Modal Dependency Parsing via Language Model Priming

Jiarui Yao[1], Nianwen Xue[1], and Bonan Min[2]

[1]Brandeis University
[2]Raytheon BBN Technologies
{jryao, xuen}@brandeis.edu, bonan.min@raytheon.com

Abstract

The task of modal dependency parsing aims to parse a text into its modal dependency structure, which is a representation for the factuality of events in the text. We design a modal dependency parser that is based on priming pre-trained language models, and evaluate the parser on two data sets. Compared to baselines, we show an improvement of 2.6% in F-score for English and 4.6% for Chinese. To the best of our knowledge, this is also the first work on Chinese modal dependency parsing.

1 Introduction

Modal dependency parsing (MDP) is the task of parsing a text into a modal dependency structure (MDS) (Vigus et al., 2019) in which each event in the text is linked to a *conceiver*, the information source of the event. An MDS is a graph in which the nodes are events and conceivers, and the edges represent the level of certainty that a conceiver holds with respect to the event. An example MDS is presented in Figure 1, where *Jeroen Weimar* is the conceiver of the event *travelled*, and is certain that the traveling event has happened, as indicated by the edge label *Pos*. Vigus et al. (2019) define 6 categories of modal strength, or levels of certainty, and they are *full positive (Pos), partial positive (Prt), positive neutral (Neut), negative neutral (Neutneg), partial negative (Prtneg)* and *full negative (Neg)*. The root node of an MDS is always the author (AUTHOR) of a document, the ultimate source of all information sources mentioned in the text.

Modal dependency parsing is thus the task of taking a text as input and parsing it into a modal dependency structure. MDP departs from previous approaches to event factuality prediction that cast it as an event classification (e.g. Saurí and Pustejovsky (2012)) or regression (e.g. Lee et al. (2015)) problem aimed at just predicting the level of certainty of an event. The level of certainty alone is insufficient in judging the factuality of an event, and knowing the information source (conceiver) is also crucially important. For example, in Figure 1, our judgment of whether the event *travelled* has happened also crucially depends on the credibility of the information source, *Jeroen Weimar*, in addition to the level of certainty the information source holds towards the event.

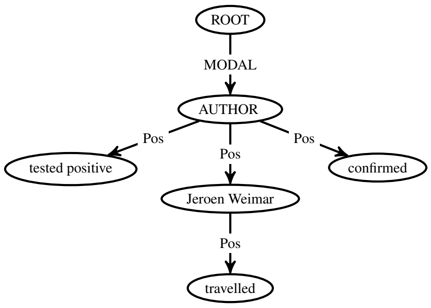

Figure 1: A modal dependency tree for "A person in Traralgon had **tested positive** to COVID-19 on Sunday. The Victorian government's COVID-19 response commander Jeroen Weimar **confirmed** 'this individual has **travelled** to Melbourne.' "

Yao et al. (2021) develop the first modal dependency parser by first separately extracting events and conceivers, then building up the MDS bottom-up with a ranking model. One shortcoming of this approach is that it fails to capture the fact that the status of an entity as a conceiver is conditioned on its being the information source of an event. For instance, in Figure 1, *a person* is an entity but is not a conceiver as it is not the source of any event. As a result, Yao et al. (2021) report relatively low conceiver extraction F-score compared to event extraction (70.4% for conceiver extraction vs. 90.8% for event extraction). Errors in conceiver extraction will propagate to the structure building stage, leading to lower overall MDS parsing accuracy.

In this paper, we describe an approach to MDP based on language model priming in which we construct a prompt with an event and use it to predict

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2913 - 2919
July 10-15, 2022 ©2022 Association for Computational Linguistics

its chain of conceivers as well as the level of certainty the conceivers hold [1]. This approach avoids the error propagation problem in Yao et al. (2021) and also takes better advantage of powerful pre-trained language models. Our experiments show that this approach outperforms previous models for both English and Chinese.

2 Approach

We approach MDS parsing by first performing event extraction, then use the extracted events to construct the prompt for the purpose of identifying their conceivers. Given a document, a language model such as BERT (Devlin et al., 2019) is used to obtain the contextualized representation for each token. A standard BIO tagging model is then applied to identify events, where B, I, O refer to the beginning, inside, and outside of an event respectively.

The next step in MDS parsing is to identify conceivers for each extracted event. More formally, given an event e_i as a child node, the task is to extract (e_i, c_i, cc_i), where c_i is the conceiver of e_i, and cc_i is the conceiver of c_i. In theory, a child event can have a chain of conceivers longer than two, but in over 96% of the cases, an event has a chain of two conceivers or less. We thus made the simplifying assumption that an event can have a chain of two conceivers at most.

Our model receives an event-specific text sequence as input, then predicts a tag from a target set {B-C, I-C, B-CoC, I-CoC, O} for each token in the sequence. B-C and I-C labels are for tokens in c_i, and B-CoC, I-CoC are for tokens in cc_i. We construct the event-specific sequence, seq_i, by concatenating a prompt and a context sequence in the form of *[CLS] a prompt [SEP] a context sequence [SEP]*. Let s_i denote the sentence containing e_i, we add token markers <EVENT>, </EVENT> before and after the event span in s_i to get the prompt for e_i. For a child event e_i, its parent conceiver can usually be found within a window surrounding s_i. Thus, the context sequence for e_i is constructed by taking the surrounding sentences of s_i in a window, followed by two special tokens <AUTHOR> and <NULL> representing the AUTHOR and NULL-CONCEIVER node[2]. Figure 2 shows an example of the input sequence seq_i with gold tags.

The input sequence seq_i is then encoded with a pre-trained language model. Let H = ($\boldsymbol{h}_1$, ..., $\boldsymbol{h}_m$) denote a sequence of contextualized representations for the input tokens in seq_i, the score for the tag of the j-th token is:

$$\hat{y}_j^{tag} = \mathrm{FFN}_1(\boldsymbol{h}_j),$$

where FFN_1 is a feed-forward neural network.

To learn the edge label between a child node and its parent node, we use a separate feed-forward neural network to map $\boldsymbol{h}_j$ to the edge label set. The edge label set includes the modal relations in the data set plus the N/A label, which is chosen when there is no relation between the child node e_i and token j, i.e. when token j is neither part of the conceiver of e_i nor part of the conceiver of the conceiver of e_i . The score for the edge label of the j-th token is:

$$\hat{y}_j^{label} = \mathrm{FFN}_2(\boldsymbol{h}_j),$$

where FFN_2 is a feed-forward neural network.

In the training phase, we minimize the following cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_l,$$

where $\mathcal{L}_t$ and $\mathcal{L}_l$ refer to the parent tagging loss and edge labeling loss respectively.

Inference In an MDS, each child node only has one parent node. To enforce a well-formed MDS, we apply two rules in the inference stage: (i) if more than one conceiver is predicted for e_i, the first prediction is taken; (ii) if a conceiver doesn't have a conceiver, by default it is attached to the AUTHOR with the majority label in the data set.

3 Experiments

3.1 Data

We evaluate our approach on an English modal dependency data set (Yao et al., 2021) and a Chinese modal dependency data set (Liu and Xue, 2022) that consists of about 300 news articles. For English, we use the same data split as in Yao et al. (2021). For Chinese, we randomly split the data set to training (train in Table 1), developing (dev) and test (test) sets. The statistics of the two data sets are in Table 1.

[1] https://github.com/Jryao/mdp_prompt

[2] The NULL-CONCEIVER node is used when the conceiver is not specified.

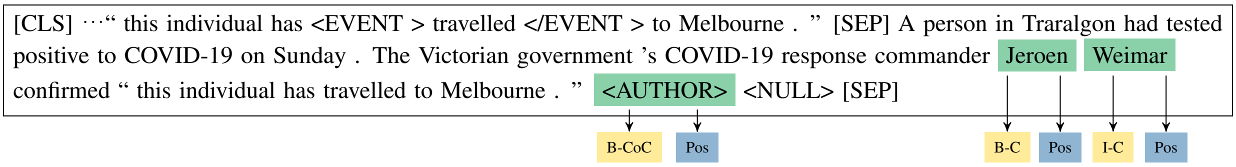
[CLS] ···“ this individual has <EVENT > travelled </EVENT > to Melbourne . ” [SEP] A person in Traralgon had tested positive to COVID-19 on Sunday . The Victorian government ’s COVID-19 response commander Jeroen Weimar confirmed “ this individual has travelled to Melbourne . ” <AUTHOR> <NULL> [SEP]

Figure 2: An example of the input sequence for language models with corresponding gold BIO tags and edge labels (O and N/A tags omitted). The child event is *travelled*. The conceiver of the event *travelled* is *Jeroen Weimar*. The conceiver of *Jeroen Weimar* is the AUTHOR. Note only the tokens after the first [SEP] token are labeled in our model. C and CoC refer to conceiver and conceiver of the conceiver respectively, Pos refers to the full positive label.

		# Doc	# Event	# Conc
English	train	289	19,541	2,344
	dev	32	2,307	298
	test	32	2,168	296
Chinese	train	237	11,679	879
	dev	30	1,464	136
	test	30	1,318	116

Table 1: Data splits for the experiments. Number of documents, events and conceivers are listed.

3.2 Baselines

When evaluating English modal dependency parsing, we compare our prompt-based model with two variants of the ranking based models described in Yao et al. (2021): a pipeline model and a joint model. The joint model uses a shared BERT encoder for both event/conceiver extraction and structure building.

As there is no existing model for Chinese modal dependency parsing, we re-implemented the joint learning variant of the ranking based model in Yao et al. (2021) to serve as our baseline, with minor modifications. We use a shared BERT encoder for the event/conceiver extraction and structure building, following Yao et al. (2021), but encode all the sentences in a document as a long sequence instead of encoding it sentence by sentence. Full details about the differences between the two models can be found in Appendix C.

3.3 Experiment Setup

We use the Hugging Face (Wolf et al., 2020) implementation of XLM-RoBERTa-base (Conneau et al., 2020) for Chinese. For English, we use BERT-large-cased (Devlin et al., 2019), same as Yao et al. (2021). When generating input sequences for the proposed prompt-based model, we use a window of 5 sentences before and 5 sentences after for English, and all the sentences before and 3 sentences after for Chinese. For the ranking baseline, we select candidate parents from the same window size as the prompt-based model, and keep at most 16 candidate parents for English, 40 for Chinese. Our window size and number of candidate parents are consistent with Yao et al. (2021) (for English), for Chinese, they cover over 99% of the cases in the Chinese development set. Full details of the hyperparameter settings can be found in the Appendix.

3.4 Main Results

Tables 2 and 3 present the experimental results. Same as Yao et al. (2021), we report the exact match scores for event identification, and micro-average F scores for all experiments. For modal dependency parsing, F scores are computed on <*child*, *parent*, *relation*> triples, with results based on system-identified events and conceivers.

Event identification In Table 2, we compare our event identification (ID) model with previous models. All models extract events using a BIO tagger. On English data, our model is slightly better than both models in Yao et al. (2021). Cross-lingually, our English event ID results are higher than Chinese results. Possible reasons are discussed in Section 3.5.

Models	English		Chinese	
	Dev	Test	Dev	Test
Yao et al. (2021)-P	92.7	90.9	-	-
Yao et al. (2021)-J	92.8	90.8	-	-
Ours	**93.2**	**91.9**	87.4	88.6

Table 2: Event identification F scores. P and J refer to the pipeline model and joint model respectively.

Overall parsing Table 3 presents a comparison of our prompt-based model with previous results and our own baseline. For both English and Chinese modal dependency parsing, our prompt-based model consistently outperforms all baselines. Our prompt-based model outperforms the pipeline model of Yao et al. (2021)-P by 3.0% on the development set and 4.4% on the test set. In addition, our own baseline is slightly better than Yao et al. (2021)-J, a ranking-based joint model, possibly because of the different encoding mechanisms the two models use (see 3.2). Lastly, compared with our own baseline, the prompt-based model achieves an improvement of 0.9% in absolute F-score on the English development set and 2.6% on the English test set. For Chinese, the improvements are even larger: 3.8% on the development set and 4.6% on the test set.

Models	English		Chinese	
	Dev	Test	Dev	Test
Yao et al. (2021)-P	69.7	67.5	-	-
Yao et al. (2021)-J	70.3	69.0	-	-
baseline (ours)	71.8	69.3	61.7	59.0
prompt-based (ours)	**72.7**	**71.9**	**65.5**	**63.6**

Table 3: Modal dependency parsing F scores.

3.5 Cross-lingual Comparison

Our experimental results show that English MDP results are in general better than Chinese. There are a few possible explanations. As discussed in Section 3.1, the English data set is larger than the Chinese data set on every count. More training data typically means higher model accuracy. A closer look at the data reveals other differences between the two data sets as well. Table 4 breaks down the types of parent a child has for the two languages. We can see that in the English data, 69% of child nodes have the AUTHOR as parent, while in the Chinese data, that percentage is 52.6%. The two data sets have similar proportion of cases when the child is in the same sentence as the parent: 23.7% vs. 27.5%. However, the Chinese data set has a much higher percentage of cases where the parent is in a different sentence from the child: 19.9% vs. 5.7%. Parents that are further away are harder to predict. There is a linguistic explanation for why in Chinese parent conceivers are further apart from the event child: Chinese allows dropped pronouns, and as a result, the conceiver is often found in a previous sentence of the event. In Table 5, 王军 (Wang Jun) in Sentence 8 is the conceiver of events in Sentence 9 because of a dropped pronoun in Sentence 9.

	AUTHOR	NULL	Same sent	Cross sent
Eng	69.0%	1.6%	23.7%	5.7%
Chn	52.6%	0.0%	27.5%	19.9%

Table 4: Statistics of parent node types: AUTHOR, NULL-CONCEIVER, parents in the same sentence, or parents in different sentences.

...[S8] 王军指出，今年是“十三五”规划收官之年，下半年各项税收工作任务异常艰巨。

...[S8] Wang Jun pointed out that this year is the end of the 13th Five-Year Plan, and the taxation tasks in the second half of the year are extremely challenging.

[S9] (王军指出) 各级税务机关既要抓好重点工作落实，努力把疫情造成的损失补回来。

[S9] (Wang Jun pointed out) Tax authorities at all levels should not only do a good job in implementing key tasks, and strive to make up for the losses caused by the pandemic.

Table 5: Examples in the Chinese data set. Tokens in parentheses are dropped in the original document.

4 Related Work

Early works cast event factuality prediction (EFP) as a classification or regression problem and have employed rule-based (Nairn et al., 2006; Lotan et al., 2013) or machine learning approaches (Diab et al., 2009; Lee et al., 2015; Saurí and Pustejovsky, 2012; Stanovsky et al., 2017). More recently, different types of neural models have been applied to this problem, such as LSTM-based RNNs (Rudinger et al., 2018), Generative Adversarial Networks (Qian et al., 2018), or graph neural networks (Pouran Ben Veyseh et al., 2019). Qian et al. (2019) and Cao et al. (2021) extended the sentence level task to document-level EFP. Our work is most closely related to that of Yao et al. (2021), which casts EFP as modal dependency parsing. However, they first extract events and conceivers and then build the MDS by ranking the candidate parents for each event. In contrast, we perform modal dependency parsing by constructing a prompt with the event to predict its conceiver parent, simplifying the pipeline. Our prompt-based approach also bears resemblance to works applying prompt-based learning to other NLP tasks, such as event extraction (Liu et al., 2020; Fincke et al., 2021), relation extraction (Li et al., 2019), named entity recognition (Li et al., 2020) and coreference resolution

(Wu et al., 2020).

5 Conclusion

In this paper, we propose a model for modal dependency parsing based on priming pre-trained language models. We evaluate the model on an English modal dependency data set, and for the first time, evaluate the model on a Chinese modal dependency data set. Experimental results show that our model consistently outperforms baselines on both data sets.

Acknowledgements

We thank the anonymous reviewers for their helpful comments.

This work is supported in part by a grant from the IIS Division of National Science Foundation (Award No. 1763926) entitled "Building a Uniform Meaning Representation for Natural Language Processing". All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation.

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600006 under the IARPA BETTER program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Pengfei Cao, Yubo Chen, Yuqing Yang, Kang Liu, and Jun Zhao. 2021. Uncertain local-to-global networks for document-level event factuality identification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2636–2645, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 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.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore. Association for Computational Linguistics.

Steven Fincke, Shantanu Agarwal, Scott Miller, and Elizabeth Boschee. 2021. Language model priming for cross-lingual event extraction.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Zhifu Liu and Nianwen Xue. 2022. A dependency structure annotation for modality in Chinese news articles. In *Proceedings of the 23th Chinese Lexical Semantics Workshop (CLSW2022)*, Fuzhou, China.

Amnon Lotan, Asher Stern, and Ido Dagan. 2013. TruthTeller: Annotating predicate truth. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–757, Atlanta, Georgia. Association for Computational Linguistics.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of the Fifth International*

Workshop on Inference in Computational Semantics (ICoS-5).

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399, Florence, Italy. Association for Computational Linguistics.

Zhong Qian, Peifeng Li, Yue Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Event factuality identification via generative adversarial networks with auxiliary classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4293–4300. International Joint Conferences on Artificial Intelligence Organization.

Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2019. Document-level event factuality identification via adversarial neural network. 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 2799–2809, Minneapolis, Minnesota. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357, Vancouver, Canada. Association for Computational Linguistics.

Meagan Vigus, Jens E. L. Van Gysel, and William Croft. 2019. A dependency structure annotation for modality. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 182–198, Florence, Italy. 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.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Jiarui Yao, Haoling Qiu, Jin Zhao, Bonan Min, and Nianwen Xue. 2021. Factuality assessment as modal dependency 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 1540–1550, Online. Association for Computational Linguistics.

A Data sets

We use a publicly available English modal dependency data set constructed by Yao et al. (2021), which consists of news articles from the following news media sources: Business Standard, Business Insider, NBC News, The New York Times, Reuters, The Guardian, The Washington Post, CNN, Fox News, Yahoo News and Wikinews. We also use a Chinese modal dependency data set constructed by Liu and Xue (2022) that consists of news articles from Xinhua newswire.

B Implementation details

We optimize our models with the BertAdam optimizer of a linear scheduler with a warmup ratio of 0.1. The learning rate is 2e-5. We apply a dropout rate of 0.1 over the last layer of the pretrained language model output to get the contextualized representations. We use a 2-layer FFN with ReLU activations for all models. The hidden unit size of the FFNs is the hidden size of the pretrained language model, i.e. 1024 for bert-large-cased, 768 for xlm-roberta-base. For the proposed prompt-based model, we use a batch size of 12, maximum sequence length of 512 for Chinese, a batch size of 6, maximum sequence length of 384 for English. Sequences that are longer than the maximum sequence length are cut to segments with a stride of 64 for both languages.

We train all the models for 30 epochs on a NVIDIA Tesla V100 (16 GB) GPU. We run all the models for 3 runs with different seeds, and report the average F-scores across runs. Each epoch takes about 45 minutes for English, 19 minutes for Chinese.

C Baselines

We give more details about the two ranking baselines: baseline (ours) and Yao et al. (2021)-J in Table 3. Given a child node, the two models first generate a candidate parent set for the child, then compute the pair score for each child-parent pair. The candidate parent with the highest pair score is selected as the parent. There are a few differences between the two models. First, Yao et al. (2021) encode a document sentence by sentence, i.e. they add a [CLS] and [SEP] token before and after each sentence and encode them with the language model. We encode all the sentences in a document together, i.e. we add a [CLS] and [SEP] token before and after each document, and encode it with the language model. If a document is longer than the maximum sequence length (T), we split it into segments and encode each segment independently. Each segment has T/2 overlapping tokens with the previous segment. The values of T are the same as the maximum sequence length values in section B. The final token representations are derived by taking the average of the token representations in each segment. Next, we obtain the node representations by simply taking the average token representations in a node, while they take the concatenation of the start token, end token and the span token vector in the node as the node representations. Lastly, even if Yao et al. (2021) propose a multi-task learning model by jointly learning node identification and structure building, they train the structure building stage with gold nodes. Our baseline is trained in an end2end fashion: the model first identifies nodes, then uses the system identified nodes as the input for the structure building stage.