Is the Understanding of Explicit Discourse Relations Required in Machine Reading Comprehension?

Yulong Wu, Viktor Schlegel and Riza Batista-Navarro
Department of Computer Science, University of Manchester
Manchester, United Kingdom
yulong-wu@outlook.com
{viktor.schlegel, riza.batista}@manchester.ac.uk

Abstract

An in-depth analysis of the level of language understanding required by existing Machine Reading Comprehension (MRC) benchmarks can provide insight into the reading capabilities of machines. In this paper, we propose an ablation-based methodology to assess the extent to which MRC datasets evaluate the understanding of explicit discourse relations. We define seven MRC skills which require the understanding of different discourse relations. We then introduce ablation methods that verify whether these skills are required to succeed on a dataset. By observing the drop in performance of neural MRC models evaluated on the original and the modified dataset, we can measure to what degree the dataset requires these skills, in order to be understood correctly. Experiments on three large-scale datasets with the BERT-base and ALBERT-xxlarge model show that the relative changes for all skills are small (less than 6%). These results imply that most of the answered questions in the examined datasets do not require understanding the discourse structure of the text. To specifically probe for natural language understanding, there is a need to design more challenging benchmarks that can correctly evaluate the intended skills[1].

1 Introduction

Machine Reading Comprehension (MRC) is concerned with the automatic extraction and generation of answers over unstructured textual data. Due to its complexity, the task is seen as suitable for evaluating Natural Language Understanding (NLU) (Chen, 2018). While neural MRC systems achieve impressive performance (Devlin et al., 2019; Lan et al., 2020), it has been revealed by some research efforts that existing MRC benchmarks might be insufficient to establish model performance, i.e., that the models are not being assessed for their capabilities to read and comprehend (Jia and Liang, 2017; Mudrakarta et al., 2018; Min et al., 2018; Sugawara et al., 2018; Feng et al., 2018; Jiang and Bansal, 2019; Min et al., 2019; Chen and Durrett, 2019; Schlegel et al., 2020; Sugawara et al., 2020). These analyses provide insights into the weaknesses of modern MRC gold standards. Nonetheless, to stimulate the development of robust MRC systems with generalisable NLU capabilities, it is necessary to investigate the strengths and weaknesses of MRC datasets on a deeper level.

In the task of MRC, it is assumed that questions test a cognitive process which involves various skills, such as retrieving stored information and performing inferences (Sutcliffe et al., 2013). Therefore, considering metrics that reflect skills required to answer questions is useful for analysing the capabilities of MRC datasets to benchmark NLU (Sugawara et al., 2020). This leads to the following intuition: if a question is solvable even after removing features (e.g., specific words) associated with an MRC skill, the question does not require the skill. Sugawara et al. (2020) examined 10 datasets with regard to multiple requisite skills for answering questions. One of the identified 12 skills is the *understanding of adjacent discourse relations*, which relies on information given by the sentence order in a passage. By randomly shuffling the order of the sentences in the context and comparing model performance on the original and the modified dataset, they concluded that most existing MRC datasets might be inadequate for benchmarking adjacent discourse relations understanding.

Discourse relations describe how two segments of discourse are logically connected to one another. Understanding them is key to answering reading comprehension questions correctly. Though the findings in Sugawara et al. (2020) are useful to understand MRC datasets with respect to discourse

[1] Our code is available at https://github.com/Yulong-W/mrcdr.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3565–3579
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

relations understanding, we argue that it is not enough to only consider inter-sentential relations as discourse relations also widely exist within sentences[2]. Furthermore, there also exist various types of relations and senses. Hence, to comprehensively assess the capacity of MRC datasets to benchmark discourse relations understanding, we assert that further research is needed.

In this paper, our aim is to provide a fine-grained analysis of the level of discourse relations understanding that is needed to answer questions in existing MRC datasets. Specifically, we focus on explicit discourse relations, which are expressed using explicit connectives. This allows us to perform analysis that goes beyond shuffling sentence order. In our work, we identify seven MRC skills that represent different aspects of understanding explicit discourse relations. With these, we examine three datasets using two strong MRC models. Our results show that these datasets might be insufficient for evaluating the understanding of explicit discourse relations. This work can potentially encourage the development of more challenging benchmarks that evaluate MRC models with respect to NLU capabilities that require discourse relations understanding.

2 Requisite Skills

As mentioned above, we identified a set of seven reasoning-related skills that require the understanding of explicit discourse relations, as shown in Table 1.

Skill s_1 is inspired by Sugawara et al. (2020), which aims to evaluate whether the understanding of adjacent explicit discourse relations is required in answering questions. Different from their proposed method (i.e., randomly shuffling the order of the sentences in a passage), we only shuffle those containing explicit connectives.

The selection of skills s_2 to s_7 is informed by the annotation scheme of the PDTB 3.0 corpus, which is annotated with information on discourse relations (Webber et al., 2019). The scheme defines 36 different senses of discourse relations. In the corpus, more than 24, 000 explicit connectives were annotated and categorised according to these senses. Based on this, we obtained a distribution of explicit connectives over the 36 senses (see Appendix A). Afterwards, we selected a subset of them (6 senses) based on the number of unique explicit connectives, total number of explicit connectives for which each sense was recorded, and the exclusiveness of these explicit connectives. The identification process is detailed in Appendix B. In the following, we provide an overview of skills s_2 to s_7.

Skills s_2 and s_3 are for the understanding of asynchronous temporal relations. Specifically, s_2 focuses on precedence while s_3 tests succession. Skill s_4 evaluates the understanding of causal relations, which are explicitly marked in the passage by connectives such as *because* and *due to*. Meanwhile, our motivation for selecting skill s_5 is to reveal whether explicit conditional reasoning is required to answer questions. Different from s_4, s_6 is for the understanding of negative causality, in which a causal relation expected on the basis of the first argument is negated by the situation described in the other. Finally, s_7 assesses expansions which provide further detail to an argument.

3 Methodology

For each of the seven identified skills, we defined an ablation method, as shown in Table 1. The

Skill	NLU Tested	Ablation Method
shuffling method		
s_1	Explicit discourse relations between adjacent sentences	Shuffle the order of the sentences that contain explicit connectives in the context
masking methods (Drop all occurrences of corresponding explicit connectives)		
s_2	Temporal reasoning (precedence)	Drop e.g. *afterward, later, . . .*
s_3	Temporal reasoning (succession)	Drop e.g. *earlier, since before, . . .*
s_4	Explicit causality reasoning	Drop e.g. *because of, due to . . .*
s_5	Explicit conditional reasoning	Drop e.g. *only if, depending on, . . .*
s_6	Negative causality reasoning	Drop e.g. *albeit, but then again, . . .*
s_7	Expansion of explicit discourse relations	Drop e.g. *additionally, moreover, . . .*

Table 1: Requisite skills and ablation methods.

[2] In the Penn Discourse Treebank (PDTB) 3.0 corpus (Webber et al., 2019), 24,369 and 29,818 tokens were annotated as connectives for intra-sentential and inter-sentential discourse relations, respectively.

design of these methods is based on the fact that explicit discourse relations are expressed using explicit discourse connectives (Webber et al., 2019). The scope of the proposed methodology hence captures only relations represented by explicit connectives, rather than all discourse relations-related features of the datasets. We assume that through shuffling the order of the sentences with connectives in the context, as well as through dropping these connectives, the corresponding relations will be broken. After applying the ablation method on the development set of an MRC dataset, if the performance of the model did not change significantly, we can say that most of the questions in the dataset are solvable even without the given skill; hence, the dataset does not sufficiently evaluate models with respect to the said skill. On the contrary, if the performance gap between the original and the modified dataset is large, we might infer that a substantial proportion of the questions require that skill. Nonetheless, should the model perform badly on the ablated dataset, we cannot take this as evidence that the model in fact acquired the investigated reasoning capabilities as the bad performance can stem from many different factors (e.g., distribution shift induced by dropping numerous words).

4 Experiments

In this section, we describe our experimental settings, present the results of our experiments and provide insights drawn from experimentation under an extreme setting whereby all explicit connectives were dropped.

4.1 Experimental Settings

Datasets. We examined three datasets with two answering styles. For span prediction datasets in which the goal is to identify a span in the passage as the answer, we used SQuAD 1.1 (Rajpurkar et al., 2016) and SQuAD 2.0 (Rajpurkar et al., 2018). For multiple choice datasets in which the correct answer is chosen from a candidate set of answers, we used SWAG (Zellers et al., 2018). We applied the ablation methods on the development set of each dataset. Sentence segmentation and tokenisation are performed as part of the pre-processing step.

Models. In the main experiment, we used the BERT-base (uncased) model (Devlin et al., 2019). Our goal is to analyse whether there exists at least one model architecture that can solve the MRC task without the understanding of explicit discourse relations; hence, it is enough to use a single model (Sugawara et al., 2020). Then, from the perspective of testing the effectiveness of the proposed MRC skills, we employed a stronger model, ALBERT-xxlarge (Lan et al., 2020). We fine-tuned the pre-trained BERT-base (uncased) and ALBERT-xxlarge model on the training set of each dataset and evaluated them on the original and the modified development sets by making use of the HuggingFace's *Transformers* library (Wolf et al., 2020). The hyperparameters of the models are reported in Appendix C.

Ablation methods. Method m_1: For the choice of explicit connectives, we used the 173 explicit connectives from the PDTB 3.0 corpus (Webber et al., 2019) (see Appendix D). We averaged the scores over five runs and report the mean and variance values in Appendix E. Methods m_2 to m_7: we list explicit connectives dropped for each sense in Appendix F. When a token is dropped, it is replaced with an [UNK] token to preserve the correct answer span. More in-depth results are reported in Appendix G.

4.2 Results and Discussion

In this section, we report the results for the skills in Table 2. In this table, for each of the ablation method used for skills s_2 to s_7, there are two versions of experimental results, shown in the white and shaded areas, respectively. Results written in the white areas were obtained by applying the ablation methods detailed in Section 4.1, i.e., by masking explicit connectives selected using the threshold-based method (see Appendix B) for which each sense was annotated. However, it can be seen in the table that except for s_7, the relative differences for s_2 to s_6 were extremely small (less than 1%) across all datasets. To further investigate whether these skills are truly not required to answer questions in the three datasets, we performed additional experiments as follows.

For the senses that represent a skill under evaluation, we dropped every explicit connective associated with those senses according to the PTDB 3.0 annotations (Webber et al., 2019). By applying these modified ablation methods, we obtained additional experimental results, shown in the shaded areas of Table 2. In the following, we discuss the observations for all the defined skills.

s_1: adjacent explicit discourse relations understanding. On all datasets, the relative changes

Skill	SQuAD 1.1 *(F1)*	SQuAD 2.0 *(F1)*	SWAG *(Acc.)*
Orig.	88.6	76.1	79.0
s_1	$86.2_{-2.7}$	$74.8_{-1.7}$	-
s_2	$88.3_{-0.3}$	$76.0_{-0.1}$	$79.0_{-0.0}$
	$88.0_{-0.7}$	$75.8_{-0.4}$	$79.0_{-0.0}$
s_3	$88.5_{-0.1}$	$76.1_{-0.0}$	$79.0_{-0.0}$
	$87.4_{-1.4}$	$75.4_{-0.9}$	$79.0_{-0.0}$
s_4	$87.9_{-0.8}$	$75.6_{-0.7}$	$79.0_{-0.0}$
	$78.6_{-11.3}$	$71.0_{-6.7}$	$78.6_{-0.5}$
s_5	$88.4_{-0.2}$	$76.1_{-0.0}$	$79.0_{-0.0}$
	$83.4_{-5.9}$	$73.8_{-3.0}$	$78.6_{-0.5}$
s_6	$88.4_{-0.2}$	$76.1_{-0.0}$	$79.0_{-0.0}$
	$87.5_{-1.2}$	$75.6_{-0.7}$	$78.6_{-0.5}$
s_7	$84.9_{-4.2}$	$74.1_{-2.6}$	$79.1_{+0.1}$
	$83.7_{-5.5}$	$73.7_{-3.2}$	$78.6_{-0.5}$

Table 2: The performance (%) of the BERT-base model with the ablation tests on the development set. Values in smaller font are changes (%) relative to the original performance of the model. For mask-related methods (m_2 to m_7), the results shown in the white areas are obtained from the initial test while the results shown in the shaded areas represent the further test, i.e., dropping all explicit connectives for which each identified sense was annotated. Acc.: accuracy as a percentage.

for s_1 were small. We do not apply m_1 to SWAG because its contexts are only one sentence long. On SQuAD 1.1 and SQuAD 2.0, the difference was hardly noticeable (less than 3% and 2%, respectively). These results indicate that most of the questions already solved in these datasets do not necessarily require the understanding of adjacent explicit discourse relations and are solvable even if the sentences appear unnaturally. This confirms the findings of Min et al. (2018), which reported that 92% of questions in SQuAD 1.1 are solvable by only looking at the sentence containing the answer.

s_2 and s_3: performing asynchronous temporal reasoning. We found that for the three examined datasets, the relative changes for s_2 and s_3 were extremely small (the biggest drop was even less than 1.5%), regardless of whether only a part or all associated explicit connectives were dropped. This indicates that these datasets might not adequately benchmark the understanding of asynchronous temporal relations.

s_4: explicit causality reasoning. In the initial experiment, the relative changes for s_4 on the three datasets were extremely small (less than 1%). However, surprisingly, after masking all explicit connectives cueing causality, except for SWAG which still featured a low drop (0.5%), the relative drops on SQuAD 1.1 and SQuAD 2.0 increased noticeably (from less than 1% to 11.3% and 6.6%, respectively). Particularly, for SQuAD 1.1, the decrease was the largest in all our experiments. Nevertheless, we cannot simply conclude that s_4 is needed to answer questions in the two datasets as the additionally dropped explicit connectives were also recorded as many other senses in the PDTB 3.0 corpus and not associated with this sense for the majority of occurrences. As we do not know exactly whether the decrease in model performance is due to this sense or any other senses, further analyses are necessary. Based on the PDTB 3.0 Annotation Manual (Webber et al., 2019), we calculated the percentage of each additionally dropped connective for this sense among the multiple senses for which it was annotated, and removed those which are rarely used for this sense from the candidate set of the dropped explicit connectives. The experiments demonstrated that the model achieved 85.1 and 74.3 (4.0% and 2.3% relative drop) F1 score on SQuAD 1.1 and SQuAD 2.0, respectively. This implies that the examined datasets might not correctly benchmark the understanding of causal relations and the reason why the relative drops were large after dropping all explicit connectives is that the other senses might be important.

s_5: explicit conditional reasoning. In the initial test, on all datasets, the relative changes were extremely small (less than 0.3%). Nonetheless, after dropping all explicit connectives describing conditional relations, except for SWAG which still showed a low drop (0.5%), the performance on SQuAD 1.1 and SQuAD 2.0 decreased by more than 3%. However, similarly to s_4, we cannot conclude whether such a decrease is due to sense representing s_5 or other senses that the explicit connectives are also associated with. As a result, we removed explicit connectives which are rarely used for this sense from the candidate set of explicit connectives and conducted further analyses. The experiments demonstrated that the model achieved 88.4 and 76.1 F1 on SQuAD 1.1 and SQuAD 2.0, respectively, both less than 0.5% relative difference. This indicates that s_5 might not necessarily

Dataset	**SQuAD 1.1** *(F1)*	**SQuAD 2.0** *(F1)*	**SWAG** *(Acc.)*
Original	94.4	87.6	89.0
Ablated	$92.4_{-2.1}$	$84.8_{-3.2}$	$86.3_{-3.0}$

Table 3: Performance of ALBERT-xxlarge on the original development set and on a version with all explicit connectives dropped. Acc.: accuracy as a percentage.

be required to answer questions in these datasets either.

s_6**: reasoning about negative causality.** On all datasets, the relative drops for s_6 were extremely small (less than 1.3%), whether with part of or all explicit connectives dropped. This demonstrates that most of the solved questions in the three MRC datasets do not necessarily require negative causal reasoning.

s_7**: recognising the expansion of explicit discourse relations.** In the initial experiment, the relative changes for s_7 on SWAG and SQuAD 2.0 were small, while that on SQuAD 1.1 was slightly larger (more than 4%). After dropping all explicit connectives for which sense *Expansion.Conjunction* was annotated, the performance of the model further decreased moderately – up to 5.5% for SQuAD 1.1, implying that compared to the other two datasets, SQuAD 1.1 might have more potential for benchmarking the understanding of the expansion of explicit discourse relations.

4.3 Further Analyses

Surprised by the moderate performance changes, we investigated the extent to which understanding of any explicit discourse relations is required by the datasets. Therefore, we dropped all explicit connectives and employed a stronger model, ALBERT-xxlarge (Lan et al., 2020) to generalise our assumption from the six specific senses to all senses. To mitigate the effect of distribution shift between training and evaluation data introduced by removing large parts of the context, we applied the ablation methods on the training set as well. The results are shown in Table 3. The performance drops no more than 3.2% for all three datasets, contributing further evidence towards the hypothesis that understanding the discourse structure of the text is hardly required to perform well on the investigated benchmarks.

5 Conclusion

In this paper, we proposed a methodology to assess the capabilities of MRC datasets to benchmark the understanding of explicit discourse relations. With seven fine-grained skills and corresponding ablation methods, we examined three large-scale datasets. The experimental results demonstrated that explicit discourse relations are not sufficiently evaluated by them, and thus there is a need to develop more challenging datasets so that their questions can correctly benchmark our defined skills. As for future work, we will develop a machine learning-based system that can recognise various senses of implicit discourse relations in the passage and further reveal whether the awareness of implicit discourse relations is required to do well on contemporary MRC benchmarks.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable suggestions. We also acknowledge the use of the Computational Shared Facility at The University of Manchester.

References

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. 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 4026–4032, Minneapolis, Minnesota. 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.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1725–1735, Melbourne, Australia. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8918–8927. AAAI Press.

Richard FE Sutcliffe, Anselmo Penas, Eduard H Hovy, Pamela Forner, Alvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2013. Overview of QA4MRE Main Task at CLEF 2013. In *CLEF (Working Notes)*.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A Senses and Their Associated Explicit Connectives

This Appendix provides a distribution of the 36 distinct senses annotated for explicit connectives (Table 4), which are calculated by referring to Appendix A of the PDTB 3.0 annotation scheme (Webber et al., 2019). For each sense, the second column lists the explicit connectives for which the sense was annotated, with counts given for each connective (in parentheses). The third column lists the total number of explicit connectives for which each sense was annotated. Discontinuous connectives are indicated with a "+" symbol between their parts.

Sense	Explicit Connectives	Total
Temporal.Synchronous	as (383), as long as (4), as soon as (9), at the same time (65), by (1), in (18), in the meantime (11), in the meanwhile (1), meantime (2), meanwhile (120), now that (2), simultaneously (6), still (1), then (4), upon (2), when (509), while (163), with (5)	1306
Temporal.Asynchronous.Precedence	afterward (6), afterwards (5), before (309), finally (13), in the end (3), later (92), later on (2), next (4), since (10), still (2), subsequently (3), then (310), thereafter (11), till (4), ultimately (15), until (143), when (4)	936
Temporal.Asynchronous.Succession	after (533), as (3), as soon as (11), before (2), by then (6), earlier (15), in the meantime (2), once (70), previously (53), since (83), since before (1), until (7), when (160)	946
Contingency.Cause.Reason	about (2), and (4), as (180), because (833), because of (12), by (10), due to (1), for (34), from (2), given (6), in (1), indeed (1), insofar as (1), not only because of (1), now that (10), on (1), since (96), ultimately (1), when (21), with (109), without (1)	1327
Contingency.Cause.Result	accordingly (5), and (5), as a result (78), consequently (10), for (1), hence (5), in the end (2), so (222), so that (10), then (7), thereby (9), therefore (26), thus (111), without (1)	492
Contingency.Cause.NegResult	—	0
Contingency.Cause+Belief.Reason+Belief	as (3), because (2), from (2), given (3), in (1), indeed (4), with (5)	20
Contingency.Cause+Belief.Result+Belief	so (1), thus (1)	2
Contingency.Cause+SpeechAct.Reason+ SpeechAct	but (1)	1
Contingency.Cause+SpeechAct.Result+ SpeechAct	and (1)	1
Contingency.Condition.Arg1-as-cond	and (22), then (1)	23
Contingency.Condition.Arg2-as-cond	as long as (13), by (2), depending on (3), depending upon (1), for (7), if (1084), if and when (2), if only (4), if+then (37), in (8), in case (6), in order (4), once (4), only if (13), so long as (4), until (17), when (116), whenever (9), where (2), with (2)	1338
Contingency.Condition+SpeechAct	because (2), if (56), if+then (1), or (2), when (12)	73

Continued on next page

Table 4 – Continued from previous page

Sense	Explicit Connectives	Total
Contingency.Negative-condition.Arg1-as-negCond	either+or (2), else (1), lest (2), or (7), otherwise (4)	16
Contingency.Negative-condition.Arg2-as-negCond	till (1), unless (98), without (9)	108
Contingency.Negative-condition+SpeechAct	—	0
Contingency.Purpose.Arg1-as-goal	—	0
Contingency.Purpose.Arg2-as-goal	and (128), for (16), if only (1), in (2), in order (51), so (44), so as (3), so that (21)	266
Comparison.Concession.Arg1-as-denier	although (206), as (7), as much as (2), by (1), despite (9), even as (2), even if (87), even though (69), even when (8), even with (2), for (1), however (5), if (6), no matter (8), regardless of (6), though (91), whatever (4), when (3), whether (7), while (203), with (2)	729
Comparison.Concession.Arg2-as-denier	albeit (1), although (105), as if (4), but (3063), but then (3), but then again (1), even so (9), even though (26), however (390), if (3), if only (1), in any case (3), in fact (4), in the end (1), indeed (1), meanwhile (2), nevertheless (32), nonetheless (25), nor (1), not only+but (1), on the one hand+on the other hand (1), on the other hand (4), only (2), or (1), regardless (2), still (115), though (128), when (1), while (2), without (19), yet (96)	4047
Comparison.Concession+SpeechAct.Arg2-as-denier+SpeechAct	and (2), but (2), if (1), or (11)	16
Comparison.Contrast	although (14), and (16), as (5), but (618), by comparison (11), by contrast (28), conversely (2), however (95), if (2), in contrast (12), in fact (7), in the end (1), like (1), meanwhile (7), neither+nor (1), nevertheless (12), nonetheless (2), not only+but also (1), on the contrary (4), on the one hand+on the other (1), on the one hand+on the other hand (1), on the other hand (32), only (1), still (75), though (16), when (1), whereas (5), while (140), with (1), yet (4)	1116
Comparison.Similarity	as (65), as though (1), as well (6), like (3), similarly (18), while (1)	94

Continued on next page

Table 4 – Continued from previous page

Sense	Explicit Connectives	Total
Expansion.Conjunction	additionally (7), along with (2), also (1736), and (6189), as much as (1), as well (12), as well as (7), besides (19), beyond (1), both+and (6), but (42), but also (1), finally (18), further (7), furthermore (12), in addition (165), in fact (36), in the end (1), indeed (54), likewise (8), meanwhile (27), moreover (103), much less (3), neither+nor (2), nor (31), not just+but (1), not just+but+also (1), not only (5), not only+also (1), not only+but (18), not only+but also (9), or (71), plus (1), separately (72), then (11), ultimately (1), while (43), with (41), yet (2)	8767
Expansion.Disjunction	alternatively (4), and then (1), as an alternative (2), either+or (36), nor (1), or (258), or otherwise (2)	304
Expansion.Equivalence	in other words (17), indeed (2), or (6), that is (2)	27
Expansion.Exception.Arg1-as-excpt	otherwise (15)	15
Expansion.Exception.Arg2-as-excpt	although (2), but (3), except (12), only (3)	20
Expansion.Instantiation.Arg1-as-instance	as if (1), in (1)	2
Expansion.Instantiation.Arg2-as-instance	as (4), for example (200), for instance (98), in fact (3), in particular (6), indeed (2), like (1), such as (2), with (7)	323
Expansion.Level-of-detail.Arg1-as-detail	as (8), in (17), in fact (1), in short (4), in sum (2), in the end (2), indeed (1)	35
Expansion.Level-of-detail.Arg2-as-detail	and (4), as though (2), by (2), for (1), in (2), in fact (34), in particular (9), in that (1), in the end (1), indeed (37), insofar as (1), only (1), specifically (10), that is (2), with (111), without (2)	220
Expansion.Manner.Arg1-as-manner	thereby (3)	3
Expansion.Manner.Arg2-as-manner	and (19), as (3), as if (1), by (174), in (13), when (2), with (6), without (62)	280
Expansion.Substitution.Arg1-as-subst	from (1), instead of (43), rather than (40)	84
Expansion.Substitution.Arg2-as-subst	alternatively (2), as much as (1), instead (112), more accurately (1), not so much as (1), rather (17), so much as (1)	135

Table 4: Senses and their associated explicit connectives annotated in the PDTB 3.0 corpus (Webber et al., 2019).

B Identification of the Senses of Explicit Discourse Relations

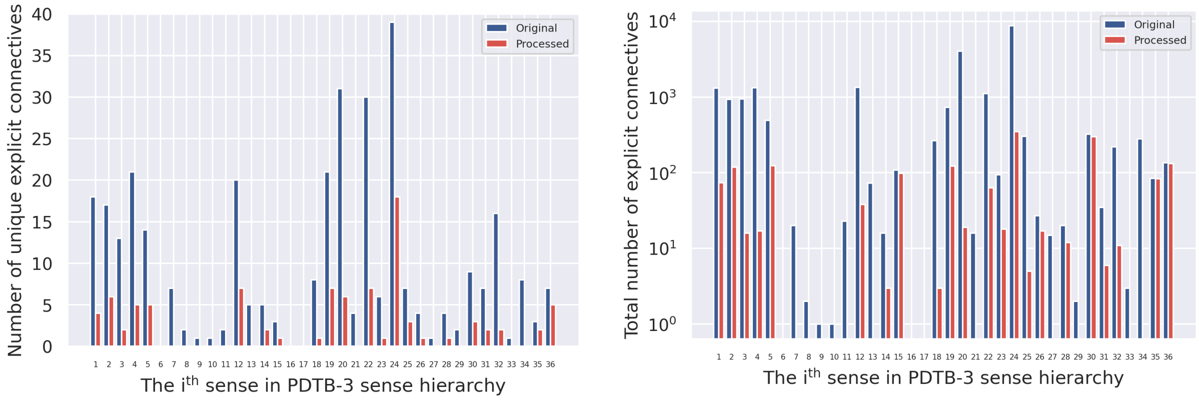
Table 4 provides the distribution of the 36 senses and their associated explicit connectives. To determine which kind of senses to focus on, we first defined two metrics: *"uniqueness"* and *"instances"* to measure the 36 senses. The first metric "uniqueness" measures the number of unique explicit connectives in each sense. For instance, as can be seen from Table 4, the sense "Contingency.Condition.Arg1-as-cond" was annotated for two unique explicit connectives: "and (22), then (1)". Therefore, its uniqueness is equal to 2. The second metric "instances" measures the total number of connectives for which each sense was annotated. Take the same example, we can see that the connective "and" was annotated 22 times as the sense "Contingency.Condition.Arg1-as-cond" and the connective "then" was annotated once as the same sense. Under this circumstance, there is a total of 23 (22+1) explicit connectives for which the sense "Contingency.Condition.Arg1-as-cond" was annotated, and thus its instances is equal to 23.

We propose that the two metrics can reflect the breadth and importance of these senses in a passage of text as Table 4 was developed from the large-scale PDTB 3.0 corpus (Webber et al., 2019), which provides a certain degree of representation. In this context, the higher "uniqueness" and "instances" a sense features, the more widely it might spreads in the context. Consequently, choosing such a sense to focus on is more likely to reveal whether the existing MRC benchmarks test the model's understanding of it.

Besides the two defined metrics, we also noticed that in the PDTB 3.0 corpus (Webber et al., 2019), many different senses were recorded for the same connective. For example, the connective "in the end" was annotated as seven types of senses. In this case, we cannot exactly examine which kind of senses the MRC datasets assessed by dropping their associated non-exclusive explicit connectives. This indicates that there is a need to consider the issue of managing explicit connectives for which multiple senses were annotated. To this end, we introduced the third metric: *"exclusiveness"*, which measures the degree of semantic overlap of explicit connectives in each sense. Ideally, to ensure that there are no overlapping explicit connectives among these senses, we can just remove all of the explicit connectives for which multiple senses were annotated and keep those that represent only one type of sense. However, after doing this, the "uniqueness" and "instances" of most senses are greatly decreased (see Figure 1a and Figure 1b). Based on this, we posit that the cost, i.e., most senses losing a considerable number of explicit connectives, is too high when attempting to retain their exclusiveness. Though the senses with only exclusive explicit connectives could meet the three metrics, they might not be enough for our data ablation purposes, as most of the explicit connectives were eliminated. Considering this, we need to find a balance between preserving the number and types of explicit connectives in each sense and maintaining its exclusiveness.

To minimise the loss in terms of "uniqueness" and "instances" of each sense while preserving "exclusiveness", we propose that if a connective C was annotated with multiple senses and it is used for sense X majority of the time, then we could include it in sense X. To identify the exact value of the "majority", we calculated the percentage of the distinct senses annotated for each non-exclusive connective and selected the sense with the highest percentage. Subsequently, we averaged these highest values and obtained the threshold, which is about 69%. Finally, we chose explicit connectives where the highest proportion of the sense for which they were annotated exceeds 69% and eliminated those below the threshold. From Figure 2a and Figure 2b, one can see that both the "uniqueness" and "instances" of the most senses with some retained non-exclusive explicit connectives increased, compared with those that only contain the exclusive connectives. This demonstrates that our method has effectively increased the number and types of explicit connectives in the most senses while maintaining their exclusiveness.

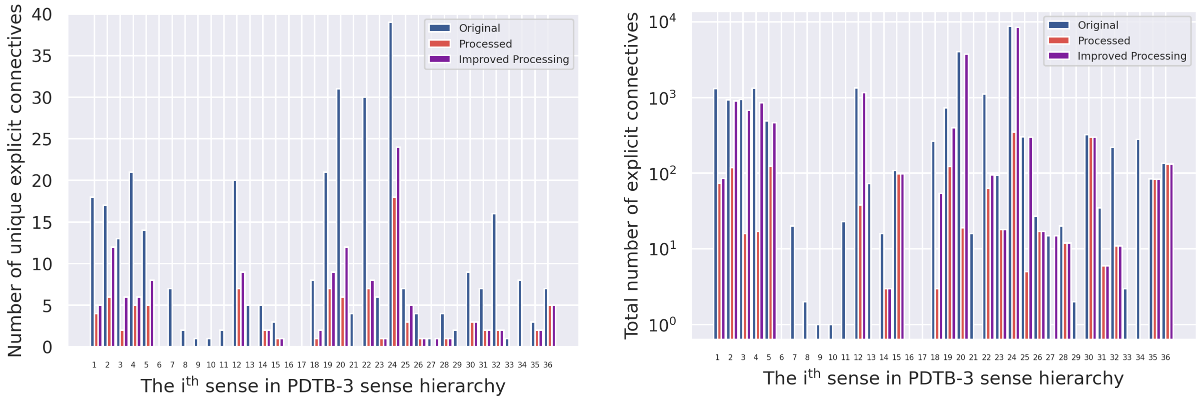
Finally, to select the candidate senses from the 36 senses, we visualised them in terms of their "uniqueness" and "instances", as shown in Figure 3a and Figure 3b, respectively. As can be seen in Figure 3a, there are a total of 12 senses with the number of unique explicit connectives above the mean value (sense 24, 2, 20, 12, 19, 5, 22, 3, 4, 1, 25, 36). Furthermore, it can be seen from Figure 3b that there are a total of 6 senses with the total number of explicit connectives larger than the average (sense 24, 20, 12, 2, 4, 3). Then, we took the intersection of these two sets of senses and obtained a sense set whereby "uniqueness" and "instances" of each sense is above the mean, and its



(a)

(b)

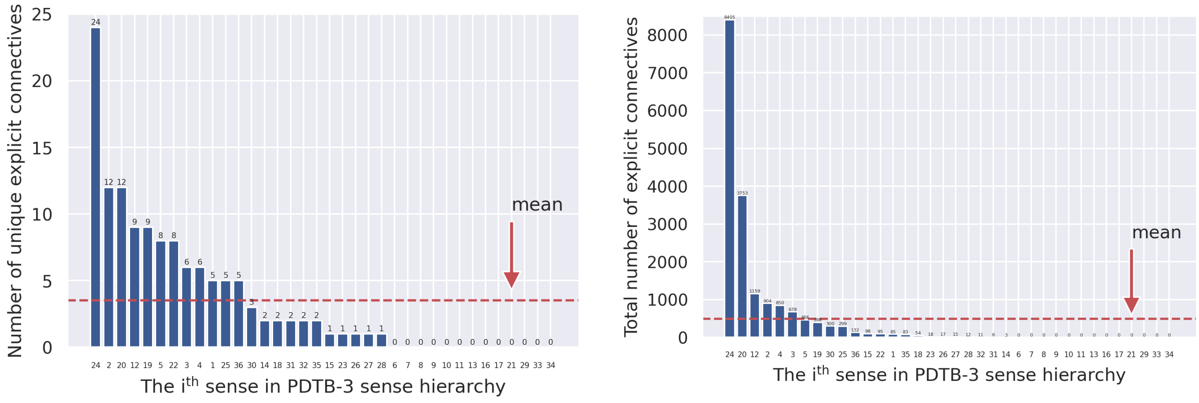
Figure 1: Visualisation of the number of unique explicit connectives and total number of explicit connectives for each sense in the PDTB 3.0 sense hierarchy (Webber et al., 2019). The blue bar represents the original senses, while the red one represents the senses after removing the non-exclusive explicit connectives.



(a)

(b)

Figure 2: Visualisation of the number of unique explicit connectives and total number of explicit connectives for each sense in the PDTB 3.0 sense hierarchy (Webber et al., 2019). The purple bar represents the senses with some retained non-exclusive connectives.



(a)

(b)

Figure 3: Number of unique explicit connectives and total number of explicit connectives for which each sense (processed) was annotated (sorted from largest to smallest).

"exclusiveness" is retained to a certain extent:[3]

- Sense 2:
Temporal.Asynchronous.Precedence
- Sense 3:
Temporal.Asynchronous.Succession
- Sense 4:
Contingency.Cause.Reason
- Sense 12:
Contingency.Condition.Arg2-as-cond
- Sense 20:
Comparison.Concession.Arg2-as-denier
- Sense 24:
Expansion.Conjunction

C Hyperparameters of the BERT-base and ALBERT-xxlarge Model

Hyperparameters used in the BERT-base and ALBERT-xxlarge model are shown in Table 5.

Dataset	d	b	lr	ep
BERT-base				
SQuAD 1.1	384	12	3e-5	2.0
SQuAD 2.0	384	12	3e-5	2.0
SWAG	80	8	5e-5	3.0
ALBERT-xxlarge				
SQuAD 1.1	384	12	3e-5	2.0
SQuAD 2.0	384	12	3e-5	4.0
SWAG	80	128	5e-5	3.0

Table 5: The hyperparameters used to fine-tune the BERT-base and ALBERT-xxlarge model on each dataset. *d* is the size of the token sequence fed into the model, *b* is the training batch size, *lr* is the learning rate, and *ep* is the number of training epochs. We used stride = 128 for documents longer than *d* tokens.

D A Set of Explicit Connectives

We list the set of explicit connectives used in this work in Figure 4.

E Performance Means and Variances in Shuffle-Based Method

We report the means and variances for the shuffling ablation method for skill s_1 in Table 6.

F The Six Identified Senses and Their Associated Explicit Connectives

Table 7 shows the six identified senses and their associated explicit connectives. For each sense, the associated explicit connectives were selected using the threshold-based method detailed in Appendix B.

G Detailed Results of SQuAD 2.0

We report the ablation results for has-answer and no-answer questions in SQuAD 2.0 in Table 8.

[3] A detailed introduction of these senses is available at https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.

Continuous explicit connectives (One token):
about, accordingly, additionally, after, afterward, afterwards, albeit, also, alternatively, although, and, and/or, as, because, before, besides, beyond, but, by, consequently, conversely, despite, earlier, else, except, finally, for, from, further, furthermore, given, hence, however, if, in, indeed, instead, later, lest, like, likewise, meantime, meanwhile, moreover, nevertheless, next, nonetheless, nor, on, once, only, or, otherwise, plus, previously, rather, regardless, separately, similarly, simultaneously, since, so, specifically, still, subsequently, then, thereafter, thereby, therefore, though, thus, till, ultimately, unless, until, upon, whatever, when, whenever, where, whereas, whether, while, with, without, yet

Continuous explicit connectives (Two tokens):
along with, and then, as if, as though, as well, because of, but also, but then, by comparison, by contrast, by then, depending on, depending upon, due to, even after, even as, even before, even if, even so, even then, even though, even when, even while, even with, for example, for instance, if only, in addition, in case, in contrast, in fact, in order, in particular, in short, in sum, in that, insofar as, instead of, later on, more accurately, much less, no matter, not only, now that, only if, or otherwise, rather than, regardless of, since before, so as, so that, such as, that is

Continuous explicit connectives (Three tokens):
as a result, as an alternative, as long as, as much as, as soon as, as well as, before and after, but then again, even before then, if and when, in any case, in other words, in the end, in the meantime, in the meanwhile, on the contrary, so long as, so much as, when and if

Continuous explicit connectives (Four tokens):
at the same time, not only because of, not so much as, on the other hand

Discontinuous explicit connectives:
both+and, either+or, if+then, neither+nor, not just+but, not just+but+also, not only+also, not only+but, not only+but also, on the one hand+on the other, on the one hand+on the other hand

Figure 4: A set of 173 explicit connectives from the annotation scheme of the PDTB 3.0 corpus (Webber et al., 2019).

Ablation Method	The ith Run	SQuAD 1.1	SQuAD 2.0		
			Has-Ans	No-Ans	Total
1. Randomly shuffle the order of the sentences with explicit connectives in the context	1	86.1	74.5	75.6	75.0
	2	86.0	74.4	75.3	74.9
	3	86.3	74.4	75.3	74.8
	4	86.2	73.9	75.7	74.8
	5	86.2	74.4	75.1	74.7
mean (variance)	—	86.2 (0.0)	74.3 (0.0)	75.4 (0.0)	74.8 (0.0)

Table 6: Ablation results with means and variances (in parentheses) for the shuffling-based method for skill s_1 over five different runs.

Sense	Explicit Connectives	Retention Rate (Uniqueness)	Retention Rate (Instances)
Temporal. Asynchronous. Precedence	afterward (6), afterwards (5), before (309), later (92), later on (2), next (4), subsequently (3), then (310), thereafter (11), till (4), ultimately (15), until (143)	70.59%	96.58%
Temporal. Asynchronous. Succession	after (533), by then (6), earlier (15), once (70), previously (53), since before (1)	46.15%	71.67%
Contingency. Cause. Reason	about (2), because (833), because of (12), due to (1), not only because of (1), on (1)	28.57%	64.05%
Contingency. Condition. Arg2-as-cond	depending on (3), depending upon (1), if (1084), if+then (37), in case (6), only if (13), so long as (4), whenever (9), where (2)	45%	86.62%
Comparison. Concession. Arg2-as-denier	albeit (1), but (3063), but then (3), but then again (1), even so (9), however (390), in any case (3), nevertheless (32), nonetheless (25), regardless (2), though (128), yet (96)	38.71%	92.74%
Expansion. Conjunction	additionally (7), along with (2), also (1736), and (6189), as well as (7), besides (19), beyond (1), both+and (6), but also (1), further (7), furthermore (12), in addition (165), likewise (8), moreover (103), much less (3), nor (31), not just+but (1), not just+but+also (1), not only (5), not only+also (1), not only+but (18), not only+but also (9), plus (1), separately (72)	61.54%	95.87%

Table 7: The six identified senses and their associated explicit connectives. Exclusive connectives are underlined. The fourth and last column provides the retention rate with respect to the "uniqueness" and "instances" of each sense, respectively.

Subset	Has-Ans	No-Ans	Total
Ablation Method	5928	5945	11873
Original dataset	78.4	73.9	76.1
1. Randomly shuffle the order of the sentences with explicit connectives	$74.3_{-5.2}$	$75.4_{+2.0}$	$74.8_{-1.7}$
2. Drop explicit connectives associated with asynchronous temporal reasoning (precedence)	$77.8_{-0.8}$	$74.1_{+0.3}$	$76.0_{-0.1}$
	$77.5_{-1.1}$	$74.2_{+0.4}$	$75.8_{-0.4}$
3. Drop explicit connectives associated with asynchronous temporal reasoning (succession)	$78.3_{-0.1}$	$73.9_{-0.0}$	$76.1_{-0.0}$
	$76.8_{-2.0}$	$74.0_{+0.1}$	$75.4_{-0.9}$
4. Drop explicit connectives associated with causality reasoning	$77.3_{-1.4}$	$73.9_{-0.0}$	$75.6_{-0.7}$
	$64.7_{-17.5}$	$77.4_{+4.7}$	$71.0_{-6.7}$
5. Drop explicit connectives associated with conditional reasoning	$78.1_{-0.4}$	$74.0_{+0.1}$	$76.1_{-0.0}$
	$71.3_{-9.1}$	$76.4_{+3.4}$	$73.8_{-3.0}$
6. Drop explicit connectives associated with negative causality reasoning	$78.1_{-0.4}$	$74.0_{+0.1}$	$76.1_{-0.0}$
	$77.0_{-1.8}$	$74.2_{+0.4}$	$75.6_{-0.7}$
7. Drop explicit connectives associated with the expansion of explicit discourse relations	$73.6_{-6.1}$	$74.5_{+0.8}$	$74.1_{-2.6}$
	$72.2_{-7.9}$	$75.1_{+1.6}$	$73.7_{-3.2}$

Table 8: Results (%) on the development set of SQuAD 2.0 for subsets with normal (Has-Ans) and no-answer (No-Ans) questions. Values in smaller font are changes (%) relative to the original performance of the model. For mask-related methods (m_2 to m_7), the results shown in the white areas are obtained from the initial test while the results shown in the shaded areas represent the further test, i.e., dropping all explicit connectives for which each identified sense was annotated.