國立臺灣大學電機資訊學院資訊工程學系
碩士論文
Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University
Master Thesis

適用於點對點中文語篇剖析的遞迴類神經網路統一架構

A Unified RvNN Framework for End-to-End Chinese Discourse Parsing

林傳恩
Chuan-An Lin

指導教授：陳信希博士
Advisor: Hsin-Hsi Chen, Ph.D.

中華民國 107 年 7 月
July, 2018

doi:10.6342/NTU201803852

doi:10.6342/NTU201803852

誌謝

首先感謝我的指導教授陳信希老師，耐心的在研究方向上提供給我許多建議，也給了我一個很好的研究環境讓我學習、研究。感謝瀚萱學長，無論在討論研究方法和投稿時都給我許多幫助。感謝祐婷學姊與重吉學長在 meeting 時給予的建議。感謝友善的同屆大神同學們：博政大神、佳文大神、宗翰大神、子軒大神、瑋柔大神、斯文大神。感謝實驗室的其他同學，自然語言處理實驗室越來越人丁興旺了。最後感謝我家人的支持。

林傳恩　民國 107 年 7 月

doi:10.6342/NTU201803852

doi:10.6342/NTU201803852

摘要

中文語篇剖析有四項子任務，包含初級語篇單元分割、剖析樹建立、主次關係識別、語篇關係辨識等。本文展示一個點對點中文語篇剖析器，並提出一套統一架構，可以對輸入之中文篇章直接產生完整的中文語篇剖析結果。我們的剖析器以遞迴類神經網路為基礎，同時對四項子任務進行學習，在中文語篇樹庫（CDTB）資料集上，達到最先進的效能。我們釋出了這個剖析器的原始碼與預先訓練完成的模型，立即可用。據我們所知，這是第一個開放原始碼的中文剖析工具集，而且這套獨立的工具集不須依賴外部資源（如句法剖析器），便於下游應用的整合。

關鍵字： 自然語言處理、中文語篇剖析、遞迴類神經網路、篇章結構、基本篇章單元

doi:10.6342/NTU201803852

doi:10.6342/NTU201803852

Abstract

This paper demonstrates an end-to-end Chinese discourse parser. We propose a unified framework based on recursive neural network (RvNN) to jointly model the subtasks including elementary discourse unit (EDU) segmentation, tree structure construction, center labeling, and sense labeling. Experimental results show our parser achieves the state-of-the-art performance in the Chinese Discourse Treebank (CDTB) dataset. We release the source code with a pre-trained model for the NLP community. To the best of our knowledge, this is the first open source toolkit for Chinese discourse parsing. The standalone toolkit can be integrated into subsequent applications without the need of external resources such as syntactic parser.

Keywords: Natural Language Processing, Chinese Discourse Parsing, Recursive Neural Network, Discourse Structure, Elementary Discourse Unit

doi:10.6342/NTU201803852

doi:10.6342/NTU201803852

Contents

doi:10.6342/NTU201803852

doi:10.6342/NTU201803852

List of Figures

doi:10.6342/NTU201803852

doi:10.6342/NTU201803852

List of Tables

doi:10.6342/NTU201803852

doi:10.6342/NTU201803852

Chapter 1 Introduction

Discourse structure appears in every articles. Parsing the discourse structure involves large-scale understanding of the text. In Chinese, discourse relations often appear in more implicit ways. With the help of neural network approach, we can learn to recognize the discourse structure and relations without the need of surface lexical and syntactic features.

In this thesis, we will discuss first how to construct discourse structure from Chinese raw text in paragraph level based on recursive neural network, we then try to recognize the discourse relations, finally develop an end-to-end Chinese discourse parsing system.

1.1 Background

1.1.1 Discourse parsing and its application

A discourse unit is a sequence of words, ranging from a single sequence, several sentences, even to the whole paragraph. As pointed out by Mann and Thompson [1988], no part in an article is completely isolated. Discourse parsing is aimed at identifying how the discourse units are related with each other, forming the hierarchical structure of an article. There are many NLP applications have been shown benefited from the information extracted by discourse parsing. Following are three examples:

summarization: Louis et al. [2010] explored how discourse structure and relations help text summarization, and concluded that the former information improves the performance more significantly.

information retrieval: In the work of Lioma et al. [2012], discourse paring is used to help model the article relevance.

doi:10.6342/NTU201803852

text categorization: Ji and Smith [2017] conduct experiments on several tasks, including article sentiment classification, news categorization and other text categorization tasks. The discourse information improve the performance in almost all tasks to a certain extent.

1.1.2 Task Definition

From a piece of text like a paragraph or an article to a discourse tree, there are a number of subtasks to deal with, including elementary discourse unit (EDU) segmentation, tree structure construction, center labeling, and discourse relation recognition. We will introduce these subtasks with the example paragraph (S1) shown in Figure 1.1, which consists of three sentences and seven EDUs, numbered as (a), (b), (c) to (g). This example is extracted from CDTB. we will discuss this corpus in next chapter.

Given (S1) as the input paragraph, our goal is to generate a discourse parse tree of (S1) as illustrated in Figure 1.2.

Following the CDT scheme Li et al. [2014b], we define the discourse parsing task to involve four perspectives as follows:

EDU Segmentation: EDUs are special discourse units (DUs) acting as the leaf nodes of a discourse parse tree. According to most Chinese discourse corpora, the EDU is limited to clause, which is segmented by some punctuation marks and containing at least one predicate that expresses at least one proposition. Besides, an EDU should be related to other EDUs with some propositional function instead of acting as a part of other EDUs.

Discourse Structure Construction: The combination of successive DUs (including both EDUs and non-leaf DUs) generates new discourse units (DUs) in a higher level of the discourse tree. As shown in Figure 1.2, the EDUs (d), (e), and (f) are combined as a new DU, which is further combined with the other EDU (g), and so on. Finally, the hierarchal structure is constructed covering all EDUs in (S1).

Sense Labeling: In CDTB, four top types of discourse relation sense are defined, including **Causality**, **Coordination**, **Transition**, and **Explanation**. Taking the relation between (a) and (b) in (S1) as an example, (a) states the premise and (b) states the outcome, so a sense of causality is the sense of discourse relation is Causality.

doi:10.6342/NTU201803852

(a) 浦東開發開放是一項振興上海，建設現代化經濟、貿易、金融中心的跨世紀工程，(Pudong's development and opening up is a century spanning undertaking for vigorously promoting Shanghai and constructing a modern economic, trade, and financial center.)

(b) 因此大量出現的是以前不曾遇到過的新情況、新問題。(Therefore, new situations and new questions that have not been encountered before are emerging in great numbers.)

(c) 對此，浦東不是簡單的採取 "幹一段時間，等積累了經驗以後再制定法規條例" 的做法，(In response to this, Pudong is not simply adopting an approach of "work for a short time and then draw up laws and regulations only after experience has been accumulated.")

(d) 而是借鑒發達國家和深圳等特區的經驗教訓，(Instead, Pudong is taking advantage of the lessons from experience of developed countries and special regions such as Shenzhen,)

(e) 並且聘請國內外有關專家學者，(by hiring appropriate domestic and foreign specialists and scholars,)

(f) 並且積極、及時地制定和推出法規性文件，(actively and promptly formulating and issuing regulatory documents.)

(g) 使這些經濟活動一出現就被納入法制軌道。(So that these economic activities are incorporated into the sphere of influence of the legal system as soon as they appear.)

Figure 1.1: Sample paragraph (S1) consisting of seven EDUs (a), (b), (c), ..., (g).

Center Labeling: The centering from semantic discriminates the focus of the two DUs joined. A discourse relation is either mononuclear or multinuclear. A mononuclear relation, which is the join of two DUs, usually has a nucleus DU and a satellite DU. The nucleus DU reflects the intention focus of the discourse and is thus more salient in the discourse structure, whereas the satellite DU presents supportive information for the nucleus. For example, In the join of (a) and (b) in Figure 1.2, (b) act as a nucleus since it stands for the outcome statement in the discourse relation of the causality sense. Since the nucleus may be the front DU, the later DU, or even both DUs, there are three centering types for a mononuclear relation: **Front**, **Latter**, and **Equal**, respectively. The multinuclear relation only occurs with the discourse sense Coordination, where multiple DUs combined in parallel, the centering relation comes out to be multinuclear like the join of (d), (e), and (f) in Figure 1.2.

doi:10.6342/NTU201803852

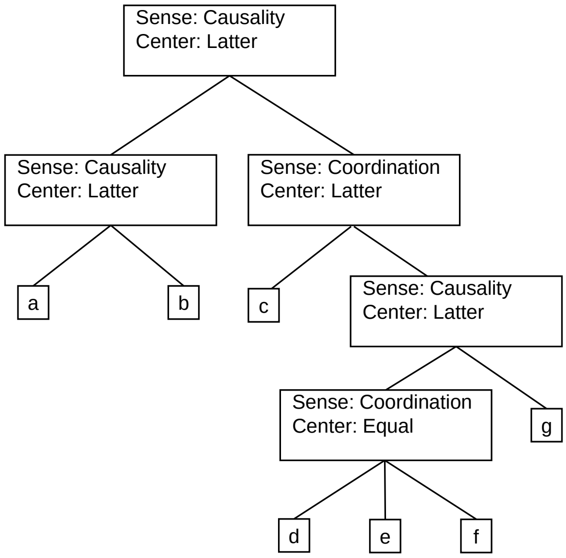

Figure 1.2: Discourse parse tree of (S1).

Besides, **connective** is also an important aspect in the original CDT scheme. Connectives or discourse markers are a set of words or phrases suggest the sense of discourse relation. In (S1), the connectives 因此 (therefore) and 對此 (In response to this) denote the discourse relation in the sense of causality; the parallel connective 不是... 而是...(is not...but...) suggests the sense of coordination. Discourse relations are divided into two types according to the existence or not of connectives. **Explicit** relations have a connective and **Implicit** relations have not. In deed, the CDT scheme views connectives as the predicates of discourse relations. However, we skip connective recognizing during our discourse parsing process. One reason is that we expect the unified neural network framework could learn to recognize connectives implicitly, and the another reason is that connectives does not appear in most discourse relations in Chinese corpora. We will discuss this issue in the next section.

1.2 Motivation

The performances of prior discourse parsing systems are limited due to several issues. First of all, most system conduct a pipeline approach. Indeed, The subtasks in Chinese discourse parsing depend on each other. In a pipelined system, there may be a severe issue

doi:10.6342/NTU201803852

of error propagation among elementary discourse unit (EDU) segmentation, connective recognition, parse tree construction, and relation labeling Kang et al. [2016].

Furthermore, there may be separate units to deal with discourse relations that has a connective (called explicit relation) and that doesn't (called implicit relation) Kang et al. [2016]. The intention is to make the best use of connective cue for the explicit relations. Unfortunately, the properties of Chinese are quite different from that of English, resulting challenging issues to build a Chinese discourse parser. Firstly, in Chinese Discourse Treebank (CDTB), the only Chinese discourse corpus with structure annotation at the paragraph levelLi et al. [2014b], 82% discourse relations have no explicit connective, comparing to 55% in Penn Discourse Treebank (PDTB), one of the largest English discourse corpus Prasad et al. [2014]. This makes a difficulty for the approach to discourse relation recognition that heavily relies on the literal cues in the text. Secondly, the annotation scheme used in CDTB is much different from that of RST-DT. Therefore, the parser constructed for RST-DT cannot be directly adapted to the CDTB dataset.

The other issue is that prior Chinese discourse parser relies on linguistic features extracted by external third party packages. For a toolkit targeting real-world applications, a standalone system is more robust and easy to deploy.

1.3 Goals

For the aforementioned reasons, in this work we propose an end-to-end Chinese discourse parser that performs EDU segmentation, discourse tree construction, and discourse relation labeling in a unified framework based on RvNN Goller and Kuchler [1996]. RvNN learns to construct the structured output through merging children nodes to parent nodes in the bottom-up fashion. Within the RvNN paradigm, recurrent neural networks (RNNs) are employed to model the representations from word segments, discourse units, to the whole paragraph. With these approaches, we are able not to rely on external parser and pass the connective recognizing part to adapt to the characteristics of Chinese text. In the prediction stage, we use the CKY algorithm to deal with both local and global information during the construction of discourse parse tree, eliminating the gap between the

doi:10.6342/NTU201803852

bottom-up approach and top-down annotation schemes.

The contributions of this work is three-fold. 1) We release a ready-to-use toolkit for end-to-end Chinese discourse parsing. To the best of our knowledge, this is the first publicly available toolkit for Chinese discourse parsing.[1] 2) We propose a unified framework based on RvNN for this task. Our model achieves the state-of-the-art performance. 3) Without the need for external resource like syntactic parser, our standalone end-to-end parser can be easily integrated into subsequent applications. The open source package can be even adapted to other languages.

1.4 Structure

The rest of this paper is organized as follows. Chapter 2 describes related works. Chapter 3 introduces the datasets. We present our system in Chapter 4 and discuss the system performance in Chapter 5. Chapter 6 concludes the remarks.

[1] https://github.com/abccaba2000/discourse-parser

doi:10.6342/NTU201803852

Chapter 2   Related Work

2.1   English Discourse Corpora

In this section, we will introduce the two commonly used English discourse corpus. The first one is the Rhetorical Structure Theory Discourse Treebank (RST-DT) Carlson et al. [2001]. RST-DT is annotated from 385 *Wall Street Journal (WSJ)* articles, which is selected from the Penn TreebankMarcus et al. [1993]. RST-DT follows the *Rhetorical Structure Theory (RST)* Mann and Thompson [1988]. In the RST framework, the discourse structure can be represented as a tree, with EDUs being the leaf nodes, and each internal node annotated by a *rhetorical relation*.

The second corpus is the Penn Discourse Treebank (PDTB) Xue et al. [2005], which is annotated on 2,159 WSJ articles selected from the Penn Treebank. PDTB adapts a predicate-argument view, one relation has two DUs as arguments. The whole discourse structure is not limited to be a complete tree. We use Example 2.1.1 and Example 2.1.2 to illustrate this aspect. Each example shows a discourse relation annotated in PDTB, and both these two relations are of **EntRel** relation sense which means ”the only relation between the two arguments is that they describe different aspects of the same entity”, as described in Zhou and Xue [2012]. We can see that the second argument of the relation in Example 2.1.1 appears again as the first argument of the relation in Example 2.1.2. This phenomenon can not appear in a hierarchically annotated corpus such as CDTB.

Example 2.1.1. [Rolls-Royce Motor Cars Inc. said it expects its U.S. sales to remain steady at about 1,200 cars in 1990.]$_{arg1}$[The luxury auto maker last year sold 1,214 cars in the U.S.]$_{arg2}$

doi:10.6342/NTU201803852

Example 2.1.2. [The luxury auto maker last year sold 1,214 cars in the U.S.]$_{arg1}$[Howard Mosher, president and chief executive officer, said he anticipates growth for the luxury auto maker in Britain and Europe, and in Far Eastern markets.]$_{arg2}$

Besides. PDTB annotate connectives and divide discourse relations to be explicit or implicit.

2.2 English Discourse Research

Since the release of RST-DT, the research on English discourse parsing has attracted attention in recent years Braud et al. [2017], Zhao and Huang [2017], Wang et al. [2017]. Li et al. [2014a] adopt RvNN from word level to the whole discourse, with a binary classifier to deal with the probability of two discourse units merging into a bigger unit, and a classifier to label the relation. Bowman et al. [2016] propose a neural network-based shift-reduce model with handcrafted features to build the parse tree in RST-DT dataset. Zhao and Huang [2017] integrates RST-DT and PDTB, develop a span-based constituency parser that jointly parses in both syntax and discourse levels

2.3 Chinese Discourse Corpora

There are fewer works on Chinese discourse corpus until recent years. The only corpus with the annotation of discourse structure at the paragraph level in Chinese is the Chinese Discourse Treebank (CDTB) dataset developed by Li et al. [2014b], which contains 500 Xinhua newswire documents from the Chinese Treebank Xue et al. [2005]. CDTB is the main training and testing data in our work. We will discuss this corpus in next chapter.

Zhou and Xue [2012] annotated another Chinese Discourse Treebank (CDTB-Zhou), which follows PDTB annotation scheme with some adaption to Chinese linguistic characteristics to annotate 890 articles of CDT. Again, this annotation scheme does not limit the discourse structure to be in hierarchical form. The paragraph in Figure 1.1 in Chapter 1 are annotated in both CDTB and CDTB-Zhou; however, in CDTB-zhou, the EDU (b)

doi:10.6342/NTU201803852

is related to EDU (a) with a **Causation** relation while (b) is also related to (c)-(d)-(e)-(f) with a **Conjunction** relation.

2.4 Chinese Discourse Research

Prior work of Chinese discourse parsing focuses on inter-sentential parsing Huang and Chen [2012] and shallow parsing Xue et al. [2016a], Wang and Lan [2016], The CoNLL 2016 Shared Task deals with shallow parsing Xue et al. [2016b]. So far, there is quite less work on complete hierarchical Chinese discourse parsing at paragraph or article level. Shih and Chen [2016]build an end-to-end parser focusing on explicit relations where a connective is presented. Kang et al. [2016] make the text propagate through different components: EDU detector, discourse relation recognizer, discourse parse tree generator, and attribution labeler, and the system is the current state-of-the-art. Both of the two are pipeline system with pure hand-crafted features.

2.5 Recurrent Neural Network

Recurrent Neural Network (RNN) is reportedly successful in learning the text representation, and we also integrate this component into our discourse parsing model. Figure

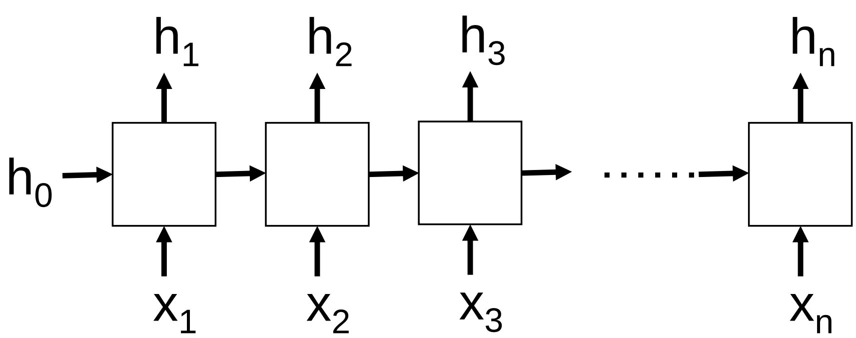

Figure 2.1: The basic RNN framework

2.1 demonstrated the framework of a basic RNN. RNN connects its unit iteratively to learn the representation of a ordering sequence. $(x_1, x_2, x_3 ... x_n)$ denotes the representations of the input sequence of RNN. For an input sentence, x_i may be the vector representation

doi:10.6342/NTU201803852

of the i^th word or character. Commonly, RNN needs another input which is the h_0 in Figure 2.1, h_0 is another initialized h-dimensioned vector. $(h_1, h_2, h_3...h_n)$ is the output representation sequence while each h_i is also the input of the next RNN unit. Figure 2.2 shows the computation flow inside a basic RNN unit. The formula of the computation in a unit may be like as follow:

$$\vec{h_i} = \sigma(W \begin{bmatrix} \vec{x}_i \\ \vec{h}_{i-1} \end{bmatrix} + \vec{b}) \tag{2.1}$$

where W is the weighting matrix, $\vec{b}$ is the bias vector, and σ is some activation function such as $tanh()$. Both W and $\vec{b}$ are parameters to be trained with some objective function. There are many variation of neural network suitable for the discourse parsing task, and the

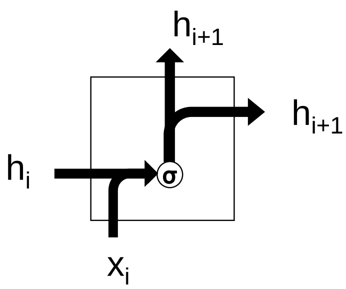

Figure 2.2: The basic single RNN unit

Long Short-term Memory (LSTM) neural network Hochreiter and Schmidhuber [1997] is a variation of RNN that makes use of memory and forgetting mechanism to keep valuable information in the processing through the sequence.

2.6 Recursive Neural Network

Recursive Neural Network (RvNN) connects its unit hierarchically to process tree structured data. Many tasks have benefited from this recursive framework, including sentence parsing Bowman et al. [2016], and sentiment analysis Socher et al. [2013]. Also, RvNN has been successfully applied on English discourse parsing task Socher et al. [2013]. Figure 2.3 shows the framework of a basic RvNN. Each x_i denotes the representation of

doi:10.6342/NTU201803852

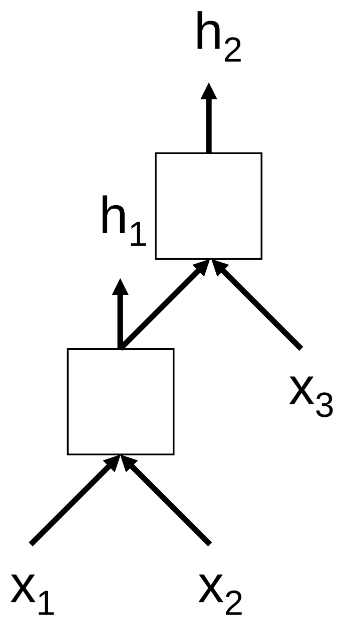

Figure 2.3: The basic RvNN framework

each element of the tree-structured input. In the discourse parsing task, an x_i can be the representation of the $i^t h$ DU in the input paragraph. Each h_i acts as the output of each RvNN unit and also the input to the next unit. 2.4 shows the computation flow inside a basic RvNN unit. The formula of the computation in a unit may be like as follow:

$$\vec{h_k} = \sigma(W \begin{bmatrix} \vec{x}_i \text{ or } \vec{h}_i \\ \vec{x}_j \text{ or } \vec{h}_j \end{bmatrix} + \vec{b}) \tag{2.2}$$

Again, W is the weighting matrix, $\vec{b}$ is the bias vector, both of which can be adjusted during the training process. σ is the activation function.

The TreeLSTM Tai et al. [2015] takes both advantages from LSTM and RvNN, generalizing the LSTM unit to take the representations of two children node in the binary tree as input, and ouput the representation of the parent node. TreeLSTM has been proved effective for predicting semantic relatedness of two sentences and sentiment classification Tai et al. [2015]. We will discuss the mechanism of TreeLSTM in Chapter 4 when introducing our model.

doi:10.6342/NTU201803852

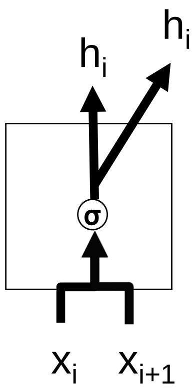

Figure 2.4: The basic single RvNN unit

doi:10.6342/NTU201803852

Chapter 3 Datasets

3.1 Chinese Discourse Treebank

As mentioned in Chapter 2, we use CDTB annotated by Li et al. [2014b] as our main dataset. Each paragraph is annotated with EDUs, connectives, discourse structures, relation senses and centers informations in CDTB. We then gives the analysis of CDTB in this section.

	Articles	Paragraphs	EDUs	Relations
Number	500	2,342	10,609	7,308

Table 3.1: Statistics of CDTB.

Table 3.1 lists the amount of articles, paragraphs, EDUs and relations in CDTB. The $7,308$ relations are categorized into four main relation sense. We give the interpretation by Shih [2015] of each relation sense below. Following each interpretation, we also give examples which are also used for latter discussion.

Coordination: Coordination is used when the arguments are descriptions on different aspects of the same things that share common features.

Example 3.1.1. [去年實現進出口總值達一千零九十八點二億美元，(Last year, the total value of imports and exports reached US$1,092,200,000,)]$_{arg1}$[占全國進出口總值的比重由上年的百分之三十七提高到百分之三十九。(which increased from 37% in the previous year to 39% in the proportion of the total import and export value of the country)]$_{arg2}$
(sense: Coordination, center: Former, type: Implicit)

doi:10.6342/NTU201803852

Example 3.1.2. [有關部門先送上這些法規性文件，(The relevant departments will send these regulatory documents first,)]$_{arg1}$[然後有專門隊伍進行監督檢查。(and then have a special team to supervise and inspect.)]$_{arg2}$
(sense: Coordination, center: Latter, type: Explicit)

Causality: Causality is used when an event in an argument causes the event in another argument. It expresses the relationship between the cause and the effect.

Example 3.1.3. [浦東開發開放是一項振興上海，建設現代化經濟、貿易、金融中心的跨世紀工程，(Pudong's development and opening up is a century spanning undertaking for vigorously promoting Shanghai and constructing a modern economic, trade, and financial center.)]$_{arg1}$[因此大量出現的是以前不曾遇到過的新情況、新問題。(Therefore, new situations and new questions that have not been encountered before are emerging in great numbers.)]$_{arg2}$
(sense: Causality, center: Latter, type: Explicit)

Example 3.1.4. [上海浦東近年來頒佈實行了涉及經濟、貿易、建設、規劃、科技、文教等領域的七十一件法規性文件，(In recent years, Shanghai Pudong has promulgated and implemented 71 legal documents covering economic, trade, construction, planning, science and technology, culture and education, etc.,)]$_{arg1}$[確保了浦東開發的有序進行。(ensuring the orderly development of Pudong.)]$_{arg2}$
(sense: Causality, center: Former, type: Implicit)

Transition: Transition is used when the arguments contrast with each other. It shows the difference between arguments.

Example 3.1.5. [數年前，北海還是北部灣一個默默無聞的小漁村，(A few years ago, Beihai was still a small fishing village in the Beibu Gulf.)]$_{arg1}$[然而三五年時間北海已建成了一個現代化都市的框架，街上客流如潮，樓房拔地而起。(However, in the past three or five years, the Beihai has built a framework of a modern city. The streets are full of passengers and buildings.)]$_{arg2}$
(sense: Transition, center: Latter, type: Explicit)

doi:10.6342/NTU201803852

Example 3.1.6. [科爾認為，俄羅斯軍隊最終全部撤離德國是"歐洲戰後歷史的終結"。(Cole believes that the final withdrawal of the Russian army from Germany is "the end of post-war history in Europe.")]$_{arg1}$[他說，１９４１年６月２２日德國進攻蘇聯是不能忘記的。(He said that, the Germany's attack on the Soviet Union on June 22 in 1941 could not be forgotten.)]$_{arg2}$
(sense: Transition, center: Former, type: Implicit)

Explanation: Explanation expresses the same concept using different wordings. It is used for arguments that try to explain the same thing in different ways.

Example 3.1.7. [建築是開發浦東的一項主要經濟活動，(Buildings are a major economic activity in the development of Pudong.)]$_{arg1}$[百家建築公司、四千餘個建築工地遍佈在這片熱土上。(Over the years, hundreds of construction companies and more than 4,000 construction sites have been scattered throughout the land.)]$_{arg2}$
(sense: Explanation, center: Former, type: Implicit)

Example 3.1.8. [外商投資企業的出口商品仍以輕紡產品為主，(The export commodities of foreign-invested enterprises are still dominated by light textile products.)]$_{arg1}$[其中，出口額最大的商品是服裝，去年為七十六點八億美元。(Among them, the largest export commodity is clothing, which was 7.68 billion US dollars last year.)]$_{arg2}$
(sense: Explanation, center: Latter, type: Explicit)

Table 3.1 shows the relation sense distribution. We can see that the distribution of the four senses is quite unbalanced. While the Coordination sense appears in more than half of all relations, Transition has only the proportion of 2.9%. Avoiding bias may be an issue when predicting the relation sense.

	Coordination	Causality	Transition	Explanation
Number	4,148	1,331	212	1,617
Proportion	56.8%	18.2%	2.9%	22.1%

Table 3.2: Distribution of discourse relation senses in CDTB.

Table 3.1 gives the relation type distribution of CDTB. About three quarters of discourse relations are of implicit type. It indicates the big challenge for Chinese discourse

doi:10.6342/NTU201803852

relation sense and center classification since implicit relations have less lexical cues to the task.

	Explicit Relations	Implicit Relations
Number	1,814	5,494
Proportion	24.8%	75.2%

Table 3.3: Distribution of discourse relation types in CDTB.

To further look into the property of discourse relations in CDTB, we use Table 3.1 to inspect the join distribution of relation sense and relation type.

We can see that most of the Coordination and Explanation relations are of Implicit type. It might be because that in Chinese writing, two neiborghing statements can often be interpreted to be about the same aspect without lexical cues. As shown in Example 3.1.1, we can easily infer that both two DUs are discussing about the total value of imports and exports. In Example 3.1.7, the latter DU is obviously the elaboration of the former. Conversely, some connectives such as " 而且 (and)", " 還 (also)"," 先... 然後 (first...and then)"，" 其中 (Among them)" which appear in these two relations can be omitted without changing the meaning. Example 3.1.2 and Example 3.1.8 can illustrate this aspect.

In contrast, Explicit relations act as the majority of Transition relations. It is much more difficut to delete a connective such as " 但是 (but)" or " 然而 (however)" in a Transition relation. Example 3.1.7 shows that we may take more efforts to recognize a Transition relation without such connective.

	Coordination	Causality	Transition	Explanation
Explicit	974	466	173	201
Implicit	3,174	865	39	1,416

Table 3.4: Distribution of discourse relation senses and types in CDTB.

Table 3.1 shows the join distribution of relation sense and relation type.

Most discourse relations of Coordination sense did not put semantic center on either arguments. Even when a Coordination relation is labeled with center Front or Latter, the central DU is not so obvious, as shown in Example 3.1.1 and Example 3.1.2.

While the latter statements are often more prominent in Transition relations, Explanation relations often forcus on the former statements which act as the main idea. Example

doi:10.6342/NTU201803852

3.1.5 and 3.1.7 can illustrate this aspect. It is also worth noticing that in the Explanation relation of center Latter shown in Example 3.1.8, the connecive " 其中 (among which)" is used to emphasize the largest export commodity of foreign-invested enterprises.

	Coordination	Causality	Transition	Explanation
Front	283	416	11	1,398
Latter	184	875	191	197
Equal	3,681	40	10	22

Table 3.5: Distribution of discourse relation senses and centers in CDTB.

Table 3.1 lists the relation distribution of different argument number. Note that only the relations of Coordination sense and Equal Center are possible to have more than two arguments. According to the tabel more than 99% relations have no more than 4 arguments and about 91% relations are binary relations. Example 3.1.9 demonstrates a special case of a 8-argument relation. It lists a series of parallel DU to describe the production value improvement.

Argument Number	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Relation Number	6653	454	135	40	15	6	1	1	0	0	1	0	1	0	1

Table 3.6: Distribution of argument number in CDTB.

Example 3.1.9. [據廣州市統計局提供的資料，去年廣州市完成國內生產總值一千四百四十五點八四億元；(According to the data provided by the Guangzhou Municipal Bureau of Statistics, last year Guangzhou completed a GDP of 144.448 billion yuan;)]$_{arg1}$[完成工業增加值五百七十三點四八億元；(completed industrial added value of 573.384 billion yuan;)]$_{arg2}$[農業增加值八十點七五億元；(agricultural added value of 80 A total of 750 million yuan;)]$_{arg3}$[固定資產投資六百五十五點四五億元；(fixed assets investment of 6.545 billion yuan;)]$_{arg4}$[社會消費品零售總額六百四十四點三二億元；(total retail sales of consumer goods reached 634.324 billion yuan;)]$_{arg5}$[外貿出口總值六十五點一三億美元；(total export value of 6.53 billion US dollars;)]$_{arg6}$[實際利用外資二十六億美元；(actual use Foreign investment was US$2.6 billion;)]$_{arg7}$[零售物價指數上漲百分之四點三。(the retail price index rose by 4.3%.)]$_{arg8}$

doi:10.6342/NTU201803852

	。	，	、	：	；	？
Number of Punctuations	4,615	9,176	2,826	229	346	27
Number as EDU Boundaries	4,584	5,416	8	62	337	25
Proportion to be EDU Boundaries	99.3%	59.0%	0.2%	27.1%	97.4%	92.5%
	！	…	）	—	”	」
Number of Punctuations	13	4	5	47	429	134
Number as EDU Boundaries	13	3	5	8	73	75
Proportion to be EDU Boundaries	100.0%	75.0%	100.0%	17.0%	17.0%	55.9%

Table 3.7: Distribution of punctuations in CDTB.

Table 3.1 lists the number of punctuations in CDTB as well as the number of the punctuations being a EDU boundary. We can see from it that some punctuations are much more clear-cut for EDU boundary detection than others. For example, almost all '。's are EDU and almost all '、's are not. In contrast, '，'s are very ambiguous when encountering while they are the punctuations which appear most frequently. So, whether the punctuation encountered is a boundary of a EDU is the main challenge when doing EDU detection.

3.2 Chinese Treebank

In the last part of our experiment, we attempt to use the syntactic structure information of each sentence when training our model. Since the articles of CDTB is selected from CTB which is a much bigger Chinese corpus annotated with syntactic information, so we also use CTB to look for this desired information.

Figure 3.1 presents a part of syntactic parsing tree annotated in CTB as an example. Similar to the case of discourse parsing, we can retrieve folowing informations from a Chinese syntactic parsing tree:

Word segmentation: A sequence of Chinese characters is first segmented into words, a Chinese word may consist of one to three characters in most cases. In Figure 3.1, the word '借鑒 (take advantage of)' consists of character '借 (borrow)' and '鑒 (refer)'. Words are the leaf nodes in a syntactic parsing tree.

POS tags: A word is labeled by its part of speech (POS) tag. In a syntactic parsing

doi:10.6342/NTU201803852

tree, the POS tag is labeled on the direct parent node of its corresponding word. In Figure 3.1, the POS tag of ’ 借鑒 (take advantage of)’ is ’VV’ which is a subclass of verb. The POS tag of ’ 發達 (developed)’ is ’JJ’ which is a subclass of noun classifier.

Syntactic structure and grammatical labels: The tree structure we see in Figure 3.1 is the syntactic structre. Each nodes except the leaf nodes and the POS nodes are tagged with grammatical labels which represent the grammatical relations of the node. For example, ’NP’ means a ’noun phrase’.

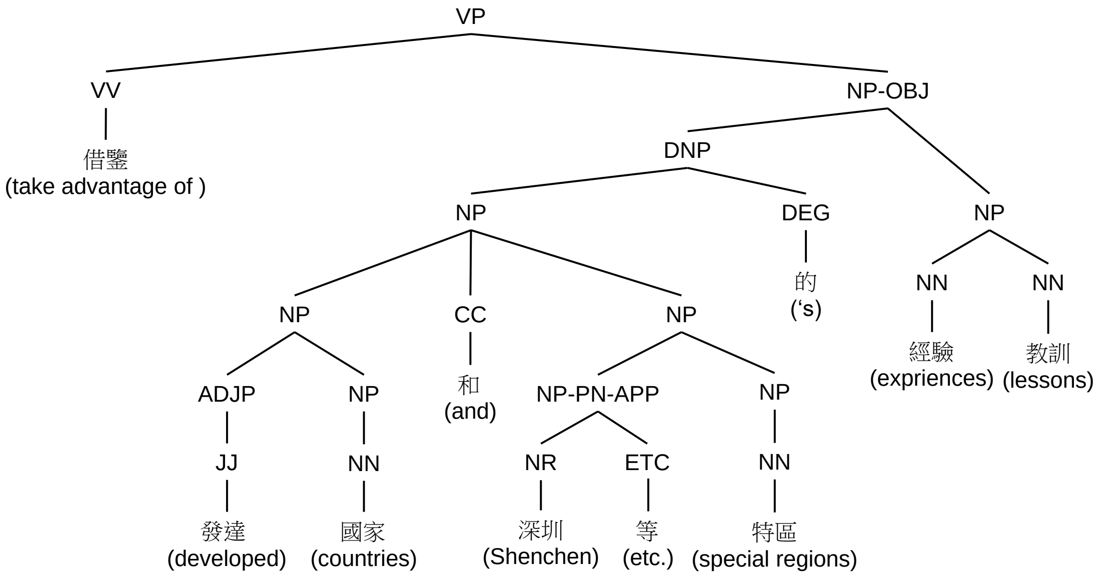

Figure 3.1: Sample syntactic tree in CTB.

The syntactic information is useful for the discourse parsing task. However, to avoid the need of external syntactic parser, our model needs to learn to extract syntactic information itself. In Section 5.4 of Chapter 5, we describe our attemption to integrate syntactic structure information into our model.

doi:10.6342/NTU201803852

Chapter 4 Methods

4.1 System Overview

The architecture of our united framework for end-to-end Chinese discourse parsing is shown in Figure 4.1. For a given text, we first segment the text into m text segments $\mathbf{w}^1, \mathbf{w}^2, \mathbf{w}^3, ..., \mathbf{w}^m$ by using punctuation marks as delimiter, where $\mathbf{w}^i = (w^i_1, \ldots, w^i_{n_j})$ forms the sequence of words in the ith text segment. The words are fed into an embedding layer, and $\mathbf{w}^i$ is then represented as $\mathbf{e}^i = (e^i_1, \ldots, e^i_{n_j})$. Then, LSTM is trained to convert $\mathbf{e}^i$ into the segment representation $\mathbf{s}^i$, and $\mathbf{s}^1, \mathbf{s}^2, \mathbf{s}^3, ..., \mathbf{s}^m$ serve as the input for the RvNN. Through the RvNN, segments are hierarchically joined to DUs in the bottom-up fashion. Finally a single discourse parse tree is constructed, and the sense and the centering relations of each join are labeled.

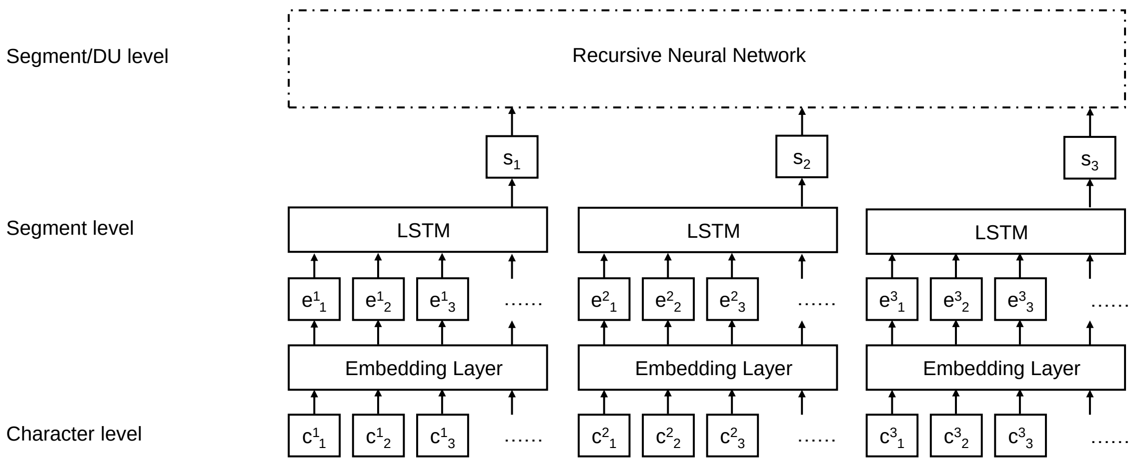

Figure 4.1: Architecture of our discourse parser.

doi:10.6342/NTU201803852

4.2 Recursive Neural Network

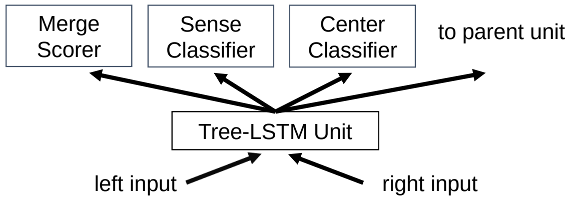

Figure 4.2: Tree-LSTM unit for discourse parsing.

Figure 2.3 shows the unit in our RvNN based on the Tree-LSTM unit [Tai et al., 2015]. Given the left and the right inputs (i.e. two text segments or two DUs), the Tree-LSTM composition function produces a representation for the new tree node. The Tree-LSTM unit generalizes the LSTM unit to tree-based inputs. Similar to LSTM, Tree-LSTM makes use of intermediate states as a pair of an active state representation $\vec{h}$ and a memory representation $\vec{c}$. We use the version similar to Bowman et al. [2016] as the formula:

$$\begin{bmatrix} \vec{i} \\ \vec{f_l} \\ \vec{f_r} \\ \vec{o} \\ \vec{g} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(W_{\text{comp}} \begin{bmatrix} \vec{h}_s^1 \\ \vec{h}_s^2 \end{bmatrix} + \vec{b}_{\text{comp}} \right) \tag{4.1}$$

$$\vec{c} = \vec{f_l} \odot \vec{c}_s^2 + \vec{f_r} \odot \vec{c}_s^1 + \vec{i} \odot \vec{g} \tag{4.2}$$

$$\vec{h} = \vec{o} \odot \tanh(\vec{c}) \tag{4.3}$$

where σ is the sigmoid activation function, $\odot$ is the element-wise product, and the pairs $\langle \vec{h}_s^1, \vec{c}_s^1 \rangle$ and $\langle \vec{h}_s^2, \vec{c}_s^2 \rangle$ are input from its two children tree nodes. The output of Tree-LSTM is the pair $\langle \vec{h}, \vec{c} \rangle$. Note that the Tree-LSTM unit is designed for binary tree. In Section 4.4, we show how to handle the multinuclear, where more than two children nodes join, in this framework.

The representation $\vec{h}$ and $\vec{c}$ produced by Tree-LSTM is taken for four usages: merge scoring, sense labeling, center labeling, and as input for the upper Tree-LSTM unit. In the

doi:10.6342/NTU201803852

prediction stage, the representation will be first sent into the merge scorer to measure the probabilities of the join of its two children tree nodes:

$$\vec{p_m} = \text{softmax}(W_m \begin{bmatrix} \vec{h} \\ \vec{c} \end{bmatrix} + \vec{b}_m) \tag{4.4}$$

The output $\vec{p_m}$ is a 2-dimensional vector, representing the probabilities of to merge and not to merge.

Similarly, the sense classifier and the center classifier compute the probability distribution $\vec{p_s}$ and $\vec{p_c}$ as follows:

$$\vec{p_s} = \text{softmax}(W_s \begin{bmatrix} \vec{h} \\ \vec{c} \end{bmatrix} + \vec{b}_s) \tag{4.5}$$

$$\vec{p_c} = \text{softmax}(W_c \begin{bmatrix} \vec{h} \\ \vec{c} \end{bmatrix} + \vec{b}_c) \tag{4.6}$$

For sense labeling, $\vec{p_s}$ consists of 6 values constituting the probabilities of six senses: Causality, Coordination, Transition, Explanation, subEDU, and EDU. Our end-to-end parser constructs the discourse parse tree from the text segments, EDUs, and to non-leaf DUs in an united framework, so we need to use the last two categories to mark which condition the current node is under EDU level.

For center labeling, $\vec{p_c}$ consists of 3 values constituting the probabilities of the three center categories including Front, Latter, and Equal. Center labeling is only performed at the DU level.

4.3 Text Segmentation

The raw text sent to our system is first divided to several segments by using specific punctuation marks as delimiter. The punctuation marks include full-stop (。), question mark (？), exclamation mark (！), comma (，and 、), semicolon (；), colon (：), right

doi:10.6342/NTU201803852

quotation mark (” and 」), ellipsis (⋯), and dash (—).

A segment forms an EDU by itself if it contains at least one predicate and expresses at least one proposition. Otherwise, multiple segments are merged to form an EDU. For example, the EDU (f) in Figure 1.1 consists of two segments “并且积极” (“and actively”) and “及时地制定和推出法规性文件” (“promptly formulating and issuing regulatory documents”).

4.4 Handling Binary Tree Structure

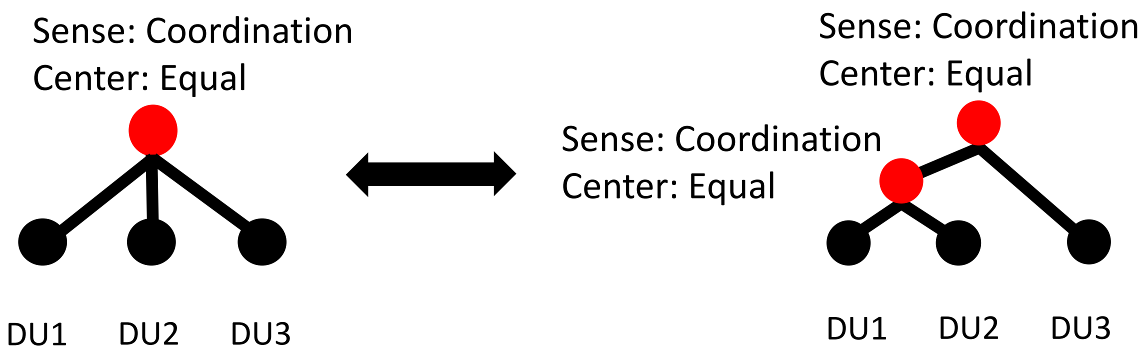

Figure 4.3: Transforming between a multi-way tree and a binary tree for the discourse relation.

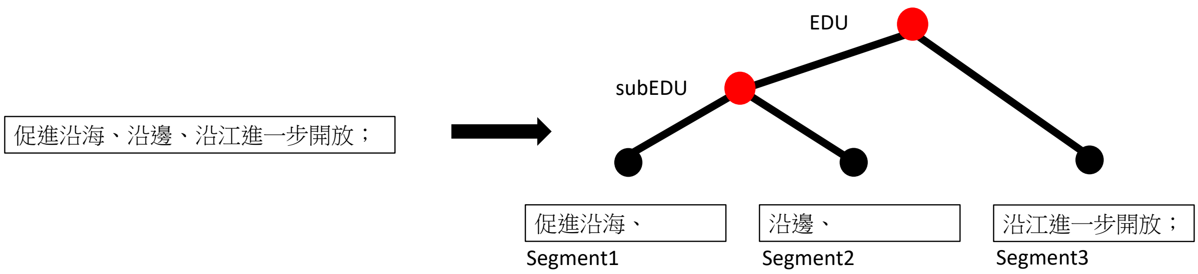

Figure 4.4: Transforming an EDU composed of multiple segments to a binary tree.

In the CDT scheme, the discourse parse tree is not limited to binary tree. As mentioned in Section 4.2, however, Tree-LSTM is modeled as binary tree. Therefore, we have represent the discourse parse tree with the structure of binary tree. In the discourse parse tree, there are two cases where a tree node has more than two children. In the first case, the tree node is an internal DU with the sense type coordination, where its all subtrees are

doi:10.6342/NTU201803852

parallel joined. As shown in Figure 4.3, we transform the tree node to two new nodes in the left-first merging scheme. In the second case, the tree node is an EDU that consists of more than two text segments as shown in Figure 4.4. Similarly, we merge the k segments into tree structure with $k - 1$ binary nodes in the left-first merging scheme. The highest node is labeled as EDU, and all the rest of the nodes are labeled as subEDU.

4.5 Parser Training

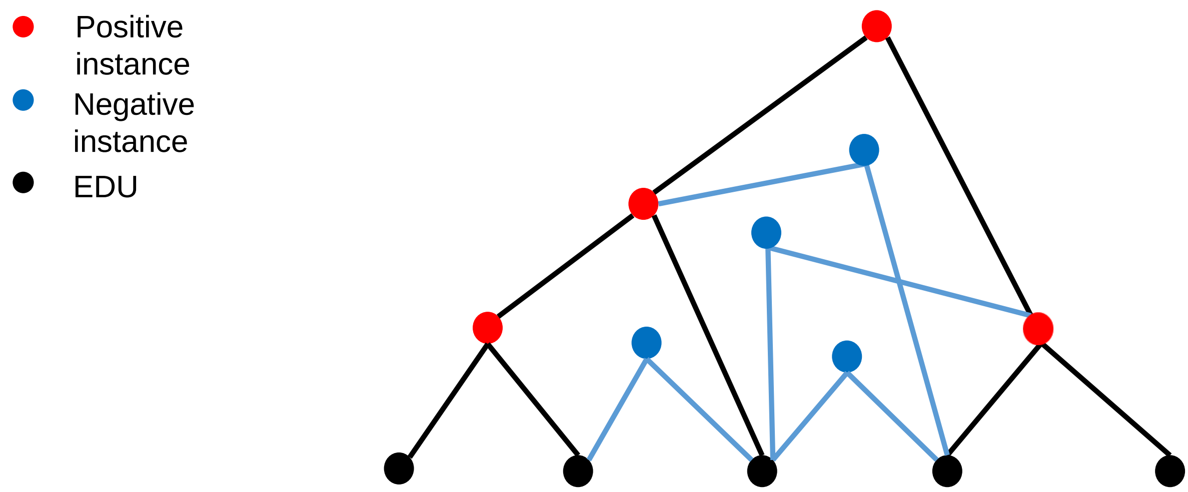

Figure 4.5: Training Instances Example

To train the RvNN, the positive instances are the tree nodes extracted from the discourse parse trees in CDTB. Figure 4.5 illustrate a parsing tree constructed from five segments(the black node). The black nodes, black lines, and the red nodes form the parsing tree itself. So, we can find four subtrees by considering one red node as the root. These four subtrees are the four positive instances we can derive from this parsing tree for training. On the other hand, we select arbitrary two neighboring subtrees and merge them into a new tree. The new tree is regarded as a negative instance if it is inconsistent with the ground-truth. We can see from Figure 4.5 that there are four possible trees as negative instances with the blue nodes as the root. The losses of the merging scorer, the sense classifier, and the center classifier, $\mathcal{L}_{\mathrm{m}}$, $\mathcal{L}_{\mathrm{s}}$, and $\mathcal{L}_{\mathrm{c}}$, respectively, are measured with cross-entropy. When training on negative instance, we don't need to estimate the performance

doi:10.6342/NTU201803852

of classification of sense and centering. In contrast, in the positive cases, we sum up the loss of the three and optimize them jointly. More formally, our loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\mathrm{m}}, \text{ if the instance is negative} \\ \mathcal{L}_{\mathrm{m}} + \mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{c}}, \text{ otherwise} \end{cases} \tag{4.7}$$

We use stochastic gradient decent (SGD) with the learning rate of 0.1 for parameter optimization.

4.6 Parse Tree Construction

In the prediction stage, we construct the discourse parse tree based on the predictions made by Tree-LSTM. We modify the Cocke–Younger–Kasami (CKY) CKY algorithm Younger [1967] to maximize the probability of the whole parse tree. The CKY-like dynamic programing algorithm simulates the recursive parsing procedure, considering local and global information jointly. In each step of the dynamic programming procedure, we consider several combinations of two neighboring trees L and R, merge them to a new tree N, and select two such Ns with higher probability $Pr(N)$ as candidates for future steps. Pr(N) is formulated as follows:

$$Pr(N) = Pr_{Merge}(L, R) \times Pr(L) \times Pr(R) \tag{4.8}$$

The $Pr_{Merge}(L, R)$ above is the output of the merge scorer in our model. Since we always stored the top two Ns in each entry of the dynamic programming table, we mark our CKY-like algorithm as **CKY2**. See Algorithm 1 for details.

doi:10.6342/NTU201803852

Algorithm 1 Discourse Parse Tree Construction with Dynamic Programming

$Probs \leftarrow$ table[][]
for $level$ from 0 to $n-1$ **do** ▷ The span range
    for col from 0 to $n - level$ **do** ▷ The start column
        $Probs[level][col] \leftarrow 0$
        $Candidates \leftarrow$ list[]
        for k from 1 to $level$ **do**
            $LeftTree \leftarrow$ GetTree($k-1$, col) ▷ Get the candidate tree for left span
            $RightTree \leftarrow$ GetTree($level - k$, $col + k$)
            $NewTree$, $MergeProb \leftarrow$ Merge($LeftTree$, $RighTree$)
                ▷ Apply RvNN unit
            $Prob \leftarrow MergeProb \times Probs[k-1][col] \times Probs[level-k][col+k]$
        $Tree, Prob \leftarrow Candidates.MaxProb()$
            ▷ Get the maximum probabilty and the tree
        SaveTree($level$,col,$Tree$) ▷ Save as the candidate tree
        $Probs[level][col] \leftarrow Prob$

4.7 Model Variation

Variations of our RvNN framework will also be tested for comparison. In the version shown in Figure 4.6, instead of running our CKY algorithm throughout the whole construction process from segment level, CKY is only adopt after EDU detection. For EDU detection, we process through the segment representations s_1, s_2, $s_3 ... s_n$ from left to right, judging based on the merge scorer whether to merge the next segment as a part of EDU or separate it to be the start of a new EDU. The intention is to fit how we construct our training instance, as mentioned in Section 4.5. We abbreviate our original model as **RvNN-CKY2**, and this modified version as **RvNN-CKY2+Seq-EDU**.

Considering each segment representation s_i output by the LSTM layer only contains informations of the segment itself. We attempt to add one bi-LSTM layer between the original LSTM and CKY part to integrate context information into each segment representation, as shown in Figure 4.7. In this version, s_1, s_2, $s_3 ... s_n$ are first fed into the bi-LSTM, and the outputs f_1, f_2, $f_3 ... f_n$ are then fed into the RvNN. We abbreviate this version of our model as **RvNN-CKY2+bi-LSTM**.

In a further attempt to integrate syntactic information into our model while still avoid the need of external syntactic parser, we generalize our model to not only construct dis-

doi:10.6342/NTU201803852

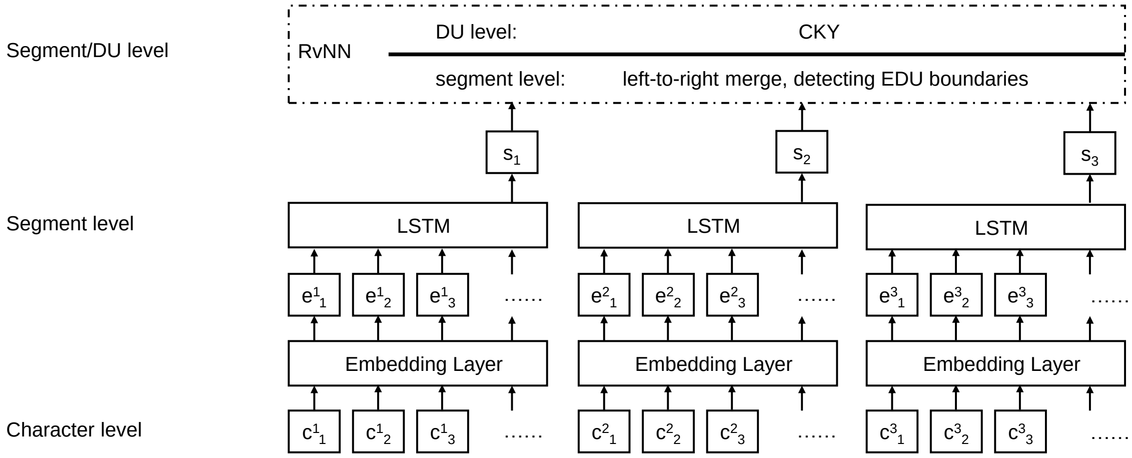

Figure 4.6: The RvNN-CKY2+Seq-EDU model.

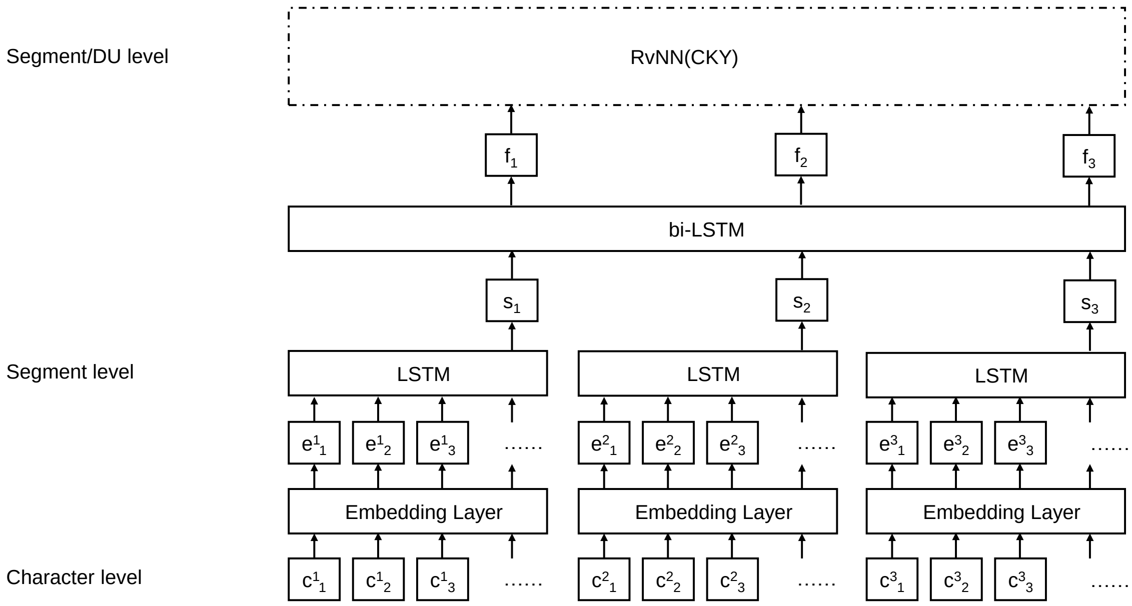

Figure 4.7: The RvNN-CKY2+bi-LSTM model.

course structure, but also to construct syntactic structure for each text segment automatically. This variation of our model is shown in Figure 4.8. Instead of feeding each character embedding e^i_j for the i^{th} segment into a LSTM layer to get the segment embedding s_i, the model processes e^i_j through another RvNN with CKY2 algorithm similar to the process of the original RvNN, resulting in a two-staged RvNN framework. We use the same training set in CDTB, and we can get the gold syntactic parsing for each articles from CTB which is the source corpus for CDTB. Note that we only use this external resource other than CDTB

doi:10.6342/NTU201803852

during the training stage. After training, our model can construct the syntactic structure automatically. We abbreviate this version of our model as **2-Staged-RvNN-CKY2**.

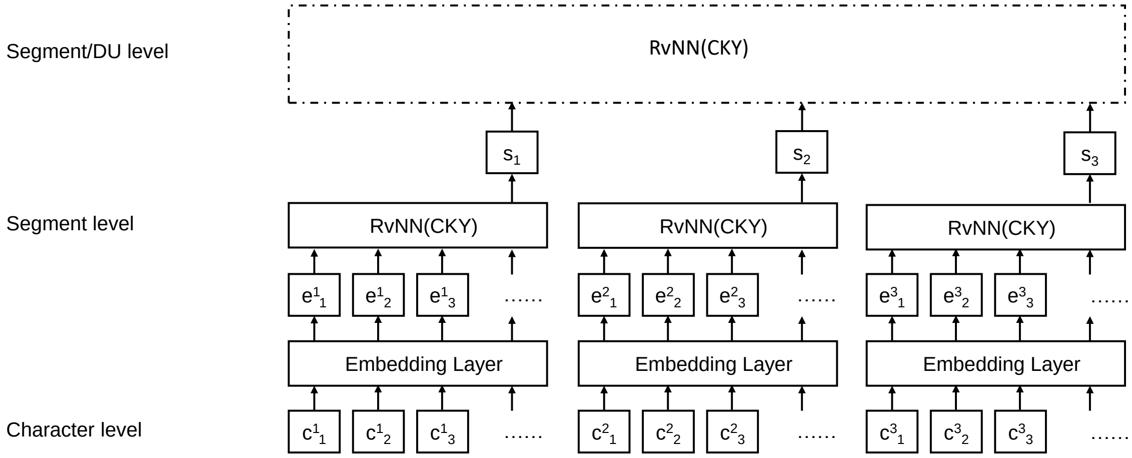

Figure 4.8: The 2-Staged-RvNN-CKY2 model.

In this version of our model, the RvNN process from characters to the whole paragraph through a deep tree structure, as shown in Figure 4.9. We then adapt sampling mechanism to reduce the huge amount of training instances. We sample the character level training instances (the subtrees rooted in a red or blue node in the character level area in Figure 4.9) to the amount as same as the amount of instances above segment level. Also, we do not merge a node above segment level with a neighboring node of character level to construct an instance.

doi:10.6342/NTU201803852

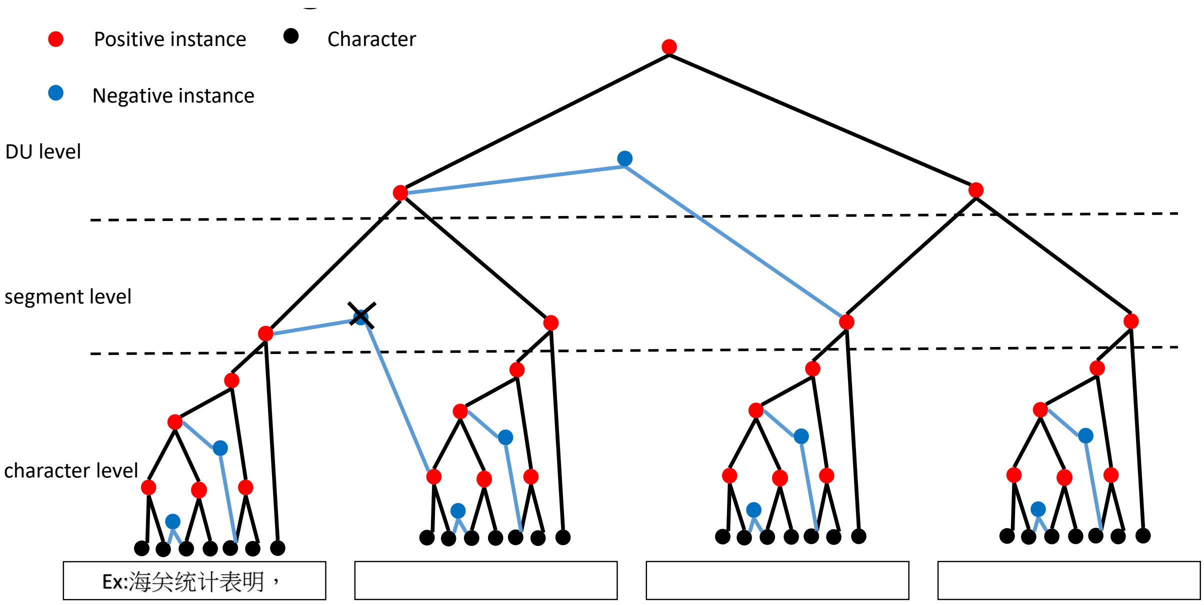

Figure 4.9: The training instances construction for the 2-Staged-RvNN-CKY2 model.

doi:10.6342/NTU201803852

Chapter 5 Experiments

In this chapter, we will describe how we conduct the experiments and discuss the results.

Our experiments have three parts. In the first part, experiments are conducted based on the gold EDU segmentation. We will compare our model with the work of Kang et al. [2016] which is so far the only complete end-to-end Chinese discourse parsing research on the corpus following the CDT scheme. They report their performance in two settings according to where the model extracting syntactic features from. The first one used the Berkeley parser, and we abbreviate this version as **Fang and Zhou Setting 1**. The second one used the gold syntactic trees, and we abbreviate this version as **Fang and Zhou Setting 2**. We will compare these baselines mentioned above with our model **RvNN-CKY2** which stores top two candidates in each entry of the dynamic programing table in our CKY-like algorithm as mentioned in Section 4.6 in Chapter 4. We also consider a simplified version called **RvNN-CKY** which only stores the top candidate, **RvNN-Greedy**, which construct the discourse parsing tree using greedy algorithm.

In the second part, experiments are conducted to construct discourse parsing tree from raw text. **RvNN-CKY2**, **RvNN-CKY2+Seq-EDU**, and **RvNN-CKY2+Bi-LSTM** as mentioned in Section 4.7 in Chapter 4 are tested for comparison. We also use the same baselines as in the first part.

In the last part, we attempt to develop a joint parsing model which learns to parse from syntactic level to discourse model, and to see whether the automatic syntactic structure may benefit the performance of downstream discourse parsing. We conduct this experiment on the **2-Staged-RvNN-CKY2** model mentioned before, and also compare it with

doi:10.6342/NTU201803852

the same baselines.

We implement our models with Pytorch Paszke et al. [2017]. We split the dataset into training and testing parts in the same way as in Kang et al. [2016]. Table 5 shows how the data is split.

	Training	Testing
Docs(#)	450	50
File List	0001-0090,0101-0190,0201-0290, 0301-0325,0400-0454,0500-0509, 0520-0554,0590-0596,0600-0647	0091-0100,0191-0200,0291-0300, 0510-0519,0648-0657
Trees(#)	2,125	217
EDUs(#)	9,616	1,017

Table 5.1: Training and testing data split

5.1 Evaluation Matrix

In this section, we introduce the evaluation matrix. Firstly, for evaluating the EDU detection performance, we calculate the F-score of whether the punctuations are correctly classified to be a boundary of an EDU or not. Secondly, the standard evaluation tool PARSEVAL Carlson et al. [2001] is performed to measure the F-score of the tree structure prediction. Following PARSEVAL, an internal node N_p of the predicted tree is considered to be a true positive if we can find another node N_g in the gold tree, such that N_p and N_g dominate the same span of leaf nodes. In the example of Figure 5.1, the internal nodes are marked red, and the node dominating the span $[1, 3]$ and the node dominating the span $[4, 5]$ are considered as true positives.

5.2 Gold EDU Experiment

Table 5.2 shows the first experimental results. The F-score of EDU segmentation, parse tree construction (Structure), parse tree construction with sense labeling (+Sense), parse tree construction with center labeling (+Center), and parse tree construction with both sense and center labeling (Overall) are reported.

doi:10.6342/NTU201803852

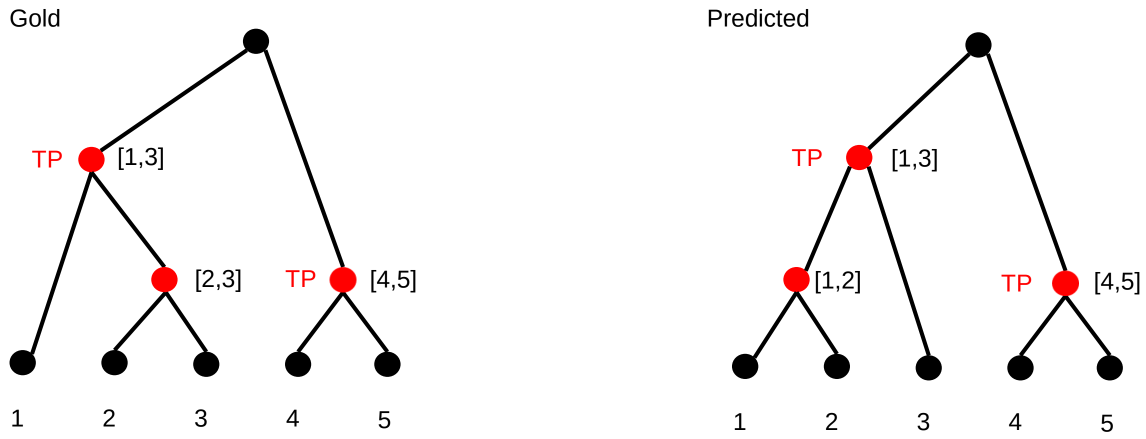

Figure 5.1: Example of the Parseval evaluation matrix.

Model	Structure	+Sense	+Center	Overall
Fang and Zhou Setting 1	52.3%	33.8%	23.9%	23.2%
Fang and Zhou Setting 2	55.6%	**34.5%**	26.5%	24.5%
RvNN-Greedy	58.2%	33.6%	27.0%	24.2%
RvNN-CKY	59.5%	33.0%	27.4%	24.2%
RvNN-CKY2	**60.0%**	34.0%	**27.6%**	**24.8%**

Table 5.2: System performances given the gold EDUs in F-score.

Our models outperform the baseline model most significantly in the **Structure** evaluation while still achieve competitive scores in other three aspects given the gold EDUs. Note that the baselines utilize syntactic information from external parsers or gold standard, so we can infer that our neural network can learn the structure information implicitly and comprehensively without external syntactic knowledge to achieve the better performance even with our simpler **RvNN-Greedy** version.

We can see from the results that the main advantage of our RvNN model is at structure construction. And in comparison with **RvNN-Greedy**, **RvNN-CKY** and **RvNN-CKY2**, we can conclude that the CKY-like algorithm indeed improves the discourse structure construction process, eliminating the gap between the bottom-up construction order and the top-down annotation scheme.

doi:10.6342/NTU201803852

Model	EDU	Structure	+Sense	+Center	Overall
Fang and Zhou Setting 1	**93.8%**	46.4%	28.8%	23.1%	20.0%
Fang and Zhou Setting 2	Not Reported	48.6%	29.0%	23.1%	21.0%
RvNN-CKY2	87.7%	49.2%	27.7%	22.6%	19.6%
RvNN-CKY2 + Seq-EDU	87.6%	**50.7%**	27.8%	**25.7%**	**22.2%**
RvNN-CKY2 + Bi-LSTM	88.9%	48.3%	24.5%	21.2%	18.0%
RvNN-CKY2 + Seq-EDU + Bi-LSTM	87.7%	47.0%	24.2%	21.1%	17.6%

Table 5.3: End-to-End system performances in F-score.

5.3 End-to-End Parsing Experiment

Table 5.3 shows the results of the end-to-end parsing experiment. The EDU detection score of **Fang and Zhou Setting 2** is not reported in the original work. We can see from the results that our models are not competitive with the baselines in EDU detection. It mainly due to the syntactic information being very useful in such task.

However, although standing on the disadvantage of the upstream EDU detection task, **RvNN-CKY2** and **RvNN-CKY2+Seq-EDU** still beat the baselines in the following discourse structure evaluation, and **RvNN-CKY2+Seq-EDU** achieves the best performance when evaluating the **overall** performance. It again shows that our model take significant advantage in discourse structure construction.

Although **RvNN-CKY2+Seq-EDU** does not perform better when predicting EDU boundaries, the left-to-right merging scheme in segments level still produces more powerful representation for latter discourse structure construction. It might due to how we formulate the training instances to be in a form of binary trees, as mentioned in Section 4.4 in Chapter 4.

With the intention to integrate context information of each segments, **RvNN-CKY2 + Bi-LSTM** and **RvNN-CKY2 + Seq-EDU + Bi-LSTM** dose not result in better performance than expected. It might due to the overfitting issue when training the model, and our CKY-like algorithm might has the equivalent effect of dealing with context information, so the use of additional bi-LSTM layer may seem clumsy.

doi:10.6342/NTU201803852

5.4 Joint Parsing Experiment

Model	EDU	Structure	+Sense	+Center	Overall
Fang and Zhou Setting 1	**93.8%**	46.4%	28.8%	**23.1%**	20.0%
Fang and Zhou Setting 2	Not Reported	**48.6%**	**29.0%**	**23.1%**	**21.0%**
2-Staged-RvNN-CKY2	86.1%	34.4%	19.7%	18.8%	16.6%

Table 5.4: System performances in F-score.

Table 5.4 shows the results of the joint parsing experiment. Our model **2-Staged-RvNN-CKY2** performs worse than the baselines and other models in Section 5.3. It might due to two reasons. Firstly, in the **2-Staged-RvNN-CKY2** version, discourse parsing is based on the segment representations resulted in the syntactic parsing performed by the former RvNN. It may suffer from error propagation when comparing to those versions that use sequential LSTM to produce segment representations. Secondly, we only integrate syntactic structure informations while omitting the labels in the syntactic parsing trees. These labels which imply the categories of each words and phrases in a sentence might be important features for EDU detection and the later discourse parsing. For example, verb phrases are important cues to detect an EDU. The **2-Staged-RvNN-CKY2** might need improvements in these two aspects in further research. Still, it is a promising direction to build a model that organizes both syntactic and discourse information, and jointly parses the whole paragraph thoroughly from character level.

5.5 Analysis

We take the **RvNN-CKY2 + Seq-EDU** model which achieve the best overall performance in the previous end-to-end discourse parsing experiment for the following analysis. We will fist analysis the loss of binary-to-multi-way transformation.

Since our framework processes on binary tree structure, but the gold discourse tree nodes may have multiple children, we must perform multi-way-to-binary transformation before training and binary-to-multi-way transformation after predicting, as mentioned in Section 4.4 in Chapter 4. loss might occurs during this process. Table 5.5 shows the eval-

doi:10.6342/NTU201803852

uation of the model on discourse tree node predicting before and after the multi-way-to-binary transformation on both predicted trees and gold trees. **B v.s. B** represents ”compare binary predicted trees with binary gold tree”, **B v.s. M** represents ”compare binary predicted trees with multi-way gold trees”, and **M v.s. M** represents ”compare multi-way predicted trees with multi-way gold trees”.

B v.s. B	**Nodes (Gold)**	**Nodes (Predicted)**	**True Positive**	**Recall**	**Precision**	**F1**
Structure	579	536	285	49.2%	53.2%	51.1%
+Sense	579	536	169	29.2%	31.5%	30.3%
+Center	579	536	155	26.8%	28.9%	27.8%
Overall	579	536	140	24.2%	26.1%	25.1%
B v.s. M	**Nodes (Gold)**	**Nodes (Predicted)**	**True Positive**	**Recall**	**Precision**	**F1**
Structure	478	536	255	53.3%	47.6%	50.3%
+Sense	478	536	139	29.1%	26.0%	27.4%
+Center	478	536	126	26.6%	23.5%	24.9%
Overall	478	536	111	23.2%	20.7%	21.9%
M v.s. M	**Nodes (Gold)**	**Nodes (Predicted)**	**True Positive**	**Recall**	**Precision**	**F1**
Structure	478	386	219	45.8%	56.7%	50.7%
+Sense	478	386	120	25.1%	31.1%	27.8%
+Center	478	386	111	23.2%	28.8%	25.7%
Overall	478	386	96	20.0%	24.9%	22.2%

Table 5.5: Evaluation of the **RvNN-CKY2 + Seq-EDU** model on discourse tree node predicting before and after the multiway-to-binary transformation.

We can see from the table that the F-score drops when transforming the binary gold trees to be multiway. It is worth noticing that the True Positive decreases by 30 in the **structure** evaluation when comparing **B v.s. B** to **B v.s. M**, and the disparity of the evaluation of **+Sense** is also 30. It indicates that all the lost 30 correctly predicted nodes are also correctly tagged with the relation senses. Recalling that the transformation only modified the nodes that not only themselves and also their parents are with the **Coordination** sense, we can infer that our model recognizes this kind of nodes very accurately.

Comparing **B v.s. M** to **M v.s. M**, we can see the gain of the transformation that fit the binary structure of predicted trees to the multiway structure of the gold trees almost equal to the loss of the transformation.

Table 5.6 shows the distribution of relation sense predicting f-scores, and it consider all the nodes in the discourse tree. Table 5.7 also shows the distribution, but it only consider

doi:10.6342/NTU201803852

the nodes whose spans are already correctly predicted by the model. We can see from the two tables that the F-scores is positively related to the instance amount of the relation sense. There is indeed a small predicting bias on the relation senses with more instances, but our model still makes a certain number prediction on each sense except the sense **Explanation** which has only 9 instances in the test dataset.

Sense	Nodes (Gold)	Nodes(Predicted)	True Positive	F1
Coordination	414	435	188	44.3%
Causality	119	48	9	10.7%
Transition	151	111	26	19.8%
Explanation	11	0	0	0.0%

Table 5.6: Distribution of relation sense predicting for all nodes.

Sense	Nodes (Gold)	Nodes(Predicted)	True Positive	F1
Coordination	249	312	188	67.0%
Causality	73	36	9	16.5%
Transition	96	79	26	29.7%
Explanation	9	0	0	0.0%

Table 5.7: Distribution of relation sense predicting for nodes whose span are correctly identified.

Table 5.8 shows the distribution of relation type (Explicit or Implicit) along with the predicting scores, and it consider all the nodes in the discourse tree. Table 5.9 also shows the distribution, but it only consider the nodes whose spans are already correctly predicted by the model. We can see from the two tables that our model doesn't show disadvantage on implicit relations. We can refer that our model did not need to rely on surface lexical cues, such as connectives, to recognize the relation senses.

Explicit/Implicit	Distribution	Recall
Explicit	157	27.4%
Implicit	538	33.5%

Table 5.8: Distribution of relation type and its recall score of sense predicting for all nodes.

doi:10.6342/NTU201803852

Explicit/Implicit	Distribution	Recall
Explicit	99	43.4%
Implicit	328	54.9%

Table 5.9: Distribution of relation type and its recall score of sense predicting for nodes whose span are correctly identified.

5.6 Case Analysis

In this section, we conduct case analysis to discuss the performance of our model. We pick a paragraph from the test set, and then compare the gold discourse parsing tree to the predicted parsing tree output by our **RvNN-CKY2** model.

Figure 5.2 lists the seven EDUs of this sample paragraph. Although the first EDU is a long sentence and consist of a lot of information, the only predicate is [指出... 救災工作。(pointed out that...pollution prevention and relief.)], so it only forms one complete EDU. However, our model mistakenly recognizes three commas in this sentence as EDU boundaries. As a result, our model splits EDU (a) into four parts (a1), (a2), (a3), (a4), as shown in Figure 5.3. It might be because that our model detects the verbs ' 聽取 (lisent to)', ' 調查 (investigate)', ' 做好 (improve)' and makes the judgment to segment an EDU. Note that our model doesn't make mistake on the comma in the segment ' 當前，(at present,)'. It indicates that our model lacks syntactic information to make correct judgments in this condition. Although our model successfully disambiguate all other punctuations, the F-score of EDU detection is only 82.4% which is lower than the overall performance 87.7%.

We then further conduct comparisons on the whole discourse parsing tree. Figure 5.4 shows the gold discourse parsing tree after the multi-way to binary transformation, and Figure 5.5 shows the output discourse parsing tree of our model. We can see from the figures that our model correctly predicts the whole subtree spanning from EDU (b) to (e), and also correctly predicts the merge of EDU (f) and EDU (g), with the relation sense predicted mistakenly. However, the EDU (a2), (a3), (a4) merge to other nodes wrongly through the construction process, resulting in many false positive nodes. The F-score of **structure** is 61.5% which is higher than the overall performance 49.5%.

doi:10.6342/NTU201803852

(a) 宋健在聽取了江蘇、安徽、山東、河南四省及淮河水利委員會的負責人關於淮河流域水污染情況和治理措施，以及國務院調查組關於赴蘇、皖、豫三省調查水污染事故情況的的後指出，當前，江蘇、安徽、山東、河南四省各級人民政府，應進一步做好抗汙救災工作。(Song Jian listened to the water pollution situation and treatment measures of the Huaihe River Basin in Jiangsu, Anhui, Shandong, Henan and the Huaihe River Water Resources Commission, and the State Council investigation team on the investigation of water pollution accidents in the three provinces of Jiangsu, Anhui and Henan. He pointed out that at present, the people's governments at all levels in Jiangsu, Anhui, Shandong and Henan provinces should further improve their work on pollution prevention and relief.)

(b) 首先要保障群眾的生活供水，(First of all, we must protect the people's living water supply,)

(c) 採取綜合措施，(take comprehensive measures,)

(d) 穩定群眾情緒，(stabilize the mood of the people ,)

(e) 保持社會安定。(and maintain social stability.)

(f) 要採取果斷措施，(It is necessary to take decisive measures)

(g) 防止淮河流域再次發生重大水污染事故。(to prevent another major water pollution accident in the Huaihe River Basin.)

Figure 5.2: Sample paragraph consisting of seven EDUs (a), (b), (c), ..., (g).

(a1) 宋健在聽取了江蘇、安徽、山東、河南四省及淮河水利委員會的負責人關於淮河流域水污染情況和治理措施，(Song Jian listened to the water pollution situation and treatment measures of the Huaihe River Basin in Jiangsu, Anhui, Shandong, Henan and the Huaihe River Water Resources Commission,)

(a2) 以及國務院調查組關於赴蘇、皖、豫三省調查水污染事故情況的的後指出，(and the State Council investigation team on the investigation of water pollution accidents in the three provinces of Jiangsu, Anhui and Henan. He pointed out that ,)

(a3) 當前，江蘇、安徽、山東、河南四省各級人民政府，(at present, the people's governments at all levels in Jiangsu, Anhui, Shandong and Henan provinces)

(a4) 應進一步做好抗汙救災工作。(should further improve their work on pollution prevention and relief.)

Figure 5.3: Partitions into four EDUs by our model of the EDU (a) in the sample paragraph

We can see from the above discussion that the advantage of our model is to recognize the Coordination relations, however, when we transform the parsing tree back to multi-way structure, this characteristic may cause significant loss. The Figure 5.6 shows the

doi:10.6342/NTU201803852

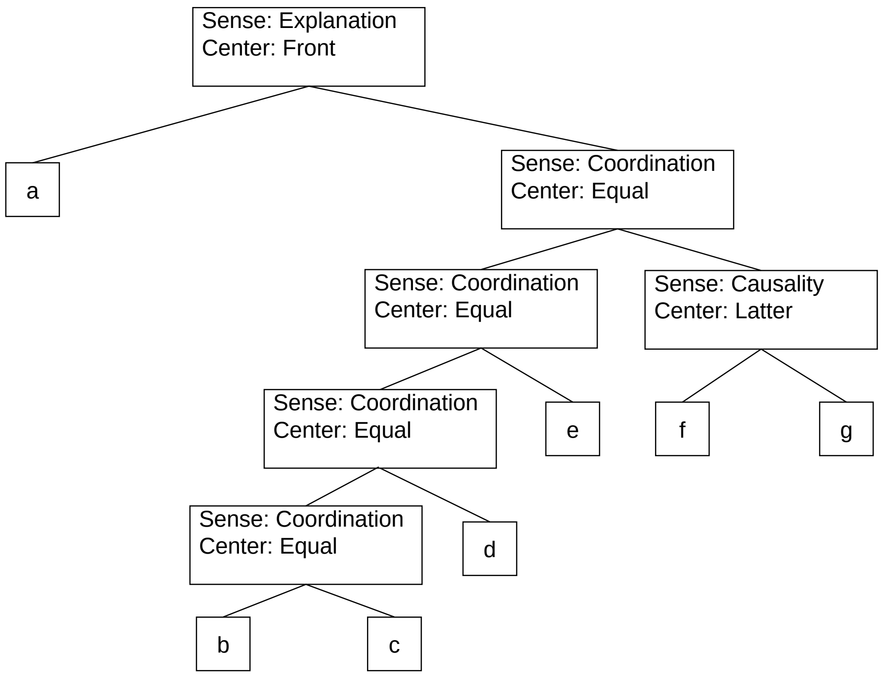

Figure 5.4: The binary form of the gold discourse parsing tree of the sample paragraph

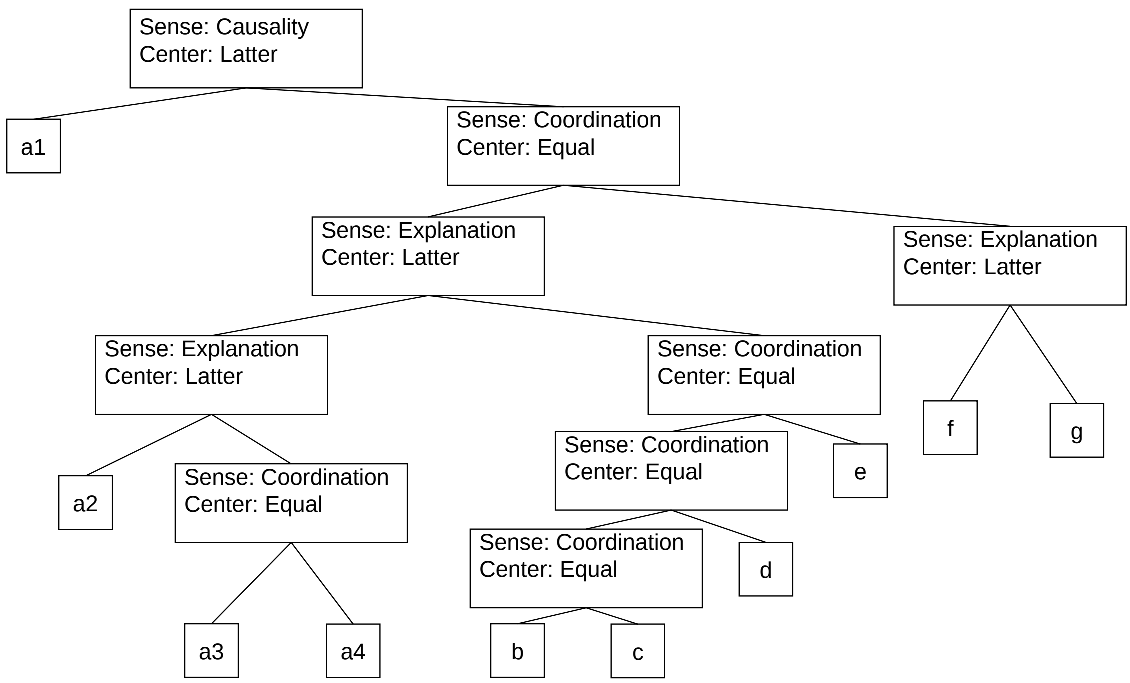

Figure 5.5: The binary predicted discourse parsing tree of the sample paragraph

original multi-way gold parsing tree, and Figure 5.7 shows the predicted parsing tree after the binary-to-multi-way transformation. After the transformation, the subtree dominating EDU (b) to (e) only counts as one true positive, causing the **structure** F-score to drop

doi:10.6342/NTU201803852

from 61.5% to 44.4%, and the **overall** F-score to drop from 46.2% to 22.2%.

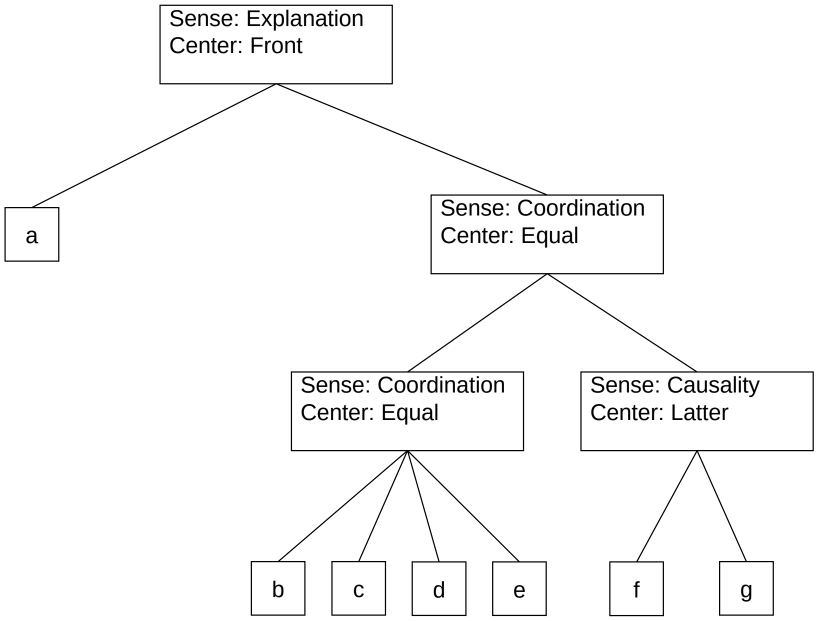

Figure 5.6: The original multi-way gold discourse parsing tree of the sample paragraph.

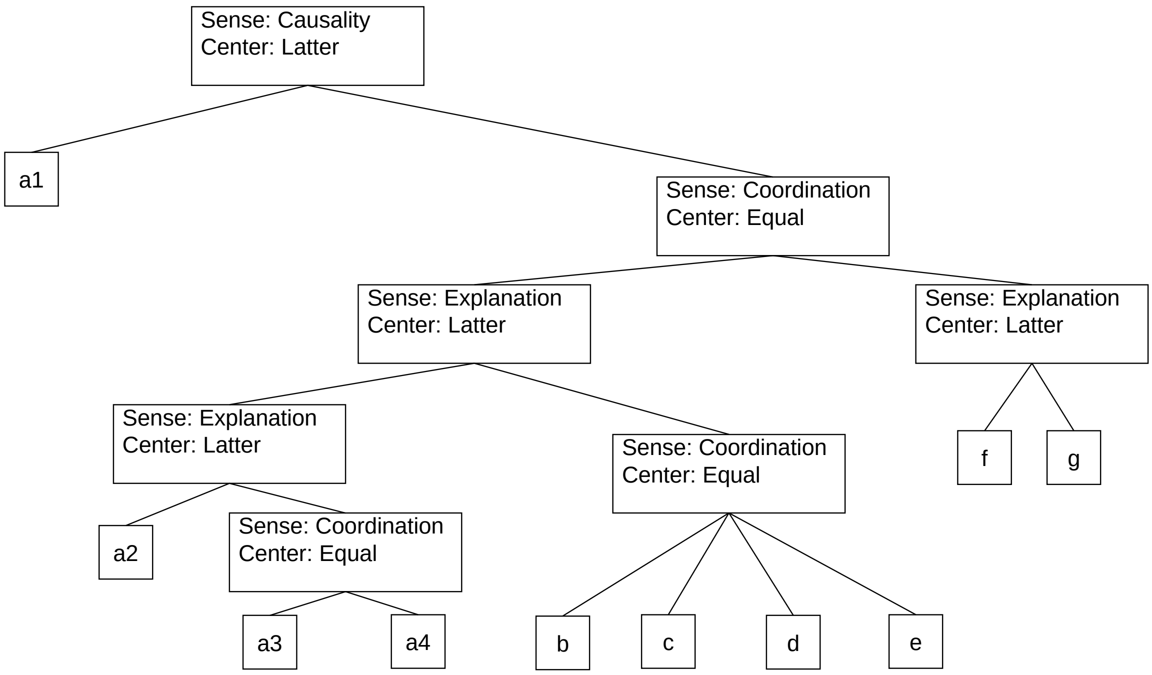

Figure 5.7: The predicted discourse parsing tree of the sample paragraph after binary-to-multi-way transformation.

doi:10.6342/NTU201803852

Chapter 6 Conclusion and Future Work

6.1 Conclusion

This paper demonstrates an end-to-end Chinese discourse parser,[1] which performs the CDT-style parsing without the need of external resources such as syntactic parser.

We propose a unified framework based on RvNN to jointly model the subtasks, avoiding error propagation comparing to previous parsing system. Experimental results show our parser achieves the state-of-the-art performance in discourse structure construction, relation center labeling and the overall performance in the CDTB dataset.

Our model take the advantage of RvNN and CKY-algorithm. The former helps to model the hierarchical discourse structure, and the latter integrate both global and local information during discourse tree construction, eliminating the gap between the bottom-up construction process and the top-down annotation scheme.

In the last part of our experiments, we attempt to modify our model to be a 2-staged RvNN framework. It learns from both syntactic structure and discourse structure of a given paragraph, and jointly parses the whole paragraph thoroughly from character level still without any external parser.

We also show the challenge issue in detecting EDU boundaries for a neural network model since this task has been found performing pretty well with syntactic features.

We release the source code of our parser with pre-trained model for the NLP community. To the best of our knowledge, this is the first toolkit for Chinese discourse parsing.

[1] http://nlg18.csie.ntu.edu.tw/cdp

doi:10.6342/NTU201803852

6.2 Future Work

6.2.1 Enhance the sense classifier and center classifier

In this research, we build an end-to-end parsing model and mainly focus on handing discourse structure. In contrast, the mechanism to label relation sense and center is so far quite simple. In fact, there has been many researches working on relation sense labeling task, as mentioned in Chapter 2. Model enhancement like attention mechanism might be quite likely to improve the performance of relation sense and center labeling.

6.2.2 Fit the parsing model to the multi-way tree structure

As shown in Section 5.5 in Chapter 5, loss occurs when performing binary-to-multi-way transformation. It mainly due to the nature of multi-way discourse structure and our binary RvNN framework. There are two possible direction to reduce this performance reduction. One is to fit our model to the multi-way structure, and it requires modification of our RvNN unit and redesign of the CKY-like algorithm. Time complexity expansion due to this generalization of structure predicting may be a considered issue. Another direction is to optimize the transformation mechanism. Effects may be made to train the model to identify the correct nodes during the transformation.

6.2.3 Integrate syntactic information to build a syntactic-discourse jointly parsing model

In the last part of our experiments, we try to build a syntactic-discourse jointly parsing model. However, the performance is not satisfying on discourse parsing tasks. So far, we only integrate syntactic structure information when training the model while omitting the labels of each internal nodes of syntactic parsing tree. It is still a promising direction to build a model that organizes both syntactic and discourse information, and jointly parses the whole paragraph thoroughly from character level.

doi:10.6342/NTU201803852

Bibliography

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 1466–1477, August 2016. URL http://www.aclweb.org/anthology/P16-1139.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. Cross-lingual rst discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 292–304, April 2017. URL http://www.aclweb.org/anthology/E17-1028.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue (SIGDIAL'01)*, pages 1–10, 2001. . URL https://doi.org/10.3115/1118078.1118083.

C. Goller and A. Kuchler. Learning task-dependent distributed representations by backpropagation through structure. In *Proceedings of the 1996 IEEE International Conference on Neural Networks*, volume 1, pages 347–352 vol.1, Jun 1996. .

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(9):1735–1780, November 1997. ISSN 0899-7667. . URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hen-Hsen Huang and Hsin-Hsi Chen. Contingency and comparison relation labeling and structure prediction in chinese sentences. In *Proceedings of the 13th Annual Meeting of*

doi:10.6342/NTU201803852

the Special Interest Group on Discourse and Dialogue (SIGDIAL'12), pages 261–269, 2012. URL http://dl.acm.org/citation.cfm?id=2392800.2392846.

Yangfeng Ji and Noah A. Smith. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 996–1005, July 2017. URL http://aclweb.org/anthology/P17-1092.

Xiaomian Kang, Haoran Li, Long Zhou, Jiajun Zhang, and Chengqing Zong. An end-to-end chinese discourse parser with adaptation to explicit and non-explicit relation recognition. In *Proceedings of the CoNLL-16 shared task*, pages 27–32, August 2016. URL http://anthology.aclweb.org/K16-2003.

Jiwei Li, Rumeng Li, and Eduard Hovy. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 2061–2069, October 2014a. URL http://www.aclweb.org/anthology/D14-1220.

Yancui Li, Wenhe Feng, Jing Sun, Fang Kong, and Guodong Zhou. Building chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 2105–2114, Doha, Qatar, October 2014b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1224.

Christina Lioma, Birger Larsen, and Wei Lu. Rhetorical relations for information retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, pages 931–940, 2012. ISBN 978-1-4503-1472-5. . URL http://doi.acm.org/10.1145/2348283.2348407.

Annie Louis, Aravind Joshi, and Ani Nenkova. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan, September 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W10/W10-4327.

doi:10.6342/NTU201803852

William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950, 2014.

Yong-Siang Shih. 中文篇章連接詞偵測、消歧、及論元辨識 = *Detection, disambiguation, and argument identification of chinese discourse connectives*. None, 2015. URL https://hdl.handle.net/11296/wsub38.

Yong-Siang Shih and Hsin-Hsi Chen. Detection, disambiguation and argument identification of discourse connectives in chinese discourse parsing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1891–1902, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL http://aclweb.org/anthology/C16-1178.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *CoRR*, abs/1503.00075, 2015. URL http://arxiv.org/abs/1503.00075.

doi:10.6342/NTU201803852

Jianxiang Wang and Man Lan. Two end-to-end shallow discourse parsers for english and chinese in conll-2016 shared task. In *Proceedings of the CoNLL-16 shared task*, pages 33–40, TOBEFILLED-Ann Arbor, Michigan, August 2016. Association for Computational Linguistics. URL http://anthology.aclweb.org/K16-2004.

Yizhong Wang, Sujian Li, and Houfeng Wang. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 184–188, July 2017. URL http://aclweb.org/anthology/P17-2029.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. The penn chinese treebank: Phrase structure annotation of a large corpus. *Nat. Lang. Eng.*, 11(2):207–238, June 2005. ISSN 1351-3249. . URL https://doi.org/10.1017/S135132490400364X.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. Conll 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19. Association for Computational Linguistics, 2016a. . URL http://www.aclweb.org/anthology/K16-2001.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. Conll 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, TOBEFILLED-Ann Arbor, Michigan, August 2016b. Association for Computational Linguistics. URL http://anthology.aclweb.org/K16-2001.

Daniel H Younger. Recognition and parsing of context-free languages in time n3. *Information and control*, 10(2):189–208, 1967.

Kai Zhao and Liang Huang. Joint syntacto-discourse parsing and the syntacto-discourse treebank. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 2117–2123, September 2017. URL https://www.aclweb.org/anthology/D17-1225.

doi:10.6342/NTU201803852

Yuping Zhou and Nianwen Xue. Pdtb-style discourse annotation of chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 69–77. Association for Computational Linguistics, 2012.

doi:10.6342/NTU201803852