An Updated Baseline System for DTC Parsing

Longyin Zhang, et al.*
zzlynx@outlook.com

Abstract

This article serves as an extension of our previous paper of *EDTC: A Corpus for Discourse-Level Topic Chain Parsing*. In our previous work, we built a corpus about discourse-level topic chain (DTC) structures and provided a bert-based baseline system. In this article, we provide a more proper baseline system that performs much better than the published one.

1 Problem Description

As the project of `ori_parser` presents, we package each sentence as a Sentence object with varied class attributions. When building the object, we accidentally took each raw annotation line as the textual content of a sentence, for example, "0 <TAB> 2 <TAB> Good grief! Charlie Brown is selling out". Under this circumstance, our baseline system took these annotated prepositive tags as powerful parsing clues, leading to its high performance in DTC parsing. Unfortunately, we did not find this technical issue until *November 22 2021* when we tried to apply the pre-trained DTC parser to the data of downstream NLP applications. In order to avoid misleading other researchers, we contribute a more powerful baseline parser in this article which performs much better than the previous one.

2 DTC Parser

This section introduces the updated baseline system. Subsection 2.1 presents the basic encoder-decoder structure. In Subsection 2.2, we integrate a dynamic information transmission process into the encoder-decoder for representation enhancing.

2.1 Encoder-Decoder Architecture

Encoding. Our corpus annotation reveals a common phenomenon that some topic chains are driven by words that are specifically related to the topic in arguments. Inspired by this, we concatenate the word embeddings and POS tag embeddings as input sequence and employ a well-behaved attention mechanism to control how much emphasis should be put on each word unit in the argument. Concretely, we use a learnable vector r to compute the weight of each word, as shown in Equation 1. After that, a bi-directional GRU (Cho et al., 2014) is used to encode the weighted input sequence as

$$k_i = \frac{r^{\mathrm{T}} w_i}{\sum r^{\mathrm{T}} w_j} \quad (1)$$

$$h_a = \mathrm{BiGRU}(w\alpha, \theta) \quad (2)$$

where α is generated by a softmax function with $k_1, \ldots, k_m$ as inputs, $w\alpha = (w_1\alpha_1, ..., w_m\alpha_m)$ is the weighted sequence, and h_a denotes the concatenation of last hidden states of the BiGRU model in both directions which serves as the representation of the argument.

Some sentences in news articles serve as *narratage* which are not argumentative and do not belong to any topic chains. Therefore, we add a fake root, $\boldsymbol{h_r}$, at the end of the argument sequence for the tails of topic chains or those narratage-like sentences to point to, obtaining $\tilde{h} = (h_1, \ldots, h_n, \boldsymbol{h_r})$. As Figure 1 shows, we input the discrete argument representations $\tilde{h}$ into another BiGRU model to generate context-aware representations, obtaining $H = (H_1, \ldots, H_{n+1})$. Moreover, the last hidden states of the BiGRU model in both directions are concatenated into d_0 for the decoder below.

Decoding. During decoding, the uni-directional GRU decoder receives as inputs the previous argument representations, $h_1, \ldots, h_n$, with d_0 as its initial state, obtaining $D = (d_1, ..., d_n)$. In this way, the encoder-decoder structure is established, and the topic chain determination is then achieved through the pointer mechanism between the decoder outputs D and the encoder outputs H.

In the pointing phase, we compute the bilinear attention scores between **observers** and encoder

*Completed in December 18, 2021.

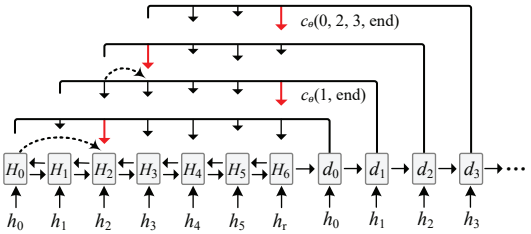

Figure 1: Architecture of the DTCP-DIT model. The arcs denote the information transmission process.

outputs corresponding to candidate successors, i.e., $H' = \{H_i | t < i \leq n+1\}$. Here, the observers at the t-th step refer to the decoder output d_t and the encoder output H_t. Following (Dozat and Manning, 2017), we apply two MLPs before the bilinear function to strip away irrelevant information. The attention score is calculated as follows

$$H'' = \text{mlp}(H') \quad (3)$$

$$o = \text{mlp}^*(d_t \oplus H_t) \quad (4)$$

$$\text{attn} = H''\mathbf{U}o \quad (5)$$

where $H'' \in \mathbb{R}^{C \times D_1}$, $o \in \mathbb{R}^{D_2}$, C is the number of candidate successors, $\oplus$ denotes concatenation, $\mathbf{U} \in \mathbb{R}^{D_1 \times D_2}$ denotes the weights of the bilinear term, and $\text{attn} \in \mathbb{R}^{C \times 1}$ refers to attention weights assigned to the candidate topic successors.

2.2 Cumulative Information Transmission

Although previously obtained argument representation is well incorporated with document context, the global context with various information could be redundant and cryptic for topic chain parsing. To tackle this problem, we propose a task-oriented method of enhancing argument representation with latent topic-dependent features. Specifically, in Figure 1 of our conference paper (Zhang et al., 2021), we aim to transmit the lexical information "coronavirus" in the first argument to the second to enhance the representation of its cryptical pronoun TO "it". Theoretically, the enhanced representation provides the second argument with more hints to determine its topical successor further. To achieve this, we integrate a dynamic information transmission process into the proposed encoder-decoder structure to well harness established local topic chains for cumulative topic information transmission, as shown in Figure 1.

Formally, given two topic-related arguments detected at the t-th time step, our goal is to transmit the topic-focused information in $\mathbf{arg}_t$ to its successor $\mathbf{arg}_m$ to enhance its representation as

$$\lambda = \text{F}_{sgm}(W_t H_t + W_m H_m + b) \quad (6)$$

$$H^*_m = H_m + \text{drop}(\lambda H_t) \quad (7)$$

where $\text{F}_{sgm}(\cdot)$ denotes the sigmoid function, H_t and H_m are hidden representations of the two arguments, and λ is a single scalar. With the lexical tradeoff among words and the contextual tradeoff between the two topic-related arguments, redundant information is well filtered by the lexical attention and the information gate λ, while useful topic-focused information directly flows to the enhanced representation H^*_m. Finally, we substitute the original representation H_m with the enhanced one for further steps.

Up to this time, the overall structure of our DTCP-DIT model permits continuously parsing articles into discourse topic chains. We implement our dynamic information transmission process in a batch transferring mode, which can ensure that the parser performs dynamic representation learning in parallel. To optimize the built model, we employ the NLL-loss in this work to maximize the probability of determining the correct topical successor at each decoding step.

$$\mathcal{L}(\Theta) = \sum_m \sum_n -\log(\hat{p}_{j,i}|\Theta) + \frac{\lambda}{2}\|\Theta\|_2^2 \quad (8)$$

where m is the batch size, n is the number of decoding steps for each document, and $\hat{p}_{j,i}$ estimates the conditional probability of selecting the correct successor argument i at the j-th time step.

3 Experimentation

We use the same evaluation metrics as Zhang et al. (2021) and report the results of the previous baseline system and the updated one.

3.1 System Settings

We experimented on the 300D GloVe (Pennington et al., 2014) and the pre-trained XLNet-large model (Yang et al., 2019). The weights of the language model were updated during training. We employed the Stanford CoreNLP toolkit (Manning et al., 2014) to obtain POS tags, and the POS embeddings were randomly initialized and then optimized during training. Notably, when using XLNet, we directly use the vector corresponding to `[cls]` as the sentence representation and omit the POS

Parameter	GloVe	XLNet
POS Embedding	32	-
BiGRU Hidden	64	192
GRU Hidden	128	384
Dropout Rate	0.33	0.33
Batch Size (Doc)	1	1
Epoch	20	20
Learning Rate	1e-3	1e-5

Table 1: Fine-tuned hyper-parameters.

Methods	Link	Chain
Zhang et al. (2021)*	32.3	16.1
Ours (GloVe)	61.5	46.1
Ours (XLNet)	63.9	50.9

Table 2: Overall performance. "*": we re-ran the system with the annotation tags omitted during bert encoding. For (Zhang et al., 2021), we experiment on bert-large.

embeddings and the lexical attention calculation process. We trained the proposed parser iteratively on the training corpus and then used the development corpus to fine-tune the hyper-parameters in Table 1. The codes and our pre-trained parser are shown in `upd_parser`.

3.2 Results

We present the replicated results of the previously introduced bert-based parser (Zhang et al., 2021) and the system introduced in this article in Table 2. From the results, we find that **(i)** Without the clues hidden within the annotation tags, the original system performs extremely weak with only 32.3 F1 in topic linking and 16.1 F1 in topic chaining. Given this, the DTC parsing task is much more challenging than we thought. **(ii)** Compared with the original system, due to the effective dynamic information transmission during parsing, our updated system achieves significantly better results in both topic linking and topic chaining. **(iii)** Comparing the results of our system over different language models, although the use of pre-trained language models can boost the performance to some extent, the performance is far from perfect.

4 Conclusion

In this article, we reported a technical issue in our previously proposed baseline system on DTC parsing. Moreover, we introduced a more powerful DTC parser where lexical and contextual topic information is exploited and transmitted along constructed DTC structures for better parsing performance. Our experiments showed that the DTC parsing task is still full of challenges that deserve further exploration.

References

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of ICLR 2017*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Longyin Zhang, Xin Tan, Fang Kong, and Guodong Zhou. 2021. EDTC: A corpus for discourse-level topic chain parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1304–1312, Punta Cana, Dominican Republic. Association for Computational Linguistics.