

# Improving Multi-Party Dialogue Discourse Parsing via Domain Integration

Zhengyuan Liu, Nancy F. Chen

Institute for Infocomm Research, A\*STAR, Singapore  
{liu\_zhengyuan, nfychen}@i2r.a-star.edu.sg

## Abstract

While multi-party conversations are often less structured than monologues and documents, they are implicitly organized by semantic level correlations across the interactive turns, and dialogue discourse analysis can be applied to predict the dependency structure and relations between the elementary discourse units, and provide feature-rich structural information for downstream tasks. However, the existing corpora with dialogue discourse annotation are collected from specific domains with limited sample sizes, rendering the performance of data-driven approaches poor on incoming dialogues without any domain adaptation. In this paper, we first introduce a Transformer-based parser, and assess its cross-domain performance. We next adopt three methods to gain domain integration from both data and language modeling perspectives to improve the generalization capability. Empirical results show that the neural parser can benefit from our proposed methods, and performs better on cross-domain dialogue samples.

## 1 Introduction

Text-level discourse parsing is to convert a piece of text into a structured format, by identifying the links and relations between Elementary Discourse Units (EDUs). Incorporating discourse information is proved beneficial for various natural language processing tasks such as machine comprehension (Narasimhan and Barzilay, 2015) and summarization (Xu et al., 2020). Since discourse parsing is involved in capturing and comprehending various semantic and pragmatic phenomena as well as understanding the structural discourse properties, it is quite challenging for machines to conduct automatic processing. There are a series of studies that provide theories and data for developing computational solutions, such as the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) with sentence-level annotation, and the Rhetorical

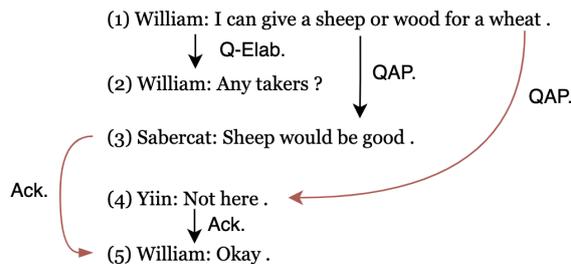


Figure 1: A multi-party dialogue example (Shi and Huang, 2019) with discourse link and relation annotation in the STAC Corpus (Asher et al., 2016). “Ack.” is short for relation “Acknowledgement”, “QAP:” for “Question-Answer-Pair”, and “Q-Elab.” for “Question-Elaboration”. The links in red form a non-projective structure (McDonald et al., 2005).

Structure Theory (RST) (Carlson et al., 2002) with document-level annotation. In RST treebanks, each processed passage is in a hierarchical constituency-based tree structure, and adjacent EDUs are merged to form larger spans<sup>1</sup> recursively (Li et al., 2014a).

Recently, the Segmented Discourse Representation Theory (SDRT) is proposed for multi-party dialogue discourse parsing (Asher and Lascarides, 2005; Asher et al., 2016), which is different from RST whose annotations are on documents. Additionally, SDRT-based annotations contain non-projective links. For example, as shown in Figure 1, a discourse structure will become non-projective when it is impossible to draw the relations on the same side without crossing (McDonald et al., 2005). In this case, the constituency-based structure is not applicable. As a result, the SDRT proposed to transform dialogue discourse trees to a dependency-based structure, where EDUs are directly linked to their precedents without forming upper-level spans.

Since manual parsing is labor-intensive and time-consuming, automatic discourse analysis under the

<sup>1</sup>The merged spans are named as complex discourse units (CDUs) in which multiple EDUs and/or CDUs are grouped together to form a single argument to a discourse relation (Asher et al., 2016)

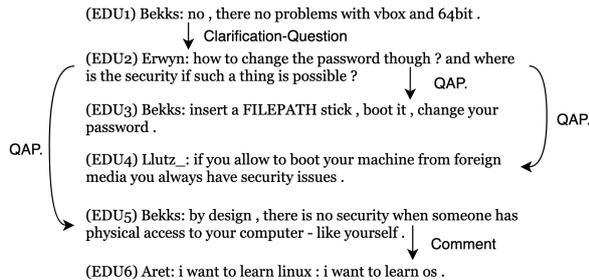


Figure 2: A multi-party dialogue example with its discourse annotation from the Molweni (Li et al., 2020).

	RST-DT	STAC	Molweni
Training Sample Size	347	1091	9000
Test Sample Size	38	100	500
Average EDU Number	56.03	10.95	8.82
Average Word Number	531.8	46.7	96.1
Annotation Scheme	RST	SDRT	SDRT
Relation Number	18	17	17
Data Domain	News	Game	Ubuntu
Conversational Data	No	Yes	Yes

Table 1: Data statistics of training samples from three text-level discourse parsing treebanks.

SDRT theory raises research interest (Badene et al., 2019). Previous models show reasonable results on benchmark treebanks (Shi and Huang, 2019), and utilizing structural information benefits follow-up applications such as dialogue summarization (Feng et al., 2020). However, domain generality is less studied yet important in practical use cases. Existing treebanks only contain limited training data (as shown in Table 1) and limited domain coverage. An SDRT parser trained on strategic game conversations (Asher et al., 2016) may not perform well on technical discussions (Li et al., 2020), and the sub-optimal parsing could further affect downstream task performance. Moreover, due to the annotation complexity, the labeled samples from various domains are not readily available for transfer learning (Yu et al., 2019).

In this paper, we evaluate and improve the cross-domain generality of neural dialogue discourse parsing: (1) we conduct a statistical analysis on existing dialogue discourse treebanks, and figure out the possible factors resulting in the gap across multiple domains from a data perspective; (2) we introduce a Transformer-based neural model for the dependency-based discourse parsing; (3) we propose three methods for better sharing the effective features across dialogue domains: utilizing prior language knowledge, cross-domain pre-training, and vocabulary refinement. Experimental results

RST-DT	1.0	0.48	0.21
STAC	0.10	1.0	0.11
Molweni	0.22	0.52	1.0
	RST-DT	STAC	Molweni

Figure 3: Word level vocabulary overlap of three text-level discourse treebanks. The vocabulary sizes of RST-DT, STAC, and Molweni are 17824, 3642, and 18936, respectively.

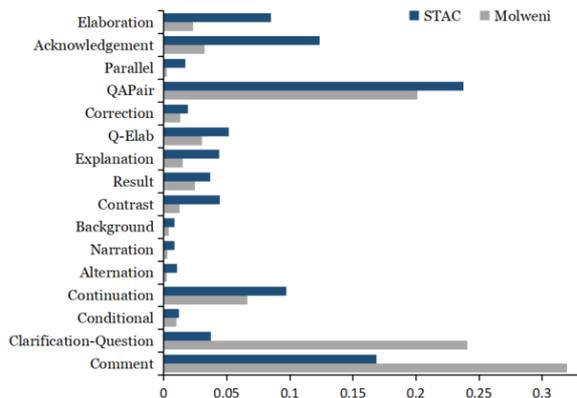


Figure 4: Discourse relation distributions of STAC and Molweni. X axis denotes the label frequency.

on STAC (Asher et al., 2016) and Molweni (Li et al., 2020) show that the parsing performance of single-domain training drops significantly on the out-of-domain samples, and it can be improved by our proposed methods.

## 2 Corpora Analysis

In this section, we conduct a statistical analysis of three text-level discourse treebanks for data-related factors that potentially affect model generality.

**RST Discourse Treebank** (RST-DT) (Carlson et al., 2002) is the first corpus for text-level document parsing, and contains articles from the Wall Street Journal (WSJ). While it is not in the dialogue domain, we include it for an extensive comparison. **STAC** (Asher et al., 2016) is the first corpus for multi-party dialogue discourse parsing, and built on 1.2k strategic conversations where participants take discussion during playing an online game.

**Molweni** (Li et al., 2020) follows the same annotation scheme as STAC, and the data (12k samples) are collected from an online forum, where people discuss technical topics about the Ubuntu system.

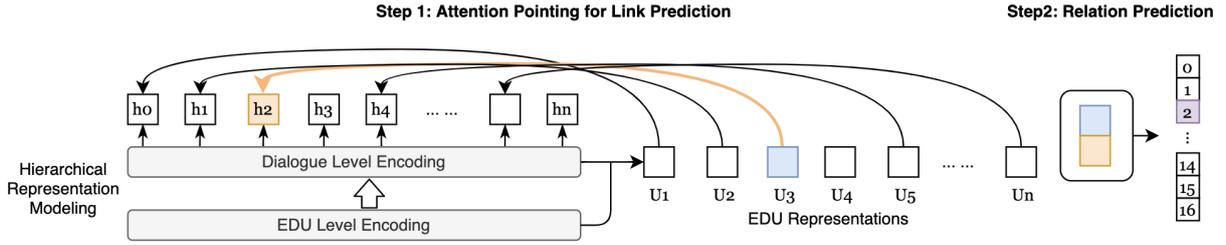


Figure 5: Overview of the dependency-based discourse parsing framework.

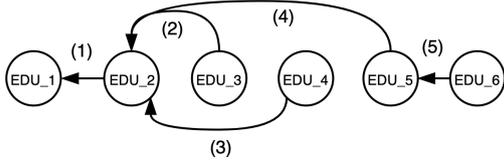


Figure 6: Illustration of parsing process on the dialogue example shown in Figure 2. Numbers in brackets denote the order of link prediction, which is in a sequential manner. This produces a dependency structure.

The data statistics are summarized in Table 1. (1) Compared with Molweni, the RST-DT and STAC have much smaller sample sizes. (2) Samples from RST-DT have a larger EDU number than STAC and Molweni, resulting in deeper parsed tree structures. The tree depth is one of the major factors that affect parsing complexity. (3) Interestingly, while the word number of Molweni is two times larger than that of STAC, no significant difference in their average EDU numbers, resulting in a similar parsing complexity from a depth perspective. (4) The lexical distributions of STAC and Molweni are significantly different sharing a small portion of common vocabulary (Figure 3), as they focus on different conversation scenarios (*Game* vs. *Ubuntu*). (5) Despite the domain distinction between STAC and Molweni, their relation distributions are similar, except that frequencies of the relation (*Clarification-Question* and *Comment*) are quite different, probably because the online technical forums contain more question-clarification and comments (Figure 4). While STAC and Molweni are annotated under the same SDRT theory, their lexical features and relation distributions are different, which we speculate will influence the domain generality.

### 3 Dialogue Discourse Parsing

#### 3.1 Task Definition

Given a dialogue that has been segmented into a sequence of EDUs  $\{u_0, u_1, \dots, u_n\}$  where  $n$  is the

EDU number, the discourse parser is applied to predict links and the corresponding relation types between the EDUs. The predicted structure constitutes a dependency tree, which is a special type of Directed Acyclic Graph (DAG). As in previous work (Shi and Huang, 2019), each EDU is only linked to one of their precedent EDUs, and there are no backward links. As shown in Figure 6, the parsing process can be conducted by a sequential scan of the EDUs. For one EDU  $u_i$ , the model predicts a dependency link by estimating a probability distribution as  $P(u_j|u_i, U_i^{pair})$  where  $0 \leq j < i$  and  $U_i^{pair} = \{(u_l, u_k, r_{l,k}) | 0 \leq l < k < i\}$  is the set of already predicted pairs before the current step  $i$ . The model then determines the relation type based on the predicted link  $P(r_{i,j}|u_i, u_j)$  where  $j < i$  and  $r_{i,j}$  is in the range of  $[0, C]$  ( $C$  is the number of relation types). Following Li et al. (2014b), we add a *root* node as  $u_0$ , and if one EDU is not linked from any preceding nodes, it is pointed to  $u_0$ .

#### 3.2 Transformer-Based Discourse Parser

In this paper, based on the sequential parsing process (Shi and Huang, 2019), we introduce a Transformer-based model for dialogue discourse parsing (as shown in Figure 5), which is comprised of the following components:

**Hierarchical Encoder.** The encoder computes EDU global representations in a hierarchical manner. A Transformer encoder (Vaswani et al., 2017) is used for token-level encoding.<sup>2</sup>

$$H_{token} = \text{TransformerEnc}([t_0, t_1, \dots, t_m]) \quad (1)$$

where  $t$  denotes token, and  $m$  denotes token number. For the  $i$ -th EDU, its local representation  $h_{edu}^i$  is obtained by averaging<sup>3</sup> its corresponding tokens hidden states. Then the local EDU representations

<sup>2</sup>Due to space limitation, refer to (Vaswani et al., 2017) for more details of the Transformer architecture.

<sup>3</sup>We also adopt *first-and-last* sum and *only-first* sum for EDU representation, and the averaging performs best.

are fed to a bi-directional GRU component (Chung et al., 2014) for dialogue-level encoding, and we get final representations  $H'$  with both local and global information.

$$h'_i = [\text{GRU}_{h_{edu}^i}^{Forward}, \text{GRU}_{h_{edu}^i}^{Backward}] \quad (2)$$

**Link Prediction.** An attentive pointer network (Vinyals et al., 2015) is used for the link prediction. For the  $i$ -th EDU, we compute a list of attentive scores with a linear layer between the current node and each candidate  $h'_j$  where  $j < i$ . Then scores are normalized by softmax function to a distribution over the previous EDUs, and we obtain the linked EDU with the largest pointing probability.

$$s_{i,j} = \text{Linear}([h'_i; h'_j]) \quad (3)$$

$$a_{i,j} = \frac{\exp(s_{i,j})}{\sum_{j=0}^i \exp(s_{i,j})} \quad (4)$$

**Relation Classification.** Given one linked pair is  $h'_i$  and  $h'_j$ , we concatenate and feed them to a relation classifier (a linear component):

$$r_{i,j} = \text{Linear}([h'_i; h'_j]) \quad (5)$$

then the output is a probability over the 17 pre-defined discourse relations. For link and relation prediction, the negative log-likelihood is adopted for the loss function.

## 4 Cross-Domain Integration

Based on the corpora analysis in Section 2, to improve the domain-level generality, we investigate three methods to encourage the neural model to utilize the shared linguistic features from different dialogue domains.

**Utilizing Language Backbone.** Large-scale pre-trained language models provide feature-rich contextualized representations (Devlin et al., 2019). In previous work, utilizing prior knowledge can boost the performance in parsing tasks, and also shows some but still limited generalization capability at domain and language level (Liu et al., 2020). Here, we select the ‘RoBERTa-base’ model (Liu et al., 2019) as the language backbone.

**Cross-Domain Pre-training.** Following Gururangan et al. (2020), we conduct the masked language modeling pre-training with the joint data of STAC and Molweni. This can fuse dialogue-related linguistic features to the language backbone, which is not pre-trained on human conversations. Moreover,

Train on Joint Data	Link	Link+Rel.
Deep Sequential Parser (Shi and Huang, 2019)		
Test on STAC	72.8	54.8
Test on Molweni	77.4	54.3
Our Proposed Parser w/ language backbone		
Test on STAC	<b>75.5</b>	<b>57.2</b>
Test on Molweni	<b>80.2</b>	<b>56.9</b>

Table 2: F1 scores of link and relation prediction with models trained on the joint data (STAC+Molweni).

Train on STAC	Link	Link+Rel.
Deep Sequential Parser (Shi and Huang, 2019)		
Test on STAC	73.1	55.7
Test on Molweni	58.6	26.2
Our Transformer-Based Parser		
Test on STAC	73.4	55.5
Test on Molweni	57.8	26.4
+ Utilizing Language Backbone		
Test on STAC	75.3 [2.5% ↑]	56.9 [2.5% ↑]
Test on Molweni	60.7 [5.0% ↑]	31.5 [19.3% ↑]
+ Cross-Domain Pre-training		
Test on STAC	75.1 [2.3% ↑]	57.1 [2.8% ↑]
Test on Molweni	62.1 [7.4% ↑]	32.6 [23.4% ↑]
+ Cross-Domain Vocabulary Refinement		
Test on STAC	75.3 [2.3% ↑]	57.1 [2.8% ↑]
Test on Molweni	63.2 [9.3% ↑]	33.1 [25.3% ↑]

Table 3: Micro-F1 scores of link and relation prediction with models trained on STAC. Values in brackets denote relative increase over the base model.

pre-training with multiple data resources can increase the domain coverage, and this step (parsing annotation is not required) can be conducted before the task-specified learning.

**Cross-Domain Vocabulary Refinement.** STIn Section 2, we observe that the vocabulary overlap between STAC and Molweni is limited (see Figure 3). Dialogues in Molweni contain a certain amount of technical-related words, whereas STAC contains more game-related words. As the model may overfit corpus-specified lexical features, a vocabulary refinement is adopted by filtering out words that are in lower frequency ( $< 20$  occurrence) and not shared by the two datasets.

## 5 Experimental Result and Analysis

### 5.1 Configuration

The proposed models were implemented using PyTorch (Paszke et al., 2019) and Hugging Face<sup>4</sup>. Learning rate was set at  $2e-5$ , and the AdamW (Loshchilov and Hutter, 2019) optimizer was ap-

<sup>4</sup><https://github.com/huggingface/transformers>

Train on Molweni	Link	Link+Rel.
Deep Sequential Parser (Shi and Huang, 2019)		
Test on STAC	42.5	18.3
Test on Molweni	77.9	54.4
Our Transformer-Based Parser		
Test on STAC	42.3	18.0
Test on Molweni	75.9	52.5
+ Utilizing Language Backbone		
Test on STAC	48.3 [14.2% ↑]	26.6 [47.7% ↑]
Test on Molweni	79.7 [5.1% ↑]	55.9 [6.4% ↑]
+ Cross-Domain Pre-training		
Test on STAC	48.8 [15.3% ↑]	28.4 [57.7% ↑]
Test on Molweni	79.6 [5.0% ↑]	55.7 [6.1% ↑]
+ Cross-Domain Vocabulary Refinement		
Test on STAC	50.5 [19.4% ↑]	28.9 [60.6% ↑]
Test on Molweni	79.5 [5.0% ↑]	55.7 [6.1% ↑]

Table 4: Micro-F1 scores of link and relation prediction with models trained on **Molweni**. Values in brackets denote relative increase over the base model.

plied. We trained each model for 20 epochs, and selected the best checkpoints based on evaluation scores. Input dialogue sequences were processed with the sub-word tokenization scheme used in ‘*RoBERTa-base*’ (Liu et al., 2019).

At the inference stage, we adopted the micro-averaged F1 score as the evaluation metric. Results of different settings are shown in Table 2-4. “*Link*” denotes link prediction, and “*Link+Rel.*” stands for a prediction that the dependency link and relation type are correct at the same time.

## 5.2 Joint Domain Evaluation

To compare performance between single-domain and joint-domain training, we obtain the upper bound parsing results on the merged data of two dialogue discourse treebanks (STAC and Molweni). As shown in Table 2, models trained on merged data achieve favorable results on both corpora, and perform slightly better than single-domain training. Moreover, our Transformer-based model with the language backbone outperforms the previous state-of-the-art baseline.

## 5.3 Cross-Domain Evaluation

To evaluate the effectiveness of the proposed domain integration methods, we conduct single-corpus training and cross-corpus evaluation (each treebank represents one dialogue domain).

For single-corpus training on STAC, as shown in Table 3, the cross-domain performance on Molweni data of all models drops significantly, especially the relation prediction. Utilizing language back-

bone brings substantial improvement. This shows that linguistic features can be shared by samples from different treebanks under the SDRT theory. Adopting cross-domain pre-training and vocabulary refinement further improve the performance, and do not affect the original domain. Combining three methods provides the parser a relative 25.3% improvement on the link+relation F1.

For single-corpus training on Molweni, as shown in Table 4, baseline models obtain low link+relation F1 scores (around 18.0) on the STAC corpus. Noteworthy, the performance decrease of *STAC(train)->Molweni(test)* is smaller than that of *Molweni(train)->STAC(test)*, we speculate that this may stem from a larger linguistic diversity in STAC data. The scores are significantly elevated by adopting language backbone, cross-domain pre-training, and vocabulary refinement, achieving a relative 60.6% improvement on link+relation F1.

## 6 Conclusion

In this paper, we investigated the domain-level generality of dialogue discourse parsing. Since existing corpora are collected from different conversation scenarios, models with single-domain training cannot perform well in other domains. The statistical analysis and experimental results suggest that domain adaptation or integration is necessary when neural parsers are applied in practical use cases, and utilizing prior language knowledge and adopting cross-domain pre-training can improve their generality.

## Acknowledgments

This research was supported by funding from the Institute for Infocomm Research (I2R) under A\*STAR ARES, Singapore. We thank Ai Ti Aw for the insightful discussions. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

## References

- Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2721–2727.
- Nicholas Asher and A. Lascarides. 2005. Logics of conversation. In *Studies in natural language processing*. Cambridge University Press.

- Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019. Weak supervision for learning discourse structure. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2296–2305.
- Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.
- Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of NAACL2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization. *arXiv preprint arXiv:2012.03502*.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.
- Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652.
- Jiwei Li, Rumeng Li, and Eduard Hovy. 2014a. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069.
- Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014b. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 25–35.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. **Multilingual neural RST discourse parsing**. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *The International Conference on Learning Representations (ICLR2019)*.
- Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 523–530.
- Karthik Narasimhan and Regina Barzilay. 2015. **Machine comprehension with discourse relations**. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1253–1262, Beijing, China. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of NeurIPS2019*, pages 8026–8037.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.
- Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS2017*, pages 5998–6008.
- Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. **Pointer networks**. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. **Discourse-aware neural extractive text summarization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.
- Chaohui Yu, Jindong Wang, Yiqiang Chen, and Meiyu Huang. 2019. Transfer learning with dynamic adversarial adaptation network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 778–786. IEEE.