Topic Segmentation in the Wild

Dhuri Shrivastava and **Sharanya Kamath** and **Harjeet Singh Kajal**
College of Information and Computer Sciences
University of Massachusetts Amherst

Abstract

Breaking down a document into multiple contiguous segments based on its semantic structure is an important and challenging problem in NLP which can assist many downstream tasks. In this work, we aim to study and improve the performance of state-of-the-art supervised text segmentation models on challenging out-of-domain datasets which are unstructured. In particular, we aim to improve the generalization of such models by devising novel training strategies.

1 Introduction

Topic Segmentation is the task of splitting the text into meaningful segments which correspond to a distinct topic or subtopic. Natural language texts, especially in unstructured formats such as chat conversations and transcripts, do not have a deliberate separation between contiguous topics. Reliable and accurate division of text into coherent segments can help in making the text more readable as well as searchable.

1.1 Task Description

This project focuses on casting the topic segmentation task as a binary classification problem i.e. designate if a sentence in a document is a start of a segment or not.

1.2 Motivation and Limitations of Existing Work

Topic segmentation of large documents is a useful problem which not only improves document understanding and readability, but more importantly helps many downstream tasks including information retrieval (Llopis et al., 2002) and text summarization.
Due to remote work becoming more common, office meetings held over video conferencing platforms are recorded and either rarely or never viewed again later. The transcripts generated by these meetings would be extremely helpful in increasing transparency and productivity in the workspace, if they are segmented to be more readable. Similar to this, chat conversations are also a kind of unstructured text which, if segmented properly, can yield better insights into the conversation. Since segmenting meeting transcripts can be much more difficult due to multiple people participation, in this project we focus on applying topic segmentation techniques on chat conversations.
There are many supervised and unsupervised learning based approaches proposed to solve this problem.

- Current state-of-the-art topic segmentation models use supervised training but they fail to generalize well on documents which are different from what they have been trained on.
- The current models can not process long sequences that occur when the document has more than four segments.

In this work we try to address these limitations and improve the performance of these models.

2 Related Work

The early research efforts in this area (Hearst, 1997) were focused on unsupervised methods which looked at lexical cohesion within small segments by counting the word repetitions.
Supervised techniques involving many neural approaches have been the focus of many recent work. (Koshorek et al., 2018) proposed a hierarchical neural model where each sentence is encoded using a Bi-LSTM over word tokens, and then a Bi-LSTM over sentence encodings is used to label each sentence as ending a segment or not. The model was trained on wikipedia dataset and showed improvement over the prior unsupervised methods.

(Lukasik et al., 2020) proposed three transformer-based architectures and compared results across different datasets and models. One of them, called the Cross-segment BERT, simply uses local context surrounding a potential segment break to make segmentation decision. The other two uses a hierarchical approach as used by (Koshorek et al., 2018), but with BERT transformers instead of LSTMs. (Solbiati et al., 2021) is another recent work which uses unsupervised approach based on BERT embeddings to segment topics in multi-person meeting transcripts.

We plan to use the hierarchical approach used by Koshorek et al. and improve its generalization on unstructured datasets. This approach consists of a hierarchy of 2 LSTM models, each treated as a separate sub-network. The first sub-network is a bidirectional LSTM architecture which generates sentence representations. A max-pooling layer at the output of this sub-network generates the final sentence representation. The second sub-network is also a bidirectional LSTM and this is used for segmentation prediction. A softmax layer at the output of this network gives us the segmentation probabilities.

3 Proposed Approach

- **Labelling LDC chat dataset**: We will be using LDC chat dataset for the testing of our model as it is an informal and unstructured dataset having conversations between people on different topics. This will be suitable to check the generalization of our model well. As the first step, we will preprocess the LDC chat dataset and synthetically label it to create documents with multiple segments. Each segment corresponds to a chat conversation/sms in the LDC dataset.

- **Implementing/Testing on baselines**: We plan to implement and test few baseline model architectures on the LDC chat dataset we create. This includes the cross segment BERT model and one hierarchical model which encodes sentences on first-level and feeds these encodings to a second-level network which predict if each sentence is a segment ending or not.

- **Fine tuning the models**: We will use a subset of the labelled LDC dataset to fine-tune our model after training it on the wiki727k dataset and compare the performance.

- **Experiment with different segment sizes**: We will check the accuracy of the models on docs with different segment sizes. The wiki-727k on which the models are pre-trained has segment size of 3.48 ± 2.23. We will check our hypothesis that they do not perform well on docs with a larger segment size. We will also compare these results across different loss functions.

- **Test across different test-sets**: To test the generalization of our model, we will collect different unstructured datasets and check accuracy on them. More about test datasets is mentioned in the dataset section.

- **Experiment with loss functions**: To improve the generalization we will experiment with using the focal loss rather than cross entropy loss. This is because cross entropy loss asks the model to be very confident about the ground truth prediction, whereas with focal loss, we do not penalize much if the ground truth probability is greater than a given threshold. We plan to compare results of focal loss vs CE loss.

3.1 Likely Challenges and Mitigations

- Current supervised models cast the topic segmentation task as a binary classification problem (i.e. designate if a sentence in a document is a start of a segment or not). Such supervised models often fail to generalize when the documents during test-time are long (e.g. greater than 4 segments per document and each segment is sufficiently long).

- The models memorize the number of segments encountered during training and fail to generalize well to documents having a greater number of segments than seen during training.

- Current work is not tested on unstructured data such as chats so we don't know if the model generalizes well to unstructured data.

- The available datasets are not structured and to test the generalization of our model, we would need more datasets.

- The P_k **Score** that is used to evaluate the performance of baseline models is sensitive to segment size.

4 Experiments

4.1 Datasets

We will be using text segmentation datasets: WIKI-727K(Koshorek et al., 2018), (Choi, 2000) to train and test our dataset. To reproduce the results and test the current baselines we used WIKI-50. For fine-tuning and testing the models, we will primarily use the BOLT SMS and Chat Data Collection(Song, 2019). To check the generalization of our model, we will also use other datasets like: ICSI Meeting Corpus(Janin et al., 2003), AMI Meeting Corpus(Carletta et al., 2006) and Amazon Topical Chat Dataset: https://github.com/alexa/Topical-Chat.

- WIKI-727K: This dataset contains more than 727,000 documents in the English language with text segmentations created according to the table of contents.

- BOLT: Linguistic Data Consortium(LDC) created this dataset by collecting SMS and Chat data in Chinese, Egyptian Arabic and English. The dataset contains 2140 Egyptian, 7844 Chinese and 9155 English conversations. We will primarily work with the english portion of the data.

- Topical Chat Dataset: This dataset contains human to human conversations in 8 broad topics. It has more than 8000 conversations. The data is split into 5 distinct groups: Train, Valid Frequent, Valid Rare, Test Frequent and Test Rare.

4.2 Evaluation metrics

As we framed the task as a binary classification problem, we report Precision, Recall and F1 measures to check performance.
Precision measures the percentage of boundaries identified by the model that are true boundaries.
Recall measures the percentage of true boundaries identified by the model.
The challenge with Precision Recall is that they are not sensitive to near misses where the prediction is off by one or two sentences. To overcome this, we primarily use the $\mathbf{P_k}$ **Score** (Beeferman et al., 1999) metric which is used by most of the existing literature and hence we will be able to compare our model's performance with the baselines. This is calculated using a sliding window based method which determines whether the two ends of the window are in the same or different segments in the ground truth segmentation. It penalizes for incorrect classification and so lower the P_k score, better is the model.

Table 1: Baseline model test P_k scores

Dataset	**P_k score**
Wiki-727K	22.13
Wiki-50	18.24

4.3 Baselines

We have selected the supervised learning model presented by (Koshorek et al., 2018) as our baseline model. The source code for the implementation of this model is publicly available at https://github.com/koomri/text-segmentation.

This model is trained on the Wiki-727K dataset (Koshorek et al., 2018). Table 1 shows the baseline model P_k results on Wiki-727K and the Wiki-50 datasets.

4.4 Software

We will be using Python to code our solution Python Libraries Frameworks: PyTorch, nltk, sklearn, glob, json, numpy, pandas, transformers, gensim.

Our plan is to extend work on the existing code given by (Koshorek et al., 2018).

4.5 Timeline

- Till 2/25: Read about cross-segment BERT and modifying attention mask in BERT. Run baseline model and reproduce the results on Wiki-727K dataset. Understanding the current methods and literature survey. Writing the formal proposal.

- Till 3/12: Build a preprocessed structured dataset from the LDC dataset. Take reference from preprocessing of Wiki-727K dataset in the supervised learning paper. Test and fine tune the baseline model on the structured LDC dataset. Brainstorm new ideas.

- Till 3/27: Try to generalize the model by modifying attention mask and implementing focal loss. Add parallel computing mechanisms (torch.DataParallel) for re-training the modified model.

- Till 4/4: Explore more unstructured datasets, and fine tune the model on them. Writing mid-term report and documentation.

Since we are in a team, we will divide the work in the following way:

- Harjeet: Preprocess the LDC dataset to make it structured like the Wiki-727K dataset. Explore more unstructured datasets.

- Sharanya: Reproduce the results of baseline model and fine-tune on LDC dataset.

- Dhuri: Implement the cross-segment BERT model and look into parallelizing it.

4.6 Contingency Plan

If the model training is not feasible and does not reach completion, or if the parallelization of cross-segment BERT model is not possible, we will try the biLSTM-biLSTM hierarchial model mentioned in the Section 4.3. We will fine-tune this model on the LDC chat dataset and test it on more unstructured datasets.

5 Acknowledgements

We like to thank Soundar Srinivasan and Samyadeep Basu for guiding us, and providing appropriate resources for this project. We also like to thank Hansi Zeng, our PhD mentor, as well as Prof. Andrew Mccallum, Rico Angell and Andrew Drozdov for helping us out with constructive feedback throughout.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. In *Machine Learning*, pages 177–210.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. *arXiv preprint arXiv:1803.09337*.

Fernando Llopis, Antonio Ferrández Rodríguez, and José Luis Vicedo González. 2002. Text segmentation for efficient information retrieval. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, page 373–380, Berlin, Heidelberg. Springer-Verlag.

Michal Lukasik, Boris Dadachev, Gonçalo Simões, and Kishore Papineni. 2020. Text segmentation by cross segment attention.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Calì. 2021. Unsupervised topic segmentation of meetings with BERT embeddings. *CoRR*, abs/2106.12978.

Dana; Strassel Stephanie; Lee Haejoong; Wright Jonathan Song, Zhiyi; Fore. 2019. Bolt english sms/chat. http://hdl.handle.net/10339/93641.