.6284>, <https://ojs.aaai.org/index.php/AAAI/article/view/6284>
14. Gong, Z., Tong, S., Wu, H., Liu, Q., Tao, H., Huang, W., Yu, R.: Tipster: A topic-guided language model for topic-aware text segmentation. In: Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part III. pp. 213–221. Springer (2022)
15. Hearst, M.A.: Multi-paragraph segmentation expository text. In: 32nd Annual Meeting of the Association for Computational Linguistics. pp. 9–16 (1994)
16. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., Li, L.: On the sentence embeddings from pre-trained language models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9119–9130 (2020)
17. Li, J., Sun, A., Joty, S.R.: Segbot: A generic neural text segmentation model with pointer network. In: IJCAI. pp. 4166–4172 (2018)
18. Lo, K., Jin, Y., Tan, W., Liu, M., Du, L., Buntine, W.: Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 3334–3340 (2021)
19. Lukasik, M., Dadachev, B., Papineni, K., Simões, G.: Text segmentation by cross segment attention. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4707–4716. Association for Computational Linguistics, Online (Nov 2020). <https://doi.org/10.18653/v1/2020.emnlp-main.380>, <https://aclanthology.org/2020.emnlp-main.380>
20. Moro, G., Ragazzi, L.: Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 11085–11093 (2022)
21. OpenAI: Gpt-4 technical report (2023)
22. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
23. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
24. Riedl, M., Biemann, C.: Topictiling: a text segmentation algorithm based on lda. In: Proceedings of ACL 2012 student research workshop. pp. 37–42 (2012)
25. Schweter, S., März, L., Schmid, K., Çano, E.: hmbert: Historical multilingual language models for named entity recognition. In: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum. pp. 1109–1129. 3180 (September 2022), <http://eprints.cs.univie.ac.at/7549/>

26. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
27. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
28. Utiyama, M., Isahara, H.: A statistical model for domain-independent text segmentation. In: Proceedings of the 39th annual meeting of the Association for Computational Linguistics. pp. 499–506 (2001)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30, p. 5998–6008. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
30. Wang, L., Li, S., Lü, Y., Wang, H.: Learning to rank semantic coherence for topic segmentation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1340–1344 (2017)
31. Xia, J., Liu, C., Chen, J., Li, Y., Yang, F., Cai, X., Wan, G., Wang, H.: Dialogue topic segmentation via parallel extraction network with neighbor smoothing. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2126–2131 (2022)
32. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200 (2020)
33. Zhang, N., Chen, X., Xie, X., Deng, S., Tan, C., Chen, M., Huang, F., Si, L., Chen, H.: Document-level relation extraction as semantic segmentation. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. p. 3999–4006 (2021)