EDTC: A Corpus for Discourse-Level Topic Chain Parsing

Longyin Zhang, Xin Tan, Fang Kong*, and Guodong Zhou
School of Computer Science and Technology, Soochow University, China
{lyzhang9,xtan9}@stu.suda.edu.cn
{kongfang,gdzhou}@suda.edu.cn

Abstract

Discourse analysis has long been known to be fundamental in natural language processing. In this research, we present our insight on discourse-level topic chain (DTC) parsing which aims at discovering new topics and investigating how these topics evolve over time within an article. To address the lack of data, we contribute a new discourse corpus with DTC-style dependency graphs annotated upon news articles. In particular, we ensure the high reliability of the corpus by utilizing a two-step annotation strategy to build the data and filtering out the annotations with low confidence scores. Based on the annotated corpus, we introduce a simple yet robust system for automatic discourse-level topic chain parsing.

1 Introduction

Topic information as a crucial auxiliary for text understanding has drawn great attention in recent decades (Wu et al., 2019; Wang et al., 2020; Sahlgren, 2020). In the literature, previous studies on topic modeling usually extract topics by introducing latent variables for tokens for topic assigning (Hofmann, 1999; Blei et al., 2003; Yishu et al., 2017). Similarly, researches on text-tilling achieve topic segments through lexical cohesion modeling (Hearst, 1997; Purver et al., 2006). Instead of lexical cohesion measuring, Rahimi et al. (2015) put their attention on evaluating the organization and cohesion of pieces of evidence and build topic chains on related text units. Besides, recent studies on argument mining explore to build links or clusters for topic-dependent arguments (Wachsmuth et al., 2018; Shnarch et al., 2018; Reimers et al., 2019). Obviously, more and more researches show that there are certain structures among topic segments that deserve deeper exploration.

In this work, we aim to explore the cohesion of topic-related text segments. Different from Rahimi et al. (2015), we show great interest in uncovering how fine-grain topics emerge, evolve, and disappear in an article, which is referred to as discourse-level topic chain (DTC) parsing. Since the DTC structure can provide relatively rich and low-noise information about certain topic aspects of articles, it is meaningful for various NLP tasks like summarization (Perez-Beltrachini et al., 2019), document similarity measuring (Gong et al., 2018), and response generation (Dziri et al., 2019).

In the literature, topic detection and tracking (TDT) (Allan, 2002) is a research area most similar to DTC parsing which aims at identifying new events and tracking how they change over time. However, the events in the TDT task refer to happenings at certain places and times which only compose a small subset of general topics. Recently, Xi and Zhou (2017) manually annotate the first Chinese DTC corpus based on the theme-rheme theory (Halliday and Matthiessen, 2004). By contrast, due to the lack of corpus, previous study on English DTC parsing usually uses unsupervised methods (Kim and Oh, 2011) to explore the structure and trends of important topics hidden within news articles. Obviously, one intractable problem facing DTC parsing is the lack of data.

This research is primarily motivated by (Polanyi and Scha, 1984; Kim and Oh, 2011) on the topic chain concept, (Xi and Zhou, 2017) on DTC corpus construction, and (Reimers et al., 2019) on topic-dependent argument linking. And our contributions mainly include two aspects: (i) building an **E**nglish corpus of **d**iscourse-level **t**opic **c**hain (EDTC) through a two-step annotation method and (ii) lunching a simple but robust Bert-based baseline system for automatic DTC parsing. Moreover, as implied in recent researches on discourse rhetorical structure (DRS) parsing (Zhang et al., 2020; Kobayashi et al., 2021; Zhang et al., 2021), discourse parsing remains challenging due to the lack of data. Under this circumstance, we annotate the

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1304–1312
November 7–11, 2021.

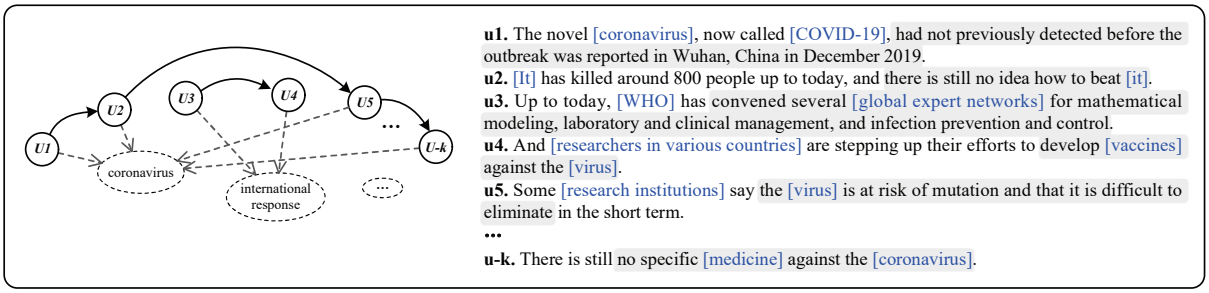

Figure 1: Example DTC structure. TOs are marked with square brackets and TEs are marked with gray background.

DTC structures for the 385 Wall Street Journal (WSJ) articles in the RST-DT (Carlson and Marcu, 2001) corpus aiming to build a bridge between discourse rhetorical structure and DTC structure for discourse researchers to utilize.

2 Corpus Annotation

Before detailing the annotation process, we give a formal introduction to the "topic" mentioned in this paper. In topic modeling, a topic is usually viewed as a probability distribution over a fixed vocabulary (Liu et al., 2016). In addition, previous studies on argument mining usually manually define some coarse-grain topic categories for either topic-dependent argument classification or clustering (Reimers et al., 2019). Different from previous work, topics in this study refer to fine-grained topic categories that fit the context. For example, given the sentence "House prices are expected to be fragile.", the coarse-grained topic label of it could be "economics" and the fine-grained label is "house price". Comparing the two kinds of labels, the first one seems more like the theme of an article which is useful in text-clustering or text-tilling, and the second one gives us more detailed description on the topic itself which is more practical in discourse-level topic chaining. For better understanding of our annotation, we present some preliminary definitions as following:

Discourse Topic Unit (DTU) refers to the elementary topic unit in our annotated DTC structure. In the literature, Xi and Zhou (2017) hold the view that each sentence is composed of multiple DTUs with different sub-topics which they refer to as elementary discourse topic unit (EDTU). Different from them, we study macro DTC structures in this work where each sentence is taken as an independent DTU[1]. It is worth mentioning that not all the DTUs are topic-bearing, there are also some sentences with no topic meaning, e.g., the sentence "Oops!".

Topic Object (TO) could be subject, object, or other noun or noun phrase in the DTU which can provide a certain basis for topic chain parsing. Usually, each TO is closely related to the topic of its DTU, and each DTU maintains an independent TO set. Notably, the "TO" mentioned here is not directly equivalent to the "entity" in co-reference resolution, the judgment of TO requires a comprehensive consideration of document context. For example, given the DTU "Drexel Burnham Lambert Inc. is the adviser on the transaction.", if the surrounding context of the DTU is mainly about the company, then we choose "Drexel Burnham Lambert Inc." as a TO; if the context is mainly about the transaction, then we choose "transaction" as a TO, and we can also select both of them if necessary. It is worth mentioning that the TOs could also be implicit ones which require human judgments.

Topic Event (TE) refers to the main phrase which most clearly expresses an event occurrence or a description of the TOs in the DTU. For the DTU u4 in Figure 1, we select "develop vaccines against the virus" as the topic event of the DTU.

With the above-mentioned definitions in mind, we argue that each DTU is composed of a set of TOs and a core TE. Based on this concept, we give the following four annotation suggestions:

- Given two adjacent DTUs in a topic chain, their TO sets should have an intersection in the topic space. For the two DTUs u3 and u4 in Figure 1, although the two corresponding TO sets, {WHO,

[1] Although we built the corpus based on RST-DT, it remains risky to directly take each elementary discourse unit (EDU) as a DTU since there are many competing hypotheses about what constitutes an EDU but without "topic" (Carlson et al., 2001). Previous work on topic-dependent argument mining usually take each independent sentence as an elementary unit, and this work is inspired by these researches.

global expert networks} and {researchers in various countries, vaccines}, have no vocabulary intersection, they are highly related in the topic space on "international response". In a sense, the relationship between TO sets is similar to that between mentions in co-reference resolution or tokens in lexical chains. The difference is that DTC parsing requires not only the correlation between TO sets but also the topic transitivity between DTUs. Therefore, for any two adjacent DTUs on a topic chain, the TE in the second DTU should evolve from the TEs in the established chain where the first DTU is located.

- Sometimes, a DTU may have topic relevance to multiple subsequent DTUs, we only opt for the **closest** and **most relevant** one for annotation. To achieve this, we follow two principles to build each arc in a topic chain: (1) For each DTU, we search its topic-related DTU from near to far; (2) We label topic links for DTUs in order and the annotated DTC structure is dynamically optimized during the human annotation process. For example, when comparing the current DTU (U-j) with previous ones, we directly replace the previously annotated arc (U-i, U-k) with (U-i, U-j) if the topic relevancy between U-i and U-j obviously surpasses that between U-i and U-k. In other words, we do not require all topic chains to be labeled, but we try to ensure the accuracy of the annotated chains as much as possible. This labeling strategy can enhance the value of this small-scale corpus to some extent.

- In news articles, many DTUs are organized in an overview-example format where similarities among the examples do exist but the evolution of topics is unseen. In this study, we do not consider simple juxtapositions like this. Taking wsj_2349 for example, "***u1:*** *The following issues were recently filed with the Securities and Exchange Commission:* ***u2:*** *American Cyanamid Co., offering of 1,250,000 common shares, via Merrill Lynch Capital Markets.* ***u3:*** *Limited Inc., offering of up to $300 million of debt securities. ...* ***u8:*** *Trans World Airlines Inc., offering of ...*". There is a certain textual structure in between the DTUs from u2 to u8 (e.g., they share the multi-nuclear relation `List` in the RST theory (Mann and Thompson, 1988)), but the topic transitivity is weak. Therefore, we do not mark any topic chains among the DTUs.

- Due to the principle of saving words and avoiding repetitions, ellipsis and co-reference occur frequently. Under this condition, we need to manually fill in the ellipsis and clarify the co-reference for better annotation.

Here we take the example in Figure 1 to illustrate the annotation process. Simply put, the annotation process is also the process of comparing the TO and TEs of the current DTU with that of the previous ones. According to the annotation instructions, we do the comparison from near to far aiming to obtain the **closest** path for two adjacent DTUs on the chain. For the DTU u1, its TO set contains two topic objects, i.e., "coronavirus" and "COVID-19", and its core topic event can be sketched as "coronavirus outbreak in Wuhan". Correspondingly, the TO set of u2 contains a pronoun object "it", referring to "coronavirus", and its core TE is manually detected as "there is still no idea how to beat it". Obviously, the two TO sets have an intersection (i.e., "coronavirus") and the TE in u2 does evolve from that in u1. Consequently, we mark a topic link between the two DTUs. For u3, both the TO set and TE do not meet our annotation requirements, so we neither link it to u1 nor u2. For u4, the TO set is relevant to that of u3 as international institutions and the two TEs are also interrelated, we therefore build a link between them. In this way, the overall vein of topic chains will be built after several rounds of comparison. Notably, from the resulting graph we find that the topic chain with u1, u2, u5, and u-k on it does provide rich and low-noise information about the evolution of COVID-19, which reflects the practical value of our annotated DTCs.

Subjective Differences in Manual Annotation. A Chinese saying about Shakespeare is that "*There are a thousand Hamlets in a thousand people's eyes*". From the above annotation process we find that one intractable problem of DTC annotation is the high subjective differences between annotators. More precisely, judging whether the temporary TE evolves from the previous one is really a very subjective problem, and it is hard to make a strict regulation for the annotators. In this case, we tackle the issue from two aspects: (i) using a well pre-trained topic model to assist manual annotation in a two-step fashion and (ii) calculating the confidence scores of the annotations for data filtrating.

Two-Step Annotation: The two-step method consists of two phases: first automatically building

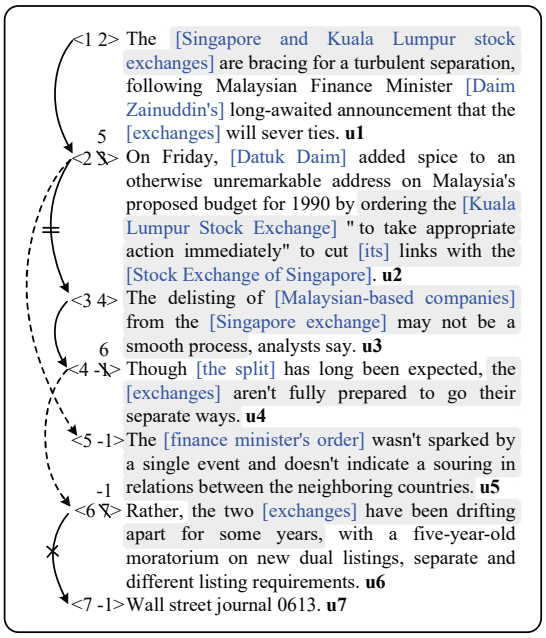

Figure 2: The two-step corpus annotation process. The TOs and TEs are marked out for reference.

topic links between topic-related DTUs[2] and then manually refining the automatic annotations for DTC structures. As depicted in Figure 2, each DTU is preceded by an index pair (i, j) according to which u-i and u-j are connected through a topic link. And u-i is an ending unit when j equals -1. The solid arcs in the example refer to the topic links generated in the first stage. On this basis, we bring in an auxiliary marker to refine the chain structures where "$\times$" means that the initial topic arcs (either machine-labeled or manually labeled links) are unreasonable and should be deleted directly, and "$=$" means that the original arcs should be replaced with more proper topic links predicted by the human annotators, e.g., the dashed arcs in the example. In this way, we can dynamically optimize the DTC structures during the human annotation process thus determining the **most relevant** DTUs for annotation. Our statistics show that around 37.4% of the automatic annotations are retained in the corpus and 62.6% of them are invalid and re-annotated by our annotators. According to this, although there is a great dissimilarity between automatic and manually annotated structures, the topic links of the pre-trained model do provide a good reference for better annotation consistency.

[2]Recently, Reimers et al. (2019) use superior contextualized language models for argument linking, which has proven to have great capabilities in aggregating arguments for unseen topics (https://github.com/UKPLab). To improve the reliability of the initial chains, we only keep the topic links with topic similarity higher than 0.9 in the first stage.

length: #	1: 715	2: 442	3: 266
	4: 159	5: 92	6+: 83

Table 1: Distribution of chain lengths.

Annotation Confidence: As stated before, considering the problem of subjective difference, it's really challenging to build a topic link between two DTUs because we're not sure if they're the most relevant. Although it is hard to strictly regulate the annotators' subjectivities, it is feasible to calculate the reliability of each annotation item. Therefore, we aim to ensure the quality of the corpus by filtering out the annotations with low confidence scores. Specifically, given the annotation results of the pre-trained topic model, (τ, ι), and that of three annotators, $(\tau, \nu), (\tau, \iota)$, and (τ, ν), on the DTU τ, we set the confidence of the pre-trained topic model to 0.5 and that of human annotators to 1, then the confidence score of each annotation on τ can be calculated as: $(\tau, \iota) \rightarrow (0.5 + 1)/3.5$, $(\tau, \nu) \rightarrow 2/3.5$. Based on the results, the annotation (τ, ν) with the highest confidence score of 0.57 is determined as the result. Following this way, we can greatly alleviate the "subjectivity" problem by retaining annotations with high confidence. According to our statistics, the averaged confidence score of each DTU annotation is around 0.73.

Data Details. The annotated corpus contains 385 news articles (7962 DTUs) from RST-DT (Carlson and Marcu, 2001). We annotate 4122 topic links corresponding to 1757 topic chains in the corpus, and the chain length distribution is presented in Table 1. Obviously, the distribution of chain langths is uneven and most chains have less than 5 topic arcs. For supervised learning, we have divided the dataset into three parts (the test corpus is consist with that of RST-DT), as shown in Table 2. Based on the test corpus, we calculate the annotation consistency with an averaged Cohen's kappa value of 0.72. Concretely, we compare three groups of manual annotations on DTUs with each other for kappa value calculation and report the average score. The data and codes are published at https://github.com/NLP-Discourse-SoochowU/DTCP.

3 Baseline

Recent years have witnessed the great effects of pre-trained language models (Devlin et al., 2019;

Corpus	Doc.	Sent.	Link	Chain
Train	313	6352	3260	1410
Dev.	34	740	403	164
Test	38	870	459	183

Table 2: Statistic results for the datasets.

Method	Link	Chain
Bert-base	89.5	78.9
Bert-large	91.7	82.1
Human-level	94.2	89.1

Table 3: Baseline performance (F1).

Yang et al., 2019; Cui et al., 2020) on natural language understanding. Following previous work, we introduce a Bert-based (Devlin et al., 2019) method in our baseline system.

Given a discourse with k-1 DTUs, we use the pre-trained Bert[3] model to encode the entire discourse where each DTU is surrounded by the `[CLS]` and `[SEP]` tokens. And we take the Bert output corresponding to `[CLS]` as our DTU representation. Following previous work, we also fine-tuned the pre-trained language model parameters during the training process. For the convenience of calculation, a zero-initialized vector u_z is added at the end of the DTU sequence for the tail DTUs of the topic chains or the isolated DTUs to point to, obtaining $\mathrm{U} = (u_1, \ldots, u_{k-1}, u_z)$. For dependency parsing, we simply build a bi-linear function between U and its duplicate to achieve it, as following:

$$\mathrm{U}_\alpha = \mathbf{W}_\alpha \mathrm{U} + \mathrm{b}_\alpha$$
$$\mathrm{U}_\beta = \mathbf{W}_\beta \mathrm{U} + \mathrm{b}_\beta$$
$$s = \mathrm{U}_\alpha^{\mathrm{T}} \mathbf{W} \mathrm{U}_\beta$$

where U_α and U_β are $(D \times k)$ matrices representing U and its duplicate, $\mathbf{W} \in \mathbb{R}^{D \times D}$ denotes the parameters of the bilinear term, and $s \in \mathbb{R}^{k \times k}$ refers to the scores for each DTU upon its candidate successor DTUs. The detailed system configuration is presented in the Appendix.

We measure the micro-averaged F1 scores of both topic links and chains for performance, and we do not take those isolated DTUs into consideration to avoid the overestimation of performance. For human performance, we asked 5 other researchers majoring in human language analysis to manually annotate the test set and took the averaged F1 scores as human performance. Experimental results in Table 3 show that fine-tuning the contextualized Bert model can achieve a great performance close to human level. By observing the model outputs (sampled in Appendix), we find that the automatically parsed chain structures are highly consistent with the manual annotations, which indicates the high reliability of our corpus. Notably, the obtained system has good generalization and robustness, and can be easily migrated to other NLP tasks for DTC structure incorporation.

[3]The pre-trained models are borrowed from https://huggingface.co/transformers.

4 Conclusion

In this research, we explored how fine-grain topics emerge, evolve, and disappear within an article. To address the lack of data, we built an English DTC corpus through a two-step annotation method, and filtered out the annotations with low confidence scores to ensure the high reliability of the corpus. During annotation, we found that each annotated topic chain does provide relatively low-noise information about a certain aspect of the article and the complete DTC structure can well describe the overall vein of topics in an article. With this in mind, we introduced a simple and robust baseline system, and the parsing model we trained can be straightforwardly harnessed in downstream topic-sensitive NLP tasks to boost performance.

It is worth mentioning that we annotated the WSJ articles in the RST-DT corpus also aim to allow the discourse researchers to explore the potential correlation between RST- and DTC-style discourse analysis in future work.

Acknowledgements

The authors would like to thank Yuqing Xing, Jialong Xie, and the other annotators for their valuable discussion and advice on this research. This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600, Projects 61876118 and 61976146 under the National Natural Science Foundation of China and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

James Allan. 2002. Introduction to topic detection and tracking. *Topic Detection and Tracking. The Information Retrieval Series*, 12:1–16.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan.

2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Hongyu Gong, Tarek Sakakini, Suma Bhat, and JinJun Xiong. 2018. Document similarity for texts of varying lengths via hidden topics. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2341–2351, Melbourne, Australia. Association for Computational Linguistics.

MAK Halliday and CMIM Matthiessen. 2004. *An Introduction to Functional Grammar*. Hodder Education.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR*.

Dongwoo Kim and Alice Oh. 2011. Topic chains for understanding a news corpus. *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science*, 6609:163–176.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. Improving neural RST parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612, Online. Association for Computational Linguistics.

Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5.

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.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. 2019. Generating summaries with topic templates and structured convolutional decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5107–5116, Florence, Italy. Association for Computational Linguistics.

Livia Polanyi and Remko Scha. 1984. A syntactic approach to discourse semantics. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 413–419, Stanford, California, USA. Association for Computational Linguistics.

Matthew Purver, Konrad P. Körding, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Sydney, Australia. Association for Computational Linguistics.

Zahra Rahimi, Diane Litman, Elaine Wang, and Richard Correnti. 2015. Incorporating coherence of topics as a criterion in automatic response-to-text assessment of the organization of writing. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 20–30, Denver, Colorado. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Magnus Sahlgren. 2020. Rethinking topic modelling: From document-space to term-space. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2250–2259, Online. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend?

blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Weishi Wang, Steven C.H. Hoi, and Shafiq Joty. 2020. Response selection for multi-party conversations with dynamic topic tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6581–6591, Online. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with topic-aware news representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1154–1159, Florence, Italy. Association for Computational Linguistics.

Xuefeng Xi and Guodong Zhou. 2017. Building a chinese discourse topic corpus with a micro-topic scheme based on theme-rheme theory. *Big Data Analytics*, 2:1–14.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Miao Yishu, Grefenstette Edward, and Blunsom Phil. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*.

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021. Adversarial learning for discourse rhetorical structure parsing. 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 3946–3957, Online. Association for Computational Linguistics.

Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6386–6395, Online. Association for Computational Linguistics.

Appendices

A. Model Configuration

We used the 768D Bert-base and 1024D Bert-large model for DTU representation. In order to prevent memory overflow, we segment each article according to the maximum length of 64, and encode the segmented text fragments in turn. We manually set the dropout rate, learning rate, L2 regularization value by 0.2, 1e-5, and 1e-5, respectively, according to their contributions to F1-score, and the number of hyper-parameter search trials was around 15. We trained the models iteratively on the training corpus for 20 rounds with the batch size set to 1 (document), and we got the best model around the 18-th round. We implemented the codes based on the PyTorch framework, and all the experiments were conducted on the NVIDIA Tesla P40 GPUs with the random seed set to 2. The number of parameters in each model and the runtime time of each system are shown in the table below.

System	Parameter scale	Runtime
Bert-base	111,553,025	270s
Bert-large	337,671,937	541s

Table 4: The parameter scale and runtime (seconds per round) of our systems.

B. Instances of DTC Parsing

Referring to our system outputs, we find that the automatically parsed DTC structures are highly consistent with human annotations. Here, we present some automatic DTC structures constructed by the Bert-large-based system for reference.

B.1. **[u1]** Moody's Investors Service said it reduced its rating on $165 million of subordinated debt of this Beverly Hills, Calif., thrift, citing turmoil in the market for low-grade, high-yield securities. **[u2]** The agency said it reduced its rating on the thrift's subordinated debt to B-2 from Ba-2 and will keep the debt under review for possible further downgrade. **[u3]** Columbia Savings is a major holder of so-called junk bonds. **[u4]** New federal legislation requires that all thrifts divest themselves of such speculative securities over a period of years. **[u5]** Columbia Savings officials weren't available for comment on the downgrade. **[u6]** FRANKLIN SAVINGS ASSOCIATION (Ottawa, Kan.) – Moody's Investors Service

Inc. said it downgraded its rating to B-2 from Ba-3 on less than $20 million of this thrift's senior subordinated notes. **[u7]** The rating concern said Franklin's "troubled diversification record in the securities business" was one reason for the downgrade, citing the troubles at its L.F. Rothschild subsidiary and the possible sale of other subsidiaries. "They perhaps had concern that we were getting out of all these," said Franklin President Duane H. Hall. "I think it was a little premature on their part." ***wsj_2375***

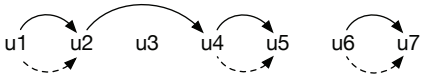

Figure 3: Human annotated (solid arcs) and automatically generated (dashed arcs) DTC structures for B.1.

B.2. **[u1]** GAF, Part III is scheduled to begin today. **[u2]** After two mistrials, the stakes in the stock manipulation trial of GAF Corp. and its vice chairman, James T. Sherwin, have changed considerably. **[u3]** The first two GAF trials were watched closely on Wall Street because they were considered to be important tests of the government's ability to convince a jury of allegations stemming from its insider-trading investigations. **[u4]** In an eight-count indictment, the government charged GAF, a Wayne, N.J., chemical maker, and Mr. Sherwin with illegally attempting to manipulate the common stock of Union Carbide Corp. in advance of GAF's planned sale of a large block of the stock in 1986. **[u5]** The government's credibility in the GAF case depended heavily on its star witness, Boyd L. Jefferies, the former Los Angeles brokerage chief who was implicated by former arbitrager Ivan Boesky, and then pointed the finger at Mr. Sherwin, takeover speculator Salim B. Lewis and corporate raider Paul Bilzerian. **[u6]** The GAF trials were viewed as previews of the government's strength in its cases against Mr. Lewis and Mr. Bilzerian. **[u7]** Mr. Jefferies's performance as a witness was expected to affect his sentencing. **[u8]** But GAF's bellwether role was short-lived. **[u9]** The first GAF trial ended in a mistrial after four weeks when U.S. District Judge Mary Johnson Lowe found that a prosecutor improperly, but unintentionally, withheld a document. **[u10]** After 93 hours of deliberation, the jurors in the second trial said they were hopelessly deadlocked, and another mistrial was declared on March 22. **[u11]** Meanwhile, a federal jury found Mr. Bilzerian guilty on securities fraud and other charges in June. **[u12]** A month later, Mr. Jefferies was spared a jail term by a federal judge who praised him for helping the government. **[u13]** In August, Mr. Lewis pleaded guilty to three felony counts. ***wsj_1331***

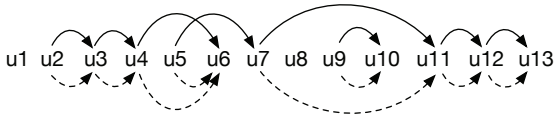

Figure 4: DTC structures for B.2.

B.3. **[u1]** MedChem Products Inc. said a U.S. District Court in Boston ruled that a challenge by MedChem to the validity of a U.S. patent held by Pharmacia Inc. was "without merit." **[u2]** Pharmacia, based in Upsala, Sweden, had charged in a lawsuit against MedChem that MedChem's AMVISC product line infringes on the Pharmacia patent. **[u3]** The patent is related to hyaluronic acid, a rooster-comb extract used in eye surgery. **[u4]** In its lawsuit, Pharmacia is seeking unspecified damages and a preliminary injunction to block MedChem from selling the AMVISC products. **[u5]** A MedChem spokesman said the products contribute about a third of MedChem's sales and 10% to 20% of its earnings. **[u6]** In the year ended Aug. 31, 1988, MedChem earned $2.9 million, or 72 cents a share, on sales of $17.4 million. **[u7]** MedChem said the court's ruling was issued as part of a "first-phase trial" in the patent-infringement proceedings and concerns only one of its defenses in the case. **[u8]** It said it is considering "all of its options in light of the decision, including a possible appeal." The medical-products company added that it plans to "assert its other defenses" against Pharmacia's lawsuit, including the claim that it hasn't infringed on Pharmacia's patent. **[u9]** MedChem said that the court scheduled a conference for next Monday – to set a date for proceedings on Pharmacia's motion for a preliminary injunction. ***wsj_2336***

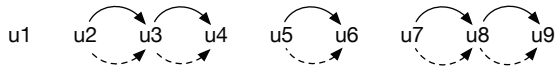

Figure 5: DTC structures for B.3.

B.4. **[u1]** ALBERTA ENERGY Co., Calgary, said it filed a preliminary prospectus for an offering of common shares. **[u2]** The natural resources

development concern said proceeds will be used to repay long-term debt, which stood at 598 million Canadian dollars (US$510.6 million) at the end of 1988. **[u3]** The company plans to raise between C$75 million and C$100 million from the offering, according to a spokeswoman at Richardson Greenshields of Canada Ltd., lead underwriter. **[u4]** The shares will be priced in early November, she said. ***wsj_1183***

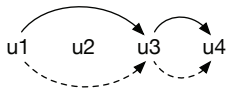

Figure 6: DTC structures for B.4.

B.5. **[u1]** Three new issues begin trading on the New York Stock Exchange today, and one began trading on the Nasdaq/National Market System last week. **[u2]** On the Big Board, Crawford & Co., Atlanta, (CFD) begins trading today. **[u3]** Crawford evaluates health care plans, manages medical and disability aspects of worker's compensation injuries and is involved in claims adjustments for insurance companies. **[u4]** Also beginning trading today on the Big Board are El Paso Refinery Limited Partnership, El Paso, Texas, (ELP) and Franklin Multi-Income Trust, San Mateo, Calif., (FMI). **[u5]** El Paso owns and operates a petroleum refinery. **[u6]** Franklin is a closed-end management investment company. **[u7]** On the Nasdaq over-the-counter system, Allied Capital Corp., Washington, D.C., (ALII) began trading last Thursday. **[u8]** Allied Capital is a closed-end management investment company that will operate as a business development concern. ***wsj_0607***

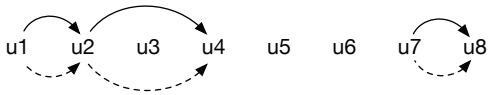

Figure 7: DTC structures for B.5.