Using a Penalty-based Loss Re-estimation Method to Improve Implicit Discourse Relation Classification

Xiao Li† **Yu Hong**†* **Huibin Ruan**† **Zhen Huang**‡
† School of Computer Science and Technology, Soochow University, 1 Shizi, Suzhou, CHN
‡ School of Computer Science, National University of Defense Technology, Changsha, CHN
†{emilyxiao0512, tianxianer, hbr416}@gmail.com; ‡huangzhen@nudt.edu.cn

Abstract

We tackle implicit discourse relation classification, a task of automatically determining semantic relationships between arguments. The attention-worthy words in arguments are crucial clues for classifying the discourse relations. Attention mechanisms have been proven effective in highlighting the attention-worthy words during encoding. However, our survey shows that some inessential words are unintentionally misjudged as the attention-worthy words and, therefore, assigned heavier attention weights than should be. We propose a penalty-based loss re-estimation method to regulate the attention learning process, integrating penalty coefficients into the computation of loss by means of overstability of attention weight distributions. We conduct experiments on the Penn Discourse TreeBank (PDTB) corpus. The test results show that our loss re-estimation method leads to substantial improvements for a variety of attention mechanisms.

1 Introduction

The goal of pairwise sentence-level discourse analysis is to determine the relation that is held by a pair of arguments (Prasad et al., 2008), where an argument generally stands for a narrative sentence. Implicit discourse relation classification is a challenging subtask. It is required to determine the relation on the condition that the explicit connective (i.e., a syntactic conjunction) is not given. For example, the arguments in Figure 1 are a pair of semantically-related arguments, where the possible connective "*but*" that may signal the `comparison` relation has been omitted.

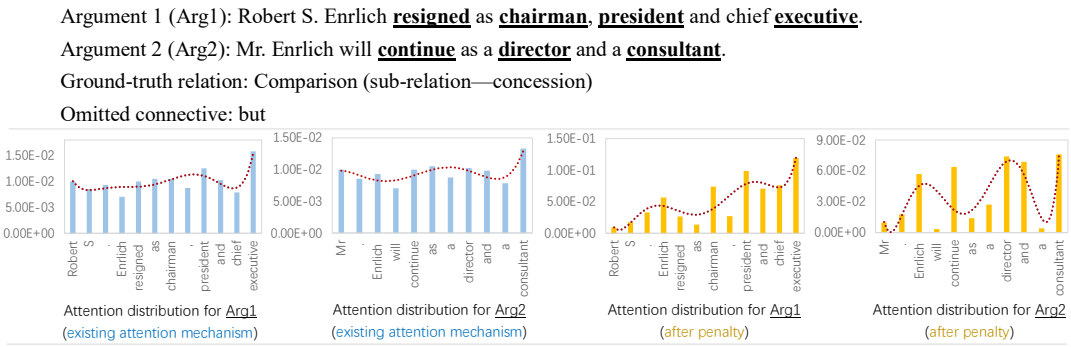

Figure 1: Example of arguments, connective and relationship, as well as attention weight distributions

Detecting the content words that imply correlations between arguments contributes to the relation determination (Marcu and Echihabi, 2002). We refer such kind of content words to attention-worthy words, such as the words shown in bold in Figure 1. The current attention mechanisms have been proven effective in recognizing and utilizing attention-worthy words. They generally assign heavier weights to attention-worthy words conditioned on either internal (Lin et al., 2017) or external context

* Corresponding author: Yu Hong (tianxianer@gmail.com)
This work is licensed under a Creative Commons Attribution 4.0 International License. License details: `http://creativecommons.org/licenses/by/4.0/`.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1513–1518
Barcelona, Spain (Online), December 8-13, 2020

Attention Models	Functions (α)	Parameters (θ)
Self	$\alpha = softmax(W_{s2}tanh(W_{s1}[H_1, H_2]^T)$	W_{s1}, W_{s2}
Interactive (for H_1)	$\alpha = softmax(tanh(H_1 W_{\eta_1} \Phi([\overline{H_2}, cls]) + b_{\eta_1}))$	W_{η_1}, b_{η_1}
Interactive (for H_2)	$\alpha = softmax(tanh(H_2 W_{\eta_2} \Phi([\overline{H_1}, cls]) + b_{\eta_2}))$	W_{η_2}, b_{η_2}
Multi-layer	$\alpha = softmax(W_{s,k}tanh(W_{a,k}H1 + W_{b,k}(M_k \otimes e)))$	$W_{s,k}, W_{a,k}$
	$M_k = tanh(W_{m,k}[R^1_{k-1}, R^2_{k-1}, R^1_{k-1} - R^2_{k-1}, M_{k-1}])$	$W_{b,k}, W_{m,k}$

Table 1: The equations of different attention mechanisms as well as parameters (where, Φ denotes the non-linear transformation and cls is a special classification token in BERT, $k(k = 1, 2, 3)$ is the attention-layer number, the memory vector M is used to preserve the information of previous layer, $M_k \otimes e$ is the operation that repeatedly expands the dimensions of encoder states).

(Ma et al., 2017). Benefiting from the positive effects of the heavily weighted attention-worthy words on representation learning, the existing attention-based neural networks obtains considerable performance gains for discourse relation classification.

However, our survey shows that some of inessential words are highlighted with heavier weights by the attention mechanisms. As a result, the attention weight distributions fall into the over-smooth transition state (as shown in Figure 1). This makes it difficult to sensitively perceive the effects of attention-worthy words or even misleads the encoder during encoding. To solve the problem, we propose to estimate attention-oriented penalty coefficients by means of overstability of attention weight distributions. On the basis, we integrate the penalty coefficients into the loss measurement process (Section 2), so as to optimize the parameters of attention mechanisms by backward propagation of penalty. Briefly, we aim to use penalty coefficients to obtain distinguishable attention weights. In Figure 1, we show the jagged attention weight distributions obtained after using our penalty coefficients. We carry out experiments on PDTB v2.0 (Prasad et al., 2008), a corpus that comprises a large-scale pairwise argument instances, along with pre-annotated implicit relation tags. The test results show that our method substantially improves the attention-based discourse relation classification (Section 3).

2 Approach

We utilize BERT (Devlin et al., 2019) as the baseline encoder, and connect it with a multi-layer perceptron (MLP) to form the discourse relation classifier. In addition, we reproduce three attention mechanisms, including self (Lin et al., 2017), interactive (Ma et al., 2017) and multi-layer (Liu and Li, 2016) attention mechanisms. On the basis, we couple them with the baseline (BERT) encoder. Similarly, they are also connected with a MLP respectively for discourse relation classification. It is noteworthy that the attention mechanisms mentioned above have been carefully studied on PDTB v2.0. Though, they were built over some slightly weak word embeddings. For fair comparison, we choose to couple them with the pre-trained BERT encoder. BERT is fine-tuned in all of our experiments.

Assume that a certain neural attention mechanism is defined as $\mathcal{F}(H_1, H_2, \theta)$, we tend to optimize θ. In the equation, H_1 and H_2 denotes the encoder states of a pair of arguments which are obtained using the pre-trained BERT, and θ stands for the shorthand of all parameters of the attention mechanism. We specify the parameters of all the considered attention-based computational models in Table 1. We optimize θ by re-estimating the loss $J(\theta)$ using penalty coefficients. The penalty coefficients are measured by means of mean deviations among different attention weights. Backward propagation is used as usual to tune the parameters in θ conditioned on the re-estimated loss. The loss $J(\theta)$ is calculated as follows:

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(y^{(i)}, \hat{y}^{(i)}) - \chi(\theta); \qquad L(y^{(i)}, \hat{y}^{(i)}) = -\sum_{j=1}^{C} P(y_j^{(i)}) \log(P(\hat{y}_j^{(i)})) \tag{1}$$

$$\chi(\theta) = \lambda[(\sigma_1)^2 + (\sigma_2)^2] \tag{2}$$

where, C denotes the relation class number, P($\hat{y}_j^{(i)}$) stands for the probability that the relation class is predicted as the i-th class, σ_1 denotes the scalar deviation value that is calculated over the attention

Systems	COM	CON	EXP	TEM
Bert (Baseline) (Devlin et al., 2019)	45.67	56.46	73.84	37.01
+ Self (Lin et al., 2017)	46.66	56.75	73.40	38.94
+ Self + Penalty	**50.45**	**58.27**	**75.45**	**39.02**
+ Interactive (Ma et al., 2017)	48.11	57.02	74.66	38.84
+ Interactive (Ours)	48.85	57.96	74.90	39.23
+ Interactive (Ours) + Penalty	**49.24**	**58.46**	**75.41**	**40.40**
+ Multi-layer (two layers) (Liu and Li, 2016)	47.02	57.97	74.96	39.61
+ Multi-layer (two layers) + Penalty	49.87	58.38	75.33	41.25
+ Multi-layer (three layers) (Liu and Li, 2016)	48.47	58.09	74.27	38.70
+ Multi-layer (three layers) + Penalty	**50.11**	**58.77**	**76.26**	**43.26**

Table 2: Test results for different attention mechanisms which are coupled with our penalty mechanism.

Model	COM	CON	EXP	TEM
Bert (Baseline) (Devlin et al., 2019)	45.67	56.46	73.84	37.01
+ Self (Lin et al., 2017)	46.66	56.75	73.40	38.94
+ Self + Interactive (ours)	49.40	58.48	75.37	39.77
+ Self + Interactive (ours) + Penalty	**50.91**	**58.88**	**76.35**	**43.51**

Table 3: The test results for the combination of attention mechanisms and shareable penalty mechanism.

weights for H_1, while σ_2 is calculated for H_2, and λ denotes a hyperparameter which is separately set for different attention mechanisms (section 3). It can be preconceived that a smaller deviation will lead to a relatively larger loss. In other words, if the attention weight distribution is smooth (corresponding to a small deviation), the loss will relatively be increased, and as a result, the attention parameters θ will be dramatically changed by backward propagation.

3 Experimentation

We evaluated our model on the benchmark PDTB v2.0 (Prasad et al., 2008). The four main relation classes are considered in the experiments, including Comparison (abbr., COM), Contingency (CON), Expansion (EXP) and Temporality (TEM). We follow the previous work (Ji and Eisenstein, 2015) to split datasets, using sections 02-20 as the training set, sections 00-01 the development set and sections 21-22 the test set. For comparison purpose, we use F1 as the evaluation metric.

We utilize BERT that outputs word embeddings with the hidden size of 768. There are 12 self-attention heads considered in BERT. The max length of the input sequence is set to 163, in which the maximum length N of an argument is set to 80. In addition, the batch size is set to 15, and gradient descent is set separately: β_1=0.9 and β_2=0.999. The learning rate is set to 5e-5 and dropout rate is set to 0.1. We set the hyperparameter λ to 1e-3 for both the multi-layer and self-attention mechanisms, and 1e-2 for the interactive attention mechanism.

The main test results are shown in Table 2, in which various attention mechanisms are coupled with our penalty-based loss re-estimation model (penalty mechanism for short). Note that the label of "Interactive (Ours)" denotes the reproduced interactive attention mechanism which introduces the special classification token "*cls*" (see Table 1) into the encoder state. It can be observed that the proposed penalty mechanism yields substantial improvements for every attention model. Besides, we combine the self and interactive attention mechanisms and utilize a shareable penalty mechanism to improve them. The performance is shown in Table 3. It can be found that the F1 scores obtained for all the four relation classes are increased further. It proves that our penalty mechanism is capable of producing shareable penalty coefficients for different attention models.

We compare our method to the state-of-the-art. As shown in Table 4, our best model (i.e., the combined attention models coupled with the shareable penalty mechanisms) outperforms the previous work for the comparison (COM) and expansion (EXP) relations. In addition, it achieves comparable performance to

Model	COM	CON	EXP	TEM
Zhang et al (2015)	33.22	52.04	69.59	30.54
Qin et al (2016)	41.55	57.32	71.50	35.43
Liu and Li (2016)	36.70	54.48	70.43	38.84
Qin et al (2017)	40.87	54.56	72.38	36.20
Lan et al (2017)	40.73	**58.96**	72.47	38.50
Dai and Huang (2018)	46.79	57.09	70.41	**45.61**
Guo et al (2018)	40.35	56.81	72.11	38.65
Bai and Zhao (2018)	47.85	54.47	70.60	36.87
Nguyen et al (2019)	48.44	56.84	73.66	38.60
He et al (2020)	47.98	55.62	69.37	38.94
Our best	**50.91**	58.88	**76.35**	43.51

Table 4: Comparison to the State-of-the-art approaches.

that of Lan et al. (2017)'s work for the contingency (CON) relation, which was being at the top of the list for years. For the temporality (TEM) relation, our method results in less severe performance reduction when it improves the performance for other three relation classes.

4 Related work

Recently, neural networks have been widely studied for argument representation learning (Zhang et al., 2015), which is admitted to be the crucial issue for discourse relation recognition. Due to the capacity of generating low-dimensional continuous representations for arguments, RNNs with Bi-LSTM are used during encoding. Chen et al (2016) couple Bi-LSTM with a gated relevance model. Liu and Li (2016) use multi-layer attention computation over the output of Bi-LSTM. Meanwhile, Liu et al (2016) build a multi-task learning framework with Convolutional Neural Network (CNN) for argument encoding. By contrast, Lan et al (2017) integrate Bi-LSTM into the multi-task framework and couple it with the attention mechanism. Guo et al (2018) utilize the interaction mechanism to weight the representations emitted by Bi-LSTM, and perform a deeper encoding by tensor network. Dai and Huang (2018) use Bi-LSTM to bring paragraph-level contextual information into argument representations.

In addition, Qin et al (2016) build a hybrid neural model which couples two gated CNNs to extract both word-level and semantic-level convolutional features. Further, Qin et al (2017) integrate generative adversarial networks into multi-task learning network. Hereafter, Bai and Zhao (2018) establish multi-task network using multi-layer gated CNNs. The network is additionally coupled with residual networks and interactive attention mechanisms. Nguyen et al (2019) enhance Bai and Zhao (2018)'s multi-layer CNNs-based multi-task learning by minimizing the divergence between connective-level embeddings and relation-level embeddings. He et al (2020) develop a joint learning architecture which updates both geometric and semantic features during encoding.

5 Conclusion

Our experiments demonstrate that the utilization of penalty coefficients for loss re-estimation can effectively strengthen the attention-based implicit discourse relation classification. Nevertheless, our survey shows that some attention-worthy words fails to be effectively perceived by the current attention mechanisms. More importantly, the semantics of such kind of attention-worthy words can be well-encoded only through the understanding of related common sense. Therefore, in the future, we will utilize commonsense knowledge graph to enhance the attention modeling method.

Acknowledgements

We are grateful for the insightful comments of reviewers. This work is supported by the national NSF of China via Grant Nos. 62076174, 61672368, 61751206 and 61672367, as well as the Stability Support Program of National Defense Key Laboratory of Science and Technology via Grant No. 61421100407.

References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 571–583. Association for Computational Linguistics.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Berlin, Germany*, pages 1726–1735.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA*, pages 141–151.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA*, pages 547–558.

Ruifang He, Jian Wang, Fengyu Guo, and Yugui Han. 2020. Transs-driven joint learning architecture for implicit discourse relation recognition. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 139–148. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Trans. Assoc. Comput. Linguistics*, 3:329–344.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark*, pages 1299–1308.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France*.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1224–1233. The Association for Computational Linguistics.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI 2016, Phoenix, Arizona, USA*, pages 2750–2756.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia*, pages 4068–4074.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 368–375. ACL.

Linh The Nguyen, Ngo Van Linh, Khoat Than, and Thien Huu Nguyen. 2019. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4201–4207. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA*, pages 2263–2270.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P. Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada*, pages 1006–1017.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal*, pages 2230–2235.