Recent Trends in Linear Text Segmentation: A Survey

Anonymous ACL submission

Abstract

Linear Text Segmentation is the task of automatically tagging text documents with topic shifts, i.e. the places in the text where the topics change. A well-established area of research in Natural Language Processing, drawing from well-understood concepts in linguistic and computational linguistic research, the field has recently seen a lot of interest as a result of the surge of text, video, and audio available on the web, which in turn require ways of summarising and categorizing the mole of content for which linear text segmentation is a fundamental step. In this survey, we provide an extensive overview of current advances in linear text segmentation, describing the state of the art in terms of resources and approaches for the task. Finally, we highlight the limitations of available resources and of the task itself, while indicating ways forward based on the most recent literature and under-explored research directions.

1 Introduction

Linear text segmentation, also known as topic segmentation, is the task of identifying topic boundaries in a text using coherence modeling and/or local cues (Purver, 2011). The attribute 'linear' derives from the fact that in this setting, which is the most popular but not the only one, topics are considered "linearly" as following one another in documents and, as such, *linear* text segmentation ignores any sub-topic or hierarchic structure and focus on finding the boundaries between the topics thus linearly defined. This is also distinguished from topic *classification*, which relates to classifying text with the correct topic class; while linear text *segmentation* is strictly tasked with identifying the part of a text in which a topic boundary occurs. Such boundaries then have a relevant role in a variety of contexts, such as finding individual news stories in a news show or podcast (Ghinassi, 2021) or even as a pre-processing step for tasks like summarization (Zhong et al., 2021).

This survey aims to give a comprehensive, yet brief overview of the field, highlighting the evolution of the approaches used to tackle the task as well as the available metric and resources and what remains to be done. Such a survey is much needed as previous surveys on the topic are mostly outdated at this point (see, e.g., Purver, 2011). Crucially, previous surveys lack an in-depth exploration of the use of language models for the task, where transformer-based language models and Large Language Models (LLMs) have now become, as in other areas of NLP, central for the task. In this survey, then, we aim to fill this gap by showing how the field has slowly shifted to use features from transformer-based language models and supervised learning as the framework of choice and how LLMs are just starting to get traction. In doing so, we will also highlight the various problems of resources and evaluation which, we argue, are central for further developments in the field. Finally, we discuss future directions.

This work is a necessary step for summarising and grounding recent research in the field, while pointing towards future developments which are worth the focus of future research. Note that this survey does not touch upon sub-areas like multi-modality and more niche domains like video-lecture segmentation: we focus on NLP and on the domains in which topic segmentation has traditionally been seen as a central task. Future research might integrate the current work with these aspects.

2 Linear Text Segmentation Approaches

2.1 Basic Units

A first step in designing a linear text segmentation system is deciding which basic unit of text to use as input to the system. Generally, linear text segmentation systems work either at the word, sentence (or pseudo-sentence), or paragraph level.

Research in discourse structure has highlighted that paragraphs usually play crucial roles in conveying different topics in written text (Halliday and Ruqaiya, 1976; Grosz and Sidner, 1986) and, as such, early literature often used the paragraph as the unit (Yaari, 1997). As the technology started being applied to domains such as multimedia content, spoken language, and, in general, text not having paragraph information, however, the role of paragraphs as preferred basic units was progressively superseded by textual features corresponding to words and sentences; or, in early literature, *pseudo-sentences*, in which an arbitrary number of words are aggregated to avoid introducing error from sentence tokenization (now largely a solved task for languages such as English). In the case of multi-speaker scenarios such as most meeting transcripts the preferred basic units are usually *speaker turns*, as segments that are usually sufficiently complete to represent coherent units or at least to convey the communicative intention shared by speaker and hearer, but systems working at the word level have been widely used as well.

Currently, the preference for using word or sentence-based methods seems to be mostly dependent on the type of features being used in end-to-end systems. Models built on word-topic probability distributions (Purver et al., 2006; Sun et al., 2008; Misra et al., 2011) or word embeddings, then, use words as basic units (Koshorek et al., 2018; Arnold et al., 2019; Yu et al., 2023), while models built on sentence embeddings employ sentences or speaker turns (Ghinassi, 2021; Ghinassi et al., 2023b; Solbiati et al., 2021).

2.2 Unsupervised Methods

2.2.1 Count-based Methods

One of the earliest unsupervised techniques for linear text segmentation, TextTiling, used two adjacent sliding windows over sentences and compared the two blocks of sentences inside these windows using cosine similarity between the relative bag-of-words vector representations (Hearst, 1994). The same algorithm has been successfully used with different, more informative sentence representations, such as TF-IDF re-scoring of bag-of-words (Galley et al., 2003). To further improve the individuation of topically incohesive adjacent windows of sentences, the C99 algorithm was proposed (Choi, 2000). This method builds on the intuitions of TextTiling but substitutes the step in which the similarities are scored with a divisive clustering algorithm, improving over the original approach.

Another early approach in topic segmentation was that of using the distance between sentence representations in a dynamic programming framework, including Hidden Markov Models (HMMs). Count-based language models (i.e. n-gram models) were proposed in this context, where the probability of different words under different topics has been used either directly in an HMM framework (Yamron et al., 1998) or using a linear dynamic programming approach as in the U00 system (Utiyama and Isahara, 2001). The most recent approach in this sense, BayesSeg, added probabilistic models of cue phrases to a count-based language model, reaching results that are still competitive (Barzilay and Lapata, 2008). The use of language models, even though in a radically different way, is at the base of the most recent segmentation systems.

2.2.2 Topic Modelling Methods

Early on, researchers combined techniques from the closely related task of topic modelling to perform topic segmentation. The use of topic models for the task falls broadly into the category of generative topic segmentation models, as it shifts the focus from discriminatively identifying areas of low cohesion and local cues, to directly modeling the underlying topics "generating" the different segments in the document (Purver, 2011).

Most early approaches in this sense build on various forms of Latent Dirichlet Allocation (LDA) as a method to automatically individuate topics in text via count-based features (Blei et al., 2003). LDA produces, among its outputs, a matrix of word-topic assignments, storing the probability of each word in the given vocabulary under different topics. Dynamic programming approaches have been widely used in this context. The MM system, for example, used such a framework in conjunction with probabilities derived from word-topic assignments to decide over the most likely topic at each word in the sequence (Misra et al., 2011).

More recently, TopicTiling used word-topic assignments from LDA models to create word vectors and, by aggregating word vectors, sentence vectors to be used as sentence representations for the TextTiling algorithm (Riedl and Biemann, 2012).

An advantage of using topic modelling as a base for topic segmentation is that such algorithms automatically yield the classification of topic segments as a by-product, as the probability associated with different topics can be aggregated at the segment level after segmentation (Purver et al., 2006). Using generative topic models also makes it easier to tackle the task in a hierarchical fashion, where the level of granularity of the topics (and therefore of the segmentation) can be directly controlled (Du et al., 2013). These are indeed properties that do not yet have a parallel in modern end-to-end systems and, as we will see, combining the two paradigms is a research direction worth pursuing.

2.2.3 Embeddings-based Methods

Another more recent strand of research has drawn from improvements in vector semantics and initially used word embeddings to determine the coherence of consecutive words in the context of topic segmentation. This concept has been variously applied in algorithms such as GraphSeg (Glavas et al., 2016), comparing consecutive sentences based on a graph of similarities between their constituent word embeddings.

More recently, the evolution of neural language models has shifted the paradigm from word-based methods to sentence-based ones, in which dense sentence representations are obtained from transformer-based language models like BERT and employed in conventional techniques such as TextTiling (Ghinassi, 2021; Solbiati et al., 2021).

2.2.4 LLM-based Methods

During last year, pioneering work has also been carried out using multi-billion parameter LLMs such as ChatGPT and prompt engineering to treat the problem as a Natural Language Generation (NLG) task (Fan and Jiang, 2023; Yu et al., 2023). The use of LLMs in a zero-shot setting can be classed as an unsupervised method, and it has been shown to outperform all other unsupervised methods after careful prompt optimization (Fan and Jiang, 2023; Jiang et al., 2023). This approach, then, is promising and it should be explored as a way forward to overcome specific limitations of the generally more effective supervised framework described below.

2.3 Supervised Methods

Supervised methods have been present since early on in the field. The surge of these methods, however, coincides with the improvements in neural language modeling and, as such, we limit our description to such methods. For an in-depth discussion of discriminative supervised methods before neural language models, we refer to (Purver, 2011).

2.3.1 Single-Task Methods

As mentioned, advances in neural language models have changed also the landscape of linear text segmentation, as they did for NLP more generally. In the context of linear text segmentation, this meant a progressive shift towards supervised end-to-end systems (typically based on neural architectures) building on strong semantic features like modern word and sentence embeddings, as well as new large datasets to train such systems.

In the supervised setting, the segmentation problem is often treated as one of sequence tagging, where a binary scheme is used to label individual units such as sentences, to individuate where a segment ends or starts.

Among the first such approaches, TextSeg (Koshorek et al., 2018) is a hierarchical LSTM model that builds on Word2Vec features and that outperformed by a large margin other methods available at the time. Following this work, other systems have been proposed similarly building on recurrent neural networks and word embeddings, with several improvements either at the embedding level (Arnold et al., 2019) and/or at the classifier level (Badjatiya et al., 2018; Sehikh et al., 2018).

As transformer-based language models changed the landscape of NLP, transformer and LSTM classifiers for linear text segmentation drawing on sentence-level BERT features started being proposed as well (Lukasik et al., 2020; Xing et al., 2020) and they have since become the norm, as they have been shown to outperform other features for the task (Ghinassi et al., 2023a). The use of pre-trained language models like BERT to extract features (generally known as transfer learning) has been shown to improve the generalization capabilities of topic segmentation systems, thanks to the general knowledge encapsulated in such encoders.

LSTM architectures building on such features have been shown to outperform Transformers for the task in certain cases, especially when not enough training data is available (Ghinassi et al., 2023b), while they perform comparatively similarly in case of bigger datasets (Lukasik et al., 2020). This evidence also reflects the tendency of such models to overfit to specific cue phrases and domain-specific features (e.g. naming a correspondent in certain news shows, Ghinassi et al., 2023a) and the use of domain adaptation has also been proposed in this context to attenuate the problem of overfitting to specific domains that come with the supervised setting (Glavaš et al., 2021).

Finally, a very recent line of research has attempted to use transformer-based language models directly as classifiers by placing a linear classification head on top of the beginning of sentence tokens. Among the limitations of transformers is the quadratic cost of self-attention that severely limits the maximum input length in terms of tokens for models like BERT. Earlier systems like Cross-segment BERT initially limited the context available to BERT by inputting just pairs of sentences (Lukasik et al., 2020) or passing sliding windows over tokens to aggregate as much context as possible (Zhang et al., 2021). More recent works have used models such as Longformer, specifically designed to deal with long contexts to overcome this problem (Inan et al., 2022).

2.3.2 Multi-task Methods

A more recent trend in linear text segmentation systems has variously adopted multi-task learning to regularise and improve end-to-end systems. Among the drawbacks of existing end-to-end systems, it has been observed how such models tend to overfit on local, domain-dependent cues that signal topic shifts (e.g. the locution "moving on" in multi-party meetings), but often do not generalize to other domains (Ghinassi et al., 2023a). In this sense, multi-task learning works similarly to transfer learning in helping the model to extract more general features, which more closely relate to modeling the underlying topical coherence.

Systems belonging to this category mostly combine topic classification and topic segmentation, both framed as supervised tasks. Topic classification in this context is framed as the task of assigning the correct topic class to each sentence or basic unit in the text, rather than identifying the basic units which are topic boundaries (i.e. linear text segmentation). This strand of research emerged mostly due to the release of datasets comprising both topic segmentation and topic identity information (Arnold et al., 2019). Among the most successful systems in this category, S-LSTM (Barrow et al., 2020) augmented the hierarchical LSTM with a system to pool sentence embeddings from extracted segments and use the pooled segment representation as input for a topic classification system. Similarly, Transformer$^2_{BERT}$ (Lo et al., 2021) used a hierarchical transformer where each contextualized sentence representation is used as input to separate topic segmentation and topic classification classifiers. In all of these cases, the addition of topic class information has been shown to improve results, sometimes quite dramatically. There could be many reasons for this, but the main rationale is that the shared representation layers in the networks are forced to learn a representation that is similar for all of the sentences sharing a topic class, therefore forcing the model not to focus solely on local cues which often lead to massive overfitting. As a result, adding topic classification in a multi-task setting has been shown to improve the generalizability capacity of topic segmentation models (Lo et al., 2021).

To achieve a similar goal, other works have directly added a secondary loss to segmentation systems, which penalize sentence embeddings belonging to the same topic segment that is too far in the embedding space (Xing et al., 2020; Yu et al., 2023). Also in this case the use of multi-task learning significantly improved segmentation results.

Another promising research direction is the one of directly injecting the notion of coherence into topic segmentation systems. Coherence modeling relates quite closely to linear text segmentation in that areas of low coherence in a document often coincide with topic boundaries. Following this reasoning, CATS (Glavaš and Somasundaran, 2020) employs a hierarchical transformer built on top of word embeddings and adds a secondary loss in the form of a binary classification where a coherence classification head is tasked with discriminating real text snippets from corrupted ones (i.e. text snippets where the sentences have been randomly shuffled). Similarly, Longformer + TSSP + CSSL (Yu et al., 2023), the current state-of-the-art in written text segmentation, uses a Longformer as a token-level classifier and adds an auxiliary loss term where a corrupted document having sentences shuffled according to a certain probability is tagged with a series of labels describing whether consecutive sentences are shuffled or not. Both techniques proved to improve results significantly.

Finally, a relative stand-alone recent attempt to combine topic modelling and topic segmentation exists in the form of Tipster (Gong et al., 2022), a model that combines neural topic modelling and neural topic segmentation by injecting information from the neural topic model into BERT sentence representations and having them as input for a classic recurrent neural network classifier for segmentation. This is an under-explored area of research that might open interesting future directions.

3 Datasets

Many datasets for topic segmentation have been released, but very few have been widely adopted. In this paragraph, we focus on domains that are arguably the most represented in the literature and we divide them in two distinct macro-domains: namely, written text and dialogue. We mostly discuss English datasets, but we will mention in the open challenges the lack of multilingual resources.

3.1 Written Text Datasets

Written text datasets have been variously proposed over the years, but few have been widely adopted.

Choi was among the first datasets being proposed (Choi, 2000) and it consists of a synthetic dataset created by randomly concatenating sections from different parts of the Brown Corpus. This dataset, however, is too simple, which is evident from the fact that an early supervised system like Cross-Segment BERT in table 2 was able to get an error already very close to 0. More recently, Koshorek et al. (2018) proposed wiki-727k, a dataset comprising 757,000 Wikipedia articles to overcome the limitations of previous datasets (especially their lack of connection with real use case scenarios) and to provide a dataset big enough to train large supervised models like neural networks. This dataset, however, is not widely used as its size makes it expensive to train a full system on it. Most works in topic segmentation, then, currently use en_city and en_disease, two English datasets in the Wikisection collection (Arnold et al., 2019), which includes four datasets divided into two categories (articles about cities and articles about diseases) and two languages (English and German); the two datasets are much smaller than wiki-727k and much more focused in terms of domain, where the en_disease dataset is both the smaller and the more specialized dataset among the two, at it includes a variety of rare medical terms. In general, datasets scraped from Wikipedia have the advantage of not needing any manual annotation, as the headings in the articles are used as topic-shifting markers. At the same time, they present specific challenges as they are composed of portions of texts often written by multiple authors, for which segmentation models might end up recognizing changes in writing style rather than in topics.

3.2 Dialogue Datasets

Another active area of research is that of Dialogue Topic Segmentation (DTS), usually in the form of transcripts from multi-party meetings, conversations, podcasts or news shows (Purver, 2011).

Initially, datasets for DTS mostly came from the meetings and news shows domains. Early examples of such datasets are the ICSI dataset (Janin et al., 2003), which includes 70 hours of audio and annotated transcripts from academic meetings, and the TDT corpus (Allan et al., 1998) including several hundreds of audio and annotated transcripts from American TV news shows. Datasets including transcripts from TV and podcast shows have since been extremely rare and even more rarely datasets were made publicly available mostly due to copyright limitations related to this specific content; TDT itself is available only on paying a fee, while it is now considered to be too *easy*, as exemplified by the results in table 3. Some recent attempts of proposing more challenging, openly available datasets in this domain exist (Ghinassi et al., 2023c), but they are limited in scope and size. QMSUM (Zhang et al., 2022) was also recently proposed to collect together different meeting datasets and it includes summary annotation, even though it is considerably smaller than written text-based datasets.

Finally, one-to-one spoken conversations datasets have been recently proposed. Among these, TIAGE was the first manually annotated dataset for one-to-one dialogue, drawing from another existing dataset for NLG (Xie et al., 2021).

Very recently, SuperDialseg was proposed as a large dataset for one-to-one DTS comprising more than 9000 dialogues which were automatically annotated via the use of dialogues that were grounded on the use of written documents in which the separation of topics is known (Jiang et al., 2023). A large meeting dataset was also recently proposed, even though smaller than SuperDialseg, but including annotations for a variety of other tasks (Zhang et al., 2023). These are indeed very promising developments that promise to close the gap between written text segmentation and DTS. Still, more needs to be done in domains such as transcripts from podcasts and TV shows, where comparable resources do not exist. Given the fact that datasets big enough are extremely recent, supervised systems for dialogue segmentation are also rare, even though they have been shown to outperform the alternatives, if enough data are available (Jiang et al., 2023). Table 3 shows how results on dialogue datasets are similar to the ones obtained on written text datasets by comparable methods; the major challenge in this context, then, is that of

Name	Domain	Language	#Documents	#Segments per Document	#Sentence per Segment
		Written Text			
choi	Random	English	920	9.98	7.4
en_city	Wikipedia	English	19500	8.3	56.7
en_disease	Wikipedia	English	3600	7.5	58.5
de_city	Wikipedia	German	12500	7.6	39.9
de_disease	Wikipedia	German	2300	7.2	45.7
wiki-727k	Wikipedia	English	727,746	3.48	13.6
		Dialogue			
ICSI	Meetings	English	25	4.2	188
QMSUM	Meetings	English	232	5.54	96.93
SuperDialSeg	Conversation	English	9468	4.20	3.09
TDT	Media	English	600*	88.75*	-
Non-NewsSBBC	Media	English	54	7.27	72.04

Table 1: Statistics of some of the datasets discussed. * denotes that the TDT corpus is measured in hours, rather than "number of".

having enough data to train supervised systems.

Table 1 shows statistics from some of the most relevant datasets discussed so far.

4 Metrics

Even though traditional classification metrics like F1 and accuracy have been used and continue to be used in the field, specific evaluation metrics for topic segmentations have been variously suggested during the years as traditional classification metrics over-penalize near misses (i.e. a topic boundary placed close to a real one), while evidence suggests human annotators tend to disagree where exactly to place topic boundaries (Purver, 2011).

Segmentation metrics can be categorised into three groups: window-based, boundary similarity-based and embedding-based metrics. Window-based metrics, exemplified by P_k (Beeferman et al., 1999) and WindowDiff (Pevzner and Hearst, 2002), employ a sliding window approach, comparing reference and hypothesis boundaries in the window. Boundary Similarity (Fournier, 2013), proposed more recently to overcome some of the problems with window-based metrics, works by representing the input sequence using the identity of the topic segment each element in the sequence belongs to. Given such a representation for both the hypothesized and reference segmentation, edit distance is used to quantify the error. Finally, reference-free embeddings use notions of embedding similarities to measure similarity within (and/or difference between) hypothesized topic segments, but they lag behind reference-based metrics (Lucas et al., 2023).

Figure 1 summarises the three different methods just described. P_k, WindowDiff, and F1 are the most used metrics in the field. P_k and WindowDiff, however, have been shown to have specific flaws related to penalizing certain types of errors more than others (Georgescul et al., 2006). Boundary Similarity, which was proposed to overcome some of the limitations, is not as popular with few works using it and most literature preferring P_k, notwithstanding its limitations (Ghinassi et al., 2023b). This is evident in figure 2 showing how popular different metrics are in the literature by the occurrences of different metrics as used in a sample of recent works (i.e. published after 2020) we cited. We also used P_k for comparisons, but we suggest that future research look into more modern metrics like Boundary Similarity to overcome well-known evaluation problems with P_k (Georgescul et al., 2006; Ghinassi et al., 2023b).

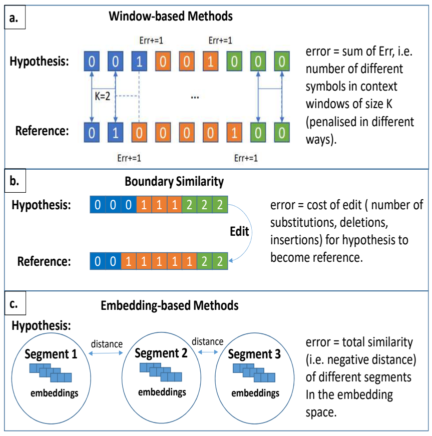

Figure 1: Segmentation metrics comparison.

5 Systems Comparison

Having described unsupervised and supervised approaches for linear text segmentation proposed during the years, table 2 and table 3 present a comparison of performance for different categories described above on some of the benchmarks described in more details in the next section.

Kind	Basic Unit	System	Choi	en_city	en_disease	wiki-727k
		Unsupervised Systems				
Count-based	sentence	TextTiling (Choi, 2000)	44	-	-	-
Count-based	sentence	C99 (Choi, 2000)	12	36.8	37.4	-
Count-based	word	U00 (Utiyama and Isahara, 2001)	9	-	-	-
Topic Modelling	sentence	TopicTiling (Riedl and Biemann, 2012)	0.95	30.5	43.4	-
Embedding-based	word	GraphSeg (Glavas et al., 2016)	7.2	-	-	-
		Supervised Systems				
Single-task	word	TextSeg (Koshorek et al., 2018)	-	24.3	19.3	22.13
Single-task	sentence	Cross-segment BERT (Lukasik et al., 2020)	**0.04**	15.4	33.9	-
Multi-task	sentence	Transformer$^2_{BERT}$ (Lo et al., 2021)	-	9.1	18.8	-
Multi-task	sentence	Tipster (Gong et al., 2022)	-	8.3	**14.2**	-
Multi-task	word	Longformer + TSSP + CSSL (Yu et al., 2023)	-	**7.4**	15.4	**13.89**

Table 2: Results of various systems described on 4 benchmarks for written text linear text segmentation. Results are reported from the works cited in the table. All results are expressed in P_k metric, the lower the better.

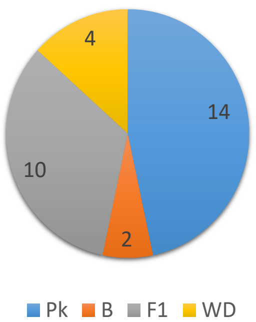

Figure 2: Number of occurrences of Pk, Boundary Similarity (B), F1 and Window Difference (WD) in cited works published after 2020.

At first glance, it can be observed how sparse the tables are: this is due to the long period considered which implies several changes of popular benchmarks over the years, but it also reflects a wider problem in the field for which benchmarks are not consistently used, especially when dealing with domains such as meetings and multimedia content. On another side, it can be seen how supervised models in table 2 largely outperform unsupervised systems. Specifically, models based on Longformer which can be trained at the word level as the one by (Yu et al., 2023) show best performance on most benchmarks. As mentioned, improvements from using multi-task settings seem consistent as most such systems outperform the alternatives, and among those Tipster (Gong et al., 2022) seems particularly promising. The reason behind such improvements is mostly related to the well-understood problem of supervised systems in topic segmentation, which tend to overfit on local cues and topic shift markers which are by their nature domain-dependent (e.g. thanking a correspondent at the end of a news story in news shows, Ghinassi et al., 2023a). As such, supervised models fail to generalize in many cases. This is even more true in domains in which scarce data is available, which is a common problem to all supervised models but seems to affect even more severely topic segmentation systems (Jiang et al., 2023). Multi-task learning, then, provides a way to direct the model away from focusing on domain-dependent local cues and to focus on properties shared by all sentences belonging to the same topic segments, as it is the case when we combine topic classification and linear text segmentation.

Given the highlighted problem of generalizability, unsupervised systems are still relevant, as the comparatively good performance of BayesSeg on the small ICSI dataset in table 3 demonstrates. The novel research on the use of LLMs, then, seems particularly relevant as the same table clearly shows how ChatGPT largely outperforms other unsupervised models on the Superdialseg dataset.

6 Conclusions: Open Challenges and Future Opportunities

The above discussion has shown how one of the major challenge in the field is the availability and the adoption of datasets (especially related to DTS). When enough data are available supervised systems can be trained for both written text topic segmentation and DTS generally showing improvements over unsupervised methods. At the same time, the large number of empty spots in our system comparison tables shows that no single dataset has ever been established as a widely recognized benchmark in the field. Such empty spots are also partly explained by the variety of different metrics for segmentation evaluation, as the lack of a single, widely recognised standard metric means that different works often use different metrics. Moreover, reported performance often does not reflect perfor-

Kind	Basic Unit	System	ICSI	TDT	SuperDialseg
		Unsupervised Systems			
Count-based	sentence	TextTiling (Solbiati et al., 2021)	38.2	-	44.1
Count-based	word	U00 (Galley et al., 2003)	31.99	**4.70**	-
Count-based	word	BayesSeg (Barzilay and Lapata, 2008)	**25.8**	-	43.3
Topic Modelling	word	HierBayes (Purver et al., 2006)	28.4	-	-
Embedding-based	sentence	TextTiling+BERT (Solbiati et al., 2021; Jiang et al., 2023)	33.6	-	49.9
LLM-based	word	ChatGPT (Jiang et al., 2023)	-	-	31.8
		Supervised System			
Single-task	word	TextSeg (Jiang et al., 2023)	-	-	**19.9**

Table 3: Unsupervised and supervised systems on benchmarks for dialogue text segmentation.

mance in real-world use cases, because of flaws of existing metrics like P_k (Georgescul et al., 2006). Future research should, in certain cases like podcast shows segmentation, propose new resources, but mostly it should establish which existing datasets and metrics are best suited to be used as benchmarks and evaluation metrics so that the numerous and rapid advances in this fast-evolving field can be compared in a fair and widely accepted setting.

Apart from resource limitations, methods for topic segmentation often assume a high level of agreement among human annotators, which isn't always the case (Purver, 2011). Identifying topics can be straightforward in domains like news shows but more challenging in contexts such as multi-party dialogue. Even when segmenting articles from Wikipedia, decisions must be made about what constitutes a significant enough topic shift (Koshorek et al., 2018). Previous research has explored hierarchical segmentation approaches, moving away from linear text segmentation (Yaari, 1997; Du et al., 2013). Recent end-to-end systems have lagged in this aspect, but the cited work combining topic segmentation and topic modelling (Gong et al., 2022) is a promising step forward to exploit knowledge about the topic structure rather than just local cues and coherence. Modern LLMs might be particularly suited to combine different tasks in a multi-task and/or zero-shot framework, as initially explored by Fan and Jiang (2023).

Our discussion primarily focused on English resources. Recently, more diverse linguistic resources have been suggested, especially for Mandarin (Zhang et al., 2023), with two German datasets also noted (Arnold et al., 2019). Few examples of datasets for other languages exist, except for the multilanguage dataset proposed by (Swędrowski et al., 2022), which remains underutilized. Multilinguality is crucial to democratize and broaden the scope of NLP research.

To summarise, in this work we have traced the various existing trends in literature for linear text segmentation within NLP and we have identified the following main challenges:

Lack of publicly available datasets: this problem affects mostly DTS (specifically the media domain) and it is crucial as recent supervised systems greatly outperform unsupervised ones. As a subset of this problem, we have also mentioned the need for standard benchmarks for the task to better track the advances in the field.

Pitfalls in existing metrics: the most popular metric, P_k has a number of well-documented shortcomings. Even though newer metrics like Boundary Similarity have been proposed, P_k is the most used even in recent works.

Low generalizability we have also discussed how the field has individuated generalizability as a key problem for the task, as many well-performing supervised systems might just be overfitting on specific cue phrases.

We suggest the following future directions as open opportunities for researchers in the field:

Use of LLMs: the rise of LLMs has already reshaped many areas in NLP, and there is similar scope in this context, especially given the problems of generalizability and the lack of resources which affect the field.

Advances in Multi-task learning: we highlight the combination of modern segmentation systems with topic modelling ones as a research direction worth developing, having deep roots in the field and narrowing the gap with hierachical segmentation, which is useful for overcoming the problem of arbitrary definition of topic granularity.

Advances in evaluation resources and metrics: we stress the importance of having a stable evaluation framework for the task. Advances in metrics are useful to deepen our understanding of a task having low human annotators agreement. Multilingual datasets, instead, can widen the reach of the available technology to less-resourced languages.

7 Limitations

Our work aimed to fill noticeable gaps in literature on topic segmentation. As previous surveys on the topic are all outdated or limited in scope, the current survey does not cover some of the many advances in the field explored in recent years. Among them, in our work we did not cover:

1. Multi-modality.
2. Topic Segmentation in nicher domains, like educational and legal text and multimedia.
3. Graph based methods for Topic Segmentation.

Another limitation of our work involves the definition of the classes for topic segmentation. In presenting an overview of available metrics, in fact, we have picked popular metrics for topic segmentation, but we have left out less used metrics that have been proposed and that might not fall neatly in the three-fold division of available methods that we have proposed.

Finally, we have mentioned the existing limitations of topic segmentation for languages other than English. Our work mostly deals with English resources, even though it mentions at least some literature dealing with other languages. This limitation is partly due to limitations within the field, which we have mentioned in our conclusions, but future work might integrate more research in this direction.

References

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic detection and tracking pilot study: Final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 169–184, Cambridge, MA. MIT Press.

Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-based neural text segmentation. In *Lecture Notes in Computer Science*, volume 10772 LNCS.

Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik. 2020. A joint model for document segmentation and segment labeling. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322, Online. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. In *Computational Linguistics*, volume 34.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. In *Machine Learning*, volume 34.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. In *Journal of Machine Learning Research*, volume 3.

F Choi. 2000. Linear text segmentation : approaches, advances and applications. In *Proc. of CLUK 3*.

Lan Du, Wray Buntine, and Mark Johnson. 2013. Topic segmentation with a structured topic model. In *Proc. 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Atlanta, Georgia. Association for Computational Linguistics.

Yaxin Fan and Feng Jiang. 2023. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. In *ArXiv*.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics*, pages 1702–1712.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

Maria Georgescul, Alexander Clark, and Susan Armstrong. 2006. An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In *Proc. 7th SIGdial Workshop on Discourse and Dialogue*.

Iacopo Ghinassi. 2021. Unsupervised text segmentation via deep sentence encoders: a first step towards a common framework for text-based segmentation, summarization and indexing of media content.

Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. 2023a. Comparing neural sentence encoders for topic segmentation across domains: not your typical text similarity task. In *PeerJ Computer Science*.

Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. 2023b. Lessons learnt from linear text segmentation: a fair comparison of architectural and sentence encoding strategies for successful segmentation. In *Proc. 14th International Conference on Recent Advances in Natural Language Processing*, pages 408–418, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. 2023c. Multimodal topic segmentation of podcast shows with pre-trained neural encoders. In

Proc. 2023 ACM International Conference on Multimedia Retrieval, ICMR '23, page 602–606, New York, NY, USA. Association for Computing Machinery.

Goran Glavaš, Ananya Ganesh, and Swapna Somasundaran. 2021. Training and domain adaptation for supervised text segmentation. In *Proc. 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 110–116, Online. Association for Computational Linguistics.

Goran Glavas, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In **SEM 2016 - 5th Joint Conference on Lexical and Computational Semantics*.

Goran Glavaš and Swapna Somasundaran. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*.

Zheng Gong, Shiwei Tong, Han Wu, Qi Liu, Hanqing Tao, Wei Huang, and Runlong Yu. 2022. Tipster: A topic-guided language model for topic-aware text segmentation. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Part III*, page 213–221, Berlin, Heidelberg. Springer-Verlag.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Michael A.K. Halliday and Hasan Ruqaiya. 1976. *Cohesion in English*. Routledge.

Marti A. Hearst. 1994. Multi-paragraph segmentation expository text. In *Proc. 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16. Association for Computational Linguistics.

Hakan Inan, Rashi Rungta, and Yashar Mehdad. 2022. Structured summarization: Unified text segmentation and segment labeling as a generation task. volume abs/2209.13759.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1.

Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. 2023. SuperDialseg: A large-scale dataset for supervised dialogue segmentation. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4086–4101, Singapore. Association for Computational Linguistics.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proc. of the Conference*, volume 2.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray L. Buntine. 2021. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In *EMNLP*.

Evan Lucas, Dylan Kangas, and Timothy Havens. 2023. A reference-free segmentation quality index (SegReFree). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2957–2968, Singapore. Association for Computational Linguistics.

Michael Lukasik, Boris Dadachev, Gonçalo Simões, and Kishore Papineni. 2020. Text segmentation by cross segment attention. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4707–4716.

Hemant Misra, François Yvon, Olivier Cappé, and Joemon Jose. 2011. Text segmentation: A topic modeling perspective. In *Information Processing & Management*, volume 47.

Lev Pevzner and Marti A. Hearst. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. In *Computational Linguistics*, volume 28, pages 19–36.

Matthew Purver. 2011. Topic segmentation. In *Spoken Language Understanding*. John Wiley & Sons, Ltd.

Matthew Purver, Konrad P. Körding, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Meeting of the Association for Computational Linguistics*, volume 1.

Martin Riedl and Chris Biemann. 2012. Text segmentation with topic models. In *Journal for Language Technology and Computational Linguistics*, volume 27.

Imran Sehikh, Dominique Fohr, and Irina Illina. 2018. Topic segmentation in asr transcripts using bidirectional rnns for change detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017 - Proceedings*, volume 2018-January.

Alessandro Solbiati, Kevin Hefferman, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised topic segmentation of meetings with bert embeddings. In *arXiv*.

Qi Sun, Runxin Li, Dingsheng Luo, and Xihong Wu. 2008. Text segmentation with lda-based fisher kernel. In *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Michał Swędrowski, Piotr Miłkowski, Bartłomiej Bojanowski, and Jan Kocoń. 2022. Multi-wiki90k: Multilingual benchmark dataset for paragraph segmentation. In *Advances in Computational Collective Intelligence*, Cham. Springer International Publishing.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proc. 39th Annual Meeting of the Association for Computational Linguistics*, pages 499–506, Toulouse, France. Association for Computational Linguistics.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1684–1690, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proc. 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 626–636, Suzhou, China. Association for Computational Linguistics.

Yaakov Yaari. 1997. Segmentation of expository texts by hierarchical agglomerative clustering. In *CoRR*, volume 3.

J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. 1998. A hidden markov model approach to text segmentation and event tracking. In *Proc. 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 333–336 vol.1.

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. 2023. Improving long document topic segmentation models with enhanced coherence modeling. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5592–5605, Singapore. Association for Computational Linguistics.

Qinglin Zhang, Qian Chen, Yali Li, Jiaqing Liu, and Wen Wang. 2021. Sequence model with self-adaptive sliding window for efficient spoken document segmentation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 411–418.

Qinglin Zhang, Chong Deng, Jiaqing Liu, Hai Yu, Qian Chen, Wen Wang, Zhijie Yan, Jinglin Liu, Yi Ren, and Zhou Zhao. 2023. Mug: A general meeting understanding and generation benchmark. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.