Queen Mary University of London

BBC

Institut "Jožef Stefan", Ljubljana, Slovenija

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

Iacopo Ghinassi, Lin Wang, Chris Newell and Matthew Purver

Topic Segmentation

- Topic segmentation is the task of automatically segmenting a sequence (usually a text) into topically coherent segments.
- In the TV domain, for example, it can be used to extract single stories from TV news shows. In general, it allows fine-grained information retrieval in unstructured databases.
- Individual stories thus extracted can then be repurposed for a variety of downstream tasks (e.g. semantic search, text summarization, etc.)

Our Contribution

- We formalise the novel class of reference-free metrics for topic segmentation.
- We thoroughly evaluate different variants within the class of reference-free metrics for topic segmentation.
- We propose our own variants, showing significant improvements over other metrics from the same class

Problem

- Metrics for topic segmentation are traditionally reference-based.

Problem

- Metrics for topic segmentation are traditionally reference-based.
- But annotated datasets for topic segmentation can be rare, especially in certain domains

Problem

LREC-COLING 2024

- Metrics for topic segmentation are traditionally reference-based.
- But annotated datasets for topic segmentation can be rare, especially in certain domains

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

Problem

LREC-COLING 2024

- Metrics for topic segmentation are traditionally reference-based.
- But annotated datasets for topic segmentation can be rare, especially in certain domains
- Reference-free metrics for this task have just very recently started being proposed…

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

Advantages

- No need for expert annotation!
- Can easily and rapidly test in new domains
- Can be used to optimise models without the need of additional reference data
- No need for expert annotation!

Disadavantages

- Usually less indicative of true performance.
- Not well formalised.
- Just one existing metric in this category: what if it’s really bad in our use case?
- Not clear the role of different crucial parameters like sentence encoder (see later)

What our work can help with?

- Not well formalised √
- Just one existing metric in this category √
- Not clear the role of certain crucial parameters √

Limitations

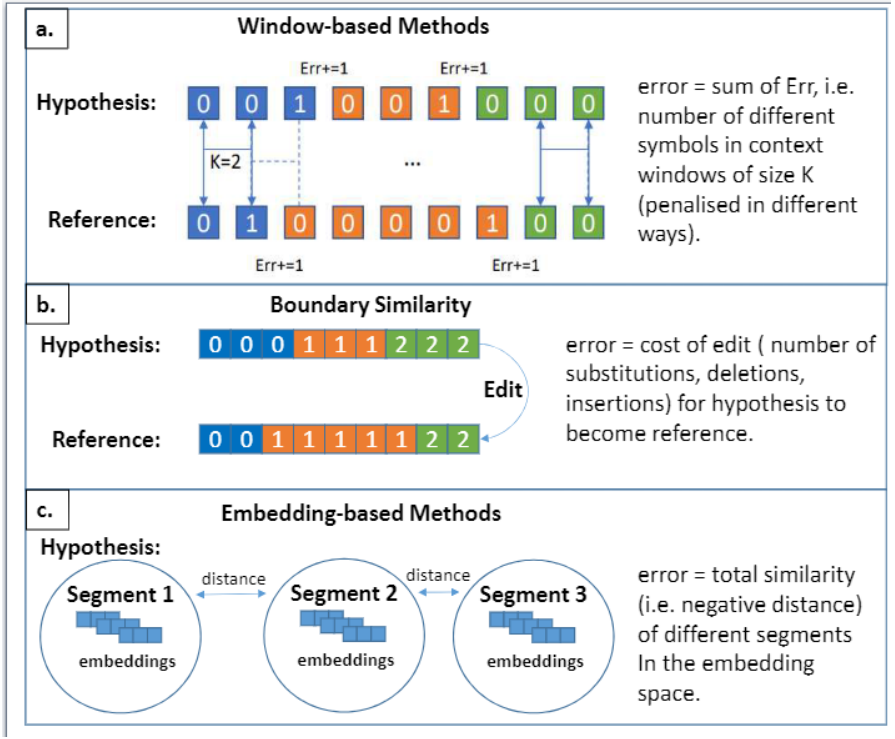
Reference-based metric will still work best, when annotated data is available

Starting from Scratch: The high-level picture

- A broad classification of existing metrics can be summarised as following:

Starting from Scratch: The high-level picture

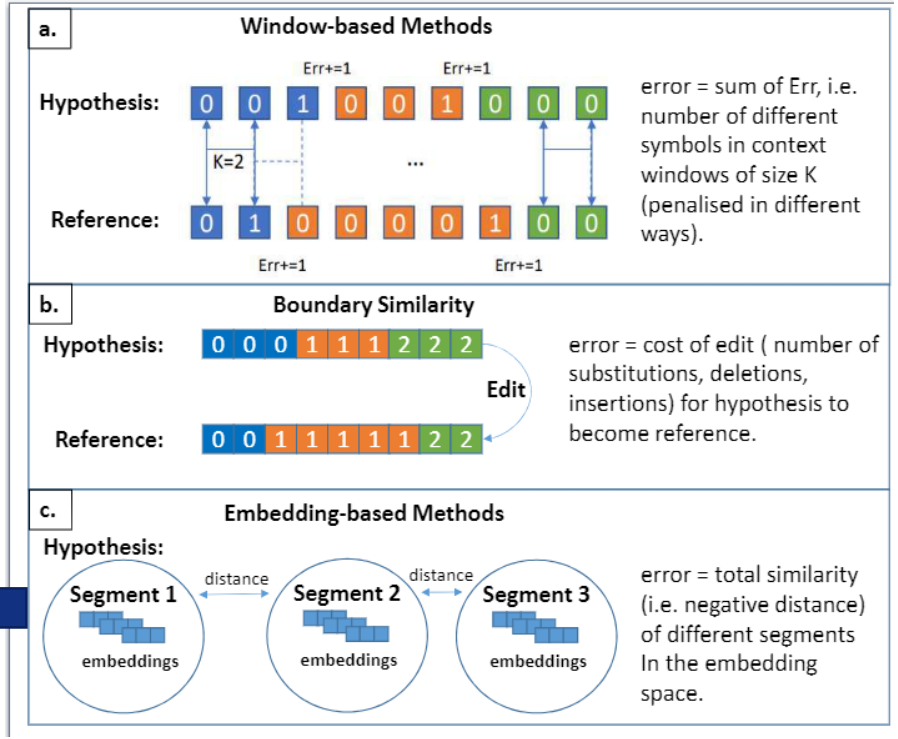
- A broad classification of existing metrics can be summarised as following:



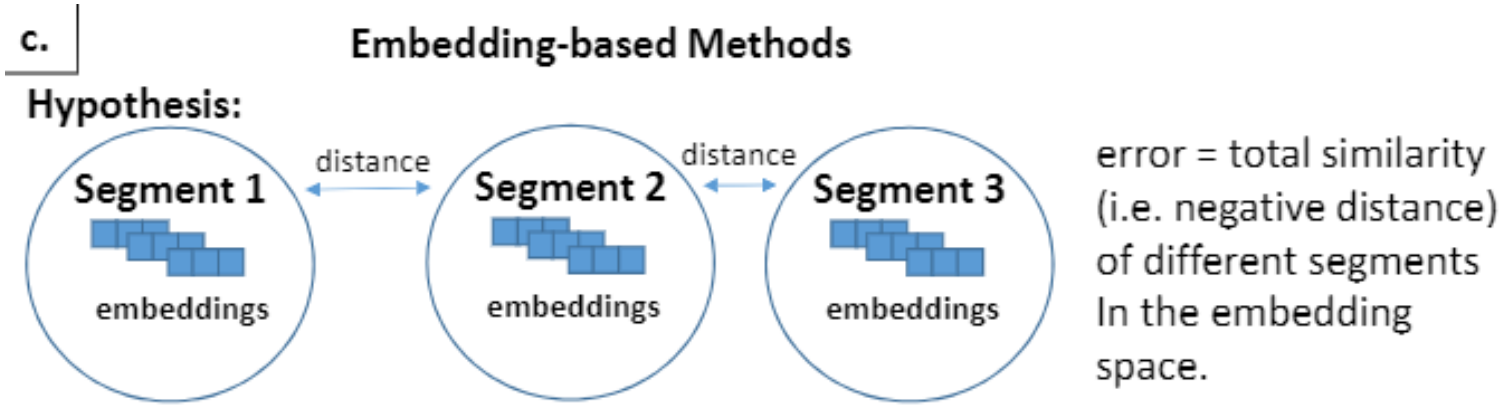
Starting from Scratch: The high-level picture

- A broad classification of existing metrics can be summarised as following:

Here is where we focus

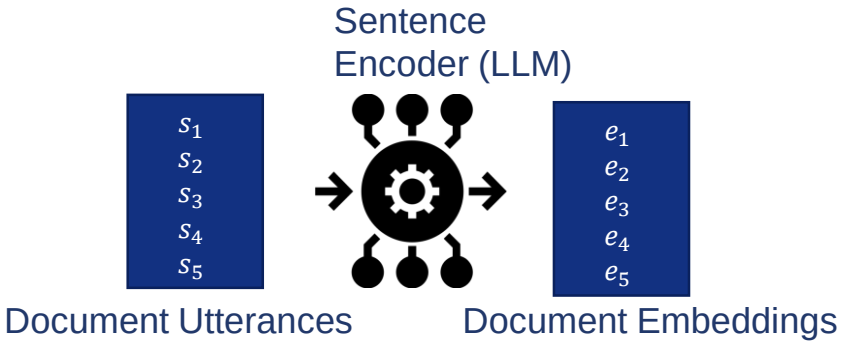


Embedding-based Methods are Also Reference-free

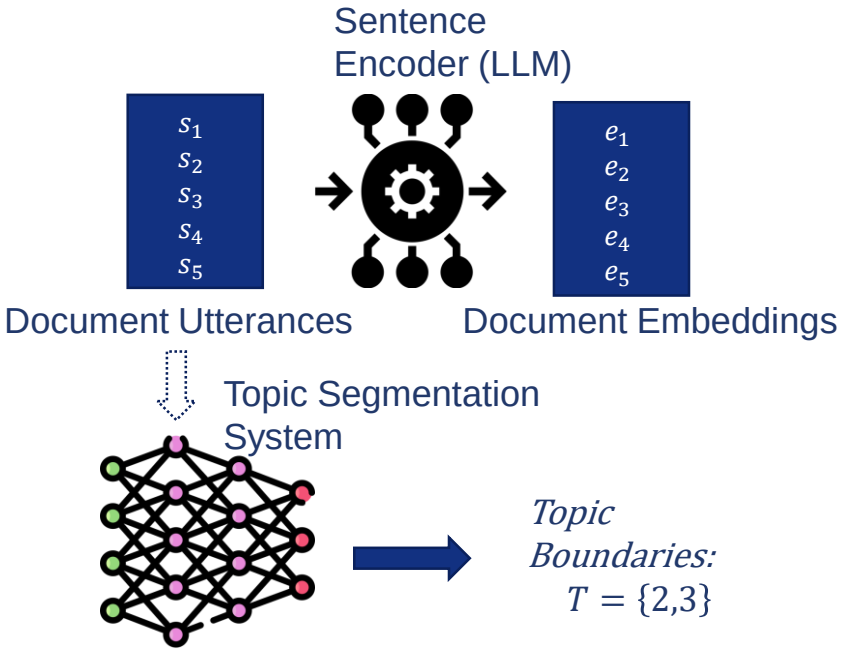


This family of metrics just rely on the inter-segmental embeddings, making it **Similar to clustering evaluation**

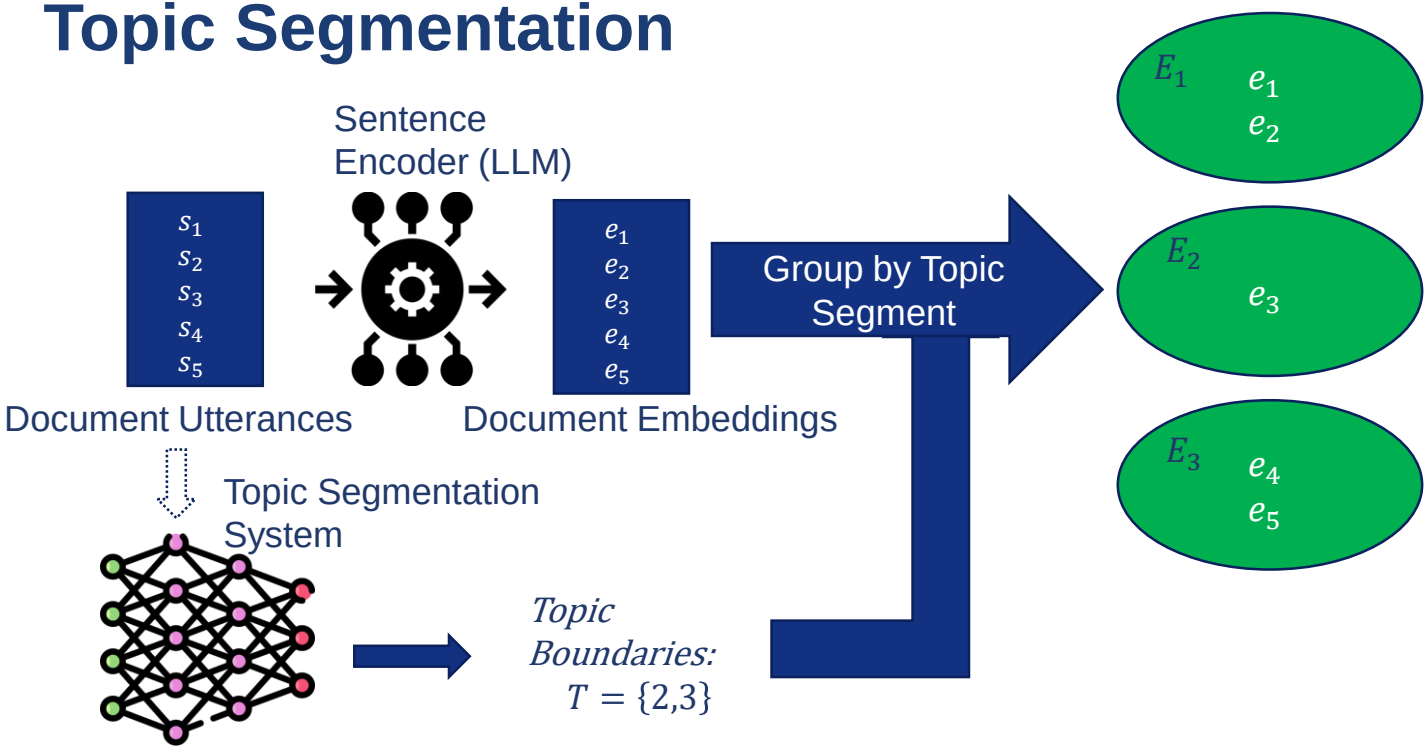
Formalising this Class of Metrics for Topic Segmentation



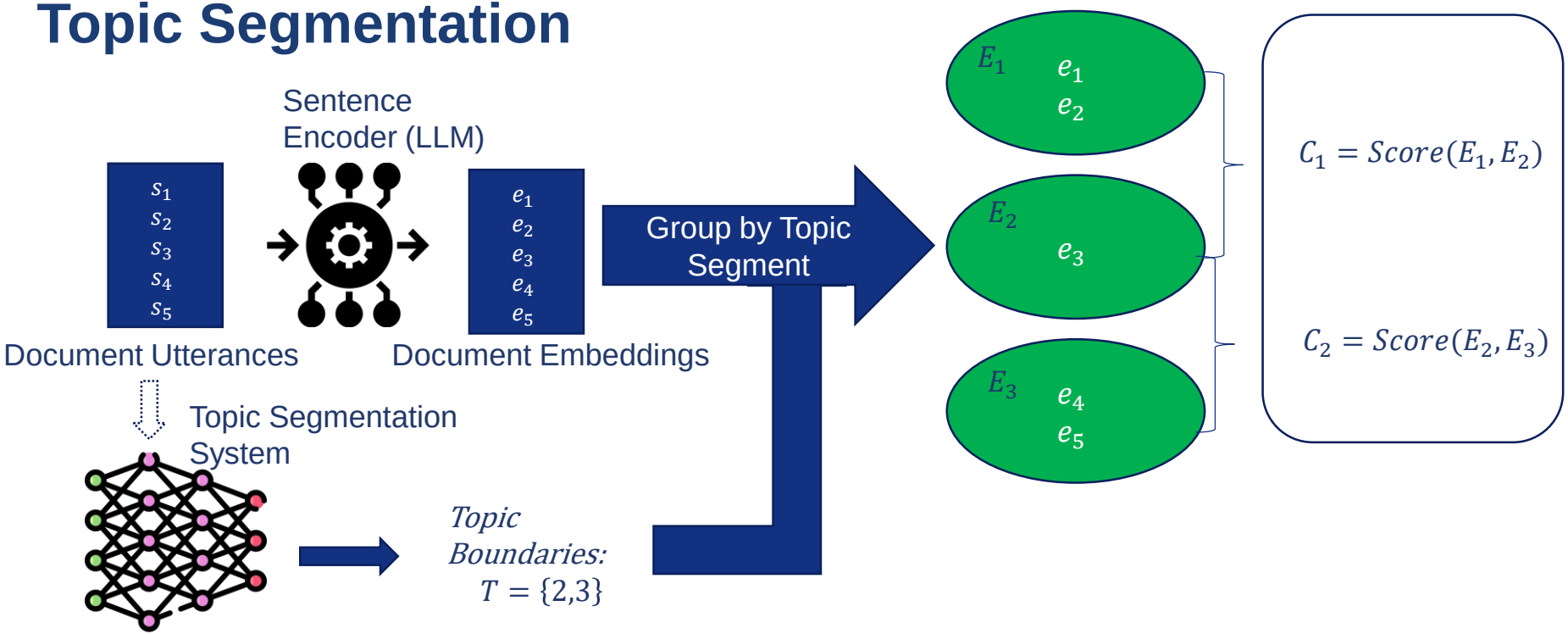
Formalising this Class of Metrics for Topic Segmentation



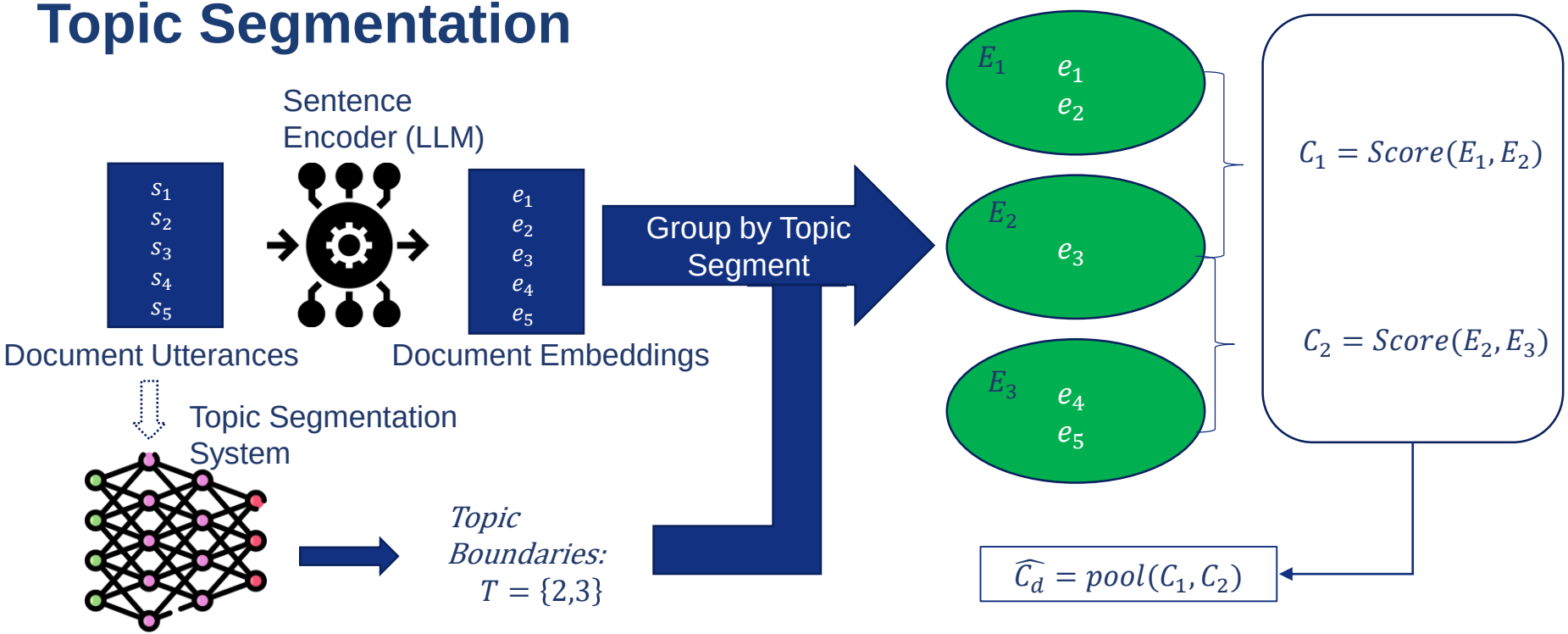
Formalising this Class of Metrics for Topic Segmentation



Formalising this Class of Metrics for Topic Segmentation



Formalising this Class of Metrics for Topic Segmentation



Few, but Crucial Parameters

- Score function
  - Determine the method used to compare grouped embeddings.
  - Some desiderata: segment length independent, lead to greater distance when segments are semantically more different, capture the local breaking cues

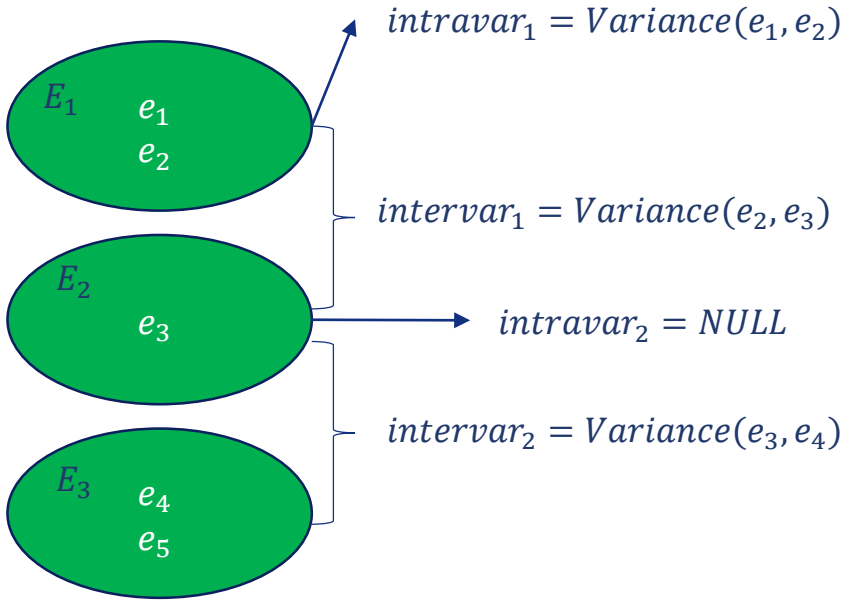
Few, but Crucial Parameters

- Score function
  - Determine the method used to compare grouped embeddings.
  - Some desiderata: segment length independent, lead to greater distance when segments are semantically more different, capture the local breaking cues

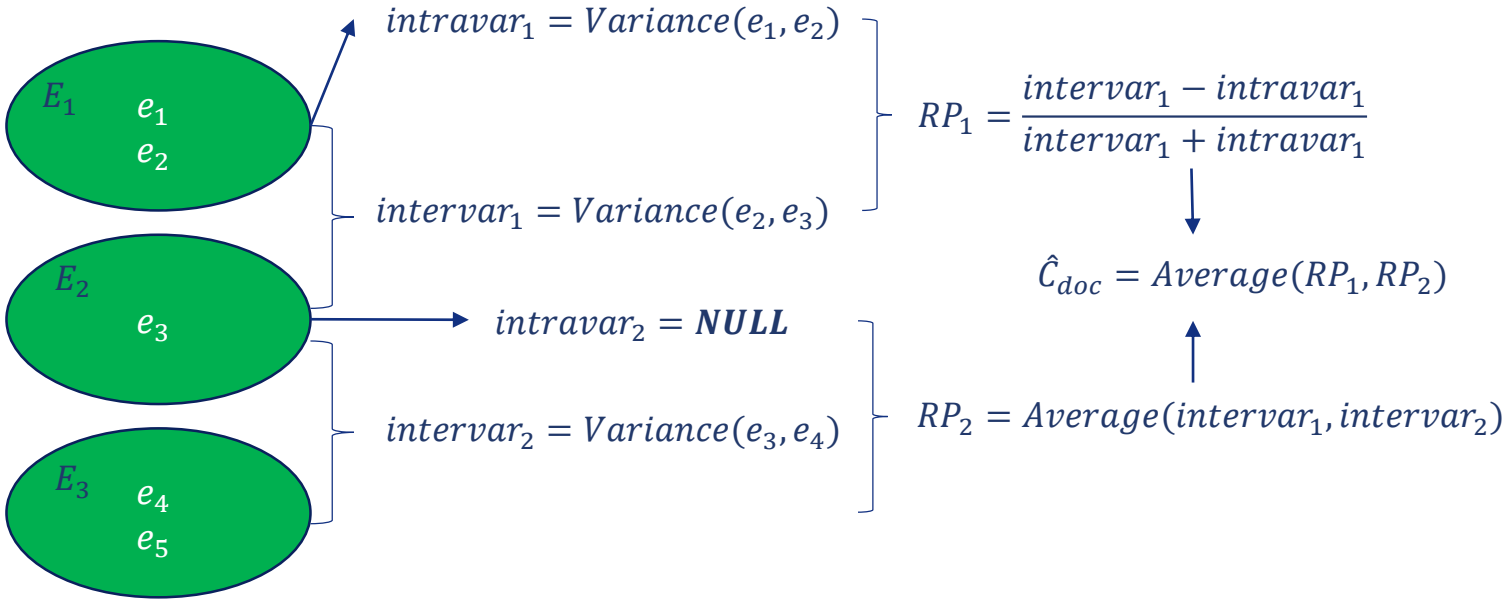
- Pool function:
  - Generally less important.
  - Simple average can be a robust baseline.

Score Functions: ARP (1)

- ARP (Average Relative Proximity):
  - Defined in Ghinassi et al. (2023)
  - It involves the comparison of inter and intra-segmental variances measured in different ways, keeping the length of the compared segments fixed.
  - Tried different functions to measure variances:
    - Standard Deviation (ARP_{std}): L2 norm of the standard deviation of the embeddings in the group.
    - Cosine Similarity (ARP_{cos}): average cosine similarity of the embeddings in the group with the group average.
    - Pairwise Cosine Similarity (ARP_{pair}): average cosine distance between all embeddings in the group.



Score Functions: ARP (2)



Score Functions: Baselines

- SegReFree: first and only embedding-based reference-free metric for topic segmentation proposed.
  - It involves the use of the Davies-Bouldin Index (Davies & Bouldin, 1979) metric for clustering evaluation with the addition of a penalty term to penalise clusters (i.e. topic segments) which are too short.
  - Unlike the original metric, it just computes the metric for adjacent clusters, i.e. segments next to each other: this is consistent with all the metrics we compare
- Silhouette Score: we also implement another traditional metric for clustering evaluation and adapt it to this context.
  - The Silhouette score (Rousseuw, 1987) is a well understood metric to evaluate clustering techniques.
  - Only modification, also in this case, is to limit the metric to compare just adjacent clusters.

Experimental Setup: Sentence Encoders

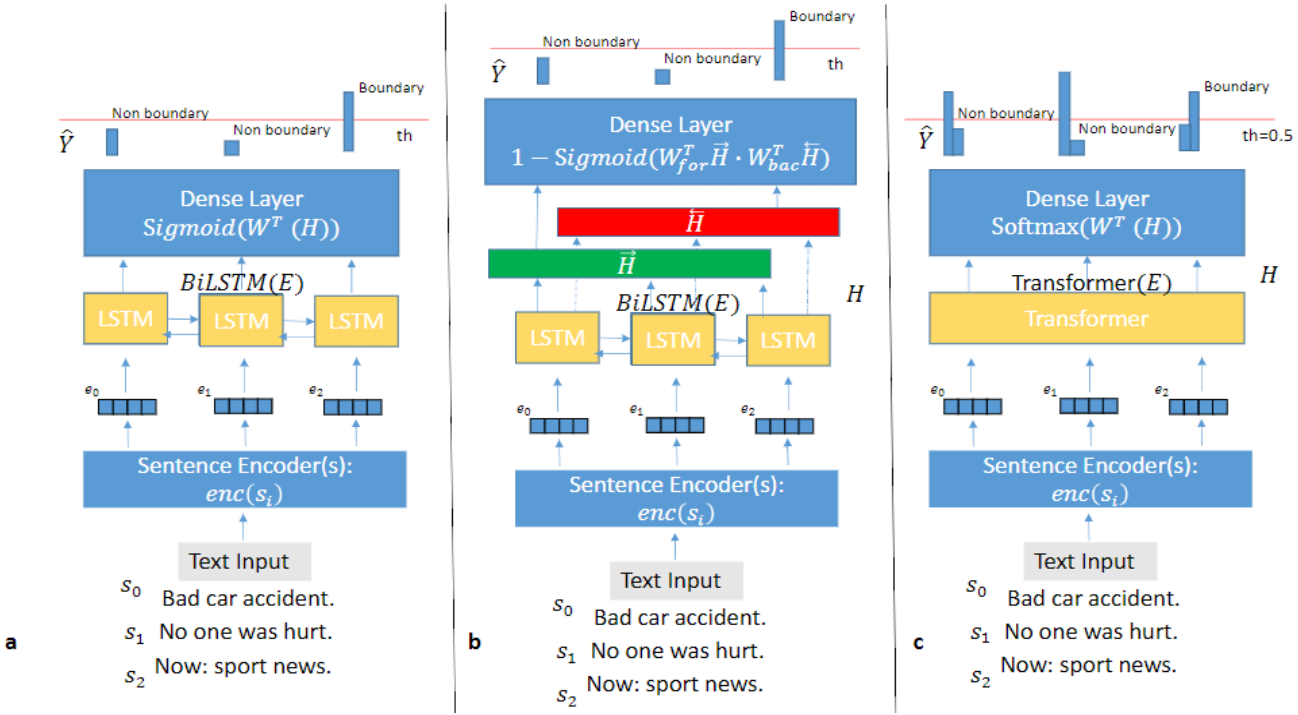
In all models settings we used, we have tried 2 different sentence representations from LLMs:

1) **RoBERTa**: sentence embeddings extracted by averaging the last layer from RoBERTa (Liu et al., 2019), abbreviated (abbr.) RoB.

2) **MPNET**: sentence embeddings generated by all-mpnet-base-v2, a general purpose sentence encoder released and benchmarked by the authors of the popular python framework sentence-transformers and previously used by Lee et al., 2023.

3) Falcon: Big LLM (13 Billion parameters) reaching high-performance in various benchmarks at the time of writing. Mean pooling over last layer.

Experimental Setup: Topic Segmentation Models



- We used three different sequence models: a normal BiLSTM classifier (**a**), a specific version of BiLSTM (Dot-BiLSTM) where the final score is given by the dot product of forward and backward states (**b**), and a Transformer encoder classifier (**c**).
- All these models were used by previous literature and using them in a consistent setting can tell us which architecture performs best for the task.

Experimental Setup: Datasets

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,

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.

Experimental Setup: Evaluation

We compute Pearson Correlation coefficients of the different reference-metrics with traditional metrics based on human annotations on the given datasets.

We include the following traditional metrics to compare with the reference-free ones:

- Pk → window-based
- Window Difference (WD) → window-based
- Boundary Similarity (B) → edit-boundary based

Real Systems Evaluation

		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**

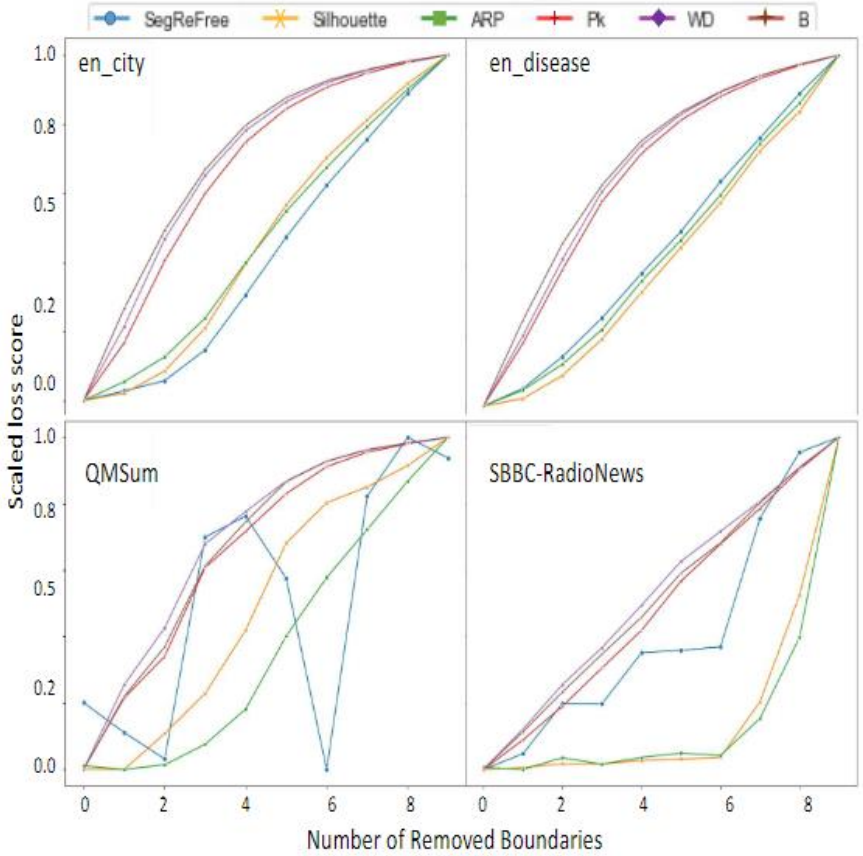
Synthetic Evaluation Results

- We additionally evaluate the metrics in two synthetically created scenarios:

1. Boundary removal: we remove k boundaries from the ground truth and compare how the reference-free segmentation metrics behave
2. Boundary transposition: we transpose boundaries by a number k of sentence from the ground truth and compare how the reference-free segmentation metrics behave

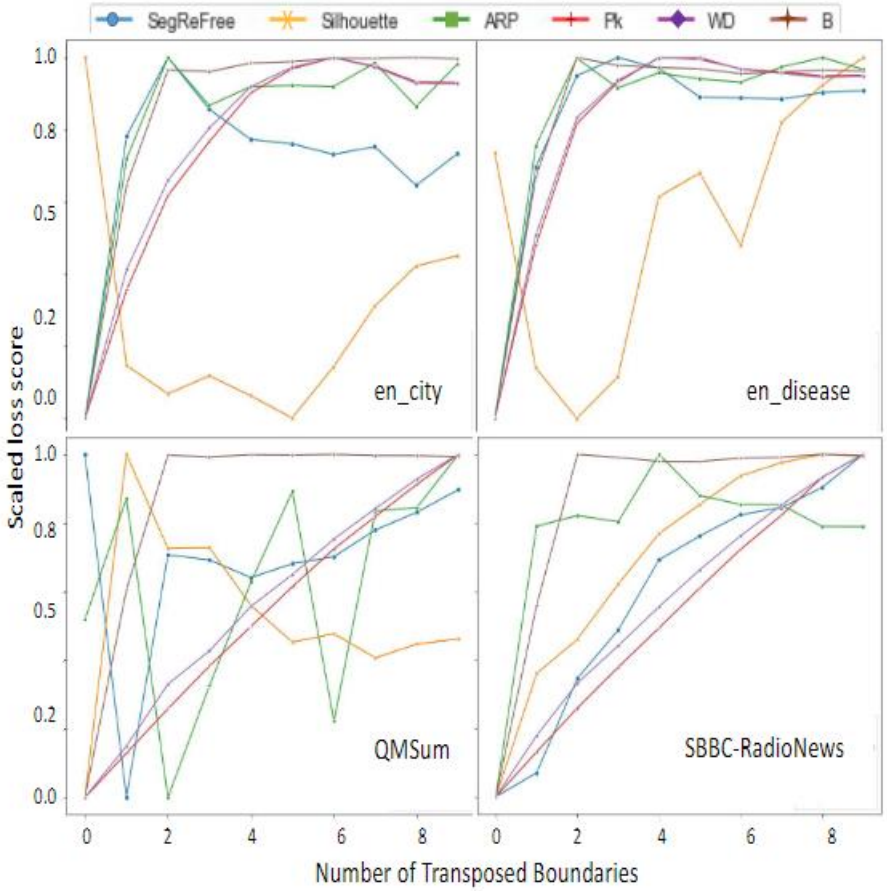
Boundary Removal

Note: scores from 0 to 1 are normalised metric-wise (i.e. 0 is the best score in the metric group and 1 the worst)



Boundary Transposition

Note: scores from 0 to 1 are normalised metric-wise (i.e. 0 is the best score in the metric group and 1 the worst)



Conclusion and Future work

We have formalised the class of reference-free, embedding-based metrics for topic segmentation.

In this class, we have proposed state-of-the-art methods, which reach very close performance to reference-based metrics.

The ARP score method, then, is a viable way to evaluate topic segmentation systems in the absence of annotated data.

Further work will need to address problems related to boundary transposition as well as addressing the edge cases (e.g. one-sentence segments).

Future advancements in sentence encoding will also benefit this approach further.

Queen Mary
University of London

Thank you very much!