Biaffine Modal Dependency Parsing

Anonymous ACL submission

Abstract

A modal dependency structure represents a web of connections in a document describing the source and epistemic strength of statements that helps to establish factuality in a given text. Obtaining such graphs defines a core task of modal dependency parsing, which involves event and source identification as well as labeling of modal relations between them. In this paper, we propose a simple yet effective biaffine modal dependency parser for English and Chinese that outperforms previous work.

1 Introduction

At a time when we find ourselves inundated with endless streams of new information and knowledge, being able to identify and trace a source of information as well as confidence with which it is conveyed is often helpful—if not sometimes critical—for better understanding the context behind a text or discourse. Modal dependency structure (MDS) (Vigus et al., 2019) is designed with such representation in mind, where the sources (formerly known as *concievers*) and events are the nodes of the graph and their edges denote (1) source of factualiy via its direction and (2) level of certainty via its label as a combination of 3 modal strengths (*Full*, *Partial*, and *Neutral*) and 2 polarities (*Affirmative* and *Negative*) based on the annotation scheme from FactBank (Saurí and Pustejovsky, 2009).

Figure 1 shows an example modal dependency tree for a sample document: ‘`Kim left to join the others. “They are probably eating,” she said.`’ Rooted by an abstract author (`author`) node whose presence is implied everywhere as the creator of the document, an MDS often shows heavy traffic through the author as a principal source of many statements. In the example, the author is responsible for claiming that the event of *Kim* having *left* and *said* occurred with full certainty, but the opposite is the case with *join*

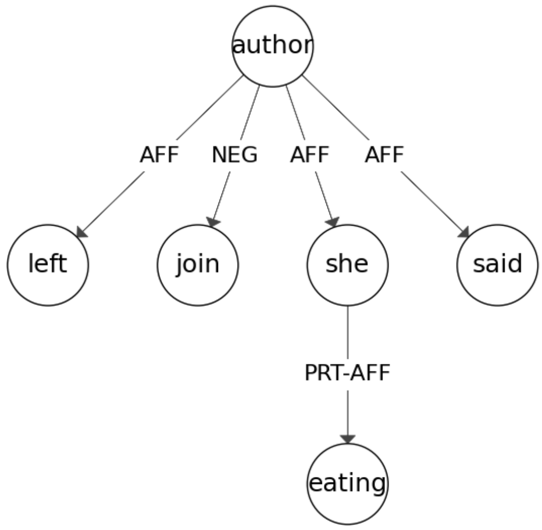

Figure 1: Example of Modal Dependency Graph for “Kim **left** to **join** the others. ‘They are *probably* **eating**,’ she **said**.” `AFF` stands for `full-affirmative`, `NEG` for `full-negative`, and `PRT-AFF` for `partial-affirmative`.

event, which is best described as a purpose behind *Kim*’s decision to leave. The author further participates in a chain of conceivers that can be seen with the `author-to-she` (coreferent to *Kim*) `full-affirmative` edge. This representation allows for a chain of sources to arise which is typical with reporting or relaying of information. In Figure 1 it is *Kim*’s judgment that *eating* **probably** (`partial-affirmative`) happened, which is then relayed with full confidence by the author (`full-affirmative`) to the audience.

In order to obtain modal dependency tree[1] from a text input, modal dependency parsing (MDP) needs to identify the events and conceivers in addition to predicting relations between them. Yao et al. (2021) first reported baseline results on MDP with

[1] In general, MDS forms a tree not a graph.

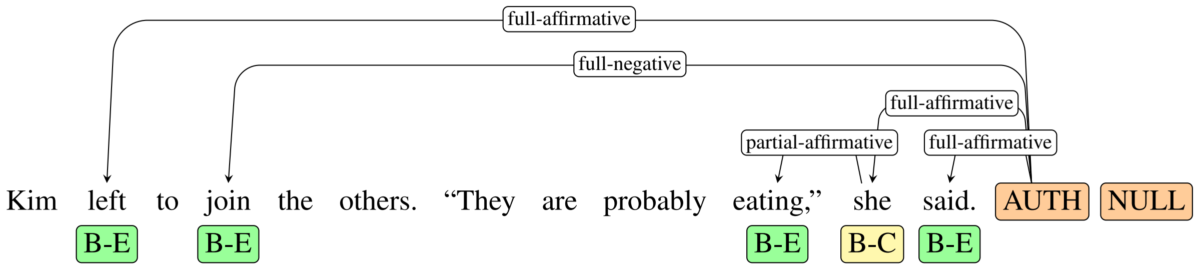

Figure 2: Example of biaffine modal dependency parsing for: ‘Kim left to join the others. “They are probably eating,” she said.’ Orange nodes indicate special abstract nodes `Author` and `Null-Conceiver`. BIO tags below are predicted by the tagger for spans of events and conceivers. Arcs and their labels above are generated by biaffine dependency parser.

a 2-stage pipeline that consists of tagging followed by ranking of parent candidates to construct a graph. Yao et al. (2022) followed up by framing the task as language model priming, in which a prompt with an event in question is provided with context from which its parent, optional grandparent as well as their modal labels are predicted by a fine-tuned language model. While this method avoids error propagation from earlier work by being trained end-to-end, the context is local in scope as determined by the number of sentences before and after the sentence in which the event in question occurs. This entails having to manually define a context window which is often arbitrary and sub-optimal.

In this work, we present a simple yet effective solution in the form of biaffine modal dependency parsing whose context scope naturally encompasses the entire document. The model consists of token-level classification for event/conceiver identification paired with biaffine module for arc generation and labeling. This approach not only avoids the error propagation of baseline ranking model but also only requires a single pass over a document owing to its global scope. Experiments show that our approach outperforms previous work in English and Chinese MDP.

2 Related Work

Traditionally, event factuality prediction (EFP) was seen as a classification or regression problem that involved rule-based (Nairn et al., 2006; Lotan et al., 2013) or statistical approaches (Diab et al., 2009; Saurí and Pustejovsky, 2012; Lee et al., 2015; Stanovsky et al., 2017). With widespread adoption of deep learning came a surge of neural models for the task, for instance based on LSTMs (Rudinger et al., 2018), GANs (Qian et al., 2018) or GNNs (Pouran Ben Veyseh et al., 2019). Yao et al. (2021) is the first work that casted EFP as modal dependency parsing and reported baseline results along with publicly available annotations in English[2]. This was followed up by prompt-based parser (Yao et al., 2022) that alleviated error propagation inherent in the pipeline approach of the baseline in addition to reporting first results on Chinese MDP. Our biaffine model further simplifies the setup while improving on model performance in both languages. This line of approach based on deep biaffine scoring mainly traces its roots to dependency parsing (Dozat and Manning, 2017, 2018; Zhang et al., 2020) but has also been explored in other areas such as NER tagging (Yu et al., 2020) and constituency parsing (Bai et al., 2021; Chen and Komachi, 2023).

3 Approach

Our approach predicts (1) event and conceiver spans via token classification and (2) arcs and relation labels via biaffine dependency parsing in a single step. These modules share a common document encoder which relies on pre-trained language model (PLM) for contextualized embeddings.

The BIO tagger is inherited from Yao et al. (2021) and Yao et al. (2022) where B, I, and O refer to beginning, inside, and outside of a span respectively. The identified events and conceivers then attempt to locate their parent via biaffine scoring mechanism in a greedy manner. Once the parent is located, the newly created edge is labeled by a separate biaffine layer.

Figure 2 shows an example of this approach, where the input text is augmented with two special tokens Author and Null Conceiver[3] at the end.

[2] `https://github.com/jryao/modal_dependency`

[3] A Null Conceiver is a special case when a conceiver is not specified.

English	Train	Dev	Test
Documents	289	32	32
Sentences	6,825	740	759
Tokens	151,487	17,308	17,177
Conceivers	2,344	298	296
Events	19,541	2,307	2,168
AFF	18,425	2,205	2,077
NEG	800	99	89
PRT-AFF	1,292	165	158
NEUT-AFF	1,368	136	140

Chinese	Train	Dev	Test
Documents	237	30	30
Sentences	3,187	398	366
Tokens	79,809	10,352	10,053
Conceivers	879	136	116
Events	11,679	1,464	1,318
AFF	10,879	1,383	1,257
NEG	331 (298*)	50 (45*)	31
PRT-AFF	919	103	101
PRT-NEG	0 (26*)	0 (5*)	0
NEUT-AFF	429	64	45
NEUT-NEG	0 (7*)	0	0

Table 1: Summary statistics of English and Chinese modal dependency datasets. Conceivers does not include Author which occurs once per document. Labels does not include Depends-on which occurs once per document. AFF stands for Affirmative, NEG stands for Negative, PRT stands for Partial and NEUT stands for Neutral. *Numbers in parenthesis in Chinese statistics denote counts of fine-grained negative values in a 6-way version of the corpus.

They serve as target index for arc generation as shown by the dependency arcs above. Colored BIO tags below indicate the spans of events and conceivers as predicted by the tagger.

The figure also highlights the core difference of our setup against that of conventional dependency parsing, where it is assumed that every token has a parent to point to. Since this approach only focuses on spans annotated by BIO tagger, it may be described as being comparatively sparse, which is partially offset by the fact that text input in MDP is generally a multi-sentence document.

Model

Formally, a document $\boldsymbol{d}$ is represented as a sequence of tokens $(\boldsymbol{t_0}, ..., \boldsymbol{t_{-1}}, \text{AUTH}, \text{NULL})$, where the surface tokens are followed by two special tokens denoting the Author and Null Conceiver.

Let $\boldsymbol{H} = (\boldsymbol{h_0}, ..., \boldsymbol{h_{-1}}, \boldsymbol{h}_{\text{AUTH}}, \boldsymbol{h}_{\text{NULL}})$ be the contextualized embedding output from PLM for the document $\boldsymbol{d}$. Tag score for $\boldsymbol{i}$th token is obtained by a feedforward layer:

$$\hat{y}_i^{tag} = \text{FFN}(\boldsymbol{h_i})$$

Arc and relation scores for $\boldsymbol{i}$th token and $\boldsymbol{j}$th parent candidate token is obtained by two independent biaffine scorers:

$$\hat{y}_{i,j}^{arc} = \text{Biaffine}_1(\boldsymbol{h_i}, \boldsymbol{h_j})$$
$$\hat{y}_{i,j}^{rel} = \text{Biaffine}_2(\boldsymbol{h_i}, \boldsymbol{h_j})$$

Our model attempts to minimize the negative log likelihood which is the sum of cross entropy losses from 3 sub-tasks:

$$\mathcal{L} = \mathcal{L}_{tag} + \mathcal{L}_{arc} + \mathcal{L}_{rel}$$

Loss signals for arc and relation are not generated from non-event and non-conceiver tokens.

Inference

Spans of events and conceivers are first identified by the BIO tagger. As we search for parent of each of these entities, non-events and non-conceivers (labeled O by the tagger) are masked out to guide decoding process.

For each span, the first token is taken as representative of the whole, and the arc generator produces a score against all of the other spans and special tokens Author and Null Conceiver, with the argmax as the most compatible head. If a parent span consists of multiple tokens, it is only required that some index within the span be predicted by arc generator in order to correctly assign the parent. The emergency fall-back behavior is to attach to the Author node to ensure the graph is connected.

4 Experiments

4.1 Data

The parser is trained and evaluated using the English (Yao et al., 2021) and Chinese (Liu and Xue, 2023) modal dependency corpora. We follow previous work on the train/eval/test splits for both languages. The summary statistics are provided in Table 1.

Models	Split	English			Chinese		
		Event	Conceiver	Parsing	Event	Conceiver	Parsing
Baseline	Dev	92.8	71.1	71.8*	-	-	61.7*
	Test	90.9	70.4	69.3*	-	-	59.0*
Prompt-based	Dev	93.2	-	72.7	**87.4**	-	65.5
	Test	91.9	-	71.9	**88.6**	-	63.6
Biaffine	Dev	**93.3**	**72.7**	**74.0**	86.7	**88.6**	**68.2**
	Test	**92.0**	**74.2**	**72.6**	87.4	**87.5**	**66.1**

Table 2: Experimental results showing Event and Conceiver identification and Parsing micro-F score. Empty values indicate unreported results. *Baseline parsing results are based on the re-implementation of Yao et al. (2022) rather than from the original publication (Yao et al., 2021).

Unlike English dataset which only offers coarse modal labels where all of negative polarity labels are merged into `full-negative` (Yao et al., 2021), Chinese dataset additionally offers a fine-grained version with `partial-negative` and `neutral-negative` annotations, albeit only a few in number. It is not explicitly stated which version is used in the experiments of Yao et al. (2022); we report results using the fine-grained version.

4.2 Setup

We use the Huggingface[4] (Wolf et al., 2020) implementation of Longformer-base (Beltagy et al., 2020) as PLM in English experiments. The choice is largely based on its context window of 4k tokens, making it a suitable choice for encoding documents compared to other variants with smaller context window such as BERT (Devlin et al., 2019). For Chinese, in the absence of a robust Longformer-equivalent for the language, we use XLM-roberta-base (Conneau et al., 2020). Similar to Yao et al. (2022), input sequences in Chinese longer than the encoder's context window are split into smaller segments using a stride which is half the size of context window. Each segment then gets encoded independently before being merged together for the output projection layers from BIO tagger and biaffine dependency parser. Biaffine layer implementation is based on SuPar[5].

4.3 Results

Table 2 shows overall parsing results on English and Chinese MDP in micro F-score as average across 3 different seeds. Our biaffine approach outperforms the prompt-based model by 1.3% on the development set, along with a modest 0.7% gain on the test set in English. The improvement is more significant with Chinese, with 2.5% increase in both development and test set despite lower tagging score for Events.

4.4 Analysis

It appears that conceiver identification still remains a major bottleneck in English MDP, although it is fundamentally tied to event identification and edge attachment. This is because a conceiver is never a terminal node in MDS; its existence always implies at least one child—another conceiver or, generally speaking, an event. Therefore, detecting conceivers with higher accuracy would always entail balanced improvement across a range of different sub-tasks to reach optimal performance.

The marked increase in Chinese MDP appears to be because of the setup used in Yao et al. (2022), where the context in the prompt-based model for Chinese includes all of the past sentences and 3 sentences after the current event. While this is presumably based on the distribution of arc lengths based on the number of sentences crossed, it greatly increases the context space which makes the problem more difficult than in English, where the context includes 5 sentences before and 5 sentences after. The advantage of our biaffine approach is that it does away with having to define an often arbitrary context window by covering the entire context naturally.

5 Conclusion

This work presents a biaffine modal dependency parser that is simple yet effective. The model is evaluated on English and Chinese datasets and in both instances show improved performance compared to previous work.

[4] https://huggingface.co/docs/transformers
[5] https://github.com/yzhangcs/parser

6 Limitations

MDP experiments remain focused on English and Chinese due to the limited availability of modal dependency annotations in other languages. However, with the adoption of modal dependency structure in Uniform Meaning Representation (UMR) (Van Gysel et al., 2021), more and more annotations for low-resource languages such as Arapaho, Cocama-Cocamilla, Navajo, Sanapaná and potentially additional languages may be prepared and released for future model fitting.

The fact that the overall loss consists of 3 different signals makes the training potentially unbalanced and slow to converge. In future work, we plan to investigate whether tagging could be absorbed as part of arc and label generation, thereby eliminating one of the loss terms at the cost of increased difficulty for the remaining tasks.

References

Xinyi Bai, Nan Yin, Xiang Zhang, Xin Wang, and Zhigang Luo. 2021. Entity-aware biaffine attention fornbsp;constituent parsing. In *Artificial Neural Networks and Machine Learning – ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I*, page 191–203, Berlin, Heidelberg. Springer-Verlag.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Zhousi Chen and Mamoru Komachi. 2023. Discontinuous combinatory constituency parsing. *Transactions of the Association for Computational Linguistics*, 11:267–283.

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.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. *Preprint*, arXiv:1611.01734.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

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.

Zhifu Liu and Nianwen Xue. 2023. A dependency structure annotation for modality in chinese news articles. In *Chinese Lexical Semantics*, pages 143–157, Cham. Springer Nature Switzerland.

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.

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.

Roser Saurí and James Pustejovsky. 2009. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

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.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. 35(3):343–360.

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.

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.

Jiarui Yao, Nianwen Xue, and Bonan Min. 2022. Modal dependency parsing via language model priming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2913–2919, Seattle, United States. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Yu Zhang, Zhenghua Li, and Min Zhang. 2020. Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3295–3305, Online. Association for Computational Linguistics.

A Corpus Details

The publicly available English dataset (Yao et al., 2021) contains newswire annotations from various news media sources (Yao et al., 2022). The Chinese dataset also consists of newswire data from Xinhua news agency.

B Implementation Details

Hyperparameter	English	Chinese
PLM	longformer-base	xlm-roberta-base
PLM Dropout	0.1	0.1
Max. Seq. Len.	4096	512
Chunk Encoding*	False	True
Batch Size	4	1
Grad. Acc. Steps	4	4
Epochs	1,000	1,000
Optim.	AdamW	AdamW
LR	5e-5	5e-5
Weigh Decay	0.01	0.01
Warmup Prop.	0.1	0.1
Arc Hidden Dim.	512	400
Arc Dropout	0.33	0.33
Rel. Hidden Dim.	128	100
Rel. Dropout	0.33	0.33

Table 3: Hyperparameters used in experiments. *Chunk Encoding refers to a document being split into 'chunks' by tokenization with stride, in order to cope with documents longer than PLM encoder's Max Seq. Length. For details, see 4.2.

C Experimental Details

All experiments were run on a single NVIDIA RTX A6000 GPU and each run takes about 6 to 8 hours with the hyperparameters in B. The number of parameters for the English model based on Longformer-base (Beltagy et al., 2020) is 149,386,249; that of Chinese model based on XLM-roberta-base (Conneau et al., 2020) is 278,770,441.