When Cohesion Lies in the Embedding Space: Embedding-Based Reference-Free Metrics for Topic Segmentation

Iacopo Ghinassi[1], **Lin Wang**[1], **Chris Newell**[2], **Matthew Purver**[1,3]
[1]Queen Mary University of London, UK [2]BBC R&D, UK [3]Institut Jožef Stefan, Slovenia
i.ghinassi@qmul.ac.uk

Abstract

In this paper we propose a new framework and new methods for the reference-free evaluation of topic segmentation systems directly in the embedding space. Specifically, we define a common framework for reference-free, embedding-based topic segmentation metrics, and show how this applies to an existing metric. We then define new metrics, based on a previously defined cohesion score, Average Relative Proximity. Using this approach, we show that Large Language Models (LLMs) yield features that, if used correctly, can strongly correlate with traditional topic segmentation metrics based on costly and rare human annotations, while outperforming existing reference-free metrics borrowed from clustering evaluation in most domains. We then show that smaller language models specifically fine-tuned for different sentence-level tasks can outperform LLMs several orders of magnitude larger. Via a thorough comparison of our metric's performance across different datasets, we see that conversational data present the biggest challenge in this framework. Finally, we analyse the behaviour of our metrics in specific error cases, such as those of under-generation and moving of ground truth topic boundaries, and show that our metrics behave more consistently than other reference-free methods.

Keywords: Evaluation Methodologies, Topic Detection & Tracking, Neural language representation models

1. Introduction

Topic segmentation is a well-established challenge in natural language processing and serves as the initial step for numerous downstream applications like topic-driven summarisation and semantic search. The task involves automatically breaking down a text into coherent units with shared topics (Purver, 2011): for instance, a lengthy transcript from a news programme can be partitioned into individual stories to assist users in retrieving more pertinent and specific information (Reynar, 1999). Similarly, a lengthy article can be divided into sections to aid reading (Hearst, 1997).

Recent research has introduced several advances in this field, but evaluation remains difficult for this task, partly due to the scarcity of expert-annotated datasets in specific domains.

To overcome some of these problems, there has been a recent surge in interest in *reference-free* metrics, designed to score a hypothesised segmentation of a document without the need to refer to any expert annotation. Initial attempts in this direction seem promising, but they are limited in scope and no formal definition in our knowledge has been outlined for this type of evaluation framework.

As such, we propose a general taxonomy of such evaluation techniques, which we name *embedding-based topic segmentation metrics*. We propose our own method within this framework, showing improvements and closely reflecting the behaviour of reference-based metrics in multiple scenarios.

Furthermore, as these techniques are based on sentence embeddings, we evaluate three different sentence encoding methods from Large Language Models (LLMs) and show that the use of very large models does not give any significant improvement over smaller, well-optimised encoders.

Finally, we show how our metric performs in synthetically created cases, so as to highlight the behaviour of our approach in specific situations.

2. Related Work

2.1. Existing Metrics

Several topic segmentation evaluation metrics have been suggested in the literature (see e.g. Beeferman et al., 1999; Pevzner and Hearst, 2002; Fournier and Inkpen, 2012; Fournier, 2013b). These metrics all rely on the use of reference topic boundaries, conventionally resulting from human annotations. Most recently, the use of reference-free metrics has seen a surge in interest in the NLP community for applications such as machine translation (Leiter, 2021) and natural language generation (Ke et al., 2022). Reference-free metrics have the obvious advantage of not needing any expert annotation, while also possibly avoiding problems related to annotator agreement. Very recently, an initial attempt has been made to devise a reference-free metric also for the task of topic segmentation (Lucas et al., 2023), but this is the only such attempt and, as such, it lacks comparison to other possible reference-free methods.

Including this recent reference-free approach, existing segmentation metrics can be categorised into

LREC-COLING 2024, pages 17525–17536

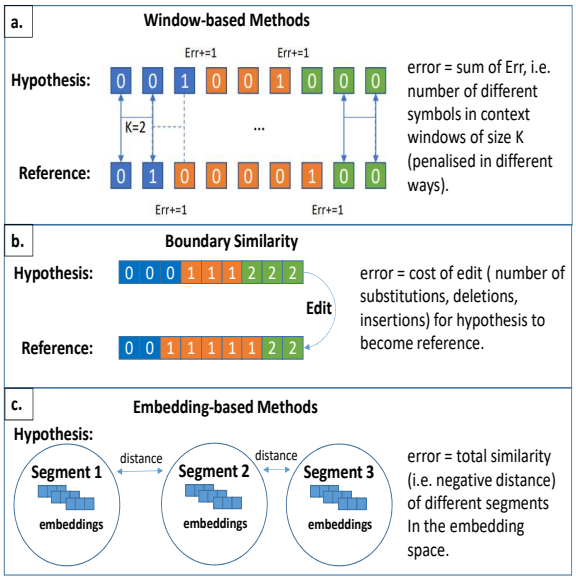

Figure 1: Illustrations of three different family of metrics for topic segmentation: window-based (a), boundary similarity (b) and the reference-free family proposed in this work, embedding based (c).

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, 2013b), proposed more recently to overcome some of the problems with window-based metrics, works by representing the input sequence by means of the identity of the topic segment each element in the sequence belongs to. Given such a representation for both the hypothesised and reference segmentation, minimum edit distance is used to quantify the error.

All together, P_k, WindowDiff and Boundary Similarity are the most used metrics in the field and, even though they do present inherent problems, many works make up for them by reporting two or all of these metrics as they present largely complimentary weaknesses (Georgescul et al., 2006).

Finally, in the context of reference-free evaluation, we can turn to notions of embedding similarities to measure similarity within (and/or difference between) hypothesised topic segments. This is the approach proposed in Lucas et al. (2023); it generalises the intuition behind a number of methods for unsupervised segmentation proposed recently, that all work by exploiting local minima in the similarities of consecutive sentence embeddings (Ghinassi, 2021; Solbiati et al., 2021; Harrando and Troncy, 2021).

Figure 1 summarises the three different methods just described. In this work, we formalise this new family of evaluation techniques — the reference-free embedding-based methods — and propose our own method within this category, which, as we show, outperforms the alternatives. In evaluating the various methods we use the three traditional metrics described before as a gold standard, as they are by far the most used in the field and they are closer to human judgements, due to the fact that they are based on human annotations.

2.2. Methods for Topic Segmentation

Early text segmentation methods like TextTiling (Hearst, 1994) used sliding windows with cosine similarity between bag-of-words representations, looking for local minima. Later approaches incorporated more informative sentence representations, including TF-IDF scoring (Galley et al., 2003) and topic probabilities (Riedl and Biemann, 2012).

More recently, neural supervised methods showed significant improvements over non-neural unsupervised methods (Koshorek et al., 2018).

Transformer-based LLMs such as BERT have also been recently used for topic segmentation by using them as sentence encoders extracting sentence-level features to be input to various neural models like Bidirectional Long-Short Term Memory (BiLSTM) networks (Xing and Carenini, 2021) or Transformers (Lo et al., 2021).

Unsupervised methods have also been recently proposed to overcome the data scarcity problem (Ghinassi, 2021; Harrando and Troncy, 2021; Solbiati et al., 2021). These methods build on the same intuition as the early methods like TextTiling, looking for local minima in similarity, but rely on the use of sentence representations from LLMs to characterise that similarity. In this, they closely resemble our general framework for reference-free evaluation, as they mostly exploit notions of similarities between sentence embeddings. This last line of research has similarities with our framework, but in our case embedding similarities are used to evaluate existing topic segmentation systems.

3. Methodology

3.1. General Framework

Here we define the general framework of embedding-based methods for reference-free evaluation of topic segmentation.

These methods need no ground truth annotation; they work by comparing sentence embedding similarities to assess whether sentences in a given topic segment are more cohesive than average (i.e. good segmentation performance) or less (bad performance). Although we use the term ‘sentences’ here, we take it to cover other units of text such as utterances, depending on the domain of application

(e.g. in conversational datasets we use utterances or speaker turns rather than sentences). Although at least one attempt to use sentence embeddings in this way exists, no formal definition of this framework for the evaluation of topic model has yet been proposed; in this section we aim to fill this gap.

Formally, we define the sentence embedding for sentence (or otherwise defined unit of text) s_i at the i_{th} position of the given document as $e_i = enc(s_i)$, where enc is a suitable encoder. Then, we define a set of topic boundary positions $\mathcal{T} = \{t_0, t_1, t_2, ..., t_i, t_N\}$ where $t_0 = 0$ and t_N equals the length of the current document (i.e. number of sentences). In this case t_i represents the sentence position in the given document where the ith topic segment ends, while the start of the same segment can be inferred as t_{i-1}. We then group the embeddings according to topic boundaries $\mathcal{T}$ and we get $\{E_1, ..., E_i, ..., E_N\}$, where

$$E_i = \{e_{t_{i-1}}, ..., e_{t_i}\} \tag{1}$$

where i ranges from 1 to N. At this point, we obtain a coherence score $\mathcal{C}$ for each consecutive segment $\{t_i, t_{i+1}\}$ as follows:

$$\mathcal{C}_i^d = Score(E_i, E_{i\pm1}) \tag{2}$$

where i ranges from 1 to $N-1$ and $\mathcal{C}_i^d$ is the coherence score for document d and topic segment i. Here, $Score$ determines how we use the embeddings to compare consecutive segments, the main choice in this framework; we propose a number of possible scoring functions below.

We perform pooling across all the sentences in a document and documents in the corpus to obtain:

$$\hat{\mathcal{C}} = pool_d(pool_n(\mathcal{C}_i^d)) \tag{3}$$

where $pool_d$ and $pool_n$ are pooling functions at the corpus and document-level respectively. In the simple case in which both those operations consists in a simple average (as it is in all of our settings), we then have:

$$\hat{\mathcal{C}} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N_d}\sum_{i=1}^{N_d}\mathcal{C}_i^d \tag{4}$$

where D is the total number of documents and N_d is the length of the given document d (simply defined as N in previous equations).

3.2. Scoring Functions

The scoring function $Score$ is the central part of our framework. Here we show the scoring functions that we use in our experiments, covering a range of popular measures for embedding distances.

3.2.1. Clustering-based

As a baseline, we use traditional clustering evaluation metrics:

SegReFree: This method was proposed in the already cited work by Lucas et al. (2023), using a measure borrowed from clustering research in order to score consecutive segments given a hypothesised segmentation. Specifically, this metric is the Davies-Bouldin Index (Davies and Bouldin, 1979) with an additional correction term. Formally, for each segment i represented as before as grouped embeddings E_i, they compute a centroid $c_i = \frac{1}{|E_i|}\sum_{e\in E_i} e$, where $|E_i|$ is the number of embeddings e in E_i. A dispersion measure S_i is then defined as the average Euclidean distance between c_i and all the embeddings $e \in E_i$. These intra-cluster distances are then modified by:

$$S_i = \frac{S_i}{1 - \frac{1}{\sqrt{|E_i|}}} \tag{5}$$

At this point, for each triplets of consecutive segments $\{E_{i-1}, E_i, E_{i+1}\}$, they compute first the Euclidean distance between the relative centroids $\mathcal{M}_{ij}, j \in \{i-1, i+1\}$, which is then used to compute a ratio of pairwise intra-cluster distances and centroid distances:

$$\mathcal{R}_{ij} = \frac{S_i + S_j}{M_{ij}}, j \in \{i-1, i+1\} \tag{6}$$

Finally the maximum value of $\mathcal{R}_{ij}$ for each segment E_i is taken:

$$\mathcal{C}_i^d = max(\mathcal{R}_{i,i-1}, \mathcal{R}_{i,i+1}) \tag{7}$$

And the final corpus-level score is computed as per equation 4 above.

This method includes a number of problems. First, the correction factor $\sqrt{|E_i|}$ implies that the algorithm can't deal with topic segments of one sentence; the authors warned that in such instances the algorithm should output a default value which we set to $\sum_{i=1}^{N_d}\mathcal{C}_i^d$, i.e. the average score in the document. If all segments in the document have just one sentence, then we set $\mathcal{C}_i^d = 10$, where 10 is a high value empirically chosen to penalise such occurrences.

Secondly, the metric is unbounded, making it difficult to compare across different use cases.

Silhouette Score: This metric is also a commonly used one in clustering evaluation and it involves the idea of comparing the average intra- and minimum inter-cluster distances of individual data points (Rousseeuw, 1987). In this context, we modify it to compare just adjacent clusters (i.e. topic segments). For each embedding $e_i \in S_n$, where S_n is the n_{th} topic segment, we compute:

$$a_i = \frac{1}{|S_n| - 1}\sum_{e_j \in S_n; j\neq i} distance(e_i, e_j) \tag{8}$$

Where $|S_n|$ is the number of embeddings in the n_{th} segment, e_i and e_j are the i_{th} and j_{th} embeddings in the same segment respectively and $distance$ can be any distance function, which here is the Euclidean distance following the original algorithm. For the same embedding e_i we then compute:

$$b_i = min_{S_m} \frac{1}{|S_m|} \sum_{e_j \in S_m} distance(e_i, e_j) \quad (9)$$

where S_m is one of the following or previous topic segment ($m \in \{n+1, n-1\}$) and e_j is the j_{th} embedding of topic segment S_m. If $|S_n| > 1$, the two quantities are then combined as:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (10)$$

If $|S_n| = 1$, instead we set $s_i = 0$. After having gathered all the scores s_i for each $e_i \in S_n$, we average them to obtain:

$$C_i^d = \frac{1}{|S_n|} \sum_{s_i \in S_n} s_i \quad (11)$$

where in this case C_i^d corresponds to the segment-level metric for segment S_n in document d with $i = n$. We finally compute the corpus-level metric $\hat{\mathcal{C}}$ following equation 4.

This metric has the advantage over SegReFree that is bounded, ranging from -1 (worst) to 1 (best). We further adjust the metric to be a loss function ranging from 0 (best) to 1 (worst) by applying the following transformation:

$$\hat{\mathcal{C}}_{loss} = 1 - \frac{\hat{\mathcal{C}} + 1}{2} \quad (12)$$

3.2.2. Our Methods

Using the method above as a baseline, we compare a number of alternative scoring functions within the general framework laid out above. All the methods we propose use as their starting point the Average Relative Proximity (ARP) score (Ghinassi et al., 2023a). In its original form, ARP was proposed as a way to compare different sentence encoders for the task of topic segmentation without needing to train a segmentation system. Here we further develop this idea, modifying the metric to work as a reference-free metric for evaluating topic segmentation systems. We employ three versions of this score, all of which are described in details below.

ARP$_{\text{std}}$: This method makes use of the original formulation of the ARP score (Ghinassi et al., 2023a), which measures the dispersion in a group of embeddings as the norm of the standard deviation across each dimension of the embeddings. Given two consecutive embedding groups corresponding to topic segments, E_i and E_{i+1}, we define E_{ter} as a set including embeddings from E_i and embeddings from E_{i+1}, while E_{tra} includes just embeddings from E_i. We compute the variance of the embeddings inside the same topic segment as:

$$intravar_i = ||std(\mathbf{E}_{tra})||^2 \quad (13)$$

Furthermore, we compute the variance of the embeddings crossing multiple topic segments as:

$$intervar_i = ||std(\mathbf{E}_{ter})||^2 \quad (14)$$

In order to overcome the standard deviation bias of towards bigger values for smaller samples, we force the intra-variance to be computed on a set of embeddings of the same size as that from which the inter-variance is computed. This is done by adding a cutting point $cut = |E_i|/2$ where $|E_i|$ is the number of embeddings in the current segment, so that $E_{tra} = E_i$ and $E_{ter} = E_i^{cut:|E|} \oplus E_{i+1}^{0:cut}$, with $\oplus$ representing concatenation.

The two scores are aggregated to obtain a single score representing the relative proximity of embeddings belonging to the same segment. We compute it as

$$RP_i = \frac{intervar_i - intravar_i}{intervar_i + intravar_i} \quad (15)$$

At this point we perform a pooling operation that also accounts for the cases in which some or every segment consists of single embeddings. We do so by applying the substitution

$$\mathcal{C}_\rangle = \begin{cases} RP_i, if |E_i| > 1 \\ \frac{1}{N} \sum_{i=1}^{N} intervar_i, if |E_i| = 1 \forall i \\ \frac{1}{N} \sum_{i=1}^{N} intravar_i, otherwise \end{cases} \quad (16)$$

Where the above ultimately makes the method default to more uncertainty ($\mathcal{C} \approx 0$) as we encounter more single-sentence segments.

We then follow equation 4 to obtain the average relative proximity score $\hat{\mathcal{C}}$.

As in this case $\hat{\mathcal{C}}$ range from -1 to 1 and it gives more cohesive segments higher values, we finally transform it into a loss function (i.e. the lower the better) in the range 0 to 1 by applying equation 12.

ARP$_{\text{cos}}$: In this version, we substitute equation 13 with:

$$intravar_i = \frac{1}{|E_{tra}|} \sum_{e \in E_{tra}} cosine(\hat{E_{tra}}, e) \quad (17)$$

where $\hat{E_{tra}} = \frac{1}{|E_{tra}|} \sum_{e \in E_{tra}} e$.

We do the same with equation 14, such that:

$$intervar_i = \frac{1}{|E_{ter}|} \sum_{e \in E_{ter}} cosine(\hat{E_{ter}}, e) \quad (18)$$

where $\hat{E_{ter}} = \frac{1}{|E_{ter}|} \sum_{e \in E_{ter}} e$.

In both cases, the notation reflects that explained above and, from here, we can follow equation 15

and 16 to obtain the final score $\hat{C}$ and 12 to transform it in a loss score.

ARP$_{\text{pair}}$: Lastly, we include a scoring method which directly computes the average pairwise distance of embeddings (rather than their dispersion from the segment centroid). We do that by substituting again equation 13 with:

$$intravar_i = \frac{1}{\frac{n(n-1)}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} cosine(e_i, e_j) \forall i < j \tag{19}$$

Where $n = |E_{tra}|$, e_i and e_j are the i_{th} and j_{th} elements of E_{tra} respectively and the intra-variance in this case represents the average of each possible pairwise cosine similarities between the members of E_{tra}, defined as before.

We do the same for the inter-variance:

$$intervar_i = \frac{1}{\frac{n(n-1)}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} cosine(e_i, e_j) \forall i < j \tag{20}$$

but this time $n = |E_1|$ and e_i and e_j are the i_{th} and j_{th} elements of E_{ter}.

We finally obtain $\hat{C}$ as per equations 15 and 16, transforming it in a loss score with 12.

3.3. Embedding Models

As the name of our framework suggests, a good embedding model is key for the success of these type of metrics. As usual in recent research in NLP, we compare three different LLMs that we use to extract sentence embeddings:

RoBERTa: We use the base version of the architecture (Liu et al., 2019), including 12 layers and 768 dimensional embeddings. To obtain sentence representations we perform a simple average of the last layer, shown to be effective in a variety of scenarios (Huang et al., 2021).

MPNET: The original architecture derives from Song et al. (2020); we use the base version having same size as RoBERTa. It is optimised for various sentence-level tasks and performs best in benchmark results (Reimers and Gurevych, 2020). Again, the average of the last layer is used as the final sentence representation.

Falcon: We include a more recent LLM, which belongs to the family of models with over a billion parameters. Such models show impressive capabilities in language generation, but they have also been shown to fall short as sentence encoders (Jiang et al., 2023). In this case we use the small version of Falcon, a 7 billion parameter LLM pretrained on selected data from the web and showing considerable improvements over comparable open-source models on a variety of tasks (Penedo et al., 2023). We use the average of the last layer to obtain sentence representations.

3.4. Evaluation Methods

To evaluate the proposed metrics we compare them to three traditional metrics which use a reference segmentation produced by human expert annotators. The metrics we use are P_k (Beeferman et al., 1999), WindowDiff (Pevzner and Hearst, 2002) and Boundary Similarity (Fournier, 2013b); we use Pearson Correlation coefficient to show the correlation between the results obtained with the reference-free metrics and the traditional metrics. To do so we collect hypothesised segmentations in two ways described below.

3.4.1. Real System Evaluation

We perform topic segmentation with a number of real segmentation systems and compute each metric on the resulting hypothesised segmentations. Specifically, we use the 9 supervised approaches described in Ghinassi et al. (2023b), including 3 sentence encoders producing the sentence-level features and for each encoder a BiLSTM classifier, a Transformer encoder classifier and a modification of the BiLSTM classifier previously proposed by Sehikh et al. (2018). To these supervised approaches we add the ground truth segmentation (i.e. results of the metrics when we use the correct segmentation from the annotators) and nine random baselines, one outputting a topic boundary at each sentence with probability $\frac{1}{k}$, where k is the average number of segments per document in the corpus, while the other eight methods output a topic boundary at each sentence with probability $\frac{1}{n}$ with n ranging from 2 to 9.

3.4.2. Synthetic Evaluation

Following previous work by Lucas et al. (2023), we also evaluate our metrics in two specific scenarios.

Boundary Removal: in this case we progressively remove a number n of topic boundaries from the ground truth, therefore generating segmentations biased towards false negatives.

Boundary Transposition: we also progressively transpose existing boundaries a number n of sentences away from their original place, to examine how lenient our metrics are in case of the predicted boundary being further and further away from the original one.

4. Datasets

To have a broad coverage of different domains, we use 4 different datasets from existing literature:

en_city (Arnold et al., 2019): this dataset from the WikiSection collection includes Wikipedia articles about cities, where the headings in the original article are used as markers, marking a topic shift.

en_disease (Arnold et al., 2019): again from the WikiSection collection, this dataset is smaller in size and the articles deal with diseases, therefore including a more specialised medical lexicon.

QMSum (Zhong et al., 2021): this dataset aggregates three smaller conversational datasets for topic segmentation from meeting transcripts, ICSI (Janin et al., 2003), AMI (Carletta et al., 2006) and a third dataset of Canadian parliamentary meetings released by the authors themselves. This dataset includes real face-to-face conversations and, as such, longer and more blurred topic segments, while many disfluencies and speech acts also make it a more complex dataset for the task.

SBBC-RadioNews (Ghinassi et al., 2023): proposed as a lightweight dataset for multimodal topic segmentation in the media domain, it includes 47 radio news shows from the BBC Sound collection.

For each dataset and in each experiment we use the default test set.

5. Experimental Setup

In all experiments we use the parameters described by Ghinassi et al. (2023b) for the training of all topic segmentation models and their optimisation.

For the Pk and WindowDiff metrics, we use the default window size $k = \frac{1}{2}K$ where K is half the average segment length in the given document. For boundary similarity metric, we keep the fixed value for the maximum transposition position at 2.

Pk, WindowDiff and Boundary Similarity were computed using the standard python library $segeval$ (Fournier, 2013a); all other reference-free metrics were implemented by us.

Finally, for cases in which no segmentation is output by a given real world system, either for a document or (in a few cases) for an entire dataset, we have skipped the document/dataset and it is not reflected in the results.

6. Results

6.1. Real System Evaluation

Table 1 shows the results on all datasets obtained by performing a correlation analysis of our four embedding-based metrics with the three traditional metrics and by using the three different sentence encoders previously mentioned.

As a general observation, it can be noticed how the correlation between the embedding-based metrics and the window-based metrics Pk and Window Difference can be quite high, often reaching over 90%. The correlation with the Boundary Similarity metric, instead, is generally lower but in some instances it still reaches over 90% as well. One of the reasons for this discrepancy might relate to specific weaknesses of the window-based metrics which tend to penalise more false positives over false negatives (Georgescul et al., 2006) and, as such, behave similarly to the reference-free metrics in penalising more cases in which more segment boundaries are output. For the random-based systems described above, then, a lower probability of outputting a topic boundary is less penalised even if the output boundaries are equally random.

Domain seems to be important as well: with QMSum, embedding-based methods struggle in correlating with reference-based metrics, reflecting the aforementioned difficulty of this dataset.

6.1.1. Scoring Functions Comparison

When we compare the scoring functions discussed in section 3.2, we see that the best correlation values are mostly exhibited by ARP-based methods, especially ARP_{pair}. SBBC-RadioNews is an exception as here the Silhouette method reaches the best correlation values (even though ARP_{pair} scores slightly higher in Boundary Similarity).

The success of ARP_{pair} can be explained by the fact that it does not compare centroids but pairs of embeddings, therefore giving a higher weight to anaphora and repetitions, often used in discourse as cohesive tools (Halliday and Hasan, 1976). This is exemplified by the much better results reached by this method in the SBBC-RadioNews dataset, where elements such as proper names in news stories are repeated throughout a topic segment. The same explains Silhouette's success on this dataset, as it compares pairs of embeddings, too.

On the other hand, this same characteristic can lead to a worse metric in cases in which such repetitions are not indicative of topic continuity. This is the case in QMSum dataset, where there are many utterances consisting in disfluencies (e.g. "Mmm") and common words (e.g. "Ok") throughout the transcripts. In this case, a method like Silhouette fail, having negative or close to null correlation with all of the metrics based on human judgement. If we look at the b_i term in equation 9, in fact, we can see that in case two sentences that are exactly the same or very similar occur in consecutive segments, then b_i will tend to be smaller. The same logic also explains the failure of SegReFree in some cases, even though there is a difference as in this case centroids rather than individual embeddings are compared. From equation 6 we can see that SegReFree heavily rely on the distance between segments' centroids as a way to scale up or down the average inter-cluster distance. This way, if two consecutive segments are very similar (e.g. by means of several repetitions) the resulting score R will be very high, tending to infinity as the centroids' distance tends to 0, which in turn skew results and might result in numerical overflow prob-

		en_city			en_disease			QMSum			SBBC-RadioNews		
		RoB	Fal	MPN	RoB	Fal	MPN	RoB	Fal	MPN	RoB	Fal	MPN
Pk	SegReFree	0.25	0.38	0.78	0.15	-0.05	0.65	0.41	-0.62	-0.66	0.26	0.28	-0.69
	Silhouette	0.13	0.81	0.36	0.83	0.85	0.88	-0.42	-0.42	-0.46	0.97	0.97	**0.98**
	ARP_{std}	0.91	0.94	0.93	0.91	0.9	0.93	0.75	0.72	0.77	0.85	0.75	0.85
	ARP_{cos}	0.91	0.93	0.93	0.91	0.9	0.93	0.77	0.73	**0.85**	0.85	0.79	0.85
	ARP_{pair}	0.93	0.95	**0.96**	0.92	0.9	**0.95**	0.64	0.44	0.38	0.89	0.75	0.93
WD	SegReFree	0.34	0.51	0.75	-0.04	-0.23	0.5	0.47	-0.77	-0.81	0.27	0.35	-0.78
	Silhouette	0.08	0.79	0.33	0.75	0.78	0.83	-0.5	-0.5	-0.53	**0.99**	0.98	**0.99**
	ARP_{std}	0.92	0.95	0.93	0.89	0.9	0.92	0.82	0.79	0.86	0.87	0.82	0.86
	ARP_{cos}	0.92	0.95	0.93	0.9	0.92	0.92	0.84	0.8	**0.92**	0.89	0.86	0.86
	ARP_{pair}	0.94	**0.97**	0.95	0.91	**0.94**	**0.94**	0.7	0.52	0.49	0.93	0.83	0.92
B	SegReFree	0.23	0.29	0.81	0.28	0.07	0.72	0.22	-0.36	-0.38	0.24	0.24	-0.63
	Silhouette	0.24	0.87	0.4	0.82	0.8	0.84	0.02	0.03	-0.01	0.94	0.93	0.96
	ARP_{std}	0.93	0.94	0.95	0.82	0.8	0.85	0.34	0.37	0.35	0.88	0.75	0.91
	ARP_{cos}	0.92	0.92	0.95	0.81	0.78	0.84	0.38	0.4	**0.47**	0.88	0.78	0.91
	ARP_{pair}	0.94	0.93	**0.97**	0.83	0.78	**0.88**	0.25	0.25	-0.04	0.89	0.7	**0.97**

Table 1: Results from our experiments with real segmentation systems. Results are expressed in terms of Pearson Correlation coefficients with regard to the reference-based metrics Pk, Window Difference (WD) and Boundary Similarity (B). For each of the four datasets we report the correlation of the relative reference-based metric and the reference-free metric computed with one of the proposed scoring methods and one of three sentence encoders: RoBERTa (RoB), Falcon (Fal) or MPNET (MPN).

lems. SegReFree is also influenced by how close in the embedding space different encoders tend to encode both similar and dissimilar sentences. Previous literature, in fact, have shown how different encoders like RoBERTa tend to express similarities between even very dissimilar sentences as very close in the embedding space, while still being able to assign higher similarity to more similar sentences (Ghinassi et al., 2023a); this however has a strong effect on SegReFree, following what explained above about the influence of very small centroid distances in the algorithm's denominator. ARP methods, instead, are more robust with respect with the choice of encoder and with respect to repetitions in different segments. On one hand, similarly to Silhouette, the denominator has merely a normalising function, while it is less dependent on repetitions than Silhouette because it uses the average inter-cluster similarity rather than the minimum distance (0 if an identical sentence appears in an adjacent segment).

6.1.2. Sentence Encoders Comparison

When we turn to compare different sentence encoders, we can immediately notice how MPNET seems to be consistently the best.

RoBERTa and, especially, Falcon lead to better correlation with reference-based metrics in few cases, but they both tend to lead to very bad results in other cases, such as when used with the SegReFree method, which varies the most under different encoders. It is interesting to notice how Falcon perform much better in all metrics when using Silhouette method on the en_city dataset. This however is an isolated case and it might originates from a more specific interaction of this method and the encoder in the given context.

In general, comparatively much smaller encoders like MPNET outperform very big LLMs such as Falcon in this task. This evidence is in line with previous observations on similar semantic similarity tasks and denote a limit of more recent LLMs, which still perform worse than smaller fine-tuned models (Jiang et al., 2023; Deshpande et al., 2023).

6.2. Synthetic Evaluation

Figure 2 shows the value of different metrics when we progressively remove true topic boundaries from the ground truth labels of our 4 datasets. The metrics' scores have all been normalised per metric with a MinMax scaler, so that they all lie between 0 and 1, where 1 is the worst score in the given metric group and 0 is the best scoring one.

The boundary removal experiments shows good results for all the reference-free metrics in most cases, confirming the results reported in Lucas et al. (2023) which showed how SegReFree is able to correctly penalise cases in which we progressively remove boundaries in the same way as traditional metrics based on human judgement do. An exception is QMSum, where ARP and Silhouette follow more closely the traditional metrics while SegReFree has an irregular pattern. On the contrary, SegReFree shows a more clear upward trend than the other two metrics on SBBC-RadioNews.

When we turn to the boundary transposition experiments shown in figure 3, instead, Silhouette seems to be the metric having the most problems, showing quite different trends from the traditional metrics for all but the SBBC-RadioNews dataset.

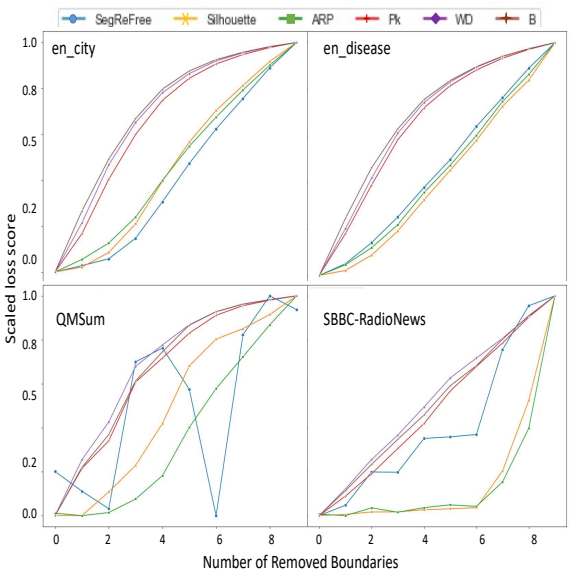

Figure 2: Boundary removal experiments showing relation between scaled loss scores output by our metrics (y-axis) and number of removed boundaries (x-axis) for the 4 datasets using MPNET.

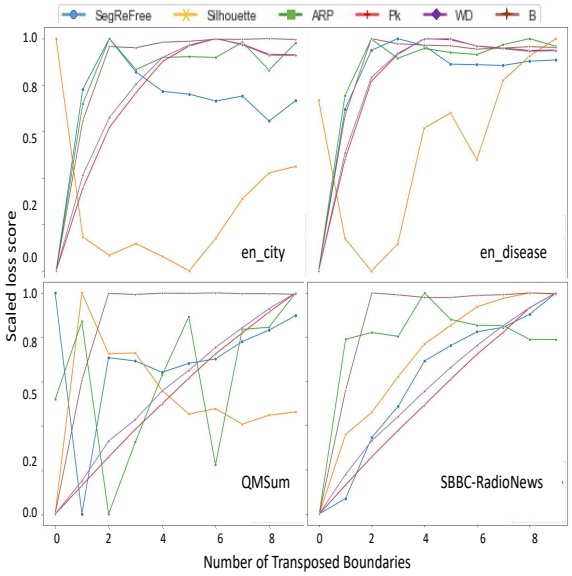

Figure 3: Boundary transposition experiments showing relation between scaled loss scores (y-axis) and the number of transposed boundaries (x-axis) for the 4 dataset using MPNET.

SegReFree seems to output better results after around 2 transposed boundaries in the WikiSection datasets, while the ARP score behaves much more closer to the Boundary Similarity metric. All reference-free metrics show an inconsistent behaviour on the QMSum dataset.

The synthetic experiments show two main things. First, this type of evaluation (especially if we look at SegReFree's performance in the boundary removal experiments) might overestimate the performance of reference-free metrics, as these metrics might perform very similar to existing metrics in some of these controlled experiments. Most embedding-based metrics, for example, show a trend similar to traditional ones in the QMSum dataset, while we can see how in the real systems' experiments Silhouette yields very different scores, negatively correlated with the scores from the same metrics.

Secondly, even though the performance of the metrics is often similar in the experiments, there are some evident failures especially in the case of Silhouette in most boundary transposition experiments. This evidence reflects the performance of such metrics in real systems and confirm that ARP scores are more robust, yielding results that are more similar to reference-based metrics.

6.3. Mean and Variance of Correlations

As a final point, it can be seen and it has been already noted that throughout the experiments there are a number of negative correlation values and, in general, cases for which there is a high variance among the correlation coefficients. It is interesting to notice that almost all the negative correlations in table 1 come from SegReFree and Silhouette. Our method, ARP, gives much stronger positive correlations across the board, which highlights its relative strength as opposed to the other two methods. The ARP method only gives a negative correlation value with one particular variant in one case (the QMSum dataset with the MPNET encoder, comparing to the Boundary Similarity metric), and even there the value is very close to zero so that it most likely represents a lack of correlation rather than a negative one.

If we look at table 2, highlighting mean and standard deviation for each method and metric, it is clear that the average correlations for ARP methods are far greater than the ones from the competing methods, with SegReFree having the lowest average correlation. On another hand, ARP methods also lead to more robust results across the experiments as the standard deviation are consistently lower for each metric and when considering all the metrics together.

7. Conclusion

In this work we have established a new framework for reference-free evaluation of topic segmentation systems, which can potentially allow self-supervised training and scoring of such systems.

We have also proposed a set of such metrics based on the ARP method, which we have shown to correlate better than existing reference-free methods with traditional metrics for topic segmentation

	Pk		WD		B		All	
	mean ↑	std ↓	mean ↑	std ↓	mean ↑	std ↓	mean ↑	std ↓
SegReFree	0.10	0.48	0.56	0.54	0.14	0.41	0.27	0.48
Silhouette	0.46	0.57	0.42	0.59	0.57	0.38	0.48	0.51
ARP_{std}	0.85	0.07	0.87	0.05	0.74	0.23	0.82	0.12
ARP_{cos}	**0.86**	**0.07**	**0.89**	**0.04**	**0.75**	**0.20**	**0.83**	**0.10**
ARP_{pair}	0.80	0.20	0.84	0.16	0.70	0.33	0.78	0.23

Table 2: Mean and Standard Deviation (std) for the correlation coefficients presented in table 1, aggregated by metric. Last column presents the statistics obtained concatenating the results of the relative method for each metric.

based on human expert annotations. We have then tested different aspects of the embedding-based metrics, such as their behaviour under different encoders and in different domains. Our experiments have shown that encoders specifically fine-tuned for sentence-level tasks mostly work better, even when compared to LLMs which are bigger by several orders of magnitude. Furthermore, we have shown how the performance of these metrics can change a lot depending on the domain of the application. In the best cases, our best method outputs scores which are almost perfectly correlated with the scores output by reference-based metrics. When the method is applied to noisier data like dialogues in meeting transcripts, however, this type of metric shows several shortcomings. Further experiments with progressively removing or transposing ground truth boundaries mostly confirmed the results with real systems, while also showing how this type of evaluation might be too lenient and it is better used together with non-synthetic results.

By looking more closely at the aggregated correlation coefficients for the different methods we have used, we have further confirmed that ARP scores are better suited across all experiments, yielding not only stronger positive correlations but also less variable correlation coefficients across encoders, datasets and metrics, therefore providing a more stable metric.

Finally, this work lays the foundation for a new type of reference-free, embedding-based metrics for topic segmentation, which originates from but could also extend to different tasks like topic modelling and text clustering more in general, where different works in this direction already exist.

Limitations still exist, especially related to dialogue data. Future work might expand the present one by proposing solutions for these cases.

8. Ethical Concerns

The main ethical concerns stemming from this work are the risk of misusing the metric, like over-relying on it for critical decisions, as well as possible losses of jobs for linguists in case the metric is used as a total substitute for expert annotators. Furthermore, potential biases could arise in its application, especially if the embedding model's training is unsuitable for the specific domain, while the environmental cost of using very large language models should be leveraged against the relative benefits when choosing the encoder for this family of methods.

9. Acknowledgements

This work received partial financial support from the UK EPSRC under the Sodestream (EP/S033564/1) and ARCIDUCA (EP/W001632/1) projects, and from the Slovenian Research Agency (ARRS) core research programme Knowledge Technologies (P2-0103) and project EMMA (L2-50070).

10. Conflict of Interests

All authors declare that they have no conflicts of interest.

11. Data and Code Availability Statements

All datasets used in this work are publicly available from the following links:

WikiSection: https://github.com/sebastianarnold/WikiSection

QMSUM: https://github.com/Yale-LILY/QMSum

SBBC-RadioNews: https://zenodo.org/records/7821475

All of the reference-free metrics described were implemented by us and they are freely accessible in the following repository: https://github.com/Ighina/ARP_Score.

12. Bibliographic References

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander

Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

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

David L. Davies and Donald W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Ameet Deshpande, Carlos E. Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen, and Karthik Narasimhan. 2023. Csts: Conditional semantic textual similarity. In *arXiv*.

Chris Fournier. 2013a. Evaluating text segmentation. Master's thesis, University of Ottawa.

Chris Fournier. 2013b. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712, Sofia, Bulgaria. Association for Computational Linguistics.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 152–161, Montréal, Canada. Association for Computational Linguistics.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, page 562–569, USA. Association for Computational Linguistics.

Maria Georgescul, Alexander Clark, and Susan Armstrong. 2006. An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In *COLING/ACL 2006 - SIGdial06: 7th SIGdial Workshop on Discourse and Dialogue, Proceedings of the Workshop*.

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. In *2nd International Workshop on Data-driven Personalisation of Television (DataTV-2021) at the ACM International Conference on Interactive Media Experiences (IMX 2021) (DataTV-2021)*.

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. *PeerJ Computer Science*, 7:e408.

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 *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2023)*, pages 408–418, Varna, Bulgaria. INCOMA Ltd.

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

Ismail Harrando and Raphaël Troncy. 2021. And cut! exploring textual representations for media content segmentation and alignment. In *2nd International Workshop on Data-driven Personalisation of Television (DataTV-2021) at the ACM International Conference on Interactive Media Experiences (IMX 2021) (DataTV-2021)*.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 9–16, USA. Association for Computational Linguistics.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. In *arXiv*.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland. 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 - Proceedings of the Conference, volume 2.

Christoph Wolfgang Leiter. 2021. Reference-free word- and sentence-level translation evaluation with token-matching metrics. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 157–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*.

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.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. In *Arxiv*.

Lev Pevzner and Marti A. Hearst. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1):19–36.

Matthew Purver. 2011. Topic segmentation. *Spoken Language Understanding*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Jeffrey C. Reynar. 1999. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, page 357–364, USA. Association for Computational Linguistics.

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

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

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. *arXiv*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS 2020*. ACM.

Linzi Xing and Giuseppe Carenini. 2021. 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*, pages 167–177, Singapore and Online. Association for Computational Linguistics.

13. Language Resource References

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

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

Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. 2023. Multimodal topic segmentation of podcast shows with pre-trained neural encoders. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 602–606, New York, NY, USA. Association for Computing Machinery.

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

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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.