discopy: A Neural System for Shallow Discourse Parsing

René Knaebel
Applied Computational Linguistics
Department of Linguistics
University of Potsdam
Germany
rene.knaebel@uni-potsdam.de

Abstract

This paper demonstrates `discopy`, a novel framework that makes it easy to design components for end-to-end shallow discourse parsing. For the purpose of demonstration, we implement recent neural approaches and integrate contextualized word embeddings to predict explicit and non-explicit discourse relations. Our proposed neural feature-free system performs competitively to systems presented at the latest Shared Task on Shallow Discourse Parsing. Finally, a web front end is shown that simplifies the inspection of annotated documents. The source code, documentation, and pretrained models are publicly accessible.

1 Introduction

Whenever people compose text, they (consciously or not) make sure that related sentences are cohesive and coherent within a common section. By using models to analyze discourse, we identify relations within a text that consist of phrases and a certain sense. These discourse relations and their understanding are important for other tasks in the NLP community such as abstractive summarization (Gerani et al., 2014), text simplification (Zhong et al., 2020), and argumentation mining (Hewett et al., 2019). The Penn Discourse Treebank (PDTB) (Prasad et al., 2008), for example, describes discourse as a set of individual relations. The following two (artificial) examples demonstrate the main types of relations annotated in their corpus:

1. While **it is raining outside**, *I clean the dishes.* (Temporal.Synchronous)

2. *Yesterday I walked through the rain*, **today I am sick**. (Contingency.Cause)

Relations consist of two *arguments* (Arg1 in italics and Arg2 in bold) and a *sense* (e.g. Temporal.Synchronous[1]) is assigned. The first example is called an *explicit* relation, because the relation is signaled by a phrase called *connective* (Conn is underlined). The second example refers to the group of *implicit* relations because of the absence of a connective that signals the relation. Additionally, if adjacent sentences have an *entity-based* relation, they are marked as EntRel.

The area of *shallow discourse parsing* (SDP) aims to build models finding aforementioned discourse structures. Started through the development of PDTB, shallow discourse parsing gained more awareness by two shared Tasks (Xue et al., 2015, 2016). Successful systems at the last competition were those of Wang and Lan (2016); Qin et al. (2016); Schenk et al. (2016); Oepen et al. (2016); Stepanov and Riccardi (2016).

This work introduces an end-to-end neural system that implements recently researched components for these tasks. Our goal is to design components that all rely on the same contextualized word embeddings as input and, thus, it avoids the necessity to train huge neural networks. We compare our results with state-of-the-art shallow discourse parsers that took part at the CoNLL Shared Task in 2016 (Xue et al., 2016) which is to our knowledge the last time full systems were published. The contributions of this paper are as follows:

1. We introduce a simple modular and easily extendable framework for shallow discourse parsing. We also provide pretrained systems by recent advances in word representations that demonstrate competitive performance compared to existing systems.

2. We provide a first version of a web front end to visualize parser outputs and, optionally, to connect a parser via REST API.

[1] Senses are ordered in a hierarchy of up to three levels. Temporal refers to the first level and Synchronous to the second, respectively.

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 128–133
November 10–11, 2021. ©2021 Association for Computational Linguistics

2 System Architecture

In our framework `discopy`, we design a clean and modular parser pipeline principle. Each component has to implement a specified interface that assumes to process a well defined document structure. Starting with a document without annotated discourse relations, each component further adds its predictions while the document is passed through the pipeline, for example one component identifies connectives and their sense and, based on this prediction, another component extracts arguments around these connectives to complete the explicit relations. This parser architecture makes the definition of a custom pipeline easy and components remain interchangeable. Together with the framework, we provide multiple components for an end-to-end neural shallow discourse parser. The simple pipeline interface (exemplified in Listing 2) with methods such as *fit*, *parse*, and *eval*, ensures pleasant user experience. The system combines recent research on various subtasks with advances in contextualized word embeddings. We make the framework and pre-trained components available under `https://github.com/rknaebel/discopy`. In the following, we briefly summarize the components implemented in our system.

```
# load pdtb documents with contextualized embeddings
docs_train = load_bert_conll_dataset(...)
docs_val = load_bert_conll_dataset(...)
# definition of the parser pipeline by list of
     components
parser = ParserPipeline([
    # add connective and its sense
    ConnectiveSenseClassifier(...),
    # use connective to add explicit arguments
    ConnectiveArgumentExtractor(...),
    # extract adjacent sentences without explicit
     relation
    ImplicitArgumentExtractor(),
    # attach implicit sense
    ArgumentSenseClassifier(...),
])
# fit custom pipeline on annotated documents
parser.fit(docs_train, docs_val)
# extract discourse relations
parses = parser.parse(docs_val)
```

Listing 1: Discopy pipeline source code example.

Contextualized Word Embeddings. We start by generating contextualized embeddings for each document. These embeddings come from a chosen BERT architecture that is provided by the Hugging Face `transformers` library (Wolf et al., 2020). Pre-tokenized sentences are processed again by a specific BERT tokenizer which results in a possibly higher number of sub-tokens as the input sequence. For each input token, we select and concatenate the last four hidden layers that correspond to the first sub-token, following the principle demonstrated by Devlin et al. (2019) that performed best using the BERT architecture as a feature extractor. Each component in our discourse parser uses the same input embeddings. This has the advantage, to only produce the computation-intense embeddings once in the beginning.

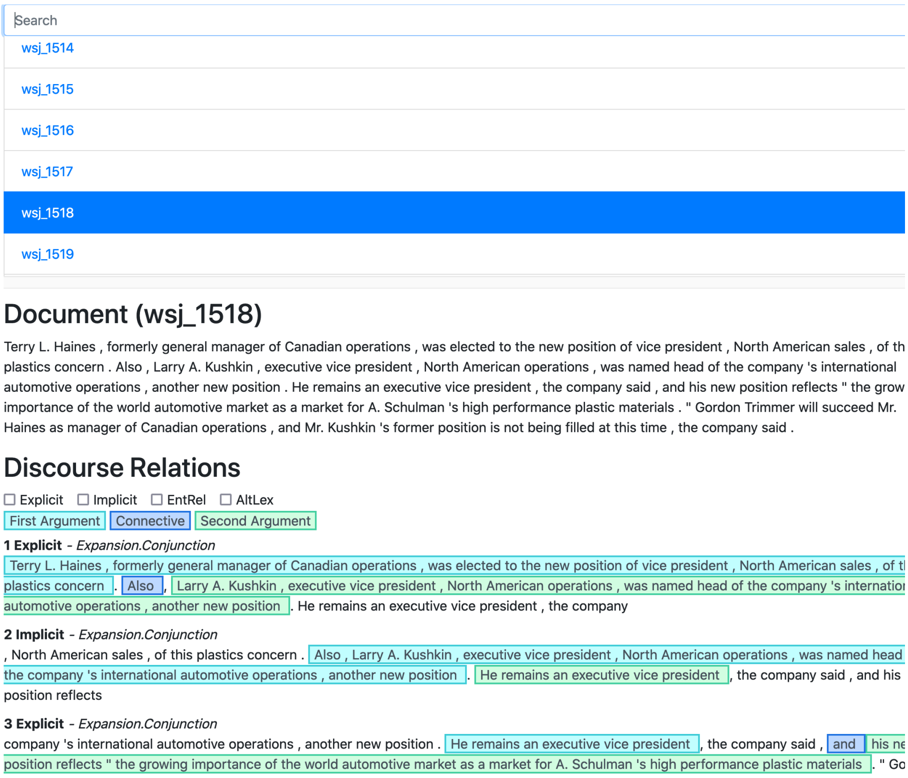
Connective Sense Classifier. The first component in the pipeline refers to the problem of connective disambiguation and explicit sense classification. This component is based on the work of Knaebel and Stede (2020) of which we use the jointly trained model to keep the number of components as small as possible. The component first extracts connective candidates based on a pattern-matching approach. Then, these candidates are classified as having a sentential meaning or a connective meaning which is indicated by a predicted sense class. The idea of using the same features for both tasks originates from Pitler and Nenkova (2009).

Explicit Connective Argument Extractor. The component that is responsible to extract explicit arguments creates a window surrounding the previously predicted connective and searches for corresponding arguments (`Arg1` and `Arg2`) within this span. We extend the connective argument extractor of Knaebel et al. (2019) and replace GloVe embeddings (Pennington et al., 2014) by our contextualized embeddings.

Implicit Sense Classifier. Implicit relations are annotated on adjacent sentences within the same paragraph that do not contain any other explicit relation. For the implicit sense classifier, we choose a simple approach and collect all implicit candidate sentence pairs. We follow previous work by Rutherford et al. (2017) which has been recently chosen as a baseline in work by Liang et al. (2020). There, embeddings for both arguments are collected and processed by a recurrent neural network. Then, both intermediate representations are combined using a maximum pooling layer and given to a fully connected layer to predict the implicit sense. We adopt this approach by using our contextualized embeddings, and we add a labels for entity-based relations (EntRel) as well as a for the absence of any relation (NoRel) to the final set of predictions.

3 Visualization

In conjunction with our system, we provide a prototypical web-based front end with several views that simplifies the inspection of parser results. It can be used to visualize annotated documents, parse docu-



Document (wsj_1518)

Terry L. Haines , formerly general manager of Canadian operations , was elected to the new position of vice president , North American sales , of th plastics concern . Also , Larry A. Kushkin , executive vice president , North American operations , was named head of the company 's international automotive operations , another new position . He remains an executive vice president , the company said , and his new position reflects " the grow importance of the world automotive market as a market for A. Schulman 's high performance plastic materials . " Gordon Trimmer will succeed Mr. Haines as manager of Canadian operations , and Mr. Kushkin 's former position is not being filled at this time , the company said .

Discourse Relations

☐ Explicit ☐ Implicit ☐ EntRel ☐ AltLex

First Argument Connective Second Argument

1 Explicit - *Expansion.Conjunction*

Terry L. Haines , formerly general manager of Canadian operations , was elected to the new position of vice president , North American sales , of tl plastics concern . Also , Larry A. Kushkin , executive vice president , North American operations , was named head of the company 's internatioi automotive operations , another new position . He remains an executive vice president , the company

2 Implicit - *Expansion.Conjunction*

, North American sales , of this plastics concern . Also , Larry A. Kushkin , executive vice president , North American operations , was named head the company 's international automotive operations , another new position . He remains an executive vice president , the company said , and his position reflects

3 Explicit - *Expansion.Conjunction*

company 's international automotive operations , another new position . He remains an executive vice president , the company said , and his ne position reflects " the growing importance of the world automotive market as a market for A. Schulman 's high performance plastic materials . " Go

Figure 1: Web-based front end overview. A selected document is shown with its discourse relations in the bottom.

ments, and inspect them interactively. Further, the visual inspection makes it easier to identify weaknesses of a developed parser, e.g. inaccurate connective identification or misplaced argument spans. The system is built on top of a simple Python web service powered by FastAPI[2] and it is designed using a reactive JavaScript library called VueJS[3]. Both sides communicate via RESTful interface with each other.

Figure 1 demonstrates the view of a selected document. At the top, there is a simple search bar for easily accessing documents. Directly below, the document's text is shown together with a list of its discourse relations at the bottom of the view. Then, each discourse relation is listed with its corresponding relation type (e.g. Explicit, Implicit, EntRel, AltLex), as well as its sense. Different parts of a relation, first and second arguments as well as the connective, are highlighted by colors. Additionally, some context tokens are given before and after the relation if possible.

4 Experiments

The **connective sense classifier** takes the connective candidate and one surrounding context tokens as input. The two hidden layers both have dimensions 256. A dropout layer follows after each hidden layer with a 0.3 drop rate. Both **argument extractors** use two bi-directional recurrent layers with a hidden size of 256 in each direction, a dropout rate of 0.2 before and after the fully connected layer having a dimension of 128. The **explicit extractor** is trained on explicit examples with a window size of 100. The **implicit sense classifier** accepts two arguments with at most 35 tokens each. The same bi-directional recurrent layer with hidden size 128 processes both arguments. The drop rate of the dropout layer that follows after concatenation is 0.25, and the size of the hidden dimension afterward is 128. All models are trained with AMSGrad Adam (Reddi et al., 2018) and a 0.001 learning rate.

5 Results and Discussion

For our evaluation, we use the same experimental design as proposed at the CoNLL Shared Task 2016 (Xue et al., 2016). We have trained multiple pipelines with varying language models used to compute the input embeddings. The model names in our experiment tables follow the naming of the word embedding model[4]. As comparison systems, we choose the overall best performing submissions

[2] fastapi.tiangolo.com

[3] vuejs.org

[4] Model names follow huggingface.co/models

Model	Explicits					Implicits			
	$F1_{parser}$	$F1_{conn}$	$F1_{A1}$	$F1_{A2}$	$F1_{A12}$	$F1_{parser}$	$F1_{A1}$	$F1_{A2}$	$F1_{A12}$
Standard WSJ Test (Section 23)									
ecnucs	40.31	93.96	51.39	76.43	44.31	22.38	64.66	66.86	50.83
oslopots	39.38	94.43	**51.99**	72.57	43.93	18.02	**69.92**	**71.45**	**53.47**
bert-base	**62.75**	95.64	51.24	**78.30**	**51.79**	25.80	45.57	46.82	39.87
bert-large	58.44	95.00	46.67	73.56	46.56	34.01	45.40	46.64	39.68
roberta-base	60.57	96.18	48.69	76.64	49.89	33.31	45.45	46.70	39.71
roberta-large	56.34	96.08	45.52	73.32	45.41	35.12	45.45	46.70	39.75
albert-base	60.57	**96.87**	47.57	76.38	48.54	**40.71**	45.69	46.94	39.99
albert-large	59.30	95.18	44.65	74.30	46.47	33.16	45.57	46.75	39.75
Wikipedia Blind Test									
ecnucs	33.94	91.34	51.05	74.20	42.84	19.54	61.05	75.83	51.15
oslopots	34.45	91.79	52.43	75.20	43.95	21.89	**64.60**	**76.40**	**52.02**
bert-base	57.80	93.88	**56.22**	74.58	53.06	23.20	44.58	55.30	41.42
bert-large	56.89	93.32	49.72	71.61	50.46	32.44	44.61	55.28	41.46
roberta-base	**59.77**	93.95	54.14	**77.40**	**56.00**	36.60	44.51	55.21	41.35
roberta-large	50.94	93.72	46.14	68.76	45.60	36.17	44.42	55.13	41.26
albert-base	55.77	**94.70**	50.46	74.77	49.91	**39.33**	44.90	55.61	41.74
albert-large	51.93	93.73	45.70	74.01	44.98	32.66	44.73	55.44	41.58

Table 1: Experimental results (**strict** F1 scores) for Section 23 of WSJ and the blind data set proposed for CoNLL Shared Task.

for each test set (ecnucs (Wang and Lan, 2016) and oslopots (Oepen et al., 2016)). Similar to the Shared Task, we compare two different thresholds: **strict** refers to an exact match of relations while **partial** counts relations as correct if their overlap is at least 70 %. Performances are measured on Section 23 of the PDTB as well as PDTB-like annotated Wikipedia data prepared for the Shared Task.

In Table 1, strict results of the studied models are summarized. In the table's left part pointing to explicit relations, neural models outperform traditional models in most categories except for Arg1 identification, but still results are pretty close. Especially the bert-base model has an overall high performance. Interestingly, albert-base performs quite good, even though it has the smallest number of trained parameters. Also, it seems not beneficial in our experimental design to use a *large* version of a model. A reason for that might be the higher number of input values per token leading the model to an unstable training. The performance for implicit relations (arguments and overall) is noticeably weak as we used I fairly simple method for recognizing arguments following Lin et al. (2014). Thereby, the overall implicit results are lowered as relations were not counted as correct during evaluation. The partial scores in Table 2 were expected to be higher compared to their strict counter parts.

6 Conclusions

In this work, we demonstrated a framework for shallow discourse parsing and integrated a web-based graphical user interface to study parser outputs. We presented an exemplary pipeline built from neural components that are based on recent models. We extend previous approaches to incorporate precomputed contextualized word embeddings. Our pipeline system performs competitively to former discourse parsers and partially outperforms them, while not using any linguistic features.

In the future, we plan to integrate more recent neural architectures into the system to improve overall scores. Further, we want to improve the graphical interface to seamlessly integrate the visualization directly into the text such as done with brat[5] for dependency parses. Thus, it will be easier to compare multiple relations in their context. Overlapping arguments are a major problem, why we decided to list relations separately. Building on that a view would be useful for comparing discourse relations across documents, e.g. comparing predictions and gold annotations or multiple parser outputs.

7 Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project 455911521.

[5] brat.nlplab.org

Model	Explicits				Implicits			
	$F1_{parser}$	$F1_{A1}$	$F1_{A2}$	$F1_{A12}$	$F1_{parser}$	$F1_{A1}$	$F1_{A2}$	$F1_{A12}$
Standard WSJ Test (Section 23)								
ecnucs	69.21	**72.16**	88.62	74.89	28.60	82.78	85.65	**86.55**
oslopots	65.96	70.75	86.90	71.27	24.36	**84.74**	**86.47**	85.85
bert-base	77.59	68.47	**88.68**	77.53	37.96	57.59	58.42	60.44
bert-large	76.11	65.78	85.67	75.33	42.74	57.39	58.15	60.21
roberta-base	76.19	69.21	87.12	**78.17**	40.97	57.28	58.25	60.19
roberta-large	73.04	64.30	86.21	74.18	43.74	57.26	58.23	60.18
albert-base	77.99	68.50	**88.67**	77.99	**47.37**	57.58	58.55	60.57
albert-large	**79.19**	67.45	85.65	77.52	42.37	57.41	58.31	60.32
Wikipedia Blind Test								
ecnucs	57.25	70.19	79.67	71.69	26.90	79.53	84.11	82.73
oslopots	56.66	71.96	81.73	71.74	33.23	**84.47**	**88.98**	**86.31**
bert-base	74.49	**73.47**	**83.86**	79.22	35.96	57.45	60.05	60.61
bert-large	72.48	68.65	78.29	76.81	40.97	57.42	60.00	60.45
roberta-base	73.49	72.93	83.72	**80.74**	45.49	57.35	59.94	60.39
roberta-large	71.15	66.79	77.56	75.40	45.87	57.27	59.75	60.43
albert-base	**75.06**	71.66	83.36	80.62	**48.67**	57.76	60.24	60.80
albert-large	74.49	68.64	81.18	77.60	41.57	57.69	60.17	60.62

Table 2: Experimental results (**partial** F1 scores with 0.7 overlap) for Section 23 of WSJ and the blind data set proposed for CoNLL Shared Task.

References

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.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

René Knaebel and Manfred Stede. 2020. Contextualized embeddings for connective disambiguation in shallow discourse parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 65–75, Online. Association for Computational Linguistics.

René Knaebel, Manfred Stede, and Sebastian Stober. 2019. Window-based neural tagging for shallow discourse argument labeling. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 768–777, Hong Kong, China. Association for Computational Linguistics.

Li Liang, Zheng Zhao, and Bonnie Webber. 2020. Extending implicit discourse relation recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147, Online. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfred Stede, Erik Velldal, and Lilja Øvrelid. 2016. OPT: Oslo–Potsdam–Teesside. pipelining rules, rankers, and classifier ensembles for shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 20–26, Berlin, Germany. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow discourse parsing using convolutional neural network. In *Proceedings of the CoNLL-16 shared task*, pages 70–77, Berlin, Germany. Association for Computational Linguistics.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291, Valencia, Spain. Association for Computational Linguistics.

Niko Schenk, Christian Chiarcos, Kathrin Donandt, Samuel Rönnqvist, Evgeny Stepanov, and Giuseppe Riccardi. 2016. Do we really need all those rich linguistic features? a neural network-based approach to implicit sense labeling. In *Proceedings of the CoNLL-16 shared task*, pages 41–49, Berlin, Germany. Association for Computational Linguistics.

Evgeny Stepanov and Giuseppe Riccardi. 2016. UniTN end-to-end discourse parser for CoNLL 2016 shared task. In *Proceedings of the CoNLL-16 shared task*, pages 85–91, Berlin, Germany. Association for Computational Linguistics.

Jianxiang Wang and Man Lan. 2016. Two end-to-end shallow discourse parsers for English and Chinese in CoNLL-2016 shared task. In *Proceedings of the CoNLL-16 shared task*, pages 33–40, Berlin, Germany. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *AAAI*.