

# Conversational Topic Segmentation with Clustering-based Intermediate Training

*Taeseung Hahn*



Master of Science  
Informatics  
School of Informatics  
University of Edinburgh  
2022

# Abstract

Topic segmentation is a fundamental NLP task that breaks down the structure of texts into semantically coherent segments. It can enhance the readability of a text and improve downstream NLP tasks such as summarization and retrieval. Recent studies have shown that it is highly effective to fine-tune a hierarchical neural segmentation model, comprising a pre-trained sentence encoder and a segment predictor, on an automatically annotated written text dataset. Nevertheless, in conversational topic segmentation, securing sufficient labels for fine-tuning is difficult because this is highly dependent on manual annotation. As a result, the segmentation model fine-tuned on conversational data performs considerably below its potential. It has been studied that aligning the tasks in two training phases can reduce the minimum amount of labels required for fine-tuning. Accordingly, this work suggests applying clustering-based intermediate training to the topic segmentation model, which bridges the gap between different tasks in pre-training and fine-tuning.

This work shows that clustering-based intermediate training can improve the segmentation performance for conversational data. The intermediate labels generated through clustering show a considerable correlation with the final task labels. In simulated low-resource situations, intermediate training is most effective when the proportion of labelled data is very small. Additionally, we investigate the effect of clustering algorithm settings on intermediate training. The effectiveness of intermediate training changes according to the hyper-parameter: the number of clusters  $k$ . When the number of clusters is excessively small, the cluster information is less informative for the segmentation task. In contrast, as the number of clusters increases several redundant clusters are generated which cause unnecessary noise. Finally, we empirically demonstrate that researchers can exploit their knowledge of the test domain to optimise the hyper-parameter  $k$ . The intermediate training was most effective when  $k$  was set based on the ground truth number of topics in the dataset.

# Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

## Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Taeseung Hahn)

# Acknowledgements

First and foremost, I would like to thank my supervisor Alexandra Birch-Mayne for her guidance and support throughout this project. This project heavily relied on her expertise within the subject area and would not have been the same without her help.

I extend my thanks to all my friends for keeping me company during the long hours in the library, providing me with constant support and encouragement. Finally, I am extremely grateful to my parents and sister for always being there, providing their love and unconditional support.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Problem Statement and Objectives . . . . .	3
1.3	Contributions . . . . .	4
1.4	Dissertation Structure . . . . .	4
<b>2</b>	<b>Background and Related Work</b>	<b>5</b>
2.1	Paucity of Labelled Data . . . . .	5
2.1.1	Manual and Automatic Annotation . . . . .	5
2.1.2	Learning Frameworks . . . . .	6
2.2	Transfer Learning and Advanced Fine-tuning . . . . .	7
2.2.1	Pre-trained Language Models and Fine-tuning . . . . .	7
2.2.2	Advanced Fine-tuning . . . . .	8
2.3	Topic Segmentation . . . . .	10
2.3.1	Unsupervised Models . . . . .	10
2.3.2	Supervised Models . . . . .	11
2.3.3	Models Mitigating Paucity of Labelled Data . . . . .	12
2.4	Clustering Algorithm: $k$ -means . . . . .	13
2.5	Evaluation of Topic Segmentation . . . . .	14
<b>3</b>	<b>Dataset and Task Overview</b>	<b>16</b>
3.1	AMI Meeting Corpus . . . . .	16
3.1.1	Structure of Meeting Scripts . . . . .	17
3.1.2	Corpus Split and the Statistics . . . . .	20
3.2	Task Overview . . . . .	20
<b>4</b>	<b>Methodology</b>	<b>22</b>
4.1	Baseline Model Architecture . . . . .	22

4.2	Experimental Design . . . . .	23
4.2.1	Overview with Training Process Setups . . . . .	24
4.2.2	Simulated Low-resource Conditions . . . . .	25
4.2.3	Clustering: Generation of Intermediate Target . . . . .	26
4.2.4	Details of Each Training Phase . . . . .	29
<b>5</b>	<b>Results and Discussion</b>	<b>31</b>
5.1	Analysis of the Relationship between Targets . . . . .	31
5.2	Analysis of the Effect of Inter-training . . . . .	36
5.2.1	Effect of Labelled Data Proportions . . . . .	36
5.2.2	Effect of the Number of Clusters . . . . .	38
<b>6</b>	<b>Conclusions and Future Work</b>	<b>39</b>
6.1	Experimental Findings . . . . .	39
6.2	Limitations and Future Work . . . . .	40
	<b>Bibliography</b>	<b>41</b>

# Chapter 1

## Introduction

### 1.1 Motivation

Topic segmentation is a fundamental NLP task that divides a given text into topically coherent pieces. This task is illustrated in Figure 1.1, where a Wikipedia text and a meeting script are divided into several segments. In the meeting script, the topic changes from the ‘opening’ to ‘participant roles’ and then to the ‘existing products’. Topic segmentation has received considerable attention due to the following benefits. First, topically segmented text can provide better readability. It is not easy to read and comprehend lengthy, unstructured text. The document structure generated by segmentation can make it easier to comprehend the overall flow of the text. In addition, it has been shown that many other downstream NLP tasks can take advantage of text segmentation including text summarisation [49, 7], information retrieval [39] and topic classification [53]. It can be distinguished from topic classification in that topic segmentation aims to find segment boundaries, whereas topic classification predicts which of the predefined topics a given text belongs to. In this project, we focus on topic segmentation rather than topic classification.

Text data can be largely divided into two different types: written text and conversational data. The written text refers to well-structured data such as books, news articles, blog posts and Wikipedia. In contrast, conversational data refer to text transcribed from conversations between two or more people. Typical examples of conversational data include scripts from business meetings or customer contact centres. Conversational data have several characteristics that are different from written text; it is not only less structured but also includes nonstandard and colloquial expressions [48]. In addition, it tends to contain fewer keywords and sentences that strongly indicate the topics. These

Wikipedia	SPK Dialoge Text
<p><b>[T1] Preface</b> [S1] Seoul is the capital and largest metropolis of South Korea.</p>	<p><b>A:</b> Welcome to the first meeting for our new television remote control. I'd welcome anybody to say something briefly about themselves.</p>
<p><b>[T2] Geography</b> [S2] Seoul is in the northwest of South Korea. [S3] It is roughly bisected into northern and southern halves by the Han River. (...)</p>	<p><b>C:</b> Great! How about starting from the left of the table and going round?</p>
<p><b>[T3] Economy</b> [S4] Seoul is the business and financial hub of South Korea. [S5] It hosts large concentration of headquarters of International companies and banks. (...)</p>	<p><b>B:</b> Ok. I'm the market research person. For this project, I'll be presenting information and statistics on what people want to get from the design, and (...)</p> <p><b>C:</b> I'm User Interface Designer. I haven't had a lot of experience in this kind of thing before but I'll be working on the design.</p> <p><b>D:</b> I'm the Industrial Designer and I'll present about (...)</p>
<p><b>[T4] Culture</b> [S6] Seoul is home to 115 museums. [S7] It includes four national and nine official municipal museums. (...)</p>	<p><b>A:</b> Would anybody like to start by giving their sort of quick views of current remote controls?</p> <p><b>C:</b> I think I find a lot of them really complicated to use with all the different buttons (...)</p>

Figure 1.1: Topic segmentation on written text (Left) and conversational data (Right), where the dashed lines indicate topic changes.

characteristics make topic segmentation for conversational data more challenging; they also make it ineffective to apply a model trained on written text to conversational data. A remarkable property of conversational data that we noted in this project is that, compared to written text, it is more difficult to obtain topic-annotated data. This is because structural elements such as headings, sections, and paragraphs in written text can be utilised as automatic topic annotation, whereas conversational data usually do not contain these structural elements without manual annotation.

Despite these challenges, conversational topic segmentation has many potential benefits in applications. For example, it can prevent the need to read the entire meeting script but provide an understanding of the flow and key points discussed during the meeting. As another example, topic segmentation of transcripts from a customer contact centre can provide an understanding of the typical patterns experienced during these calls, which can improve customer experience [18]. Furthermore, an increasing amount of conversational data has become available. As remote work becomes more common, video conferencing platforms have experienced significant increases in daily participants [25], many of which have recording and transcribing functions. Additionally, the market value of contact centre software with similar features is predicted to expand at a compound annual growth rate of 23.2% [17]. With the potential benefits of application and the growth of available data, there is an imminent need for research on methods that improve conversational topic segmentation models by overcoming the paucity of labelled data.



## 1.2 Problem Statement and Objectives

Despite the potential advantages of conversational data, previous studies have primarily focused on segmenting written text. As mentioned above, it is ineffective to apply the models trained on written texts to conversational topic segmentation because of their characteristic differences. Additionally, training a topic segmentation model on conversational data may result in limited performance due to the paucity of labelled data. This is because the annotation of conversational data depends primarily on hand labelling due to the lack of structural elements such as headings and section titles.

This project aims to verify that the conversational topic segmentation model can be improved by introducing clustering-based intermediate training. Recent studies have demonstrated that intermediate training, inter-training in short, between pre-training and fine-tuning can enhance the performance in the final task. Shnarch et al. showed that clustering-based inter-training can improve classification performance at the document level, especially when the document labels are scarce [38]. However, it remains unclear whether inter-training can improve classification tasks at the sentence level, which is an important part of topic segmentation. In addition, the research showed that inter-training was effective for document classification tasks on topical datasets but less effective on non-topical datasets. Although topic segmentation involves topics, it is a task that predicts segment boundaries, not the topics themselves. Therefore, it is also not certain whether clustering-based intermediate training will also be effective for this task. This project aims to explore these unaddressed research gaps. Additionally, we examine the effect of clustering settings on intermediate training. Finally, we also explore the usefulness of prior knowledge to maximize the effectiveness of intermediate training. In summary, this project aims to answer the following research questions:

1. Can the clustering-based intermediate training improve performance on the segmentation task for conversational data?
2. Can we exploit our knowledge of the test domain during clustering to generate better intermediate labels?

The detailed meaning and underlying assumptions that need to be verified are clarified again after introducing the methodology of this project.

### 1.3 Contributions

With the results of designed experiments, this work showed that clustering-based intermediate training can help improve topic segmentation performance for conversational data. The intermediate labels generated through clustering showed a considerable correlation with the final task labels. In the simulated low-resource conditions, the inter-training was more effective when the proportion of labelled data is small. In the experiment, the evaluation metric  $P_k$  has decreased by 8% from 0.2969 to 0.2731, when the proportion of labelled data is 10%.

In addition, the experiments showed that the effectiveness of intermediate training can vary depending on the hyper-parameter setting in clustering. When the number of clusters  $k$  was small, the cluster information was less informative on the segmentation task. In contrast, when the number of clusters increased, distinguishing certain several clusters did not provide useful information for the final task. These redundant clusters caused unnecessary noise and hurt the final performance. Finally, we empirically demonstrated that researchers can leverage their knowledge of the test domain to optimise the hyper-parameter. Specifically, we set the number of clusters based on the ground truth number of topics, which showed the best performance compared to other settings.

### 1.4 Dissertation Structure

Chapter 2 provides a high-level overview of the background knowledge within the field of topic segmentation as well as the methods used to introduce intermediate training. Chapter 3 illustrates the dataset used in this project and clarifies the terms and the topic segmentation task. Chapter 4 describes our experimental design, including the baseline model architecture and the training process setups in detail. Chapter 5 presents the results obtained from the experiment and provides an analysis of the results. Finally, we conclude by summarising the findings and limitations and suggest future work.

# Chapter 2

## Background and Related Work

This chapter presents a high-level overview and background knowledge concerning the topic segmentation and intermediate training. First, we describe the paucity of labels for conversational data and the general frameworks that have been proposed to address this problem. Then, we describe the standard transfer learning setup, and advanced fine-tuning, which is proposed to reduce the minimum amount of labels required for fine-tuning. Subsequently, we review previous unsupervised, supervised, and semi-supervised topic segmentation models. Finally, we examine the evaluation and metrics for topic segmentation.

### 2.1 Paucity of Labelled Data

The lack of labelled data is a critical bottleneck when improving topic segmentation performance, especially for conversational data. This paucity is the reason that an additional method to utilize unlabelled samples is required for the task. In this section, we investigate why it is difficult to obtain large amounts of labelled samples from conversational data and explore general learning frameworks to mitigate this problem.

#### 2.1.1 Manual and Automatic Annotation

Annotations can be divided into two types: manual annotation and automatic annotation. As its name implies, manual annotation involves human annotators. The label for each sample in the dataset is tagged by annotation experts, which is costly and time-consuming. In contrast, automatic annotations generate labels by using meta-data within the dataset. For example, during the text segmentation task, structural information such

as headings, sections, and paragraph titles can be used as labels to distinguish the topics that each textual unit discusses. To ensure sufficient labelled data for topic segmentation of written text, various datasets including Wiki727k [26] and WikiSection [3] have been proposed, most of which are based on large corpus such as Wikipedia. Recent neural network-based topic segmentation models heavily depend on these automatically annotated training datasets. It has been demonstrated that models trained on these large labelled datasets have made significant performance improvements.

However, automatic annotation is not generally available for conversational data. This is because conversational data are primarily obtained through automatic speech recognition (ASR) and do not contain any structural information that can be used as labels. In addition, segmenting conversational data with models trained on written text data is ineffective due to the differences between the conversational and written text [48]. Therefore, a method that improves the performance with only a small amount of labelled conversational data or a method that leverages unlabelled data to improve the performance is greatly needed in conversational topic segmentation.

### 2.1.2 Learning Frameworks

Various learning frameworks have been proposed to alleviate the paucity of labelled data, including self-supervised, semi-supervised, and weakly-supervised learning.

- Self-supervised learning: a learning framework in which the model trains itself to learn one part of the input from another part of the input [36]. In this learning framework, an unsupervised problem can be transformed into a supervised problem with auto-generated pseudo labels.
- Semi-supervised learning: a learning framework that is used when a fraction of the data is labelled [28]. The pseudo label for unannotated samples can be generated through a weak classifier trained on annotated samples or an unsupervised method such as clustering.
- Weakly-supervised learning: a learning framework in which the given labels are noisy [22]. The noisy labels can be either generated in a self-supervised manner or separately obtained from another source.

Each approach has its own advantages and disadvantages. Although self-supervised learning does not require labelled data, it also cannot leverage labels even when they

are available. Semi-supervised learning provides a way to use not only labelled but also unlabelled data. Weakly-supervised learning is often more robust when using low-quality labels, which can be more easily obtained. Notably, several learning frameworks can be jointly employed to develop a model that addresses a specific downstream task. For example, the pre-training described in the next section is based on self-supervised tasks, and the pre-trained language model can then be further updated for another task in a semi-supervised manner.

## 2.2 Transfer Learning and Advanced Fine-tuning

### 2.2.1 Pre-trained Language Models and Fine-tuning

It is difficult to secure sufficient labelled data for a number of NLP tasks in practical settings. A recent common approach for addressing this issue is to adopt a pre-trained language model, e.g., ELMo [34] or BERT [11], which has shown success in various tasks. A pre-trained language model is a model trained on a large generic corpus to learn general linguistic knowledge. It can be fine-tuned for other downstream tasks with task-specific data. The paradigm of pre-training and fine-tuning is described in the next paragraphs, taking BERT as an example.

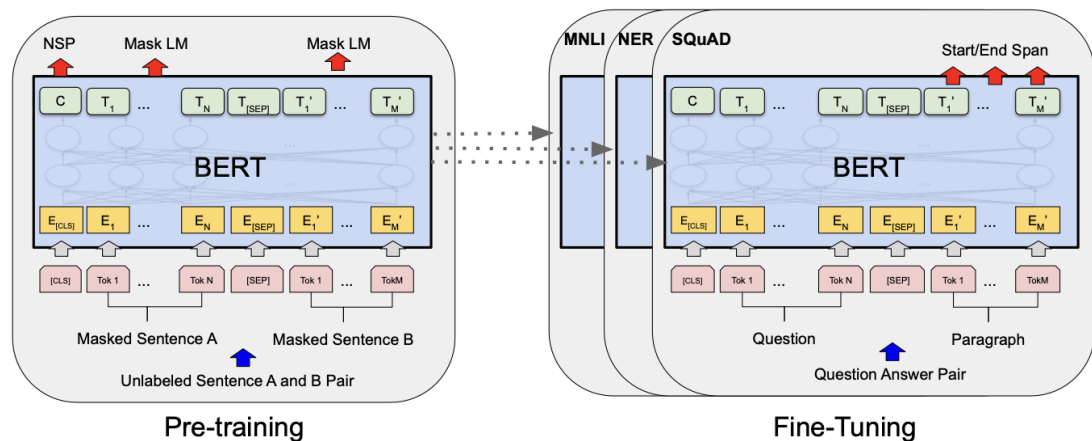


Figure 2.1: Pre-training and Fine-tuning of BERT [11]. (Left) BERT is trained on Masked Language Modelling and Next Sentence Prediction tasks in pre-training. (Right) BERT is trained on downstream tasks with the task-specific data in fine-tuning.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a transformer-based neural network language model trained on BooksCorpus [52]

and English Wikipedia. As illustrated at the left of Figure 2.1, it is trained on two different, but related tasks: Masked Language Modelling(MLM) and Next Sentence Prediction(NSP). In MLM, 15% of the tokens in the input are randomly masked and the objective of the task is to predict these hidden tokens. The objective of NSP is to predict whether two given sentences come from the consecutive text or not. Notably, these two pre-training tasks do not require manually labelled data. In other words, they can be considered tasks that are based on self-supervised learning. It has been demonstrated that, with MLM and NSP tasks, a pre-trained language model can learn a rich hierarchy of general linguistic information such as syntax, semantics and contextual relations between words [23]. Regarding the field of topic segmentation, Solbiati et al. showed that the knowledge learnt from pre-training can also benefit topic segmentation models [41] and many other previous studies introduced pre-trained language model into their model architectures [26, 5, 51, 31, 50, 30].

Although a pre-trained language model can learn general linguistic knowledge, the final target task is different from those of pre-training in many downstream tasks. For example, the segment boundary classification task in text segmentation is different from MLM or NSP in BERT pre-training. In addition, the distribution and characteristics of the final target dataset may be different from those of the large generic corpora used in pre-training. Therefore, it is necessary for the pre-trained model to learn task-specific knowledge. To this end, the pre-trained weights can be further updated on task-specific data and objectives. This training process is called fine-tuning, which is illustrated at the right of Figure 2.1. Notably, unlike pre-training, fine-tuning may require labelled data. This paradigm, which consists of pre-training and fine-tuning, is referred to as a standard transfer learning setup, in that it seeks to transfer knowledge obtained from the source task to the target task. The word “standard” is used to distinguish it from the setup with advanced fine-tuning, which is described in the next section.

## 2.2.2 Advanced Fine-tuning

Recent studies have found that the more different the tasks and data distribution in pre-training and fine-tuning are, the less effective the standard transfer learning setup is [21, 46]. This phenomenon becomes more pronounced when the labelled data for fine-tuning is not sufficient. In other words, if the data distributions and tasks are similar, the model can be optimised for the final task with only a small amount of labelled data. Hence, advanced fine-tuning have been proposed to bridge the gap between the

tasks and data distributions of each training phase. It includes adaptive fine-tuning and behavioural fine-tuning, which are presented in the next paragraphs.

Adaptive fine-tuning is a method for bridging the different data distributions in pre-training and fine-tuning. In this training setup, before the fine-tuning, the pre-trained model is further trained on task-specific data with pre-training objectives such as MLM. This additional training, i.e., intermediate training, does not require task-specific data to be labelled because the pre-training task is performed in a self-supervised manner as described earlier. In contrast, behavioural fine-tuning is performed to bridge the different tasks in two training phases. During the behavioural fine-tuning, the pre-trained model is trained with labelled data for a task similar to the final target task in fine-tuning. The labelled data may be obtained separately from the target dataset. Alternatively, pseudo labels can be generated from the target dataset in a self-supervised manner.

The effectiveness of diverse intermediate training has been studied on various NLP tasks. Glavaš and Vulić showed that explicitly injected syntax via intermediate dependency parsing training can benefit downstream language understanding tasks [16], e.g., natural language inference, paraphrase identification, and commonsense reasoning. They utilised Universal Dependencies Treebank dataset for intermediate training. Garg et al. showed a Transformer model with intermediate behavioural fine-tuning improves stability [13]. They used the ASNQ dataset for intermediate training, which was generated by transforming Natural Questions Corpus [27] to the target task. These studies focused on improving the performance and stability of the fine-tuned model without increasing the model size, utilising separately obtained large datasets in the intermediate training. However, it is scarcely possible to additionally obtain a large labelled dataset concerning the conversational topic segmentation.

Conversely, Arase and Tsujii generated a phrasal paraphrases dataset for intermediate training, by applying the alignment method [1] on the target dataset [2]. They showed that BERT can generate better representations for semantic equivalence assessment by explicitly injecting phrasal paraphrase relations via intermediate training. Shnarch et al. demonstrated that advanced fine-tuning can improve the performance of document classification tasks [38]. They experimented with adaptive fine-tuning via additional Masked Language Modelling on target data and behavioural fine-tuning with pseudo-label. The pseudo-label was generated by applying a clustering algorithm to the target dataset. They showed that BERT with clustering-based intermediate training can significantly improve document classification performance in topical datasets.

In the context of conversational topic segmentation, we note clustering-based inter-

mediate training for the following reasons:

- It does not require an additional large dataset related to the target task.
- It generates intermediate labels via a clustering algorithm. Clustering is a general method that can be applied to most datasets, including conversational data.

We expect that the topic segmentation model can be improved through clustering-based intermediate training, based on the assumption that the cluster label classification task is related to the segment boundary classification task. That is, we assume cluster labels are related to the segment labels. In addition, unlike the experiments from Shnarch et al., the clustering is performed at the utterance level rather than the document level because the objective of the task is to predict if each utterance is a segment boundary. The definition of an utterance and document are presented in Section 3.1.1.1. The objectives of this project include verifying if clustering at the utterance level can generate intermediate labels correlated with the segment boundary labels.

## 2.3 Topic Segmentation

Topic segmentation is the task of dividing a given text into a linear sequence of topically coherent units. A wide variety of methods from different perspectives have been proposed to address this task, and they can be roughly divided into supervised and unsupervised approaches. More recently, several models for mitigating the paucity of labelled data have been proposed.

### 2.3.1 Unsupervised Models

Because of the paucity of labelled training samples, earlier studies on topic segmentation focused on unsupervised approaches, and they usually involved exploiting text similarity measures. A very well-known earlier model is TextTiling [19], which utilises lexical co-occurrence and distribution. It consists of three parts: Tokenisation, Lexical scoring, and Boundary identification. Each of the parts is described as follows:

- Tokenisation divides the given input text into individual lexical units.
- Lexical scoring computes lexical scores based on lexical similarity within the adjacent sentences.
- Boundary identification involves identifying boundaries based on depth scores. The depth scores indicate how rapidly the lexical cohesion between two adjacent sentences decreases.



That is, TextTiling is based on a simple assumption that when the topic changes, the lexical distribution changes as well. C99 uses cosine similarity to calculate the inter-sentence semantic similarity [9] and LCSEg employs lexical chains to segment texts [12]. Despite minor differences in the way similarity is calculated, these methods are common in that they exploit text similarity measures to identify segment boundaries.

Despite the advantages of not requiring labelled data, unsupervised models have reported limited performance compared to supervised models [3, 5, 51]. The critical drawback of these approaches is that labelled samples cannot be utilised even when they are available.

### 2.3.2 Supervised Models

Recent studies in the field of NLP have demonstrated that formulating problems as a supervised learning task for large amounts of labelled data is significantly effective compared to non-supervised or heuristic-based methods [26]. Several studies on topic segmentation have shown remarkable progress in the same manner, especially by devising supervised neural models. Accordingly, there have been many efforts to secure sufficient labelled data [9, 12, 8, 14]. The most recent datasets are Wiki-727k [26] and WikiSection [3], which were derived from a large corpus, namely, Wikipedia.

Based on these large datasets, many supervised topic segmentation models have been proposed. Most of the recent remarkable models are based on hierarchical neural networks and formulate topic segmentation as a binary classification problem. These models generally consist of two sub-networks: The lower-level sub-network, a sentence encoder, aims at generating contextualised sentence representation from word representations. The upper-level sub-network is a label predictor, that predicts whether the given sentence is a segment boundary based on the generated sentence representation.

Many of these models have adopted long short-term memory (LSTM) [20] and its variants in their architecture. Koshorek et al. used a bidirectional LSTM and max pooling to obtain sentence embeddings for the sentence encoder [26]. At approximately the same time, Badjatiya et al. approached topic segmentation in a similar manner, but they used attention-based Bi-LSTM to make the model better consider the context of each sentence [4]. Following these studies, Xing et al. proposed a coherence-related auxiliary task to better model the context [51]. Additionally, they expanded the idea of restricted self-attention proposed by Wang et al. [47] from the word level to the sentence level so that the model absorbs more information directly from adjacent sentences. This

restriction can enhance topic segmentation because segment boundaries do not greatly depend on long-distance content and thus, long-distance signals may cause unnecessary noise. In addition, a pre-trained BERT encoder was also added to the sentence encoder to improve the model's generality.

The Transformer architecture [44] has been widely adopted for sub-networks in more recent models. Glavaš et al. proposed a model known as Coherence-Aware Text Segmentation (CATS) [15], in which a two-level Transformer is devised to obtain transformed sentence representations instead of Bi-LSTM. Similar to prior studies, they also noted that textual coherence is inherently tied to text segmentation. In addition, to make the model better consider textual coherence, the segmentation objective is augmented with the coherence-based objective. They demonstrated that the two-level Transformer-based model outperforms LSTM-based models and that auxiliary coherence modelling can further improve the performance. Then, Lo et al. proposed Transformer<sup>2</sup>, which has similar architecture to CATS [30]. It uses Transformer blocks on both the bottom-level sentence encoder and upper-level segmentation model. The augmented loss for predicting segment labels and topic labels is used to train the segmentation model. In other words, Transformer<sup>2</sup> predicts not only the segment boundaries but also the topic of the segments.

Although these models have demonstrated remarkable performance, they require large amounts of labelled data for training. Accordingly, these models have limited performance in conversational topic segmentation, for which labelled data is scarce. Therefore, additional methods for alleviating this issue are required.

### 2.3.3 Models Mitigating Paucity of Labelled Data

The learning frameworks described in Section 2.1.2 can help to address the paucity of labelled data in conversational topic segmentation. Wang et al. proposed an intuitive strategy that automatically generates a pseudo training dataset for supervision [45]. Similarly, Xing and Carenini proposed a strategy to create a training dataset for the utterance-pair coherence scoring task based on two assumptions [50]. They assumed that adjacent utterances and utterances in the same segment should have higher coherence. In other words, they formulated a coherence scoring task in a self-supervised manner without labelled data and utilised the coherence score to finally identify segment boundaries. Soleimani and Miller proposed semi-supervised multi-label topic models for document classification and sentence labelling [42]. Although this model primarily

aims to extract textual units that are relevant to a specific topic from an entire document, topic inference at the sentence level can also be conducted. The topic segmentation can be performed based on the predicted topic for each sentence. More recently, Takanobu et al. proposed a neural topic segmentation and labelling model based on reinforcement learning, in which segmentation is formulated as a weakly-supervised learning task [43].

The model proposed in this project can be considered another approach to alleviating the paucity of labelled data, by bridging the gap between pre-training and fine-tuning tasks via intermediate training. It is based on a semi-supervised framework in that it leverages not only labelled but also unlabelled samples in the intermediate training, and then trains the model on labelled samples in the fine-tuning phase.

## 2.4 Clustering Algorithm: $k$ -means

Clustering is the task of dividing a set of objects into several coherent groups so that similar objects are contained in the same group. There are several algorithms that perform this task because clustering is a general grouping task.  $k$ -means clustering is one of the most common clustering algorithms, which has many variants. It is notable that clustering algorithms, including  $k$ -means, are usually unsupervised methods, which means that the clustering process can be applied not only to labelled samples but also to non-labelled samples.

$k$ -means clustering is an iterative algorithm that aims to find a local optimum in each iteration. After the initialisation of centroid locations, it iterates the following two processes until a specific termination condition is met.

1. Assign each data point into the cluster whose centroid is closest to the data point.
2. After assigning all data points into the clusters, the centroids of every cluster are re-computed.

The result of  $k$ -means clustering can be changed depending on the settings including centroid initialization strategy and the number of clusters. In the experiment, we use ‘k-mean++’ as our default initialization strategy. In terms of cluster counts  $k$ , the objectives of this project include investigating how  $k$  changes the effectiveness of the inter-training. The detailed settings for clustering in this project are presented in Section 4.2.3.

## 2.5 Evaluation of Topic Segmentation

In this work, topic segmentation is defined as a binary classification problem; more precisely, it is a sequence labelling task at the utterance level. There are a few classical metrics that are widely used for classification problems, including accuracy, precision, recall, and F1 score. However, these metrics have limitations in that they only can measure if the predictions are correct but cannot measure how close incorrect predictions are to the ground truth. In topic segmentation, predictions near the true segment boundary should be regarded as better predictions than those far from the ground truth.

To avoid the limitations of classical metrics, Beeferman et al. proposed  $P_k$  score [6], which is used as the evaluation metric for topic segmentation in this project.  $P_k$  is a sliding window-based probabilistic metric, based on the notion that one segmenter is better than another if it can better identify when two sentences belong to the same topic. It measures how many mismatched pairs of sentences exist between the predicted segmentation and the ground truth.  $k$  indicates how far apart the two sentences in a pair are. The detailed meaning of  $P_k$  is described in the next paragraph, taking the segmentation illustrated in Figure 2.2 as an example.

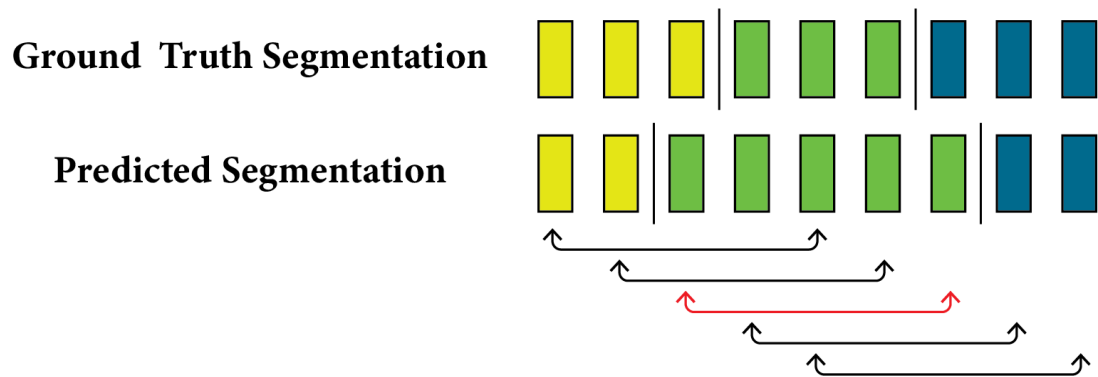


Figure 2.2:  $P_k$  is a window-based error metric. It measures how close the predicted segmentation is to the ground truth segmentation.

In Figure 2.2, the rectangles and their colour indicate sentences and topics. The black vertical bars indicate topic changes. There are nine sentences with three topics in the figure. Therefore, we can obtain five sentence pairs with  $k=4$ . The total number of pairs, five, becomes the denominator of  $P_k$ . The number of mismatched pairs between the predicted segmentation and the ground truth then becomes the numerator. For example, in the ground truth, each sentence in the third pair denoted in the red arrow belongs to different segments. On the other hand, they belong to the same segment in

the predicted segmentation. This result is called a mismatch between the ground truth and the prediction, and it increases the numerator of  $P_k$ . In this example, there are no additional mismatches in the other four pairs. Therefore, the  $P_k$  is finally calculated as  $1/5 = 0.2$ .

As we've seen in the example,  $P_k$  measures the number of mismatched pairs out of the total number of pairs. Thus, it has a value in the range of  $[0, 1]$ . Values closer to zero indicate that the predicted segmentation is closer to the ground truth because  $P_k$  is increased by the mismatch between the ground truth and the predicted segmentation.

In this project, the window size  $k$  is set to half of the average ground truth segment length as suggested by Pevzner and Hearst [35]. This setting is the same as that of many previous studies [37, 14, 26, 51]. In our experiment, the first segment boundary of each script is excluded from the evaluation, because it is always the first sentence of the script.

# Chapter 3

## Dataset and Task Overview

This chapter presents detailed information about our conversational dataset, the AMI Meeting Corpus [32], which was recorded and collected to support multi-disciplinary research. This corpus was chosen because (1) it is topic-annotated natural conversations that occurred in the real world, (2) it involves conversations in a practical domain: business meetings, and (3) it provides a higher degree of variability in speaking patterns due to the large proportion of meetings in which non-native English speakers participated.

Then, this chapter provides a brief overview of the topic segmentation task in the context of the chosen dataset. In this project, topic segmentation is regarded as a sequence labelling task. The objective of this task is to predict whether each utterance given as input is a segment boundary or not. Again, the final goal of this project is to verify if clustering-based inter-training can improve the model’s performance on this task.

### 3.1 AMI Meeting Corpus

The AMI Meeting Corpus contains 100 hours of meeting records that consist of 171 scripts. Out of the 171 scripts, 139 are annotated with the topic information. Although the entire corpus was collected in two different parts, elicited scenario data and non-scenario data, all of the scripts with topic annotation involve elicited scenario data. Therefore, in this project, we used 139 scenario meeting scripts as our dataset.

In the elicited scenario meeting, there are typically four participants. These participants play four different roles as employees in an electronics company. In the elicited scenario, the company has decided to develop a new type of television remote control because those found in the market are not user-friendly, and are also unattractive and

old-fashioned.

### 3.1.1 Structure of Meeting Scripts

#### 3.1.1.1 Definition of Terms and Hierarchical Structure

Topic	Spk	Uttr	Sent	Caption
evaluation of project's process	A	U1	S1	Okay.
			S2	Uh so let's talk about our bonuses and ...
	B	U2	S3	Mm.
	A	U3	S4	Right, right.
	C	U4	S5	That's it.
			S6	um I think another couple of days holiday ...
equipment issues	C	U5	S7	Let's see if I can get this bloody thing to ...
			S8	Whoops.
	A	U6	S9	Uh maybe we should start cleaning up the ...

Table 3.1: An example of a meeting script from the AMI Meeting Corpus. It illustrates the hierarchical structure of segments, utterances, sentences, and words. The column names 'Spk', 'Uttr', and 'Sent' denote 'Speaker', 'Utterance', and 'Sentence', respectively.

In written text data, a 'word' is defined as a set of alphabetical characters separated by spacing, and a 'sentence' is defined as a set of words separated by specific special characters, e.g., period or question mark. These definitions for word and sentence are the same in conversational data. However, conversational data has one characteristic that is different from the written text: conversations typically take place in the form of two or more participants talking alternately. Accordingly, we need to define the concept of 'utterance' when working with conversational data. In other words, one participant can utter several sentences consecutively at once, which is referred to as 'utterance'. Therefore, a meeting script has a hierarchical structure that consists of segments, utterances, sentences, and words, as illustrated in Table 3.1. The formal definitions of the terms are as follows:

- Script (Document): A set of segments that involve several different topics.
- Segment: A set of multiple utterances.
- Utterance: A set of sequential sentences in which one speaker utters.

- Sentence: A set of words separated by specific special symbols in an utterance.
- Word: A set of alphabetical characters separated by spacing in a sentence.

The dashed line in Table 3.1 represents a topic segmentation. The topic is changed from evaluation of the project's process to equipment issues. One remarkable thing is that topic segmentation in our dataset takes place at the utterance level but not at the sentence level. In other words, segmentation always occurs after the final sentence of an utterance and does not occur in any earlier sentence spoken in the utterance. Therefore, in this project, the minimum unit when identifying segmentation is an utterance.

### 3.1.1.2 Predefined Topics

The AMI Meeting Corpus has 24 topics that were predefined during the annotation phase. For example, each segment may belong to a topic such as 'cost', 'trend watching', or 'equipment issues'. Segments that do not belong to one of the 23 predefined topics were annotated as 'others'. This predefined topic information is utilised to set the candidate numbers of clusters in the  $k$ -means. Considering that the 'others' category can contain several different topics, 30 is chosen as one of the experimental numbers of clusters. 10 and 50 are additionally selected to investigate how the performance changes as the number of clusters increases or decreases.

### 3.1.1.3 Segment Granularity

It is notable that segments can be nested. One segment may include multiple segments of different topics, which means that the granularity of each segment may be different. This structure is described in Table 3.2, which uses the ES2015d script as an example. The segmentation and topic information are denoted by the delimiter '====='. The numbers following the segment delimiter indicate different levels of segment granularity. The larger the number is, the higher the granularity is. In this example, an equipment issue occurred while the participants were discussing the cost. After the equipment issue was fixed, they continued to talk about the cost. Thus, we can confirm that the 'equipment issues' segment is nested by the 'cost' segment.



---

**Speaker: Utterance**


---

=====  
 =====, 1, cost.

PM: Alright, now with that over and done with, our next step is to see if we are ...

=====  
 =====, 2, equipment issues.

PM: And um my computer's frozen. And now it's now (...)

ME: Sorry. Are you gonna do that? Okay.

PM: I'm going to um steal a cable. Um it's it's um

(...)

PM: Okay here we go. So we need to tally up how much our product will cost (...)

ID: Yeah. Yeah.

---

Table 3.2: Segment granularity: the segment of higher granularity 'equipment issue' is nested by the segment of lower granularity 'cost'.

Identifying segment boundaries with a high-level granularity can be more challenging because they are more obscure. In addition, as we have seen in the previous examples, high-level topic changes may occur unexpectedly and disrupt the overall flow of conversation. Thus, considering high-level topics can cause unnecessary noise when identifying low-level segment boundaries, which would hurt the overall performance. Therefore, in this work, we only considered the bottom-level segment boundaries which indicate the overall flow of a conversation.

#### 3.1.1.4 Participant Roles

As mentioned earlier, each participant plays a different role during a meeting. Each of the roles is described as follows [32]:

- **Project Manager (PM):** The project manager is responsible for the overall coordination. The PM also ensures that the project is carried out within the scheduled time and allocated budget.
- **Marketing Expert (ME):** The marketing expert identifies the user requirements, watches market trends, and evaluates prototypes.
- **User Interface Designer (UI):** The user interface designer is responsible for the user interface and technical functions.
- **Industrial Designer (ID):** The industrial designer is in charge of designing how the remote control works including the componentry.

The prior studies showed that considering the speaker’s information, especially the speaker’s role, helps detect segment boundaries [24, 48]. According to these findings, we also utilise speaker information in this project. The speaker’s role is added to each utterance in the script.

### 3.1.2 Corpus Split and the Statistics

Table 3.3 shows the statistical details of the AMI Meeting Corpus. In the experiment of this project, the entire corpus is split into training, development, and test sets in accordance with the division suggested by the official AMI Meeting Corpus website. Therefore, each split contains 95, 24 and 20 scripts, respectively. The size of the training set can vary according to the simulated low-resource conditions, which is explained in Section 4.2.2.

AMI Meeting Corpus	Counts
Total number of meetings	59
Total number of scripts	139
Total number of segments	1,107
Total number of utterances	29,215
Total number of sentences	50,161
The average number of segments in a script	7.96
The average number of utterances in a segment	26.39
The average number of sentences in an utterance	1.71

Table 3.3: The detailed statistics for AMI Meeting Corpus.

## 3.2 Task Overview

In this work, we cast conversational topic segmentation as a sequence labelling problem. In other words, a meeting script is considered to be a set of consecutive utterances, each of which is an input for the segmentation model. The objective of our task is to correctly predict whether each utterance is the start of a new segment or not. This task is illustrated in Table 3.4. It describes the same script shown in table 3.1; however, the speaker information and sentence index have been removed, and a beginning of a new segment has been marked.

Topic	Uttr	Caption	New Segment
evaluation	...	...	...
of project's process	U1	Okay. Uh so let's talk about our bonuses and the ...	0
	U2	Mm.	0
	U3	Right, right.	0
	U4	That's it. um I think another couple of days ...	0
equipment issues	U5	Let's see if I can get this bloody thing to work ...	1
	U6	Uh maybe we should start cleaning up the clay.	0

Table 3.4: Topic segmentation task can be cast as a sequence labelling problem. The column name 'Uttr' denotes 'Utterance'.

The beginning, rather than the end, of each segment, is labelled with 1. This setting is due to the result from our preliminary experiment, which shows that the clustering algorithm can better capture the similarity between utterances indicating a new segment. This setting is different from that used in Xing et al. [51] in which the baseline model of our experiment was incurred. However, the same setting was used in many other studies [3, 31, 41].

# Chapter 4

## Methodology

This chapter illustrates the methodology we used to test our hypothesis and answer our research questions. First, we illustrate the architecture of our baseline model. Then, the detailed settings of the experiments concerning each research question are presented in the experimental design section.

### 4.1 Baseline Model Architecture

The favourable model architecture in recent studies on topic segmentation is a hierarchical neural network that consists of two sub-networks. In the experiment, we use Enhanced Hierarchical Attentional Bi-LSTM Network (HAN) [51] as our baseline model. It was chosen as a baseline because (1) it has the architecture that has been preferred in recent studies, and (2) it is not too complex to verify the effectiveness of inter-training. In addition, this model demonstrated competitive performance not only in the intra-domain evaluation but also in the domain transfer and multilingual evaluation.

This model was proposed to address topic segmentation as a sequence labelling task in a supervised manner. It consists of two sub-networks: an utterance encoder and a label predictor, which are illustrated in Figure 4.1. The utterance encoder generates utterance embeddings from two different pre-trained word embeddings: word2vec [33] and BERT embeddings [11]. In the utterance encoder, self-attentional Bi-LSTM [29] takes the word2vec word embeddings as input and generates utterance embeddings. It is concatenated with BERT utterance embedding and then used as a final utterance embedding. Each of these utterance embeddings is devised for a specific purpose. Self-attentional word2vec embeddings are used to obtain task-specific utterances representations and BERT embeddings are used to better deal with unseen text in the test data, and thus

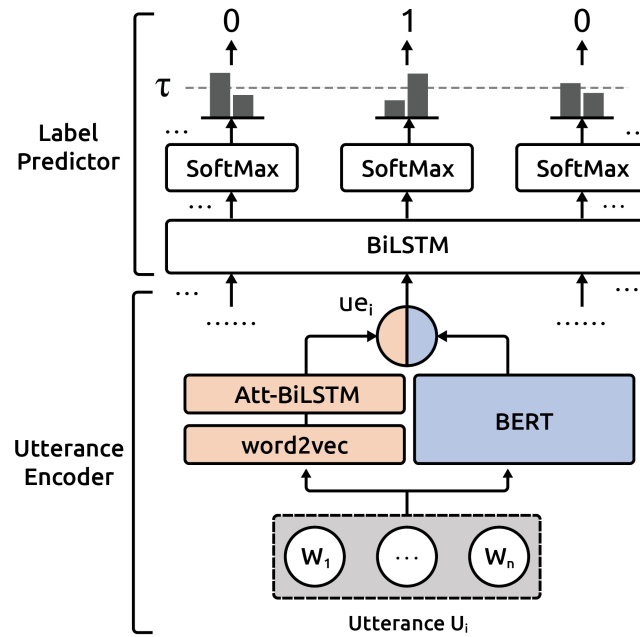


Figure 4.1: The architecture of the baseline model [51]. It consists of two sub-networks: utterance encoder and label predictor.

improve the model’s generality. The concatenated final utterance embeddings are given as input to the label predictor.

The label predictor is a Bi-LSTM network that takes concatenated utterance embeddings as its input and then outputs the probability that the corresponding utterance is a segment boundary. The utterances whose probability is over a threshold  $\tau$  are predicted as segment boundaries, which are denoted as 1 in Table 3.4. The hyper-parameter  $\tau$  is optimized on the development set during the training. To clarify, in the experiment, no modifications were made to the baseline model architecture since this project focuses on the effectiveness of the intermediate training, not a new model with different architecture. Instead, two different training processes were applied to the baseline model, and then their results were compared. These two different training process setups are illustrated in the next section.

## 4.2 Experimental Design

This section presents a high-level overview of the experiments. Subsequently, it provides the details of the experiments and clarifies what research question each setting of the experiment aims to answer. Additionally, the underlying assumptions that need to be verified in the experiment are presented.

### 4.2.1 Overview with Training Process Setups

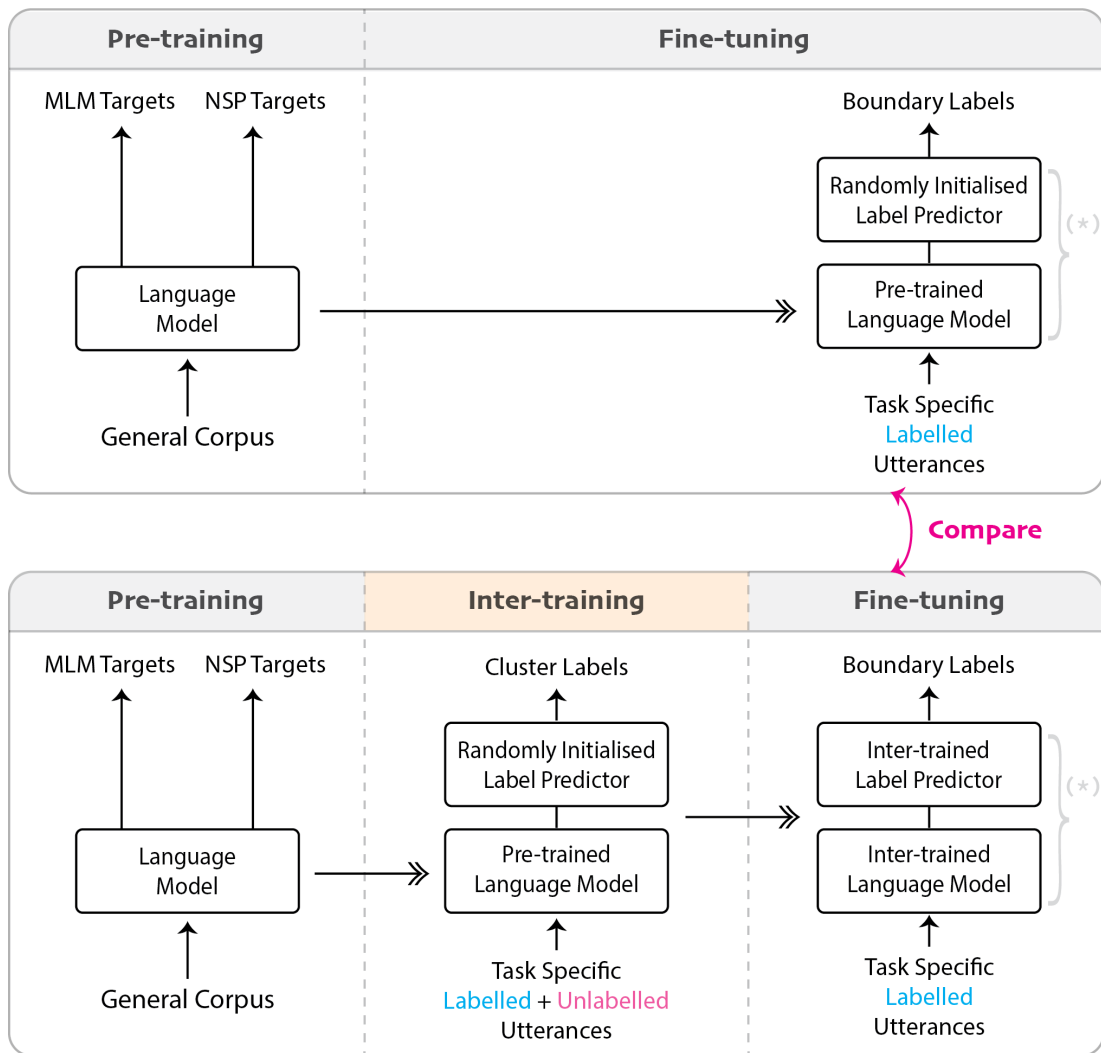


Figure 4.2: Standard fine-tuning setup (Top) and advanced fine-tuning setup (Bottom). Advanced fine-tuning includes intermediate training and standard fine-tuning. (\*) denotes the segmentation model which consists of two sub-networks.

Figure 4.2 illustrates two different training process setups: standard fine-tuning setup and advanced fine-tuning setup. The former is the existing training process for standard transfer learning described in Section 2.2.1. We suggest applying an advanced fine-tuning setup to conversational topic segmentation, which comprises pre-training, clustering-based intermediate training, and fine-tuning. The intermediate training can be considered a type of behavioural fine-tuning described in Section 2.2.2. The primary objective of our experiment is to compare the performance of the baseline model trained through each setup and hence to answer the following research questions:

*RQ*<sub>1</sub> Can the clustering-based intermediate training improve the performance of segmentation for conversational data?

As mentioned earlier, the paucity of labelled data is common in conversational data. To examine the effect of inter-training in practical settings, we simulated low-resource conditions. That is, we repeatedly compared the performances of the baseline model, trained with standard and advanced fine-tuning setups, in the settings of different proportions of labelled data. The detailed settings for low-resource conditions are described in Section 4.2.2.

*RQ*<sub>2</sub> Can we exploit our knowledge of the test domain during clustering to generate better intermediate labels?

Prior knowledge of the test domain was used to set the hyper-parameter of the clustering algorithm. That is, we set the number of clusters  $k$  based on the number of predefined topics in the AMI Meeting Corpus. The segmentation performance was compared with those of models inter-trained with different cluster counts. Additionally, the effect of an increased or decreased number of clusters on the segmentation performance was analysed.

In the following section, the simulated low-resource conditions are depicted. Subsequently, we clarify the terms concerning the labels and illustrate the process of generating intermediate targets through clustering. Finally, we present the detailed settings of pre-training, fine-tuning, and inter-training phases with respect to the research questions.

#### **4.2.2 Simulated Low-resource Conditions**

Although every sample in our dataset is topic-annotated, in a practical setting, it is more common that only a fraction of the dataset has labels. Accordingly, we simulate low-resource conditions in our experiment. To be specific, we fine-tune the model with the different proportions of labelled samples. Six different settings of proportion are used in the experiment, each of which has roughly 100%, 50%, 30%, 20%, 15% and 10% of the labelled scripts. Table 4.1 represents the detailed statistics for six different proportions of labelled data. Notably, the inter-training is not affected by these settings because it is based on the unsupervised clustering algorithm.

<b>Proportion</b>	<b>100%</b>	<b>50%</b>	<b>30%</b>	<b>20%</b>	<b>15%</b>	<b>10%</b>
Total number of scripts	95	50	30	19	15	8
Total number of segments	731	390	264	157	143	64
Total number of utterances	19139	9333	5938	3465	2017	1041

Table 4.1: Detailed statistics for different proportions of labelled data. The proportion is approximately based on the number of scripts, not the number of segments or utterances.

## 4.2.3 Clustering: Generation of Intermediate Target

### 4.2.3.1 Description for Various Labels

To clarify the explanation, three different labels are described as follows:

- Segment Boundary Label

This label indicates whether each utterance is the start of a new segment or not. It can be shortened as ‘Boundary Label’. This is the final target in the topic segmentation task and was used in fine-tuning phase.

- Cluster Label

This label is generated through clustering. It indicates to which cluster each utterance was assigned in  $k$ -means clustering. It was given as a target in the intermediate training phase, thus it can be referred to as an intermediate target.

- Topic Label


This label indicates to which predefined topic each segment belongs. The predefined topics in the AMI meeting corpus are described in Section 3.1.1.2. It was not directly used in our experiment but was used to analyse the relationship with the cluster label.

### 4.2.3.2 Generation of Cluster Label

As mentioned earlier, clustering labels are generated by applying the  $k$ -means algorithm to the utterance embeddings. The utterance embeddings can be obtained by applying various pooling strategies to the pre-trained word embeddings. In our experiment, following previous studies [51, 48], utterance embeddings were obtained by applying mean pooling to the second to last hidden layer of BERT.



k-mean Clustering



Script Index	Utterance Index	Utterance Embedding	Topic Label	Boundary Label	Cluster Label
1	1	(0.45 7.21 ... 0.71)	3	1	C4
	2	(0.12 0.45 ... 0.15)	3	0	C4
	3	(1.42 1.37 ... 6.31)	18	1	C12
	4	(0.94 0.43 ... 0.22)	18	0	0
	5	(0.44 2.71 ... 0.17)	18	0	0
	⋮	⋮	⋮	⋮	⋮
	$m_1-2$	(0.55 2.15 ... 0.11)	5	0	C7
$m_1-1$	(9.32 0.29 ... 7.84)	7	1	C12	
$m_1$	(0.21 0.85 ... 0.42)	7	0	C19	
2	1	(0.24 1.43 ... 0.27)	4	1	C28
	2	(1.52 0.38 ... 3.31)	4	0	C11
	⋮	⋮	⋮	⋮	⋮
	$m_2-2$	(9.32 0.29 ... 2.84)	9	0	C6
	$m_2-1$	(0.81 1.21 ... 0.71)	15	1	C12
$m_2$	(0.75 2.17 ... 1.15)	15	0	C3	
⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	1	(0.55 1.15 ... 0.81)	N/A	N/A	C12
	2	(0.92 0.41 ... 0.21)	N/A	N/A	C8
	⋮	⋮	⋮	⋮	⋮
	$m_{n-1}$	(0.44 2.71 ... 0.17)	N/A	N/A	C21
n	1	(1.42 1.35 ... 3.31)	N/A	N/A	C7
	2	(2.94 0.13 ... 0.22)	N/A	N/A	C2
	$m_n$	(5.12 0.25 ... 4.15)	N/A	N/A	C15

Annotated Samples

Unannotated Samples

Figure 4.3: Generation of cluster labels.

The generation of cluster labels is illustrated in Figure 4.3. According to the simulated low-resource conditions, only a portion of training data has topic labels and boundary labels. Cluster labels can be generated regardless of whether the sample is annotated or not. Accordingly, the inter-training can take not only annotated but also unannotated samples as its training data. In contrast, only annotated samples can be used in the fine-tuning phase. Therefore, the clustering-based inter-training gives the segmentation model a chance to leverage unannotated samples.

#### 4.2.3.3 Settings to Identify Relationships between Labels

We expect the cluster label classification task is related to the segment boundary classification task. Specifically, the first research question ‘ $H_1$  : Can the clustering-based

intermediate training improve the performance of segmentation for conversational data?’ is based on the following assumptions:

- Assumption 1: The intermediate targets (cluster labels) are correlated with the final targets (segment boundary labels).
- Assumption 2: Given that Assumption 1 is true, the information obtained while learning to classify intermediate targets can help better classify final targets.

In the first assumption, we are assuming that the clustering algorithm can assign semantically similar utterances to the same cluster. To verify the assumption, we perform the  $\chi^2$  independence test between (1) the cluster labels and segment boundaries, and (2) the cluster labels and topic labels. Additionally, if the cluster labels are correlated with either segment boundaries or topics of segments, we calculate Cramer’s V statistic to estimate how much correlation exists. Cramer’s V statistic is a measure of association between two nominal variables, giving a value between 0 and 1. The corresponding results can be found in Section 5.1.

#### 4.2.3.4 Settings for Investigating the Effect of Prior Knowledge

In  $k$ -means clustering, the number of clusters  $k$  is a hyper-parameter that we need to set manually. In our experiment, to answer the second research question ‘ $H_2$  : Can we exploit our knowledge of the test domain during clustering to generate better intermediate labels?’, we set  $k$  based on our prior knowledge of the test domain, namely, the number of predefined topics. The AMI meeting corpus has 24 different ground truth topics, including ‘others’. Considering that ‘others’ can indicate several different topics, we set the default value of  $k$  to 30. In order to compare the performance of the default setting, 10 and 50 were selected as candidate values of  $k$ . 50 came from the previous study [38] and was chosen to examine the effect of bigger cluster counts. Similarly, 10 was selected to investigate the effect of smaller cluster counts.

In summary, we trained the baseline model through the standard fine-tuning setup without inter-training. For comparison, we trained the baseline model through the advanced fine-tuning setup with inter-training three times separately. In each inter-training, cluster labels generated with different  $k$  (10, 30, and 50) were used. The performances of trained models were compared based on the  $P_k$  metric. These training and performance comparisons were repeated with six different proportions of labelled data, from 100% to 10%. The corresponding results are presented in Section 5.2.

## 4.2.4 Details of Each Training Phase

### 4.2.4.1 Pre-training Phase

The baseline model utilizes word2vec embeddings trained on the GoogleNews dataset and BERT embeddings trained on Wikipedia and Books corpus. Pre-training refers to the phase where BERT and word2vec models were trained on these large, generic corpus. There was no modification to the standard pre-training phase in our experiments.

### 4.2.4.2 Fine-tuning Phase

In fine-tuning phase, the model was trained on a labelled task-specific dataset, taking utterance embeddings as input features and segment boundary labels as targets. The cross-entropy loss was used as a cost function. In this work, task-specific dataset means AMI meeting corpus. Notably, unannotated samples cannot be used to fine-tune the model because they do not have segment boundary labels. In other words, in Figure 4.3, only the samples denoted in blue can be used in fine-tuning, while the samples marked in red cannot be leveraged.

Fine-tuning is different in each training setup. In the standard fine-tuning, the size of the top layer in the label predictor was set to 2, which indicates whether the input is a segment boundary or not. In the advanced training process, the fine-tuning started from the inter-trained model. The inter-trained model already had the top layer of size  $k$ , which indicates the number of clusters. In this case, the inter-trained top layer was replaced with a randomly initialised layer whose size is 2. The other inter-trained layers in the label predictor remained and were further updated in fine-tuning phase.

In our baseline model, the weights not only in the sentence encoder but also in the label predictor were updated in this phase. In other words, the bottom-level sub-network in the model was trained to obtain utterance embeddings which are more suitable for our task. At the same time, the weights in the label predictor were also updated to better predict the segment boundaries with those utterance embeddings. For convenience, this entire training process is referred to as fine-tuning in our experiments.

### 4.2.4.3 Intermediate Training Phase

The inter-training can be considered to be an approach to bridge the gap between different tasks: MLM and NLP in pre-training and segment boundary classification in fine-tuning. To this end, inter-training was performed between these two phases. To be

specific, before training the model to predict the segment boundary, the final target of our task, we trained the model with cluster labels generated through clustering. As same as fine-tuning, inter-training also used cross-entropy loss as a cost function. The top layer of the segmentation model had the size of  $k$  indicating the number of clusters.

As mentioned earlier, we expect that the model is able to obtain useful information for classifying segment boundaries, while learning how to classify cluster labels, especially from the samples that the model cannot see in fine-tuning phase. These samples are denoted in red in Figure 4.3. In detail, we expect different benefits of inter-training on utterance encoder and label predictor. First, we expect that the utterance embeddings generated after inter-training are contextualized across the cluster labels, which are correlated with segment boundary labels. Secondly, we expect that inter-training can give a better parameter initialization to the label predictor.

# Chapter 5

## Results and Discussion

This chapter summarises research questions that were broken down in the previous chapter. Afterwards, the experimental results and interpretations corresponding to each research question are presented. The detailed research questions and related assumptions can be summarised as follows:

*RQ<sub>1</sub>* Can the clustering-based intermediate training improve the performance of segmentation for conversational data?

*RQ<sub>1-1</sub>* Are the intermediate targets correlated with the final targets?

*RQ<sub>1-2</sub>* Given that *RQ<sub>1-1</sub>* is true, can the information obtained while learning to classify intermediate targets help better classify final targets?

*RQ<sub>1-3</sub>* Does the effectiveness of inter-training change with the proportion of labelled data?

*RQ<sub>2</sub>* Can we exploit our knowledge of the test domain during clustering to generate better intermediate labels?

*RQ<sub>2-1</sub>* When we set the hyper-parameter utilising the knowledge of the test domain, does the model outperform those with different values of hyper-parameter?

*RQ<sub>2-2</sub>* How does the segmentation performance change with an increased or decreased number of clusters?

### 5.1 Analysis of the Relationship between Targets

The section aims to answer the research question ‘*RQ<sub>1-1</sub>* : Are the intermediate targets correlated with the final targets?’ More specifically, this question can be divided into two assumptions to be verified as follows:

- Intermediate targets are directly correlated with final task targets.
- Intermediate targets are correlated with topic labels, which are associated with final targets. In other words, intermediate targets are indirectly correlated with the final task targets.

Table 5.1 presents the result of  $\chi^2$  independence test between (1) cluster labels and boundary labels, and (2) cluster labels and topic labels, in the first and second row, respectively. It is clear that cluster labels are dependent on both segment boundary labels and topic labels.

$\chi^2$ (p-value)	Cluster Labels		
	k=10	k=30	k=50
Boundary Labels	836.38 (<0.0001)	2034.26 (<0.0001)	2030.89 (<0.0001)
Topic Labels	1386.75 (<0.0001)	4030.06 (<0.0001)	5897.22 (<0.0001)

Table 5.1:  $\chi^2$  statistics and p-values obtained by the independence test.

Based on the independence test results, Cramer's V statistics were obtained to investigate how strong the correlations between labels are, and the corresponding results are presented in Table 5.2. Cramer's V measures the association between two nominal variables, giving a value between 0 and 1. It is generally accepted that the range of (0, 0.2], (0.2, 0.6] and (0.6, 1] indicates a weak, moderate and strong association between categorical variables, respectively. Cluster labels showed the greatest correlation with boundary labels and topic labels when  $k$  was 30 and 50, respectively.

Cramer's V	Cluster Labels		
	k=10	k=30	k=50
Segment Labels	0.2090	<b>0.3260</b>	0.3257
Topic Labels	0.0897	0.1053	<b>0.1273</b>

Table 5.2: Cramer's V statistics between the labels.

However, we need to note that the result between topic labels and cluster labels when  $k = 50$  is less reliable. This is because the basic assumption of the test was not satisfied, which is that 'the expected value of each cell should be greater than 5'. In addition,

Cramer’s  $V$  statistics tend to increase as the number of cells in the contingency table between two variables increases, without strong evidence of a meaningful correlation [10]. Despite such tendency, the greatest correlation between boundary and cluster labels was observed when  $k$  was 30, which implies the correlation was strongest when  $k$  was set based on the prior knowledge of the test domain.

A natural explanation for the contribution of inter-training to the final performance is that cluster labels are informative with respect to target task labels. In other words, if the segment boundary distribution varies depending on the clusters, the cluster information can be useful when predicting the boundaries. To examine this, we investigated whether that conditional probability  $P(\text{Boundary} = 1 | \text{Cluster})$  changes according to the clusters. On the other hand, even when some specific clusters have distinct boundary distributions from others, the usefulness of cluster information may be limited, if only a few segment boundaries belong to those clusters. This is because, in this case, cluster information helps predict only a few segment boundaries. Accordingly, we also explore how many segment boundaries belong to the clusters with a relatively high conditional probability.

Tables 5.3, 5.4, and 5.5 show the conditional probabilities and the frequencies of the boundary labels when  $k=10$ , 30 and 50, respectively. These tables are sorted in descending order according to the second column, conditional probability  $P(\text{Boundary} = 1 | \text{Cluster})$ .

<i>Cluster C</i>	$P(\text{Boundary}   \text{Cluster})$		Counts (Cumulative Prop.)			
	Boundary=1	Boundary=0	Boundary=1	Boundary=0		
C2	0.16	0.84	186 (0.25)	962	(0.05)	
C7	0.08	0.92	197 (0.52)	2373	(0.18)	
C4	0.06	0.94	168 (0.75)	2701	(0.33)	
C8	0.04	0.96	61 (0.84)	1592	(0.41)	
C3	0.01	0.99	28 (0.88)	2078	(0.53)	
C6	0.01	0.99	33 (0.92)	2926	(0.69)	
C0	0.01	0.99	23 (0.95)	2152	(0.80)	
C1	0.01	0.99	15 (0.97)	1443	(0.88)	
C9	0.01	0.99	17 (0.99)	1825	(0.98)	
C5	0.01	0.99	3 (1.00)	356	(1.00)	

Table 5.3: Conditional probability and the counts of segment boundaries ( $k = 10$ ).

Table 5.3 shows the result when  $k = 10$ . Using the first row as an example, the total number of utterances belonging to cluster C2 is 1148, and there are 186 segment boundaries and 962 non-segment boundaries. The total number of segment boundaries across the clusters is 731, and thus, those belonging to C2 account for 25% ( $=186/731$ ) of the total number of segment boundaries. The conditional probability  $P(\text{Boundary} = 1 | \text{Cluster} = C2)$  is computed as  $186/1148=0.16$ ; similarly,  $P(\text{Boundary} = 0 | \text{Cluster} = C2)$  is calculated as  $962/1148=0.84$ .

From the second column of the table, it is clear that the conditional probability  $P(\text{Boundary} = 1 | \text{Cluster})$  for cluster C2, which is 0.16, is significantly higher than that of the other clusters. Thus, the segmentation model can assign a higher probability of segment boundary to the utterances in cluster C2. In addition, the segment boundaries in cluster C2 account for 25% of the total number of segment boundaries. Clusters C7, C4 and C8 also have a relatively higher probability, considering that the remaining six clusters have a probability of 1%. The segment boundaries within these four clusters make up 84% of the total number of segment boundaries. Therefore, we can expect that the cluster information help predicts many segment boundaries.

<i>Cluster C</i>	$P(\text{Boundary}   \text{Cluster})$		Counts (Cumulative Prop.)			
	Boundary=1	Boundary=0	Boundary=1		Boundary=0	
C23	0.35	0.65	135	(0.18)	248	(0.01)
C9	0.20	0.80	114	(0.34)	453	(0.04)
C13	0.18	0.82	63	(0.43)	282	(0.05)
C20	0.12	0.88	101	(0.56)	737	(0.09)
C0	0.04	0.96	45	(0.63)	1023	(0.15)
C4	0.04	0.96	20	(0.65)	466	(0.17)
...	...	...	...	...	...	...
C27	0.00	1.00	3	(0.99)	742	(0.96)
C19	0.00	1.00	1	(1.00)	688	(0.99)
C26	0.00	1.00	0	(1.00)	8	(1.00)

Table 5.4: Conditional probability and the counts of segment boundaries ( $k = 30$ ).

Table 5.4 represents the results and analysis when  $k = 30$ . These results make it increasingly clear that the information learnt while training to distinguish clusters can help predict segment boundaries. For convenience, the clusters with a probability of



more than 10% are referred to as clusters with a high probability. Clusters C23, C9, C13 and C20 have significantly higher  $P(\text{Boundary} = 1|\text{Cluster})$  than other clusters. In the case of cluster C23, the conditional probability is 35%, which is remarkably high. Moreover, segment boundaries belonging to these four clusters account for 56% of the total segment boundaries. Therefore, when the model assigns a higher probability of segment boundary to the utterances belonging to the four clusters, it is expected that the segmentation model's overall performance will improve.

<i>Cluster C</i>	$P(\text{Boundary} \text{Cluster})$		Counts (Cumulative Prop.)			
	Boundary=1	Boundary=0	Boundary=1	Boundary=0		
C16	0.39	0.61	74 (0.10)	114 (0.01)		
C17	0.35	0.65	97 (0.23)	184 (0.02)		
C24	0.21	0.79	83 (0.35)	319 (0.03)		
C46	0.15	0.85	21 (0.38)	119 (0.04)		
C49	0.13	0.87	41 (0.43)	276 (0.05)		
C23	0.10	0.90	43 (0.49)	376 (0.08)		
C26	0.10	0.90	27 (0.53)	254 (0.09)		
C11	0.07	0.93	22 (0.56)	300 (0.11)		
...	...	...	...	...	...	...
C32	0.00	1.00	2 (1.00)	643 (0.97)		
C42	0.00	1.00	0 (1.00)	218 (0.98)		
C39	0.00	1.00	0 (1.00)	140 (0.99)		
C10	0.00	1.00	0 (1.00)	2 (0.99)		
C0	0.00	1.00	0 (1.00)	169 (1.00)		

Table 5.5: Conditional probability and the counts of segment boundaries ( $k = 50$ ).

Table 5.5 shows the results when  $k = 50$ . Similar patterns observed in 5.3 and 5.4 are found here as well. However, we can find several trends suggesting that the number of clusters is unnecessarily large. For example, clusters C42, C39, C10, and C0 all contain zero segment boundaries, and thus it is unnecessary to accurately distinguish these four clusters for our final purpose. In addition, more clusters are required to cover the same proportion of segment boundaries compared to when  $k=30$ . Specifically, when  $k=30$ , four clusters with a probability of greater than 10% account for 56% of the total number of segment boundaries. In contrast, when  $k = 50$ , eight clusters with a

probability of greater than 7% are required to account for the same proportion of the segment boundaries.

In summary, there is a correlation between clustering-based intermediate targets and segment boundary targets. In addition, the result showed the greatest correlation when  $k = 30$ , which was set based on the ground truth number of predefined topics. With the same value of  $k$ , the largest amount of segment boundaries benefitted from cluster information. Finally,  $P(\text{Boundary} = 1 | \text{Cluster})$  was significantly high in specific clusters. These results suggest that if the segmentation model can predict the cluster labels, the cluster information can be leveraged to identify the segment boundaries.

## 5.2 Analysis of the Effect of Inter-training

### 5.2.1 Effect of Labelled Data Proportions

This section aims to answer two research questions:  $RQ_{1-2}$  and  $RQ_{1-3}$ . The performance of the baseline model with and without inter-training is presented in Table 5.6 and illustrated in Figure 5.1. It is worth reminding that  $P_k$  is an error-based metric as described in Section 2.5. Therefore, the lower  $P_k$  indicates better, increased performance.

Proportion of Labels	$P_k$				Performance Gain ( $\Delta$ )
	Fine-tuning	Inter-training $\rightarrow$ Fine-tuning			
		$k=10$	$k=30$	$k=50$	
100%	0.2115	<b>0.2091</b>	0.2153	0.2138	0.0024
50%	0.2121	0.2181	<b>0.2020</b>	0.2159	0.0101
30%	0.2280	0.2260	<b>0.2203</b>	0.2238	0.0077
20%	0.2576	<b>0.2481</b>	0.2487	0.2499	0.0095
15%	0.2742	0.2667	<b>0.2600</b>	0.2715	0.0142
10%	0.2969	0.2843	<b>0.2731</b>	0.2831	0.0238

Table 5.6: Summarised result of the experiment. The best performance in each row is marked in bold. The column ‘Performance Gain’ designates the difference in  $P_k$  between the standard fine-tuned model and the best model.

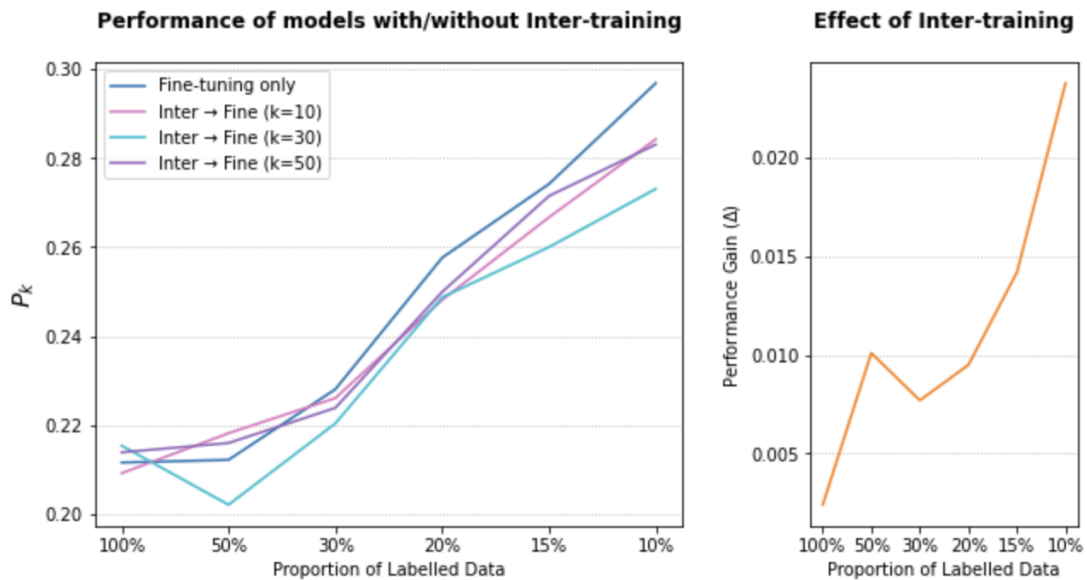


Figure 5.1: (Left) The performance changes of four models according to the proportions of labelled data. The blue line denotes the performance of the model trained through a standard fine-tuning setup. The other three lines designate that of the models trained through advanced fine-tuning setup, with different values of  $k$ . (Right) Performance improvement of the advanced fine-tuned model compared to the model without inter-training, with respect to the proportion of labelled data.

$RQ_{1-2}$  Can the information obtained while learning to classify intermediate targets help better classify final targets?

In Table 5.6, each row denotes the change in the proportion of labelled data. Regardless of the proportions, the best performance of each row is observed from a model with inter-training. Among others, the advanced fine-tuned model with  $k=30$  generally performed best.

From the left plot in Figure 5.1, it is clear that the models' performances decrease as the proportion of labelled data drops. According to the performance without inter-training,  $P_k$  remained nearly the same when the proportion of labelled data was reduced from 100% to 50%. However, it decreased significantly as the proportion decreased to 30%, 20%, 15% and 10%. Similar patterns were found in the models with inter-training, but the decreases were relatively low. That is, the model with inter-training showed relatively better performance as the labelled data decreased.

$RQ_{1-3}$  Does the effectiveness of inter-training change with the proportion of labelled

data?

The right plot of Figure 5.1 represents the effect of inter-training with regard to the proportion of labelled data. An increasing trend is observed, which suggests that the clustering-based inter-training is more effective when the labelled data is scarce.

## 5.2.2 Effect of the Number of Clusters

This section aims to answer the research question ‘ $RQ_2$  : Can we exploit our knowledge of the test domain during clustering to generate better intermediate labels?’. In the experiment, the number 30 was set to the default value of hyper-parameter  $k$ , based on the ground truth number of predefined topics in the AMI Meeting Corpus. With answers to the following two questions, we conclude that the effect of inter-training changes with the number of clusters  $k$ , and the prior knowledge can be leveraged to optimise it.

$RQ_{2-1}$  When we set the hyper-parameter utilising the knowledge of the test domain, does the model outperforms those with different values of hyper-parameter?

Figure 5.1 showed that the model with inter-training outperformed the model without inter-training regardless of the number of clusters. Among others, as suggested in the analysis of the relationship between the boundary labels and cluster labels, the inter-trained model with  $k = 30$  generally performed best. That is, the model utilising knowledge of the test domain outperformed others.

$RQ_{2-2}$  How does the segmentation performance change with an increased or decreased number of clusters?

When the number of clusters  $k$  is increased or decreased from the default setting, the final performance has decreased. The decrease in performance was similar in both cases. These decreases were expected from the result of Section 5.1, because the cluster information was the most informative with respect to the segment boundary label when  $k = 30$ : when  $k = 10$ , the cluster with the highest  $P(\text{Boundary}|\text{Cluster})$  was less distinguishable from others; when  $k = 50$ , some of the clusters were redundant to classifying boundary labels, which may cause unnecessary noise.

# Chapter 6

## Conclusions and Future Work

This work studied the effectiveness of clustering-based intermediate training to alleviate the paucity of labelled data in conversational topic segmentation. In the topic segmentation, the paradigm of pre-training and fine-tuning has been a common practice to make the segmentation model more robust and further improve performance. However, recent studies have found fine-tuning a large pre-trained model often performs considerably below its potential, when the tasks and data distributions in the two training phases are significantly different. It becomes more pronounced when a sufficient amount of labelled data for fine-tuning is not available because the model cannot be fully optimised on the final task. Hence, research is greatly needed to bridge the gap between the tasks of two training phases in conversational topic segmentation, where the labelled data is scarce due to the high dependency on manual annotation.

Advanced fine-tuning, e.g., adaptive fine-tuning and behavioural fine-tuning, has been proposed to address the gap between pre-training and fine-tuning. Nevertheless, the research on applying these methods to conversational topic segmentation is limited. Hence, we suggested applying clustering-based inter-training to this task and then experimented with its effectiveness in various experimental settings. Additionally, we investigate whether the knowledge of the test domain can be leveraged to set an optimal hyper-parameter for clustering.

### 6.1 Experimental Findings

The experimental findings of this dissertation showed that the model with inter-training performs better on topic segmentation. It implies intermediate labels generated through clustering can capture the signal of topic changes. It also suggests that other classifi-

cation tasks at the sentence level can take advantage of clustering-based inter-training. Additionally, we investigated the effectiveness of inter-training depending on the proportion of labelled data. The result showed that inter-training was more effective when the labelled data was scarce. Furthermore, the experiment showed that the effect of inter-training changed according to the number of clusters  $k$ , and we empirically confirmed that the knowledge of the test domain can be used to optimise  $k$ . Specifically, we showed that the segmentation model performed best when  $k$  was set based on the number of predefined topics in the AMI Meeting Corpus.

## 6.2 Limitations and Future Work

There have been attempts to simultaneously address text segmentation and topic classification tasks in a single model [3, 43, 30]. In this project, we focused on topic segmentation, for which it was more difficult to discern whether inter-training would be effective, and did not address topic classification. However, according to the experimental results investigating the relationship between labels, the proposed inter-training is expected to be effective for the topic classification task as well. Accordingly, it is encouraged for future work on the topic classification with inter-training.

In addition, although it has been studied that various pre-trained models can enhance the topic segmentation model's performance [30], the scope of the project is limited to using BERT. Similarly, it has been demonstrated that more sophisticated clustering algorithms, such as sequential Information Bottleneck (sIB) [40], can lead to further improvement on similar tasks [38]. However, we also limited the clustering algorithm to the most intuitive algorithm:  $k$ -means. These limitations were applied to maintain focus on the primary research questions and simplify the experiments. At the same time, they also imply potential performance improvements when using other pre-trained models and clustering algorithms.

Finally, our experiment empirically showed that prior knowledge of the test domain can be exploited to optimise the hyper-parameter of clustering, and thus result in better segmentation performance. However, the selection of other candidates for hyper-parameter,  $k = 10$  and  $50$ , was heuristic and further research on the relationship between the ground truth number of topics and the optimal  $k$  is required. In addition, the effect of inter-training can change depending on other settings of clustering, e.g., the centroid initialization strategy. Therefore, it is encouraged that future work explores more sophisticated methods to find better settings for clustering.

# Bibliography

- [1] ARASE, Y., AND TSUJII, J. Monolingual phrase alignment on parse forests. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Copenhagen, Denmark, Sept. 2017), Association for Computational Linguistics, pp. 1–11.
- [2] ARASE, Y., AND TSUJII, J. Transfer fine-tuning: A BERT case study. 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) (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 5393–5404.
- [3] ARNOLD, S., SCHNEIDER, R., CUDRÉ-MAUROUX, P., GERS, F. A., AND LÖSER, A. SECTOR: A neural model for coherent topic segmentation and classification. Transactions of the Association for Computational Linguistics 7 (2019), 169–184.
- [4] BADJATIYA, P., KURISINKEL, L. J., GUPTA, M., AND VARMA, V. Attention-based neural text segmentation. In Advances in Information Retrieval (Cham, Switzerland, 2018), G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds., Springer International Publishing, pp. 180–193.
- [5] BARROW, J., JAIN, R., MORARIU, V., MANJUNATHA, V., OARD, D., AND RESNIK, P. A joint model for document segmentation and segment labeling. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (July 2020), Association for Computational Linguistics, pp. 313–322.
- [6] BEEFERMAN, D., BERGER, A., AND LAFFERTY, J. Statistical models for text segmentation. Machine Learning 34 (1999), 177–210.

- [7] BOKAEI, M. H., SAMETI, H., AND LIU, Y. Extractive summarization of multi-party meetings through discourse segmentation. Natural Language Engineering 22, 1 (2016), 41–72.
- [8] CHEN, H., BRANAVAN, S., BARZILAY, R., AND 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 (Boulder, Colorado, June 2009), Association for Computational Linguistics, pp. 371–379.
- [9] CHOI, F. Y. Y. Advances in domain independent linear text segmentation. In 1st Meeting of the North American Chapter of the Association for Computational Linguistics (Mar. 2000), Association for Computational Linguistics, pp. 26–33.
- [10] COHEN, J. Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge, 1988.
- [11] DEVLIN, J., CHANG, M.-W., LEE, K., AND 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) (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.
- [12] GALLEY, M., MCKEOWN, K. R., FOSLER-LUSSIER, E., AND JING, H. Discourse segmentation of multi-party conversation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (Sapporo, Japan, July 2003), Association for Computational Linguistics, pp. 562–569.
- [13] GARG, S., VU, T., AND MOSCHITTI, A. TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection. Proceedings of the AAAI Conference on Artificial Intelligence 34, 05 (Apr. 2020), 7780–7788.
- [14] GLAVAŠ, G., NANNI, F., AND PONZETTO, S. P. Unsupervised text segmentation using semantic relatedness graphs. In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (Berlin, Germany, Aug. 2016), Association for Computational Linguistics, pp. 125–130.



- [15] GLAVAŠ, G., AND SOMASUNDARAN, S. Two-level transformer and auxiliary coherence modeling for improved text segmentation. Proceedings of the AAAI Conference on Artificial Intelligence 34, 05 (Apr. 2020), 7797–7804.
- [16] GLAVAŠ, G., AND VULIĆ, I. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (Online, Apr. 2021), Association for Computational Linguistics, pp. 3090–3104.
- [17] GRAND VIEW RESEARCH. Contact center software market size report. Tech. rep., 2022.
- [18] GUPTA, V., ZHU, G., YU, A., AND BROWN, D. E. A comparative study of the performance of unsupervised text segmentation techniques on dialogue transcripts. In Systems and Information Engineering Design Symposium (SIEDS) (2020), pp. 1–6.
- [19] HEARST, M. A. Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics 23, 1 (Mar. 1997), 33–64.
- [20] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. Neural Computation 9, 8 (Nov. 1997), 1735–1780.
- [21] HOWARD, J., AND RUDER, S. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 328–339.
- [22] HUANG, F., AHUJA, A., DOWNEY, D., YANG, Y., GUO, Y., AND YATES, A. Learning representations for weakly supervised natural language processing tasks. Computational Linguistics 40, 1 (Mar. 2014), 85–120.
- [23] JAWAHAR, G., SAGOT, B., AND SEDDAH, D. What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 3651–3657.

- [24] JOTY, S., CARENINI, G., MURRAY, G., AND NG, R. Supervised topic segmentation of email conversations. Proceedings of the International AAAI Conference on Web and Social Media 5, 1 (Aug. 2011), 530–533.
- [25] KARL, K., PELUCHETTE, J., AND AGHAKHANI, N. Virtual work meetings during the covid-19 pandemic: The good, bad, and ugly. Small Group Research 53, 3 (2022), 343–365.
- [26] KOSHOREK, O., COHEN, A., MOR, N., ROTMAN, M., AND BERANT, J. Text segmentation as a supervised learning task. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (New Orleans, Louisiana, June 2018), Association for Computational Linguistics, pp. 469–473.
- [27] KWIATKOWSKI, T., PALOMAKI, J., REDFIELD, O., COLLINS, M., PARIKH, A., ALBERTI, C., EPSTEIN, D., POLOSUKHIN, I., DEVLIN, J., LEE, K., TOUTANOVA, K., JONES, L., KELCEY, M., CHANG, M.-W., DAI, A. M., USZKOREIT, J., LE, Q., AND PETROV, S. Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics 7 (2019), 452–466.
- [28] LIANG, P. Semi-supervised learning for natural language. M.Eng. Thesis, Massachusetts Institute of Technology, Massachusetts, USA, 2005.
- [29] LIU, Y., SUN, C., LIN, L., AND WANG, X. Learning natural language inference using bidirectional lstm model and inner-attention. arXiv preprint arXiv:1605.09090 (2016).
- [30] LO, K., JIN, Y., TAN, W., LIU, M., DU, L., AND 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 (Punta Cana, Dominican Republic, Nov. 2021), Association for Computational Linguistics, pp. 3334–3340.
- [31] MARAJ, A., MARTIN, M. V., AND MAKREHCHI, M. A more effective sentence-wise text segmentation approach using BERT. In Proceedings of International Conference on Document Analysis and Recognition (ICDAR) (Cham, Switzerland, 2021), Springer International Publishing, pp. 236–250.

- [32] MCCOWAN, I., CARLETTA, J., KRAAIJ, W., ASHBY, S., BOURBAN, S., FLYNN, M., GUILLEMOT, M., HAIN, T., KADLEC, J., KARAIKOS, V., KRONENTHAL, M., LATHOUD, G., LINCOLN, M., LISOWSKA, A., POST, W., REIDSMA, D., AND WELLNER, P. The AMI Meeting Corpus. In Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research (Aug. 2005), Noldus Information Technology, pp. 137–140.
- [33] MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. Efficient estimation of word representations in vector space. In 1st International Conference on Learning Representations (Scottsdale, AZ, USA, May 2013), International Conference on Learning Representations.
- [34] PETERS, M. E., NEUMANN, M., IYYER, M., GARDNER, M., CLARK, C., LEE, K., AND ZETTEMAYER, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (New Orleans, Louisiana, June 2018), Association for Computational Linguistics, pp. 2227–2237.
- [35] PEVZNER, L., AND HEARST, M. A. A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics 28, 1 (2002), 19–36.
- [36] RAINA, R., BATTLE, A., LEE, H., PACKER, B., AND NG, A. Y. Self-taught learning: Transfer learning from unlabeled data. In Proceedings of the 24th International Conference on Machine Learning (New York, NY, USA, 2007), Association for Computing Machinery, p. 759–766.
- [37] RIEDL, M., AND BIEMANN, C. TopicTiling: A text segmentation algorithm based on LDA. In Proceedings of ACL 2012 Student Research Workshop (Jeju Island, Korea, July 2012), Association for Computational Linguistics, pp. 37–42.
- [38] SHNARCH, E., GERA, A., HALFON, A., DANKIN, L., CHOSHEN, L., AHARONOV, R., AND SLONIM, N. Cluster & tune: Boost cold start performance in text classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Dublin, Ireland, May 2022), Association for Computational Linguistics, pp. 7639–7653.
- [39] SHTEKH, G., KAZAKOVA, P., NIKITINSKY, N., AND SKACHKOV, N. Applying topic segmentation to document-level information retrieval. In Proceedings of

- the 14th Central and Eastern European Software Engineering Conference Russia (New York, NY, USA, 2018), Association for Computing Machinery.
- [40] SLONIM, N., FRIEDMAN, N., AND TISHBY, N. Unsupervised document classification using sequential information maximization. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 2002), SIGIR '02, Association for Computing Machinery, p. 129–136.
- [41] SOLBIATI, A., HEFFERNAN, K., DAMASKINOS, G., PODDAR, S., MODI, S., AND CALI, J. Unsupervised topic segmentation of meetings with BERT embeddings. arXiv preprint arXiv:2106.12978 (2021).
- [42] SOLEIMANI, H., AND MILLER, D. J. Semi-supervised multi-label topic models for document classification and sentence labeling. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (New York, NY, USA, 2016), Association for Computing Machinery, p. 105–114.
- [43] TAKANOBU, R., HUANG, M., ZHAO, Z., LI, F., CHEN, H., ZHU, X., AND NIE, L. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (July 2018), International Joint Conferences on Artificial Intelligence Organization, pp. 4403–4410.
- [44] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. U., AND POLOSUKHIN, I. Attention is all you need. In Advances in Neural Information Processing Systems (2017), vol. 30, Curran Associates, Inc.
- [45] WANG, L., LI, S., LV, Y., AND WANG, H. Learning to rank semantic coherence for topic segmentation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Copenhagen, Denmark, Sept. 2017), Association for Computational Linguistics, pp. 1340–1344.
- [46] WANG, N., AND LU, J. Aligning the pretraining and finetuning objectives of language models. arXiv preprint arXiv:2002.02000 (2020).

- [47] WANG, Y., LI, S., AND YANG, J. Toward fast and accurate neural discourse segmentation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Brussels, Belgium, Oct.-Nov. 2018), Association for Computational Linguistics, pp. 962–967.
- [48] WEI, H. Topic segmentation for conversational data. M.Sc. Dissertation, University of Edinburgh, Edinburgh, UK, 2021.
- [49] XIAO, W., AND CARENINI, G. Extractive summarization of long documents by combining global and local context. 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) (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 3011–3021.
- [50] XING, L., AND CARENINI, G. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (Singapore and Online, July 2021), Association for Computational Linguistics, pp. 167–177.
- [51] XING, L., HACKINEN, B., CARENINI, G., AND TREBBI, F. Improving context modeling in neural topic segmentation. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Suzhou, China, Dec. 2020), Association for Computational Linguistics, pp. 626–636.
- [52] ZHU, Y., KIROS, R., ZEMEL, R., SALAKHUTDINOV, R., URTASUN, R., TORRALBA, A., AND FIDLER, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. 2015 IEEE International Conference on Computer Vision (ICCV) (12 2015), 19–27.
- [53] ZIRN, C., GLAVAŠ, G., NANNI, F., EICHORTS, J., AND STUCKENSCHMIDT, H. Classifying topics and detecting topic shifts in political manifestos. In PolText 2016: The International Conference on the Advances in Computational Analysis of Political Text : proceedings of the conference (Zagreb, 2016), University of Zagreb, pp. 88–93.