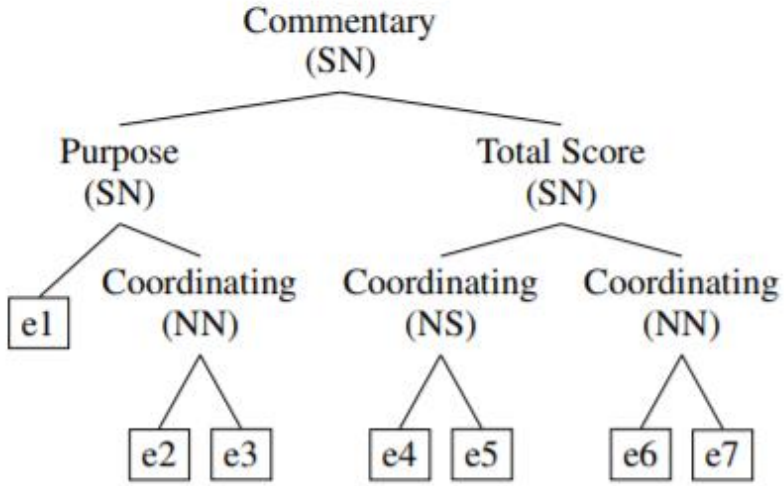
A Top-Down Neural Architecture towards Text-Level Parsing of Discourse Rhetorical Structure

Longyin Zhang[1,2], **Yuqing Xing**[1,2], **Fang Kong**[1,2]*, **Peifeng Li**[1,2], **Guodong Zhou**[1,2]
1. Institute of Artificial Intelligence, Soochow University, China
2. School of Computer Science and Technology, Soochow University, China
{lyzhang9,20184227019}@stu.suda.edu.cn
{kongfang,pfli,gdzhou}@suda.edu.cn



e1: 西藏银行部门积极调整信贷结构，/ Bank of Tibetan actively readjusts credit structure
e2: 以确保农牧业生产等重点产业的投入，/ Ensuring the investment of key industries such as husbandry production
e3: 加大对工业、能源、交通、通信等建设的正常资金供应量。/ Increase the normal supply of funds for industrial, energy, transportation, communications
e4: 去年新增贷款十四点四一亿元，/ Last year, the newly increased loan was 1.441 billion yuan
e5: 比上年增加八亿多元。/ an increase of more than 800 million yuan compared to the previous year.
e6: 农牧业生产贷款（包括扶贫贷款）比上年新增四点三八亿元；/ The loans (including aid the poor loan) for agricultural and livestock production newly increased by 438 million yuan compared to the previous year
e7: 乡镇企业贷款增幅为百分之六十一点八三。/ The increase in loans to township enterprises was 61.83%

Figure 1: An example for DRS parsing, where the text consists of 3 sentences containing 7 EDUs.

That is, adjacent EDUs are recursively combined into high-level larger text spans by rhetorical relations to form a final discourse tree in a bottom-up way. In this paper, we justify that compared with a bottom-up approach, a top-down approach may be more suitable for text-level DRS parsing from two points-of-view,

- From the computational view, only local information (i.e., the constructed DRS subtrees and their context) can be naturally employed to determine the upper layer structure in the bottom-up fashion. Due to the overwhelming ambiguities at the discourse level, global information, such as the macro topic or structure of the discourse, should be well exploited to restrict the final DRS, so as to play its important role. From the computational view, a top-down approach can make better use of global information.

- From the perceptive view, when people read an article or prepare a manuscript, they normally go from coarse to fine, from general to specific. That is, people tend to first have a general sense of the theme of the article, and then go deep to understand the details. Normally, the organization of the article is much limited by its theme. For text-level DRS parsing, a top-down approach can better grasp the overall DRS of a text and conform to the human perception process.

Previous work

- Bottom-Up: Hernault et al., 2010; Joty et al., 2013; Feng and Hirst, 2014; Ji and Eisenstein, 2014; Heilman and Sagae, 2015; Li et al., 2016; Braud et al., 2017; Yu et al., 2018.

- Top-Down, Sentence level: Lin et al. (2019) and Liu et al. (2019)

- Statistics on the RST-DT corpus show each sentence only contains 2.5 EDUs on average while each document contains 55.6 EDUs on average.

EDU Encoder

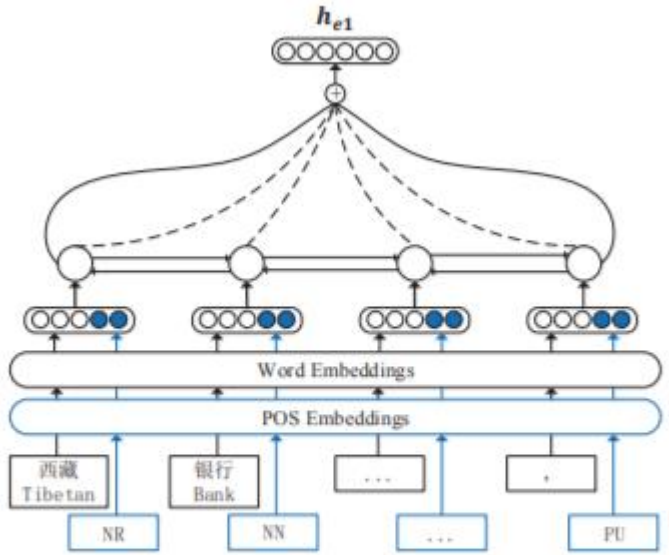


Figure 2: Architecture of the EDU encoder.

$$w_i = \frac{q^T h_i}{\sum q^T h_j} \tag{1}$$

In this way, we can achieve the encoding h_{ek} of the kth EDU in given discourse D.

$$h_{ek} = \begin{bmatrix} h_{\vec{s}} \\ h_{\overleftarrow{s}} \end{bmatrix} + \sum w_i h_i \tag{2}$$

Split Point Encoder

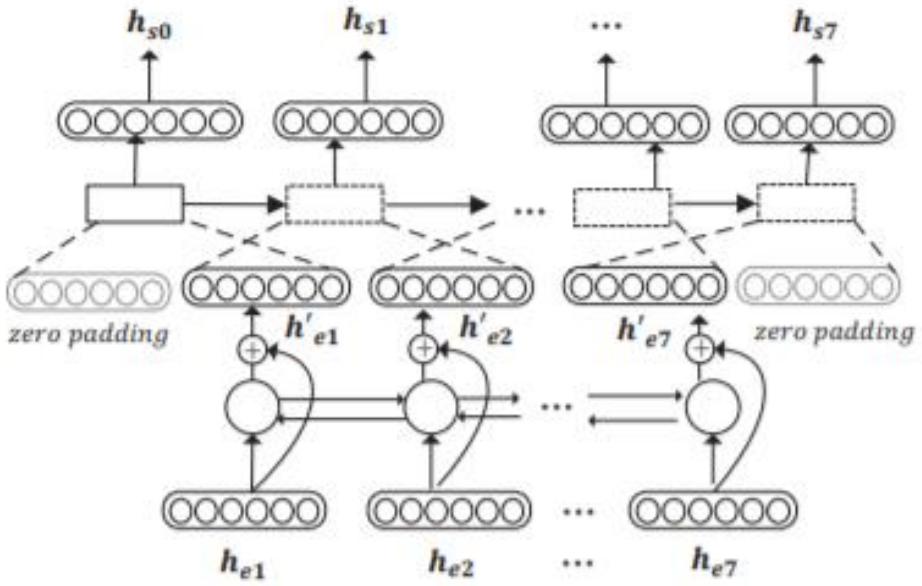


Figure 3: Architecture of the split point encoder.

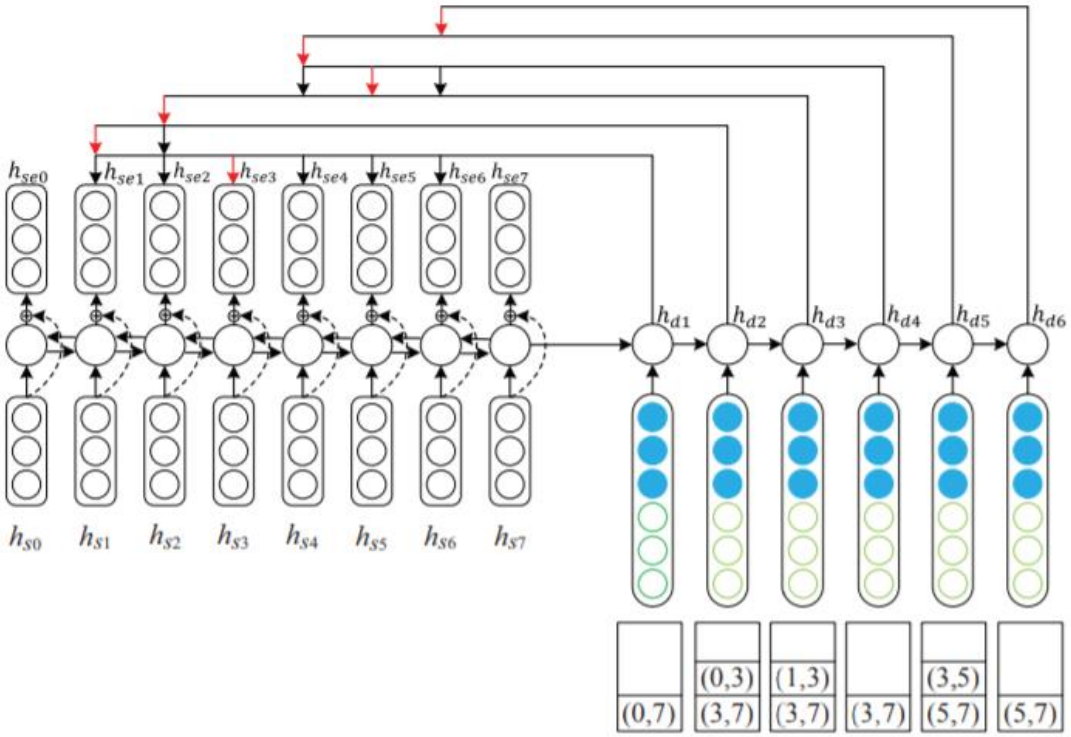
In this paper, we call the split position between any two EDUs the split point. A discourse containing n EDUs has $n-1$ split points. For example, Figure 1 contains 7 EDUs and 6 split points. The split point encoder is responsible for encoding each split point. In our model, we use the both EDUs on the left and right sides of the split point to compute the split point representation.

- GRU
- zero padding
- CNN(convolution kernel is set to 2)

Attention-based Encoder-Decoder on Split Point Ranking

Initially, the stack contains only one element, i.e., the index pair of the first and the last split points of the complete discourse $(0, N)$. At each decoding step, the index pair of the boundary split points is first popped from the top of the stack. Suppose the index pair is (l, r) at the jth step. Then, the encoding output h_{sel} and h_{ser} are concatenated to form the input of the decoder. While the decoder output at the jth step represented by h_{dj}. After that, we adopt the Biaffine Attention mechanism to the encoder output corresponding to the split points between the boundary split points (i.e., $h_{sem}, \forall m, l \leq m \leq r$) and the decoder output h_{dj}. Finally, the split point with the largest score is selected as the final result of this time. If there are still unselected split points for the new text spans formed by this decision, they are pushed onto the stack for following steps.

Attention-based Encoder-Decoder on Split Point Ranking



complete text span at the very beginning, and we feed the concatenated vector $[h_{e0}; h_{e7}]$ into the decoder to achieve the output h_{d1}. Then, the weight is computed using h_{d1} and the results of the encoder corresponding to the 6 split points between the number 0 and the number 7, i.e., $h_{se1} \ldots h_{se6}$. In this example, since the split point 3 has the largest weight, the text span is split into two parts, i.e., $(0, 3)$ and $(3, 7)$. Because there are still unselected split points in the text span $(0, 3)$ and $(3, 7)$, we push them onto the stack. In this way, we get one split point at each step. After six iterations, the discourse tree is built.

Figure 4: A parsing example of the attention-based encoder-decoder.

Previous work

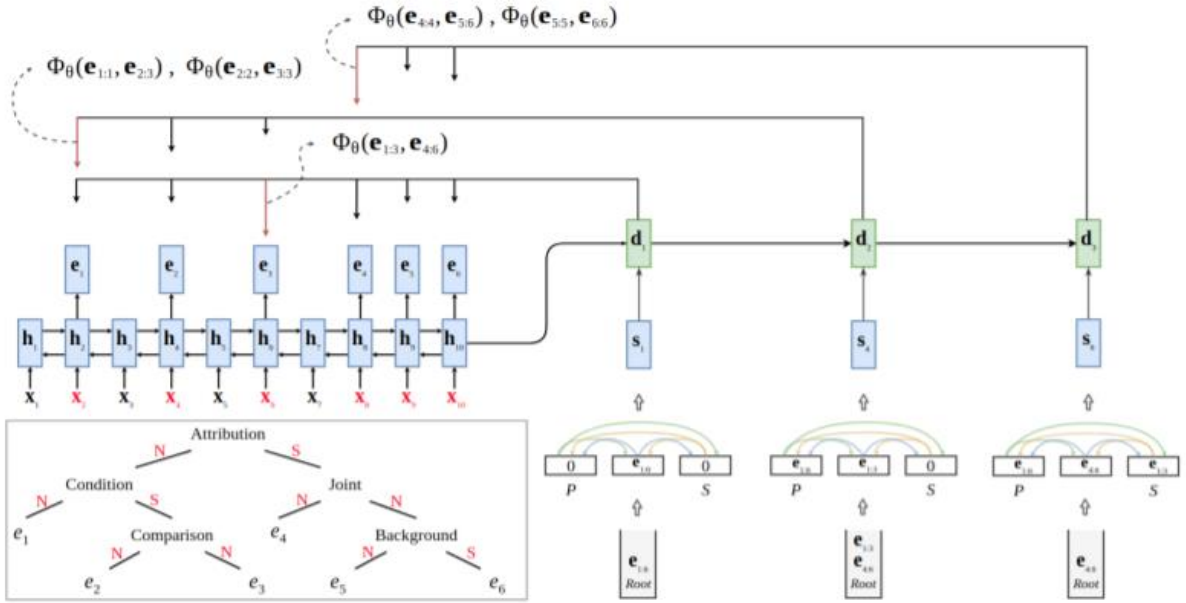


Figure 2: Our discourse parser along with the decoding process for a synthetic sentence with 10 words and 6 EDUs. EDU boundaries are marked in red color. For the inputs to the decoder at each step, P and S indicate the parent and sibling representations, respectively. $\Phi_\theta(\boldsymbol{e}_{i:k}, \boldsymbol{e}_{k+1:j})$ denotes the relation-nuclearity classifier employed by the parser to find the nuclearity and relation labels for the newly created spans, $\boldsymbol{e}_{i:k}$ and $\boldsymbol{e}_{k+1:j}$.

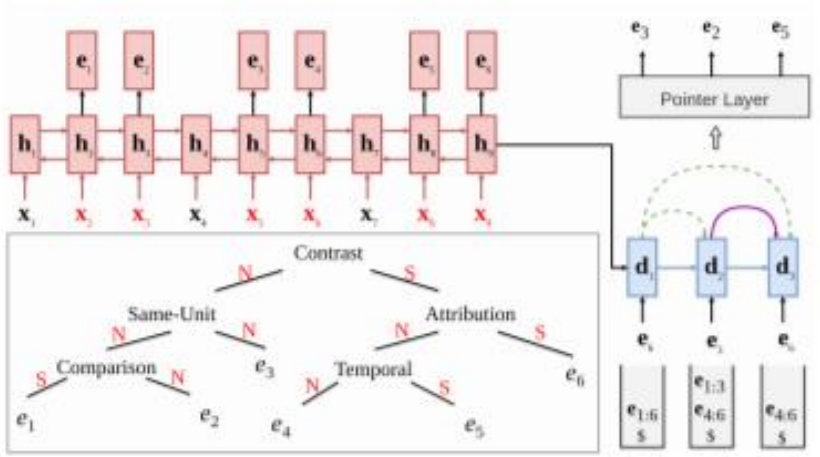


Figure 3: H-PtrNet for discourse parsing. The input symbols in red $(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_8, \mathbf{x}_9)$ represent last words of the respective EDUs. To avoid visual clutter, we do not show the attention distributions over the EDUs, rather only show the decisions. Green dash lines indicate parent connections and purple solid lines denote the immediate sibling connections.

Biaffine Attention on Text-level DRS Parsing

$$s_j^i = {h'_{sei}}^T W h'_{dj} + U h'_{sei} + V h'_{dj} + b;$$
$$W \in \mathbb{R}^{m \times k \times n}, U \in \mathbb{R}^{k \times m}, V \in \mathbb{R}^{k \times n}, s_j^i \in \mathbb{R}^k \quad (3)$$

where W, U, V, b are parameters, denoting the weight matrix of the bi-linear term, the two weight vectors of the linear terms, and the bias vector, respectively, s_j^i means the score of the ith split point over different categories, and the k denotes the number of categories (for split point determination, $k = 1$; for nuclearity determination, $k = 3$; for discourse relation classification, $k = 18$ in English and $k = 16$ in Chinese). In this way, we can determine the split point, nuclearity and discourse relation jointly.

Model Training

$$L_s = \sum_{batch} \sum_{steps} -\log(\hat{p}_i^s|\theta)$$

$$\hat{p}_i^s = \frac{s_{i,j}^{split}}{\sum s_i^{split}}$$

$$L = \alpha_s L_s + \alpha_n L_n + \alpha_r L_r$$

Experimentation

	Systems	Bare	Nuc	Rel	Full
EN	Top-down(Ours)	67.2	55.5	45.3	44.3
	Ji&Eisenstein(2014)$^{+}$	64.1	54.2	**46.8**	**46.3**
	Feng&Hirst(2014)$^{+}$	**68.6**	**55.9**	45.8	44.6
	Li et al.(2016)$^{+}$	64.5	54.0	38.1	36.6
	Braud et al.(2016)	59.5	47.2	34.7	34.3
	Braud et al.(2017)*	62.7	54.5	45.5	45.1
CN	Top-down(Ours)	**85.2**	**57.3**	**53.3**	45.7
	Sun&Kong(2018)(Dup)	84.8	55.8	52.1	**47.7**

Table 2: Performance Comparison.(Bare, bare DRS generation. Nuc, nuclearity determination. Rel, rhetorical relation classification. Full, full discourse parsing. The sign $^{+}$ means the systems with additional handcrafted features including syntactic, contextual and so on, * means with additional cross-lingual features.)

[2]We evaluate the discourse parsers proposed by Lin et al. (2019) and Liu et al. (2019) in text-level discourse parsing. However, their achieved performances are much lower than the state-of-the-art systems. The main reason is that their proposed encoders are tailored to small text spans in sentence-level discourse parsing and are not suitable for large text spans in text-level discourse parsing. In following experiments, we no longer compare our system with them.

Detailed Analysis

Height	Std	Bare		Nuc		Rel		Full	
		↓	↑	↓	↑	↓	↑	↓	↑
1	385	339	321	251	221	233	215	213	200
2	220	183	184	117	115	116	111	94	101
3	139	119	122	71	82	71	73	59	71
4	88	75	78	52	58	44	42	39	40
5	44	34	37	17	21	16	21	10	16
6	26	18	21	13	13	6	9	6	9
7	18	16	18	7	8	6	9	2	5
>= 8	13	11	10	0	0	0	0	0	0
Overall	933	795	791	535	521	497	486	426	445

Table 4: Performance over different DT levels. ("↓"- Top down approach, "↑"- Bottom up approach)

Detailed Analysis

Approach	NN	NS	SN
↓	67.0	42.2	33.7
↑	67.6	35.4	24.5

Table 5: Performance on nuclearity determination.

	EDU Num	Bare	Nuc	Rel
↑	1–5	94.8	57.9	52.0
	6–10	87.0	60.7	58.6
	11–15	78.0	50.1	45.4
	16–20	56.2	25.0	25.0
	21–25	68.9	47.0	42.4
	26–30	65.4	26.9	11.5
↓	1–5	97.0	67.1	56.6
	6–10	86.0	57.3	59.9
	11–15	75.2	50.3	41.4
	16–20	56.2	25.0	25.0
	21–25	76.6	57.7	40.8
	26–30	69.2	42.3	19.2

Table 6: Performance over different EDU numbers.