Extractive summarisation for German-language data: a text-level approach with discourse features

Freya Hewett [1,2]
[1]AI & Society Lab
Humboldt Institute for Internet and Society
Berlin, Germany
firstname.lastname@hiig.de

Manfred Stede [2]
[2]Applied Computational Linguistics
University of Potsdam
Potsdam, Germany
lastname@uni-potsdam.de

Abstract

We examine the link between facets of Rhetorical Structure Theory (RST) and the selection of content for extractive summarisation, for German-language texts. For this purpose, we produce a set of extractive summaries for a dataset of German-language newspaper commentaries, a corpus which already has several layers of annotation. We provide an in-depth analysis of the connection between summary sentences and several RST-based features and transfer these insights to various automated summarisation models. Our results show that RST features are informative for the task of extractive summarisation, particularly nuclearity and relations at sentence-level.

1 Introduction

Extractive summarisation involves directly using select phrases and sentences from a text as a summary, which still remains a strong method for producing summaries despite its simple nature (Huang et al., 2020). In this study, we examine the link between facets of Rhetorical Structure Theory (RST) and the selection of content for extractive summarisation. RST is a framework which posits that every part of coherent text has a role and a function and represents texts in a hierarchical tree structure (Taboada and Mann, 2006). The RST framework consists of the segmentation of the text into Elementary Discourse Units (EDUs), which are then grouped into bigger segments – depending on the way they relate to each other – which forms the hierarchical structure of the text. The relations between these segments are defined, and a nucleus-satellite status is given. This nucleus-satellite allocation stems from the observation that within the majority of relations which hold between two segments, one segment tends to be more important than the other. This notion of importance seems to have an inherent link to summarisation and various studies have examined this link (see Section 2). In this study we look at the role that these various aspects play in extractive summarisation for German-language texts and transfer these insights to different types of models. In this context, we introduce a new dataset of extractive summaries, analyse the RST-based features of these, collating the best features proposed over the last 20+ years and introducing a new document-based sentence embedding and use these in both linear and nuclear models. Our results compare favourably to those from closely related work on English (Louis et al., 2010). We elaborate on this study and other related work in Section 2, before describing our German-language extractive summarisation dataset in Section 3. In Section 4, we provide a detailed analysis of various RST-based features and examine how these features are distributed in the summaries. We describe our summarisation models and the results of our experiments in Section 5 before discussing the results and providing some concluding remarks in Sections 6 and 7.

2 Related work

Since the RST framework was proposed in the late 1980s (Mann and Thompson, 1988), various studies have examined the link between discourse structure and summarisation, building on even earlier concepts of text-level structural analysis ('macrostructures') (van Dijk and Kintsch, 1983). A study from Marcu (1999) built on ideas proposed by Ono et al. (1994) and empirically analysed the link between the nuclearity aspect of RST and extractive summarisation on a small sample of five texts. Marcu (1999) concluded that there is a strong correlation between nuclei and what readers perceive to be the most important units in a given text, and implemented an automatic summarisation system using RST trees. Louis et al. (2010) conducted an analysis of the RST-DT corpus which contains newspaper articles with various annotation layers including RST and other discourse structures. Louis et al.

Proceedings of the 29th International Conference on Computational Linguistics, pages 756–765
October 12–17, 2022.

(2010) specifically evaluated a subset of the corpus which consists of 150 extractive summaries, for which annotators were asked to select the most important EDUs. They analysed structural features derived from RST trees, such as the depth of a segment, in comparison to other discourse (e.g. the semantics of PDTB relations) and non-discourse features and found that the structural RST features were most useful for automatically selecting sentences for summaries. Zhong et al. (2020) used automatically parsed discourse features for the task of sentence deletion for text simplification, a task which shares many similarities with extractive summarisation. They used RST-based features – such as local nuclearity, relations and the position of the sentence in the tree – in both a linear model and a neural model.

Other studies instead looked specifically at the role that discourse segmentation plays in extractive summarisation. Li et al. (2016) automatically parsed a corpus of English language newspaper articles and compared the RST segmentation to manually annotated *summary content units*. Molina Villegas et al. (2011) also looked at discourse segmentation of Spanish-language texts for the task of sentence compression, which they consider to be sentence-level summarisation.

Neural approaches have also been proposed for combining discourse structure with English-language summarisation. Xu et al. (2020) created a BERT-based model which takes discourse units as input (as opposed to sentences) and also encodes automatically parsed RST trees in a CNN layer in the network, which resulted in an improvement to the state-of-the-art for English-language extractive summarisation. Liu and Chen (2019) experimented with three different neural architectures and compare using sentences as input to discourse segments for the task of extractive summarisation. The models which use discourse segments score higher in an automatic evaluation.

The link between summarisation and discourse structure is further explored by Xiao et al. (2021). They hypothesise that the link between the two may be bidirectional and analyse if summarisation can inform discourse structure by generating RST trees from the inner layers of a Transformer-based summarisation model.

3 Data

As far as we know, the Potsdam Commentary Corpus (PCC) is the only German-language dataset with RST annotations (Stede, 2004). The corpus consists of 176 commentaries from a German regional daily newspaper, the *Märkische Allgemeine Zeitung*. The commentaries have various layers of annotation, including part-of-speech, syntax and discourse structure but does not yet have summaries; we therefore created these ourselves. Each extractive summary consists of 3 key sentences; we use the term ‘key sentence’ throughout this paper to refer to the sentences selected to be part of the summaries. The commentaries have an average length of 11.4 sentences, so the summaries represent ca. 26% of the average length (in sentences). The extractive summaries available for the RST-DT corpus, used for example in the study by Louis et al. (2010), are of a similar size (the square root of the number of EDUs for each text) which makes the corpora easier to compare. We make the summaries publicly available[1], and a sample annotated text can be seen in Table 1.

3.1 Annotation task

To produce these summaries, annotators were asked to choose 3 sentences from each text that represent the core of the text and rank these in order of importance. The task description specified that ‘important’ in this context refers to the suitability of the sentence for a summary. Any anaphoric elements should be ‘mentally’ replaced by that what they are referring to. This annotation task is inherently subjective as the notion of importance can be interpreted in different ways. We discuss this point in more detail in the following Sections (3.2 and 3.3). On average, a sentence has 14 tokens in our corpus, which rises to 18 tokens on average for the annotated sentences. These three key sentences consist of 5 EDUs, on average.

3.2 Inter-annotator agreement

For a sample of 30 texts, two sets of annotations were gathered which resulted in a Cohen’s Kappa score of .32 when comparing the three selected key sentences (without the ranking, (Cohen, 1960)). Similar (yet lower) scores have been reported in other annotation studies of this kind: .28 when selecting 20% of ‘salient sentences from each comment which summarize it’ from online debates

[1] https://github.com/fhewett/pcc-summaries

(Sanchan et al., 2017), an average of .30 when selecting 8% to 16% of 'summary sentences that are informative and can preserve discussion flow' from meeting transcripts (Liu and Liu, 2008), and .23 when selecting sentences with relevant 'nuggets' (clauses containing one verb and one noun that are semantically 'important in the context of the given topic') in different genres of German text (Benikova et al., 2016).

The relatively low agreement score reflects the subjective nature of our task; however, all texts have at least one sentence in common in both sets of annotations (again, disregarding the ranking). The annotations for these 30 texts were harmonised using a scoring system: the highest ranked sentence was equivalent to 3 points, the second to 2 and the third to 1. The sentence with the most points was then deemed the highest ranked sentence in the harmonised annotation, and so on. Any tied scores were resolved by randomly selecting one of the sentences in the tie.

Some questions were raised by annotators: In some texts there are sometimes multiple sentences which contain the same information, and there is no clear way to rank them. The task description also mentions anaphoric entities, but for phrases such as 'that's why' (as in segment no. 9 in Table 1), it is not clear if these should also be replaced by what they are referring to. The genre of a newspaper commentary also poses a specific challenge: is the journalist's opinion or are the objective facts more important? Due to these comments, we adapted the task description to specify that the chosen three sentences should ideally contain both the objective information as well as the opinion of the journalist, and in cases where this is not possible, then the objective information should be prioritised. If there are multiple sentences which contain highly similar content, then the first sentence should be chosen. As this task description however did not lead to an improved inter-annotator agreement we gave the annotators the first (shorter) task description for the remaining texts. The remaining texts were annotated by four annotators (including the first author), who also all annotated part of the subset of the 30.

3.3 Semantic similarity

We also evaluated the inter-annotator agreement using ROUGE scores (Lin, 2004) and word mover's distance (WMD; Kusner et al. 2015). This is due to an observation that the annotated sentences were often semantically similar and contained the same information even when they were not identical. ROUGE measures the overlapping n-grams in a source and reference summary. We used all three sentences in their ranked order as a summary and used one set of annotations as the reference summary and the other set as the source summary. This resulted in a ROUGE-1 F1 score of .587, which compares to a baseline of .380, which was calculated by comparing one set of annotations to a summary consisting of 3 randomly selected sentences from a text. In a recent survey on automatic summarisation, the ROUGE-1 scores for the most recent extractive systems were between .338 and .414 (Fabbri et al., 2021). WMD calculates the shortest distance between the word embeddings of two sentences or sets of multiple sentences. Again, we used all three sentences combined together and the WMD between the two annotators' summaries was .512, compared to a baseline of three random sentences with a distance of .827. These scores show that whilst the aforementioned IAA score is relatively low when using a strict Kappa measure, the two sets of annotations do have semantic similarities, which indicates commonalities in annotators' choices of content.

4 Analysis

To investigate a potential link between RST and extractive summarisation for German-language texts, we examine various RST-based features of the sentences that were chosen in the annotation task. This builds on the studies by Louis et al. (2010) and Zhong et al. (2020) (as outlined in Section 2).

4.1 Non-discourse features

We first look at non-RST features; the average length of sentences and the (relative) position of sentences in the whole text. The average length of non-summary sentences is 13 tokens, whereas the key sentences have an average length of 18 tokens. Figure 2 shows at what position in the text the key sentences occur.

4.2 Local nuclearity

Local nuclearity refers to the nuclear-satellite relationship within a relation. We analyse the nuclear-satellite relationship at the sentence level. For example, the first sentence in Table 1 consists of 3 EDUs and is the satellite of an interpretation, as can

1	Election results which both candidates are happy with –	
2	what a rare occurrence!	
3	But that was the case yesterday evening for both mayoral candidates.	
4	***Elisabeth Herzog-von der Heide (SPD) was happy about the voting***,	1
5	***that resulted in more than 60 percent for her,***	1
6	***and Hans-Jürgen Akuloff (PDS) was pleased that he brought in the best ever result for the PDS so far.***	1
7	*Of course, it was even easier this time, as there was only one opponent.*	
8	*Luckenwalde has a broad range of diverse parties.*	
9	That's why SPD or PDS supporters should not get ahead of themselves.	
10	**Yesterday's outcome is the result of the two candidates,**	2
11	**after they eliminated the third candidate in the first round.**	2
12	Nothing more and nothing less.	
13	After the first round of Skat, the second round of Mau Mau was won by Herzog, the Queen of Hearts.	
14	It's neither a surprising nor a particularly phenomenal result but it's most definitely a clear outcome.	
15	***There was no real tension anyway,***	3
16	***after the outcome was pretty much decided in the first round of voting,***	3
17	***which has simply been confirmed in this final round.***	3

Table 1: Example text from the dataset (text ID: maz-14654). The sentences **in bold** are those that have been selected for the summary in our gold annotation. The sentences *in italics* are those chosen by our best model (FFN). The column on the left contains the ID for the segments, the column on the right the rank given to the sentence by the annotator.

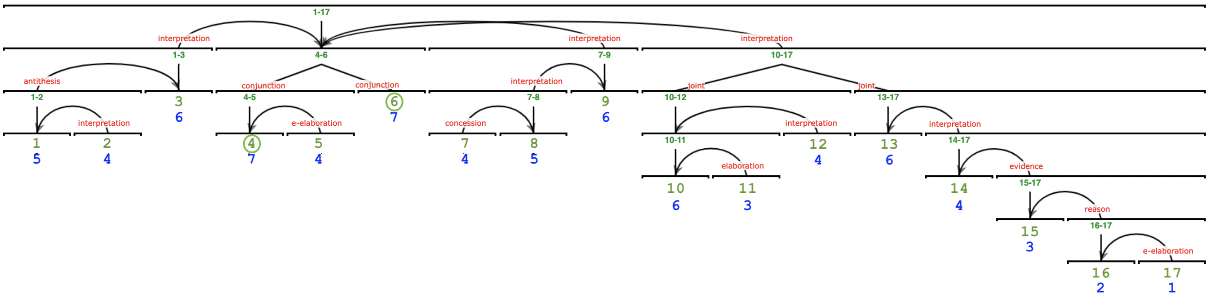

Figure 1: The RST tree for the example text (maz-14654, created using RSTWeb (Zeldes, 2016)). The circled EDUs are the most-nuclear, the blue numbers are the depth scores.

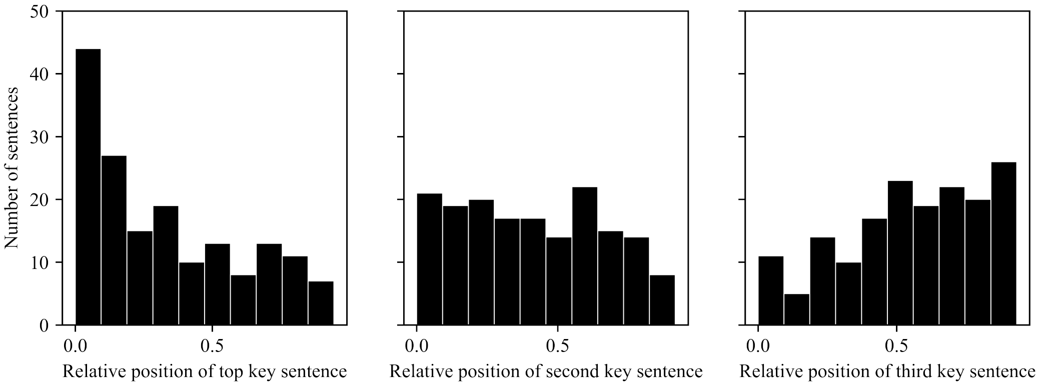

Figure 2: The positions of the three sentences in each text.

be seen in Figure 1. This analysis reflects RST's 'deletion test': the nuclearity assignment can be considered to be correct if once the satellites are deleted, the remaining EDUs still convey the main message(s) of the text (Mann and Thompson, 1988). Although simply considering the sentence-level nuclearity status does not take the rest of the tree structure into account, we feel it is still a fruitful aspect to analyse due to the "strong compositionality criterion" or strong nuclearity principle (Marcu, 2000): if a relation holds between two spans then the relation also holds between the nuclei of these spans, therefore the assignment of nuclearity status at a terminal level is fundamental as the importance is propagated up the tree.

Of all the sentences annotated as being key, 70% were nuclear. This compares to 61% of sentences in the whole dataset.

4.3 Global nuclearity

We use the term global nuclearity to refer to nuclearity with respect to the whole tree and not just at a local level. Huber et al. (2021) examine the loss of information that a binary nuclearity assignment can lead to and highlight this loss with reference to downstream tasks such as summarisation. To counteract this, we look at the **depth score** of EDUs, which allows for a more nuanced approach to nuclearity, and the **most nuclear** EDUs, which takes the whole tree into account. We use the term **depth** to refer to what Marcu (1999) terms importance score, which has also been implemented in more recent studies (Louis et al. 2010, who also use the term depth; Huber et al. 2021).

This score is calculated using the nuclearity of units and their relation to other nodes in the RST tree. We adapt the original equation (by removing the part that refers to parenthetical units, as these do not feature in our corpus) for calculating the importance score *s(u,D,d)* of a unit *u* in a discourse tree *D* with depth *d* as follows:

$$s(u,D,d) = \begin{cases} d, & \text{if } u \in prom(D) \\ max(s(u, & \\ C(D), d-1)), & otherwise \end{cases} \tag{1}$$

where C(D) refers to the child subtree and *prom(D)* the promotion set of a node: if the node is a leaf node then the promotion set is simply the leaf itself, if the node is internal then the promotion set is the union of the salient units of its immediate nuclear children. The score is simply the depth in the tree where the leaf units first occur in a promotion set. For example, we can see in Figure 1 that the depth score of the fourth EDU is 7, as it belongs to the promotion set of the root node, and the whole discourse tree has a total depth of 7: if we follow the nuclear links from the root node, we get the promotion set of EDUs 4 and 6. The depth score of EDU 3 is 6, as it belongs to the promotion set of the subtree which is one level below the root. We refer to the units with the highest depth scores as **most-nuclear** as proposed by Mann and Thompson (1988): the most-nuclear nodes can be determined by following nuclear links from the root node to the (leaf) EDU nodes. Depth scores are normalised (as otherwise the length of the text would influence the scores). As the depth score is calculated on an EDU-level (and not sentence-level), we define the depth score for a sentence as the maximum of the depth scores of the EDUs that it contains. In the same vein, a sentence is considered most-nuclear if it contains a most-nuclear EDU.

25% of the key sentences are also most-nuclear EDUs or contain a most-nuclear EDU. This corresponds to about 38% of all most-nuclear EDUs in the corpus.

65% of the texts have at least one key sentence which contains or corresponds to a most-nuclear EDU. When comparing the sentences with the three highest depth scores (taking the maximum score for sentences which contain more than one segment) to the three key sentences, 46% match.

4.4 Relations

We examine the relations of the annotated sentences: in the example in Table 1, the first two EDUs constitute one sentence and therefore the relation that we consider in our analysis would be *Antithesis* (satellite, cf. Figure 1). We consider relations to be important in the context of this analysis because annotation guidelines for different RST corpora pre-define the nuclearity assignment for relations: for example, for the relation *Purpose*, the underlying goal of the activity is the satellite and the activity itself is the nucleus (RST-DT guidelines (Carlson and Marcu, 2001); PCC guidelines (Stede et al., 2017)). Stede (2008) states that this can be problematic: by pre-defining the 'activity itself' as the nucleus, this does not allow for any flexibility for the case where actually the 'goal of the activity is more important'. Marcu (1998) suggests that to improve his proposed nuclearity-based summarisation method, one should also "exploit the semantics of rhetorical relations", citing the relation *Exemplification* as an example of a relation where even the nuclei are probably not relevant for a summary. We therefore also consider nuclearity in connection with relations as they are intertwined with each other and could give some quantitative evidence for the problems highlighted here.

Figure 3 shows which relations the key sentences have, with and without the nuclearity assignment. As the selected sentences represent 26% of total sentences, any relation ratio above this is above average. *Evaluation-s* nuclei, *Background* satellites, and *Evidence* nuclei feature the most in the key sentences.

5 Summarisation models

We use these RST features and non-discourse features in linear and neural models to predict which sentences should be kept for an extractive summary. We frame extractive summarisation as a binary classification task, predicting whether each sentence should be included in a summary or not. Our dataset consists of 167 texts and 1894 sentences; we use 30% of the dataset as a test set.

5.1 Features

We combine and adapt the feature sets used in the studies by Louis et al. (2010) and Zhong et al. (2020) and also introduce a new way of creating sentence embeddings. We also perform a detailed feature ablation to see which RST-based feature(s) are most useful for extractive summarisation. The features we use in comparison to related work can be seen in Table 2. As Zhong et al. (2020) looked specifically at content selection in the context of simplification, we do not use some of their features such as *topic* or *readability score*, as they do not seem relevant for the task of summarisation. We do not use features related to specific words (*content words, topic words*) as we anticipate that this information is implicitly encoded in the embeddings we use. Our focus is on RST-based features and a more fine-grained analysis of these on an individual basis, and so we do not include PDTB related features. The features *nucleus-satellite penalty* and *promotion score* are used to measure global nuclearity: we decide to solely use *depth* as it was shown to achieve the better results than the other two varia-

Features	**Louis et al. (2010)**	**Zhong et al. (2020)**	**Present study**
Non-discourse			
Sentence length	X		X
Document length (in sentences and tokens)		X	
Topics		X	
Position of sentence	X	X	X
Sentence embeddings		∼	X
Embeddings with document context			X
Readability scores		X	
Content words, topic signature words	X		
Discourse			
PDTB connectives	X	X	
Local nuclearity		X	X
Depth	X	∼	X
Nucleus-satellite penalty	X		
Promotion score	X		
Relations		∼	X
Most-nuclear			X

Table 2: Features used in the present study and in related work. The ∼ signifies that the features are not fully identical.

Model	**F1**	**R**	**P**	**Acc.**
Baseline	.225	.188	.281	.658
Louis et al. (2010)	.442	.344	.619	.789
LR	.488	.583	.422	.677
FFN	.506	.588	.448	.692
bi-LSTM	.440	.635	.341	.729

Table 3: Results. Highest scoring feature set for each model. We report F1 scores for the minority class (key sentence). R, P and Acc. stand for recall, precision and accuracy respectively.

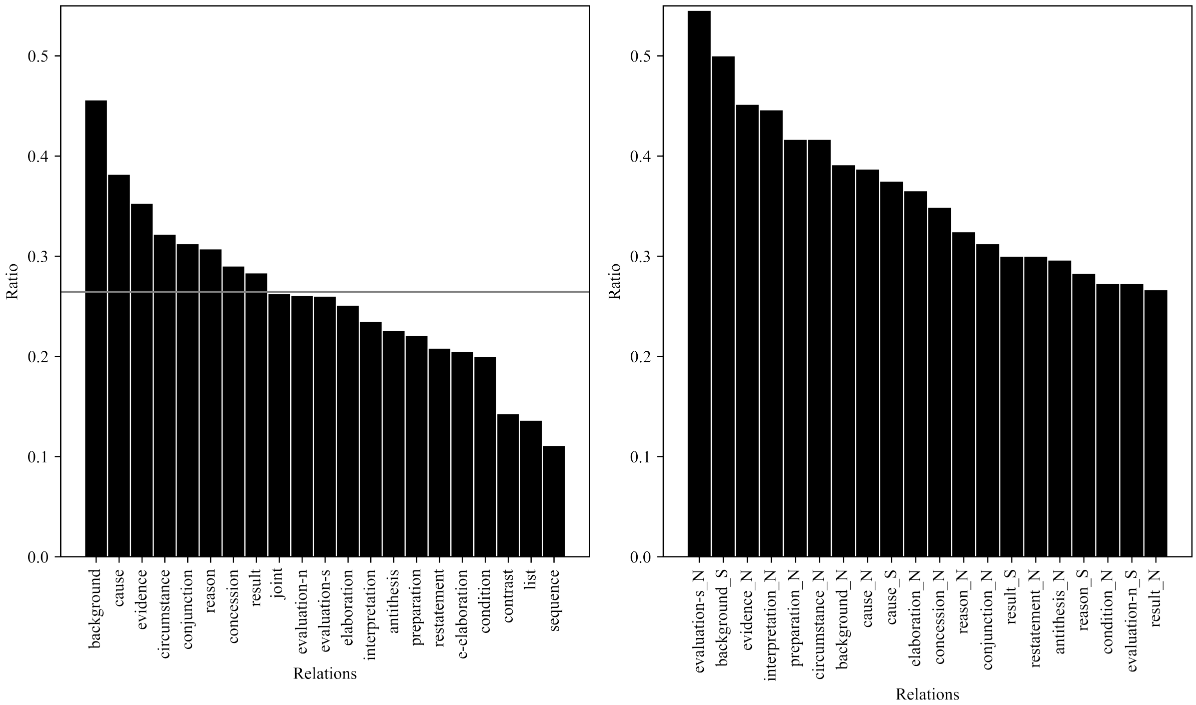

Figure 3: The ratio of relations of the selected key sentences out of the total amount of relations. Any relation ratio above 26% (the line on the graph) is above average. The figure on the right hand side shows the relations and nuclearity assignment of the key sentences; any ratio below 26% is excluded for better visibility. Relations occurring less than a total of 10 times are excluded.

tions in the context of summarisation (Marcu, 1999) and has also been used in more recent work (Huber et al., 2021). Zhong et al. (2020) use the term depth to refer to the level that the sentence occurs at in the tree and do not encode any extra information regarding promotion sets or nuclearity in this feature (in our example in Figure 1, the EDUs 1 and 2 would both have a score of 3).

The **discourse features** are the nuclearity of the sentence, the relation that the sentence belongs to, the depth of the sentence, and whether the sentence is most-nuclear or not. If the sentence contains more than one segment, we take the maximum depth score or most-nuclear score of the segments. For example, the first sentence in Table 1 would have the following features: *satellite*, *interpretation*, *depth score 6/7*, *most-nuclearity 0*. The **non-discourse features** are the length of the sentence (in tokens) and the relative position of the sentence. We also use **sentence embeddings** and an adapted variation of these, in an attempt to incorporate more document-level information. We use S-BERT sentence embeddings which are created by taking the mean of token embeddings of the input (which is usually a single sentence) from an adapted BERT model trained on multi-lingual paraphrase data (Reimers and Gurevych, 2020). We also adapt this by taking the whole text as input and then simply taking the average of the token embeddings between the sentence boundaries, thus producing an embedding for each sentence, instead of just one embedding for the whole text. The intuition is that these sentence embeddings may capture more document-level information by having seen an even larger context whilst producing the token embeddings.

5.2 Setup

As a baseline, we select the first, middle and last sentence to be in a summary. We implement a Logistic Regression model[2], a feed-forward neural network (with two hidden layers; FFN), and a bidirectional LSTM model (with one layer). For the bi-LSTM model, each text is fed as a sequence, so the input dimensions are *batch size, text length, feature dimension*. For all models, we weight the classes to counteract the class imbalance, which is skewed towards the 'non-summary sentence' class.

[2] Whilst we did also experiment with other non-neural methods, we only report on Logistic Regression as it performed the best and is also directly comparable to the work by Louis et al. (2010).

In total we have four discourse features, two non-discourse features and two types of sentence embeddings. We run models with all possible combinations of features. We experiment with different hyper-parameters, these can be found in the Appendix and we make our code publicly available.[3]

5.3 Results

The results of our experiments can be seen in Table 3. The FFN achieved the best F1 score with nuclearity, relations, most-nuclearity, sentence length and position, and sentence embeddings. The Logistic Regression (LR) model also had the best F1 score with this feature set, minus sentence length and position. The bi-LSTM achieved the best F1 score with nuclearity, relations and sentence position, in combination with sentence embeddings. All best-performing models have nuclearity, relations and sentence embeddings in common as features.

In Table 4, the best performing combinations of features can be seen as well as the performance of individual features; whilst our proposed sentence embeddings with additional document context are not the individual feature with the worst F1 score, they generally do not improve results greatly across all types of models. We report F1 score as we are interested in the minority class (key sentences).

The summary produced by the FFN model for our example text can be seen in italics Table 1. Overall, the summary reads well and includes two of the gold key sentences (shown in bold). However, the sentence which corresponds to segment number 8 does not make much sense without the following sentence (which corresponds to segment 9). The ROUGE-1 F1 score for our test set as predicted by the FFN model is .690. The ROUGE-1 recall score is .612 as compared to .479 reported by Louis et al. (2010).

Features	**F1**
Top 3 combinations	
Nuclearity, relations, most-nuclearity, sentence embeddings	.488
Nuclearity, relations, sentence embeddings	.486
Nuclearity, relations, most-nuclearity, sent. length	.486
Ranked individual features	
Sentence length	.462
Sentence embeddings	.452
Local nuclearity	.425
Depth	.391
Embeddings with document context	.385
Position	.376
Relations	.368
Most-nuclearity	.302

Table 4: Feature ablation for the Logistic Regression model. We show the best 3 combinations, as well as the individual features ranked in descending order according to the F1 score achieved when using said feature as the sole input.

6 Discussion

The results of the analysis, particularly Section 4.4 and Figure 3, show that simply looking at nuclearity in isolation may not be sufficient for some downstream tasks, as we have *Background* satellites and *Cause* satellites featuring heavily in the summaries, for example. This is reflected in the results: the best combination of features for both the FFN and LR model is nuclearity (local and global, with most-nuclearity measuring global nuclearity) in combination with relations (and sentence embeddings). The bi-LSTM model and our sentence embeddings with additional document context both perform worse than other models and features: this goes against our intuition that more text-level context is beneficial for such a task as extractive summarisation. We leave it to future work to examine if these embeddings could work in other settings, perhaps with larger datasets or pre-trained in a different manner. Simple models (such as LR) perform well, which suggests that RST features (particularly nuclearity) are strong indicators of importance. Combinations of RST features perform even better than sentence embeddings on their own. As can be seen in Table 3, our F1 scores (and our ROUGE scores, see Section 5.3) are in fact higher than those reported in (Louis et al., 2010). Whilst our recall is higher, our precision and overall accuracy are slightly lower. It is also worth noting that our results are on commentaries, a type of argumentative text, and so the results are not directly comparable (Louis et al. work with the RST-DT, articles from the Wall Street Journal); we leave this to future work to investigate more thoroughly.

[3] https://github.com/fhewett/rst-features

7 Conclusion

In this study we have introduced a new set of extractive summaries for the Potsdam Commentary Corpus; texts which have already been annotated with various linguistic features such as discourse structure, co-reference and syntax. We have shown the connection between RST-based features and sentences chosen for extractive summaries and have transferred these to various models. Our feature ablation experiments could provide useful insights for research on RST parsers for specific downstream tasks: by finding the aspect of RST which is most useful for a task such as summarisation, the parsers can be streamlined and will have less room for error. For example, one of our top models has sentence-level nuclearity and relation information, showing that in this context, the additional RST tree structure is potentially not necessary. We hope that our dataset will enable research on the link between summarisation and other linguistic features. Our analysis using manually annotated discourse structure also provides the necessary evidence to foster research on automatically parsed discourse structure and downstream tasks, such as summarisation.

Acknowledgements

Many thanks to the anonymous reviewers for their valuable comments. This research is funded by the German Federal Ministry of Education and Research (BMBF).

References

Darina Benikova, Margot Mieskes, Christian M Meyer, and Iryna Gurevych. 2016. Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1039–1050.

Lynn Carlson and Daniel Marcu. 2001. Discourse Tagging Reference Manual.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 391–409.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What Have We Achieved on Text Summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469.

Patrick Huber, Wen Xiao, and Giuseppe Carenini. 2021. W-RST: Towards a Weighted RST-style Discourse Framework. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3908–3918.

Matt J Kusner, Yu Sun, Nicholas I Kolkin, and Kilian Q Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML '15, pages 957–966.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The Role of Discourse Units in Near-Extractive Summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. volume Text Summarization Branches Out, pages 74–81. Association for Computational Linguistics.

Fei Liu and Yang Liu. 2008. What are meeting summaries?: an analysis of human extractive summaries in meeting corpus. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue - SIGdial '08*, page 80, Columbus, Ohio. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2019. Exploiting Discourse-Level Segmentation for Extractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 116–121, Hong Kong, China. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 1998. To build text summaries of high quality, nuclearity is not sufficient. In *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 1–8.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. In *Advances in Automatic Text Summarization*, pages 123–136. The MIT Press.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. A Bradford book. MIT Press, Cambridge, Mass.

Alejandro Molina Villegas, Juan-Manuel Torres-Moreno, Eric Sanjuan, Iria da Cunha, Gerardo Sierra, and Patricia Velázquez-Morales. 2011. Discourse Segmentation for Sentence Compression. In *MICAI 2011: Advances in Artificial Intelligence*, pages 316–327.

Kenji Ono, Kazuo Sumita, and Seiji Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th Conference on Computational Linguistics*, volume 1, pages 344–348, Kyoto, Japan. Association for Computational Linguistics.

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.

Nattapong Sanchan, Ahmet Aker, and Kalina Bontcheva. 2017. Gold Standard Online Debates Summaries and First Experiments Towards Automatic Summarization of Online Debate Data. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Budapest, Hungary, April 17-23, 2017, Revised Selected Papers, Part II*, volume 10762 of *Lecture Notes in Computer Science*, pages 495–505.

Manfred Stede. 2004. The Potsdam commentary corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation - DiscAnnotation '04*, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Manfred Stede. 2008. RST revisited: Disentangling nuclearity. In Cathrine Fabricius-Hansen and Wiebke Ramm, editors, *Studies in Language Companion Series*, volume 98, pages 33–58. John Benjamins Publishing Company, Amsterdam.

Manfred Stede, Maite Taboada, and Debopam Das. 2017. Annotation Guidelines for Rhetorical Structure.

Maite Taboada and William C. Mann. 2006. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Teun A. van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Academic Press.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2021. Predicting discourse trees from transformer-based neural summarizers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152, Online. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-Aware Neural Extractive Text Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Amir Zeldes. 2016. rstWeb - a browser-based annotation interface for rhetorical structure theory and discourse relations. In *Proceedings of NAACL-HLT 2016 System Demonstrations*, pages 1–5, San Diego, CA.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse Level Factors for Sentence Deletion in Text Simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(5), pages 9709–9716.

A Appendix

We experimented with the values in brackets, the values in bold are the parameters used in the final models.

Hyper-parameters for FFN

Batch size [8,12,16,32,**64**]
Epochs [5,10,**15**]
Learning rate [**0.5**,0.8,1]
Class weights [(0.3, 2), **(0.3, 2.5)**, (0.3, 6)]

Hyper-parameters for LSTM

Batch size [**8**,12,16,32,64]
Epochs [5,10,**15**]
Learning rate [0.5,**0.8**,1]
Class weights [(0.3, 4), **(0.3, 5.5)**]