The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 388-397

利用 Attentive 來改善端對端中文語篇剖析遞迴類神經網路系統

Using Attentive to improve Recursive LSTM End-to-End Chinese Discourse Parsing

王育任　Yu-Jen Wang
國立中央大學資訊工程學系
Department of Computer Science & Information
National Central University
leowang@g.ncu.edu.tw

張嘉惠　Chia-Hui Chen
國立中央大學資訊工程學系
Department of Computer Science & Information
National Central University
chiahui@g.ncu.edu.tw

摘要

篇章剖析，可以幫助我們以不同角度來理解文句之間的關係與連結，但篇章剖析資料結構目前仰賴人工標記，使這項技術無法直接利用在任意篇章中。因此至目前為止，有許多研究著手於讓電腦能夠自動對篇章進行篇章剖析，並建構出一個完整的剖析樹。以中文語料庫 CDTB 來說，欲建立完整的篇章剖析程式，其問題主要可以被分成四項，分別是子句分割、剖析樹建立、子句關係辨識與中心關係辨識。由於深度學習近幾年發展快速，因此針對篇章剖析的建構方法也從傳統的 SVM, CRF 等方法，進展到目前以遞迴類神經的方式來建構剖析篇章程式。在本篇論文中，我們使用了許多目前最新的深度學習技術，例如 Attentive RvNN、self-attentive、BERT 等方法，來提高模型的準確度。最後，我們成功將每一項任務的 F1 都提高了近 10% 左右，達到目前我們所知研究中最好的效能。

Abstract

Discourse parser helps us understand the relationship and connection between sentences and sentences from different angles, but the tree structure data still need to rely on manual marking, which makes this technology unable to be directly used in daily life. So far, there have been many research and studies on how to automatically construct the complete tree structure on the computer. Since deep learning has progressed rapidly in recent years, the construction method for discourse parser has also changed from the traditional SVM, CRF

method to the current recursive neural network. In the Chinese corpus tree library CDTB, the parsing analysis problem can be divided into four main problems, including elementary discourse unit (EDU) segmentation, tree structure construction, center labeling, and sense labeling. In this paper, we use many state-of-the-art deep learning techniques, such as attentive recursive neural networks, self-attentive, and BERT to improve the performance. In the end, we succeeded in increasing the accuracy by more than 10% of F1 in each task, reaching the best performance we know so far.

關鍵詞：深度學習, 篇章剖析, 注意力機制, 遞迴類神經網路

Keywords: Deep Learning, Discourse Parsing, Attention, RvNN

一、緒論

篇章剖析主旨在分析文本之間的關係和結構信息，最終能建構成完整的剖析樹，至目前為止，篇章剖析語料庫有許多不同的體系，例如對目前研究影響最深的 Rhetorical Structure Theory (RST)體系[1]，擁有更多標記資料且結構更為自由的 Penn Discourse Treebank (PDTB) 體系[2]，或是以中文篇章結構為主的 Chinese Discourse Treebank (CDT) 體系[3]，每一種體系在標記的定義與見解皆不同，所建立起來的關係結構也截然不同，因此相同的篇章在不同體系之間，資料可以視為一種互補的關係。由於篇章剖析中豐富的關係標記，使得篇章剖析資料運用在許多不同的研究上，皆獲得更好的準確率，但篇章剖析語料庫仰賴人工標記，使得篇章剖析的許多研究無法直接使用在未經結構化的資料上，因此我們的研究以 CDT 體系為主，目的是要建立一個中文的剖析程式，將串列結構的篇章轉換成剖析樹結構的篇章資料，並標註出結構關係。

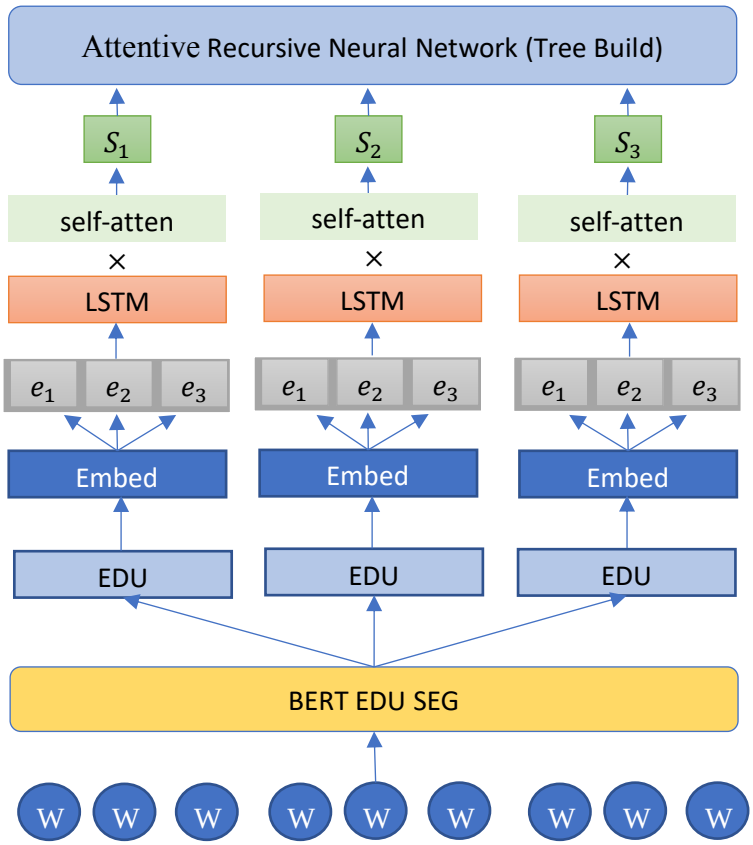
目前已經有許多針對篇章剖析進行的研究，例如 SVM Base Parser [4]、DCRF Base Parser [5]、Recursive Deep Learning [6]、Transition-Based Dependency Parser [7]，在這些研究中，將篇章剖析主要分成兩種問題，分別是子句分割與剖析樹結構建立，並將這兩項問題視為分類問題，但這些研究有以下問題無法直接應用在我們的研究上，其一是這些研究皆是針對英文進行剖析程式的建立，研究中設計了許多英語的文字特徵難以複製到中文剖析研究上，其二是大部分的研究不做子句分割的問題，而是以正確的子句當作輸入來進行訓練，而剖析模型中子句分割的好壞很重要，這會直接影響後續所有問題的效能表現，因此忽略子句分割的剖析程式，不但無法實際運用在現實生活中，也不符合我們的要求。

2018 年 Lin 提出了以中文為主的 RvNN Base Parser [8]，此研究使用中文語料庫 CDTB 來進行建構，不需要擷取額外的文字特徵，且包含完整端對端剖析結構，此研究是據我們所知，第一個針對中文的端對端篇章剖析開源程式，且達到目前我們所知中文剖析程式中最高的準確率，但研究中還有許多不足的地方，因此受到他們的啟發，我們希望以此研究架構為主軸，在效能與程式結構上進行優化，使之準確率更貼近人類的標準。Lin 於這份研究中，將篇章剖析分為四個步驟，分別是子句分割、樹結構構建、主次關係標示與篇關係辨識，在研究中使用標點符號對子句進行分割，接著輸入 Recursive neural network (RvNN) 對節點子樹資訊進行擷取，並對不同問題進行預測，最後使用 CYK 演算法做剖析樹的建構。我們針對 Lin 的架構做出以下四種優化：(1) 在訓練階段 Lin 並未使用任何預先訓練的 word embedding，但好的 word embedding 對模型的訓練極其重要，因此我們加入了 FastText embedding 來幫助訓練。(2) 一個子句中，每個字詞之間的重要程度皆不同，如果能找出字詞之間的權重資訊將能有效幫助模型的訓練，在程式中我們加入了 self-attentive[9]，讓模型能在不同順序點上學習文字或句子的重要程度。(3) 在 2016 年 Zhou 等人於 COLING 發表一篇論文[10]，將 Attentive 機制加入 RvNN 裡，幫助 RvNN 學習節點下每個子樹的重要程度，並且實驗在四個不同問題中，相較於加入 Tree Attention 前，獲得更好的效能。因此我們相信，一個樹的向量推導並不需要依賴於所有的子樹，樹狀結構中一定會有某個部份的子樹相對於其它子樹影響更大，我們並不需要依賴全部子樹來得到節點的訊息，因此我們將 attentive 機制也加入 RvNN 中，讓 RvNN 在訓練過程中找出相對重要的子樹，幫助模型進行訓練。(4) Lin 在模型中以標點符號作為子句分割的依據，但此方法的效果不理想，子句分割的 F1 比 Lin 實驗中的 baseline 還要低，因此我們將此問題看作是 sequence labeling problem，以目前做此類問題效能最好的 BERT [14] 進行子句分割實驗，並達到目前子句分割最好的效能。

二、模型設計

在這章節，我們會先針對模型的主要架構作解釋，隨後會再針對模型中個別使用到的深度學習技術做介紹。本論文架構主體以 RvNN 來進行訓練，目的是建構出一個端對端的中文剖析程式，我們程式的架構和資料準備方式參照了 Lin 所提出的方法，並在此架構上進行更進，針對我們更進的部分，我們會在次章節中進行詳細介紹。

Lin 對輸入的句子使用標點符號做子句分割，而後輸入 LSTM 擷取子句訊息，最後輸入 RvNN 內進行節點資訊提取，並建立剖析樹。我們與之不同，如圖一，將輸入的句子使用預先訓練好的 BERT 模型做子句分割，子句長度為 n，使用 FastText 做為文字的 embedding，每個字會映射到一個大小為 300 的陣列，每個子句會以字為單位轉換成大小為$300 \times n$的矩陣，則每個子句則可以表示為 $e^i = (e_1^i, \dots, e_n^i)$。之後將子句輸入 LSTM 學習，同時使用 self-attentive layer 算出每段子句文字資訊的重要程度，對文字依據不同的權重做加權，得到計算過後的子句資訊S_i，此資訊作為我們 Attentive RvNN 的輸入，並且以下到上的方式，使用 CYK 演算法針對子句做結合，最後組建出一顆完整的剖析樹。



圖一、篇章剖析訓練結構

我們採用與 Lin 相同的方式計算 RvNN，針對剖析樹的建構、主次關係與篇關係辨識等問題，會將 RvNN 對目標節點所輸出的隱藏層 $\vec{h}$ 與狀態層 $\vec{c}$，使用 SoftMax 分類器來算出不同問題選項的機率，並選取機率最高的選項做為答案，機率公式如下

$$\vec{p} = softmax\left(W_s\left[\begin{matrix}\vec{h}\\ \vec{c}\end{matrix}\right] + \overrightarrow{b_s}\right) \quad (1)$$

使用了三個 SoftMax 分類器來針對三種不同的問題做輸出，並同樣使用 CYK 演算法對剖析樹進行建構。

（一）子句分割

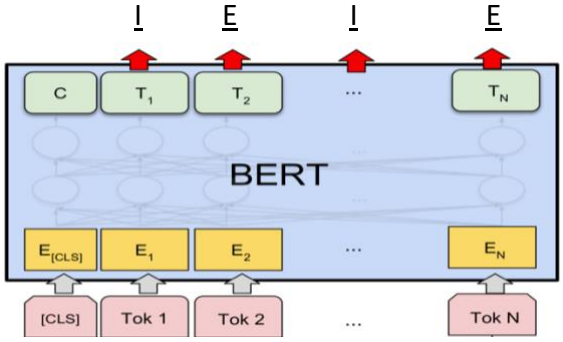
建構 CDTB 中，子句分割極為重要，子句分割點一定是位於標點符號上，但並非只要出現標點符號，句子就必須做分割。如下圖表一的句子，第一個逗號為子句分割點，但第二個逗號並不需要做子句分割。因此我們將子句分割獨立出來，視為一個序列標記問題，並且分別使用 BERT Fine-Tuning 此種在序列標記上非常熱門的方法來做訓練。

訓練中我們以 BERT 語言模型作為基礎，訓練資料字為單位做分割，使用 I 代表的 inside 與 E 代表的 End 兩種標籤來做標記，標記方式如表一，子句分割段落以 E 做標記，其餘皆以 I 來做標記。

表一、序列資料標記範例

句子	他	非	常	認	真	，	所	以	，	把	重	要	任	務	交	給	他	。
標籤	I	I	I	I	I	E	I	I	I	I	I	I	I	I	I	I	I	E

在 BERT Fine-Tuning 模型當中，我們使用 Google 所提供 12-layer, 768-hidden, 12-heads, 110M parameters 預先訓練好的中文模型作為我們的初始模型，並在此之上學習序列標記問題，我們會將輸入的文字轉換成 BERT 所定義的 Token，再將此 Token 輸入模型中進行訓練，模型架構如圖二



圖二、BERT 模型架構

（二）Self-Attentive

在 2017 年 Lin, Feng 等人於 ICLR 提出了 Self-Attentive 的機制，Self-Attentive 機制不需依靠外界的額外資訊，即可在句子內部進行 Attentive 的學習，尋找序列內部資訊的聯繫。因此，在我們的模型中，也加入了此機制，使用二维矩陣來表示句子，矩陣的每一行則表示句子的不同關係。公式如下：

$$s = (w_1, w_2 \ldots . w_n) \tag{2}$$

$$h_t = LSTM(w_t,\ h_{t-1}) \tag{3}$$

$$H = (h_1, h_2\ \ldots h_t) \tag{5}$$

$$a = softmax(W_2 \tanh(W_1 H^T)) \tag{6}$$

模型的輸入資料 s 含有 n 個資訊的序列，其中 w 代表序列的 word embedding，我們會將此資訊 s 輸入一個 LSTM，得到一個可用於計算的 h_t，LSTM 的隱藏單元為 u，則 $h_t \in R^u$，而 H $\in R^{n*u}$ 則表示為所有隱藏狀態h_t的集合，之後將 H 帶入 self-attentive 層中，最終使用 SoftMax 對每一段文字做歸一化，$a \in R^{1*n}$，此時 a 的每一維度可以認為是對應位置文字的 attentive，至此 Self-Attentive 完成。

（三）Attentive RvNN

在樹狀結構中，子樹之間彼此的資訊是有關聯的，因此將 Attentive 機制加入 RvNN 中，能讓模型更好的學習不同子樹之間的重要性。

對於我們 RvNN 輸入的隱藏層$\overrightarrow{h_s^1}, \overrightarrow{h_s^2}$，我們在計算新的隱藏層之前，會針對 $\overrightarrow{h_s^1}, \overrightarrow{h_s^2}$ 做 Attentive ，公式如下：

$$m_{k1} = \tanh\left(W^{(m_{k1})} h_k + U^{(m_{k1})} S\right) \tag{7}$$

$$m_{k2} = \text{relu}\left(W^{(m_{k2})} m_{k1}\right) \tag{8}$$

$$m_{k3} = \text{relu}\left(W^{(m_{k3})} m_{k3}\right) \tag{9}$$

$$a_k = \frac{\exp(w^t m_{k3})}{\sum_{j=1}^{n} \exp(w^t m_{k3})} \tag{10}$$

$$g = \sum_{1 \le k \le n} a_k h_k \tag{11}$$

h_k為我們要學習的任一子樹的隱藏層，S 為一個額外資訊，定義為子句經過 LSTM 後得到的資訊，我們將子樹的資訊h_k與 S 做訓練得到子樹之間的權重 a，最後將 a 乘回相

對應的子樹完成 Attentive。而此公式與 Zhou 等人在原始論文提出的公式略有不同，由於在原始方法中只使用了一個 m_{k1} 做學習，但我們發現只使用一個 m_{k1} 效能不盡理想，為了增強注意力機制的學習成果，我們增加了多層神經層 m_{k2}, m_{k3} 到模型中幫助訓練，且在這個階段，我們的資訊 S 就不參與學習。

三、實驗

在實驗中，由於我們模型的原架構來至 Lin，為了更好的與 Lin 的成果做比較，我們使用了與 Lin 相同的實驗方式，使用 standard evaluation tool PARSEVAL [11] 來作為樹狀結構 F1 的計算方法。在 Lin 的實驗中，將 Kang 等人於 2016 年提出來的實驗結果作為 Baseline，我們則將 Lin 與 Kang 的結果一起當作 Baseline 與我們的模型做比較。

實驗一中，我們針對完整的端對端篇章剖析程式做測試，並對樹狀結構建構、主次關係標示與篇章關係辨識做測試，實驗中我們的模型皆加入 FastText Embedding 和 Self-Attentive Layer，並針對是否使用 Attention RvNN 與 BERT 子句分割做比較。

實驗結果如表二，我們的模型使用模型在使用 Attentive-RvNN 後，準確率反而相較於未使用 Attentive-RvNN 來得低，而模型在加入 BERT 做子句分割後，在子句分割問題(EDU)上獲得最高的準確率，受惠於此，各項問題在後續的表現上也相繼獲得準確率的提升，得到所有研究中最好的準確率。

表二、端對端篇章剖析效能

Mod	EDU	Structure	Sense	Center	all
Kang	93.8%	46.4%	28.8%	23.1%	20.0%
Lin	87.6%	50.7%	27.8%	25.7%	22.2%
RvNN	88.6%	54.3%	35.3%	34.2%	30.8%
Atten RvNN	87.8%	55.1%	34.4%	33.1%	30.1%
BERT RvNN	**94.6**%	**57.7**%	**37.2**%	**36.0**%	**31.9**%

在前實驗一中，模型在使用 Attentive-RvNN 後，準確率反而相較於未使用 Attentive-RvNN 來得低，我們可以從下列表三來理解。Lin 的程式架構是以二元樹為建構目標，但 CDTB 本身為多元樹的結構，使用原始的多元樹來當作測試資料，並無法真實體現 Attentive-RvNN 所帶來的學習效能，因此我們將測試資料轉成二元樹後可發現，使用 Attentive-RvNN 的模型在二元樹的測試資料中，的確能幫助模型提高準確率，但模型的

效果並不能顯現在未經訓練過的多元樹結構上，因此效能在第一個實驗中被低估。

表三、標準子句二元樹與多元樹篇章剖析實驗

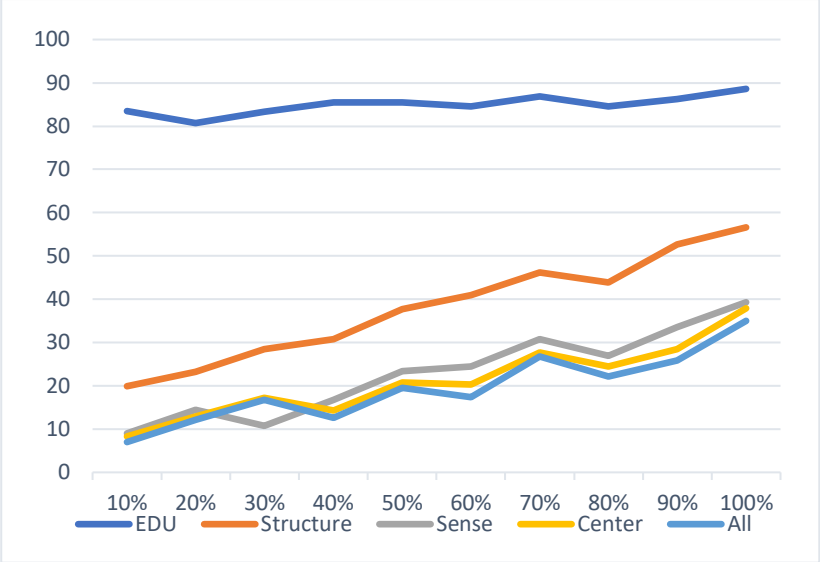
Tree Type	Model	Structure	Sense	Center	Overall
多元樹	RvNN	**63.4%**	**40.7%**	**40.5%**	**35.1%**
	Attentive-RvNN	62.4%	39.9%	38.1%	33.7%
二元樹	RvNN	69.2%	48.8%	**48.5%**	**44.2%**
	Attentive-RvNN	**69.9%**	**49.3%**	47.6%	**44.2%**

在下表四中，我們模型針對四種子句關係的準確率做比較，並將 Lin 的準確率列在最右側做參考，可以看到在我們的模型中，並列類與解說類獲得最好的準確率，而表中轉折類擁有所有類別中占比最高顯性連接詞，但卻因為訓練資料少於其他類別，所以效能表現明顯不足。

表四、子句關係分析表

Sense	Node (Gold)	Node (Pred)	True Positive	F1	Lin_F1
並列類	414	369	192	49.0%	67.0%
因果類	119	44	14	17.1%	16.5%
轉折類	151	146	54	36.3%	29.7%
解說類	11	11	8	72.7%	0.0%

最後，我們將訓練資料平分成十等份，每次訓練皆增加 10%的訓練資料，並在二元樹的測試資料下進行測試。從圖三中可以看出，子句分割在不使用 BERT 的情況下，準確率已經達到極限無法再增長，但樹狀結構建置、關係標記與中心標記還是成持續增長的狀態，因此我們認為模型目前尚未達到效能的極限，若能標記更多訓練資料，此模型的準確率會有更好的成果。



圖三、學習曲線

四、結論

本篇論文中，我們針對 RvNN Base 的中文端對端篇章剖析程式進行改進，並使我們的剖析程式達到目前最好的效能。我們加入的 word embedding 能有效幫助模型針對語言詞向量做學習，且可以避免使用過多人為產生的特徵來影響模型的訓練。而我們使用了 self-Attentive layer 與 Attentive-RvNN，在實驗中也證實，此方法能有效地幫助程式對於文句或子樹不同部分的重要程度做學習，以提高模型準確率。最後我們使用 BERT 的方法改善子句的分割，使模型的整體準確率有顯著的提升。以本實驗學習曲線的結果來看，此模型的效能受限於訓練資料，尚未達到最好的準確率，倘若能夠繼續增加訓練資料，模型將會有更好的表現。

參考文獻

[1] University of Chicago, *The Chicago Manual of Style Online*. Chicago, IL: University of Chicago Press, 2006. [Online].

[2] Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation.

[3] Li Yancui, Feng Wenhe, Sung Jing, Kong Fang, Zhou Guodong. Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure[C]. In Proceedings of the 2014 conference on Emporical Methods in Natural Language Processing, pages 2105–2114.

[4] David A. duVerle , Helmut Prendinger .A Novel Discourse Parser Based on Support Vector Machine Classification. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 665–673

[5] Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), pages 486–496

[6] iwei Li, Rumeng Li and Eduard Hovy. Recursive Deep Models for Discourse Parsing. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2061–2069

[7] Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with LSTMs. In EMNLP'15, Lisbon, Portugal. 349–359.

[8] Chuan-An Lin. A Unified RvNN Framework for End-to-End Chinese Discourse Parsing. Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 73–77 Santa Fe, New Mexico, USA, August 20-26, 2018.

[9] Zhouhan Lin , Minwei Feng, Cicero Nogueira dos Santos, Mo Yu,Bing Xiang, Bowen Zhou& Yoshua Bengio. A S TRUCTURED S ELF-ATTENTIVE S ENTENCE EMBEDDING. Published as a conference paper at ICLR 2017.

[10] Yao Zhou, Cong Liu, Yan Pan. Modelling Sentence Pairs with Tree-structured Attentive Encoder. 10 pages, 3 figures, COLING2016.

[11] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in theframe work of rhetorical structure theory. In Proceedings of the Second SIGdial Workshop on Discourse and Dialogue (SIGDIAL'01), pages 1–10.