Predicting Simplified Thematic Progression Pattern for Discourse Analysis

Xuefeng Xi[1], Victor S. Sheng[1, 2, *], Shuhui Yang[3], Baochuan Fu[1] and Zhiming Cui[1]

Abstract: The pattern of thematic progression, reflecting the semantic relationships between contextual two sentences, is an important subject in discourse analysis. We introduce a new corpus of Chinese news discourses annotated with thematic progression information and explore some computational methods to automatically extracting the discourse structural features of simplified thematic progression pattern (STPP) between contextual sentences in a text. Furthermore, these features are used in a hybrid approach to a major discourse analysis task, Chinese coreference resolution. This novel approach is built up via heuristic sieves and a machine learning method that comprehensively utilizes both the top-down STPP features and the bottom-up semantic features. Experimental results on the intersection of the CoNLL-2012 task shared dataset and the CDTC corpus demonstrate the effectiveness of our proposed approach.

Keywords: Discourse topic, theme-rheme theory, thematic progression.

1 Introduction

Derived mainly from the systemic-functional grammar [Halliday, Matthiessen and Halliday (2004)], theme and rheme are two static entities representing the way in which information is distributed in a simple sentence. While theme indicates the given information serving as the departure point of a message, which has already been mentioned somewhere in discourse or shared as mutual knowledge from the immediate context, rheme is the remainder of the message in a sentence in which theme is developed. From the view point of discourse structure analysis, sequences of thematic and rhematic choices construct certain thematic patterns instead of the actual individual choices of themes or rhemes, which named Thematic Progression Patterns [Danes (1974); Fries (1983); Zhu (1995)].

Over the last few years, discourse structure has been widely studied and proven to be a critical cohesive element at the text level [Carlson, Marcu and Okurowski (2003); Prasad,

[1] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, 215009, China.

[2] Department of Computer Science, University of Central Arkansas, Conway, AR 72035, USA.

[3] Department of Mathematics, Statistics, and Computer Science, Purdue University Northwest, Hammond, IN 46323, USA.

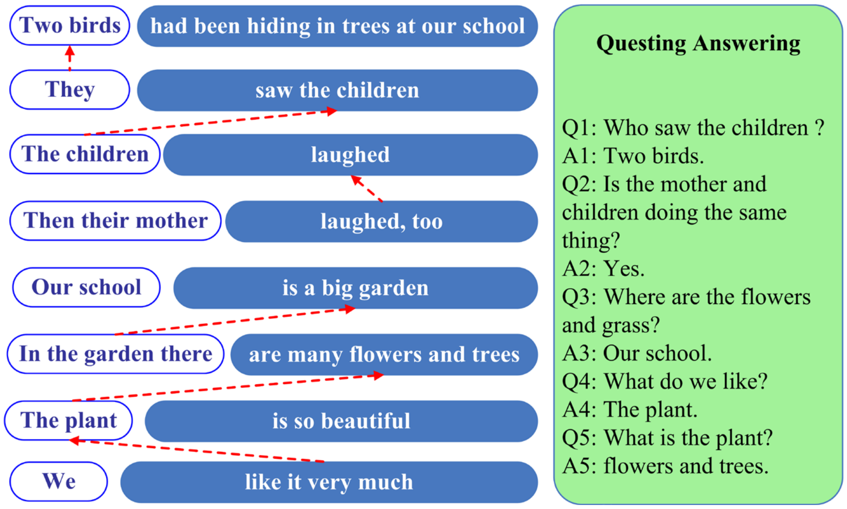
[*] Corresponding Author: Victor S. Sheng. Email: ssheng@uca.edu.

Received: 22 April 2019; Accepted: 29 April 2019.

CMC. doi:10.32604/cmc.2020.06992

Dinesh, Lee et al. (2008); Weischedel, Pradhan, Ramshaw et al. (2011); Song, Jiang and Wang (2010); Zhou and Li (2013); Rutherford and Xue (2015); Lan, Wang, Wu et al. (2017); She, Jian, Zhang et al. (2018)]. A linear segmentation of texts into proper discourse structures may reveal valuable information on, for instance, not only the theme of segment but also the overall thematic progression structure of the text. For example, the thematic progression structure of Example (a) is shown in Fig. 1, and it can subsequently be applied to various discourse analysis tasks, such as question answering, text summarization, information retrieval, etc. [Salton, Singhal, Buckley et al. (1996); Du, Buntine and Johnson (2013); Du, Pate and Johnson (2015); Wang, Li, Lyu et al. (2017); Meng, Rice, Wang et al. (2018)]

Example (a): (1) [Two birds]$_{T1}$ [had been hiding in trees at our school.]$_{R1}$ (2) [They]$_{T2=T1}$ [saw the children.]$_{R2}$ (3) [The children]$_{T3=R2}$ [laughed.]$_{R3}$ (4) [Then their mother]$_{T4}$ [laughed, too.]$_{R4=R3}$ (5) [Our school]$_{T5}$[is a big garden.]$_{R5}$ (6) [In the garden there]$_{T6=R5}$ [are many flowers and trees.]$_{R6}$ (7) [The plant]$_{T7=R6}$ [is so beautiful.]$_{R7}$ (8) [We]$_{T8}$ [like it very much.]$_{R8=T7.}$



Q1: Who saw the children ?
A1: Two birds.
Q2: Is the mother and children doing the same thing?
A2: Yes.
Q3: Where are the flowers and grass?
A3: Our school.
Q4: What do we like?
A4: The plant.
Q5: What is the plant?
A5: flowers and trees.

Figure 1: Thematic progression structure of example (a) and its application in QA

As shown in Fig. 1, each simple sentence contains a *theme node* and a *rheme node*. With the help of arrow links, the semantic relationship of nodes between contexts is expressed. When all links are connected, we get a semantic relation chain. This chain plays a very important role in some NLP applications, e.g., questing answering.

However, because of its poor computability due to a complex definition in linguistics, it hinders the further application of discourse analysis. To address this issue, we simplify the presentation of thematic progression patterns and study some highly competitive computational methods. Subsequently, we use these methods to automatically extract textual thematic progression pattern features. These features are used to implement a major task of discourse analysis, coreference resolution. Our experimental results show that these thematic progression pattern features have a good contribution to discourse analysis.

The rest of the paper is organized as follows. Firstly, we introduce the related works in Section 2. Then, we describe a formal representation of the discourse relationship and how to extract the discourse-level features automatically from this representation in Section 3. After that, we propose a model for the basic task of discourse analysis, coreference resolution, with the help of the above discourse-level features in Section 4. Furthermore, we describe our experiments and show our experimental results in Section 5, Finally, we conclude our work and make a discussion of future work in Section 6.

2 Related works

In natural language processing, coreference resolution is a major task of discourse analysis that identifies which noun phrases or mentions (called anaphors) refer to the same real-world entity or concept (called the antecedent). Traditionally, most of the popular coreference resolution methods based on learning techniques or rules mainly extract the features from different levels of words, phrases and sentences using a bottom-up strategy rather than paying attention to characteristics at the discourse level [Aone and Bennett (1995); Vilain, Burger, Aberdeen et al. (1995); Raghunathan, Lee, Rangarajan et al. (2010); Lee, Peirsman, Chang et al. (2011); Chen and Ng (2012); Bjorkelund and Kuhn (2014); Chen and Ng (2015); Clark and Manning (2016); Lu and Ng (2017); Lee, He and Zettlemoyer (2018); Kundu, Sil, Florian et al. (2018)]. However, in reality, the coreference relationship also reflects the discourse cohesion relationship. If a feature from the discourse level is missing, the anaphora resolution process is incomplete.

Machine Learning Methods Based on the Mention Model The mention-pair model is the first classifier used for machine-based learning for anaphora resolution [Aone and Bennett (1995); Vilain, Burger, Aberdeen et al. (1995)]. This method first identifies mentions from the text, and then forms mention-pairs and extracts relevant features. In its next step, a supervised learning method is used to train a classifier from the feature vectors. Finally, mentions that point to the same entity are clustered into a coreference chain. At present, the mention-pair model is one of the most influential learning-based coreference resolvers.

Heuristic Learning Method Based on Hierarchical Sieves Most anaphora resolution approaches based on the mention-pair model use a single discriminant function. However, this approach does not distinguish the priority order of features, which can lead to misjudgments. For example, such misjudgments occur when the anaphora resolution of B should belong to class A. However, it is instead misclassified into class C, because it first satisfies the characteristics of class C. In order to solve this problem, Raghunathan proposed a new model of anaphora resolution based on a multi-pass sieve framework [Raghunathan, Lee, Rangarajan et al. (2010)]. Lee extended this framework by increasing the number of sieves and adding a post-processing module. This approach achieved the first place in the English anaphora resolution of the CoNLL-2011 shared task [Lee, Peirsman, Chang et al. (2011)].

Chinese Anaphora Resolution Based on Machine Learning and Heuristic Methods In contrast to the CoNLL-2011 shared task, which was intended for developing natural language processing only in the English language, the CoNLL-2012 version of the shared task involved English, Chinese and Arabia simultaneously, which greatly enhances the difficulty of coreference resolution. Inspired by the multi-pass sieve framework model, Chen proposed a joint learning method and applied it to the three different languages and achieved a very good performance [Chen and Ng (2012)]. In particular, this approach won the individual champion of Chinese coreference resolution in the CoNLL-2012 shared task. Subsequently, researchers have also adopted this approach [Bjorkelund and Kuhn (2014); Chen and Ng (2015); Clark and Manning (2016); Lee, He and Zettlemoyer (2018); Kundu, Sil, Florian et al. (2018)].

3 Simplified thematic progression pattern

To explore the discourse relationship, we had proposed a micro-topic scheme (MTS) [Xi and Zhou (2017)] to represent the discourse cohesion structure according to the theme-rheme theory [Halliday, Matthiessen and Halliday (2004)]. Furthermore, we define a simplified thematic progression pattern (STPP) to describe the dynamic association between the contextual discourses.

The MTS scheme can be formalized as a triple as: $MTS=(S_n,S_{n+1},\delta_n)$, where $S_n \in T \cup R$, $S_{n+1} \in T \cup R$, *T* represents the set of the *themes* and *R* represents the set of the *rhemes* in the entire discourse. Here, $\delta_n \in L$, where *L* is the set of cohesion dynamic relationships of MTS between EDTUs, each called a Micro-Topic Link (MTL). To illustrate our proposed MTS, we provide the following example.

Example (b): (1) [[浦东]Satellite 开发开放]$_{\text{T1}}$ [是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，]$_{\text{R1}}$ (2) [<ZeroA>$_{\text{Nucleus}}$(因此)大量出现的]$_{\text{T2(Nucleus)=T1(Satellite)}}$[是以前不曾遇到过的新情况、新问题。]$_{\text{R2}}$ (3) [(对此)，浦东]$_{\text{T3=T2(Nucleus)}}$ [不是简单的采取“干一段时间，等积累了经验以后再制定法规条例”的做法，]$_{\text{R3}}$ (4) [<ZeroA>]$_{\text{T4=T3}}$[而是借鉴发达国家和深圳等特区的经验教训，]$_{\text{R4}}$ (5) [<ZeroA>]$_{\text{T5=T4}}$[<并且>聘请国内外有关专家学者，]$_{\text{R5}}$ (6) [<ZeroA>]$_{\text{T6=T5}}$[<并且>积极、及时地制定和推出法规性文件，]$_{\text{R6}}$ (7) [<ZeroA>]$_{\text{T7=T6}}$[使这些经济活动一出现就被纳入法制轨道。]$_{\text{R7.}}$

(1) [Pudong’s development and opening]$_{\text{T1}}$ [is a century-long undertaking to vigorously promote Shanghai and construct a modern economic, trade, and financial center.]$_{\text{R1}}$ (2) [Consequently,<during the process of [Pudong’s]$_{\text{Satellite}}$ development and opening, >]$_{\text{T2=T1}}$ [huge numbers of previously unencountered new situations and questions have emerged.]$_{\text{R2}}$ (3) [In response, Pudong]$_{\text{T3=T2}}$ [does not simply adopt an approach of “work for a short time and then draw up laws and regulations only after experience has been accumulated.”]$_{\text{R3}}$ (4) [Instead, Pudong]$_{\text{T4=T3}}$[is capitalizing on lessons learned from the experiences of developed countries and special regions such as Shenzhen,]$_{\text{R4}}$ (5) [<ZeroA>]$_{\text{T5=T4}}$ [by hiring appropriate domestic and foreign specialists and scholars,]$_{\text{R5}}$ (6) [<ZeroA>]$_{\text{T6=T5}}$ [to actively and promptly formulate and issue regulatory documents]$_{\text{R6}}$. (7) [<Based on these documents,>]$_{\text{T7=T6}}$ [economic activities are incorporated into the legal system’s sphere of influence as soon as they appear.]$_{\text{R7.}}$

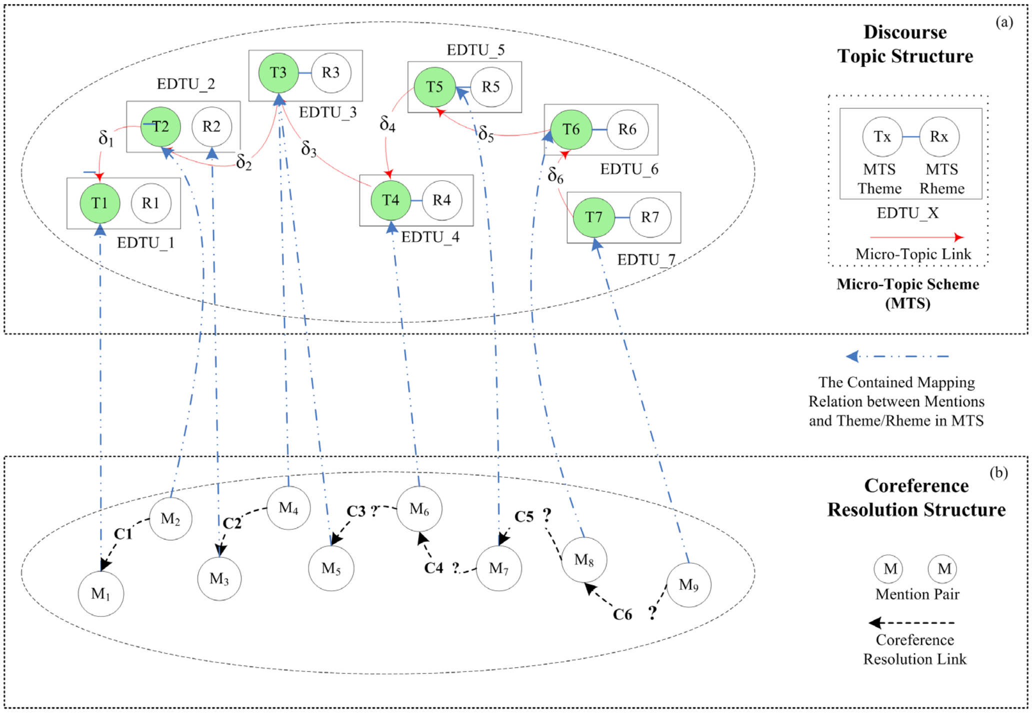


Figure 2: Discourse micro-topic scheme, mention-pair for coreference resoltuion

Part (a) in Fig. 2 gives an example of an MTS representation corresponding to Example (b) as shown above. It consists of 7 clauses, excerpted from chtb0001 which is from the OntoNotes corpus [Weischedel, Pradhan, Ramshaw et al. (2011)]. Here, a clause is equivalent to an EDTU (see Subsection 3.1) constituted by a *theme* and a *rheme* (see Subsection 3.2) and denoted by Tx and Rx, respectively. For instance, "In spite of the fact that the regulatory documents the Pudong new region" stands for the theme in the first clause (a), and the *rheme* occupies the rest of this clause, "has formulated". According to the theme-rheme theory (see Subsection 3.2), there is a reference relationship between the *theme* or the *rheme* of the current EDTU and the previous EDTU. Fig. 2 uses an arrow to indicate this reference by pointing to the *theme* or the *rheme* in the EDTU, for example, T2=T1, T3=T2, T4=T3, T5=T4, T6=T5 and T7=T6.

3.1 Elementary discourse topic unit

Inspired by the Rhetorical Structure Theory, an elementary discourse topic unit (EDTU) is defined as the basic unit of discourse topic analysis, which is limited to clauses. Specifically, an EDTU should contain at least one predicate and express at least one proposition. Moreover, an EDTU should be related to other EDTUs with some propositional function. Finally, an EDTU should be punctuated. In Example (c), (i) consists of a single sentence with a serial predicate, while (ii) is a complex sentence with two EDTUs (clauses).

Example (c)

(i) She started the car. (single sentence, serial predicate, one EDTU)

(ii) She started the car and drove off. (complex sentence, two EDTUs)

3.2 Static entity of the micro-topic scheme

Derived mainly from the systemic-functional grammar Halliday et al. [Halliday, Matthiessen and Halliday (2004)], the *theme* and the *rheme* are two static entities that represent the way in which information is distributed in a clause. The *theme* indicates given information that serves as the departure point of a message that has already been mentioned elsewhere in a text or is shared as mutual knowledge from the immediate context, while the *rheme* is the remainder of the message in a clause in which the theme is developed.

To improve the computational performance, we assume that the *theme structure* is the left part of the predicate in the EDTU in Chinese, and that the remainder is the *rheme structure*.

From a discourse analysis aspect, we are more interested in the fact that sequences of the *thematic* and *rhematic* choices create certain kinds of thematic patterns than in the actual individual choices of the *themes* or the *rhemes*. Therefore, our scheme regarding the notion of the *theme* is discourse-oriented; that is, we are most concerned with the role the *theme* fulfills in constructing and developing a dynamic relationship in a discourse as opposed to its role in individual sentences.

3.3 Simplified thematic progression pattern

Previous studies Fries [Fries (1983); Zhu (1995)] have claimed that the way in which lexical strings and reference chains interact with the *theme* and the *rheme* is not random; instead, the interaction patterns form what they refer to as a thematic progression of a text. In order to reduce computational complexity, we have simplified the presentation of thematic progression and constructed the *Simplified Thematic Progression Pattern* (STPP). Fig. 3 shows this five major dynamic relationship of STPP proposed by us as followed:

-CosTP represents the **TS** of next sentence is associated with the **TS** of the previous sentence.

(a) [Two beggars]$_{T1}$*[were hiding.]*$_{R1}$ *(b) [They]*$_{T2=T1}$*[saw the money.]*$_{R2}$

-SimTP represents the **TS** of next sentence is associated with the **RS** of the previous sentence.

(a)[Our school]$_{T1}$*[has a big garden,]*$_{R1}$ *(b) [in which]*$_{T2=R1}$*[many flowers grow.]*$_{R2}$

-CrsTP represents the **RS** of next sentence is associated with the **TS** of the previous sentence.

(a) [The exhibition]$_{T1}$*[was good.]*$_{R1}$ *(b) [I]*$_{T2}$*[liked it very much.]*$_{R2=T1}$

-CenTP represents the **RS** of next sentence is associated with the **RS** of the previous sentence.

(a) [The children]$_{T1}$*[laughed.]*$_{R1}$ *(b) [Then, their mother]*$_{T2}$*[laughed too.]*$_{R2=R1}$

-**NonTP** represents there are no any relationships between the previous sentence and the next sentence.

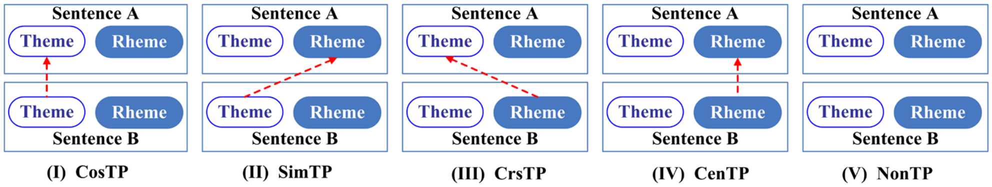


Figure 3: The visual structure of STPP

3.4 Identifying the STPP

In order to automatically extract these features, we define a task of identifying the STPP. This task primarily consists of three subtasks: EDTU Detection, Theme-Rheme Structure Detection, STPP Classification. It is summarized in Fig. 4. Previous studies have shown that the first two subtasks can be converted to the common tasks and have a good performance [Xi and Zhou (2017)], so we focus on the third subtask, STPP Classification.

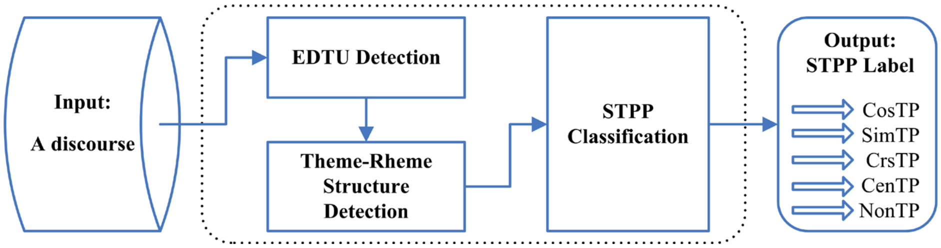


Figure 4: Identifying the STPP

We had studies some classical classifiers for the STPP classification task with features extracting from a corpus annotated ourselves.

3.4.1 Classical classifiers

We applied Scikit-learn 0.19.0[4] to build five classification benchmark experiments using the Liner SVM, Decision Tree, Random Forest, AdaBoost and GBRT (Gradient Boost Regression Tree) algorithms. Prior to this, using imbalanced-learn 0.3.02[5] to achieve resampling in order to obtain balanced data.

3.4.2 Corpus

Based on that STPP Scheme shown in Section 2, we annotated a Chinese discourse topic corpus (CDTC) with 500 discourses from OntoNotes corpus Chinese datasets (chtb0001-chtb0325, chtb0400-chtb0657). Tab. 1 illustrates the inter-annotator consistency

[4] Scikit-learn:http://scikit-learn.org/stable/index.html

[5] Imbalance-learn:http://contrib.scikit-learn.org/imbalanced-learn/stable/index.html

specifically of the CDTC. It is also used for our experiment as dataset. In order to complete the third module task-MTL Classification, we need to use the features of Theme-Rheme Structure, which is annotated as shown in Tab. 2.

Table 1: Inter-annotator consistency in CDTC corpus

Item	Agreement (%)	Kappa
EDTU	96.0	0.91
Theme-Rheme Struucture	92.0	0.83
MTL	89.0	0.86

Table 2: Feature of theme-rheme structure annotated

Name	Value	Description
ID	Integer[1-N]	Identification number
TYPE	[Entity|Event]	Type of Theme-Rheme Structure
POSITION	[Theme|Rheme]	A theme or a Rheme
LOCATION	[Root|NotR]	Is it the first one?
KEY	[Comp.|State.|Nuc.]	Coverage of Theme-Rheme Structure
RTYPE	[NotZ|Zero]	Is empty of Theme-Rheme Structure?
LINKID	Integer[1-N]	ID of the previous associated node
USETIME	Integer[1-N]	Time for one annotated process

3.4.3 Dataset

A total of 9,623 experimental data from CDTC are available, and the number of various types is shown in Tab. 3. It can be seen that there is an unbalanced dataset among various types, with the largest number of classes accounting for 53.75%, and the smallest class accounting for only 0.91%, from Tab. 3. The classifier only focuses on the smallest error of the data and neglects the distribution of the data. In order to balance the various types of data and reduce over-fitting, the SMOTE (Synthetic Minority Oversampling Technique) [Du, Buntine and Johnson (2013)] pair is used. The trained 80% of data is randomly oversampled, making the total number of data for all classes reach 4138. No oversampling is performed on the verified 20% data, but the proportion of each category in the original data set is retained when the data is segmented, i.e., the number of 20% CenTP, CosTP, CrsTP, NonTP and SimTP type data is 18, 701, 17, 1034 and 155 shown in Tab. 4, respectively.

Table 3: Dataset A of simplified thematic progression pattern

	Types of Simplified Thematic Progression Pattern					Total
	CenTP	CosTP	CrsTP	SimTP	NonTP	
Unbalanced	88	3504	85	774	5172	9623

Trained (80%)	70	2803	68	619	4138	7698
Tested (20%)	18	701	17	155	1034	1925

Table 4: Dataset B of simplified thematic progression pattern

	Types of Simplified Thematic Progression Pattern					Total
	CenTP	CosTP	CrsTP	SimTP	NonTP	
Balanced (80%)	4138	4138	4138	4138	4138	20690
Tested (20%)	18	701	17	155	1034	1925

3.4.4 Evaluation metrics

Because it is a multi-classification problem of unbalanced datasets, the experiments used Weighted Macro-averages Precision (WM-Precision), Macro-averages Recall (WM-Recall) and Weighted Macro-averages F1 (WM-F1) as evaluation indicators. The weighted Macro-averages take into account the proportion of categories in the data set. The calculation formula is as follows, where *N* represents the total number of test data sets and *support*ᵢ represents the number of instances of each type in the data set.

$$\text{WM - Precision} = \sum_{i}^{5} \frac{TP_i}{TP_i + FP_i} \frac{\sup port_i}{N} \tag{1}$$

$$\text{WM - Recall} = \sum_{i}^{5} \frac{TP_i}{TP_i + FN_i} \frac{\sup port_i}{N} \tag{2}$$

$$\text{WM - F}_1 = 2 \times \frac{\Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \tag{3}$$

$$Accuracy = \frac{1}{5} \sum_{i}^{5} \frac{TP_i}{TP_i + FP_i + FN_i + TN_i} \tag{4}$$

3.4.5 Results

After various types of balance were achieved, multiple classification experiments were performed. At the same time, F1 value, recall rate, and accuracy of the experiments on two datasets are compared. Using the 5-fold cross validation method, all results in the Tab. 5 are the average of multiple results obtained by running those classifiers for 10 times. The number (+/-) preceded is the average of the cross-validation result followed by the variance of the one.

From the Tab. 5 there are other four classifiers that have achieved better performance except Liner SVM, among which GBRT has outperformed the other systems on various indicators. This may indicate that learning method based on tree structure is more suitable for natural language discourse structure tasks than other ones. Secondly, based on the comparison between Dataset A (unbalanced) and dataset B (balanced), the results of Decision Tree and Random Forest on the balanced data set are slightly better than those of unbalanced data, and the opposite is true for AdaBoost and GBRT. This seems to indicate that the data balanced brought by the oversampling technique is conducive to the training of Decision Tree and Random Forest, but the accompanying data noise also brings trouble to the training of AdaBoost and GBRT at the same time. Finally, four benchmark systems all achieved performance over. 80 on four types of indicators, such as Weight Macro-average Precision, Recall, and Macro-F1. This shows that our proposed *simplified thematic progression pattern* (STPP) has a competitive advantage in computability, which provides a good foundation for the further use of STPP for discourse analysis.

Table 5: Automatically predicting results on dataset A and dataset B

Classifier	DataType	WM-Precision	WM-Recall	Macro-F1
Liner SVM	Dataset A	0.5588(+/-)0.043	0.5405(+/-)0.001	0.3848(+/-)0.002
	Dataset B	0.4706(+/-)0.167	0.0182(+/-)0.002	0.0175(+/-)0.003
Decision Tree	Dataset A	0.8166(+/-)0.007	0.8165(+/-)0.008	0.8164(+/-)0.007
	Dataset B	0.8223(+/-)0.006	0.8134(+/-)0.007	0.8174(+/-)0.006
Random Forest	Dataset A	0.8272(+/-)0.007	0.8397(+/-)0.006	0.8281(+/-)0.006
	Dataset B	0.8410(+/-)0.007	0.8310(+/-)0.008	0.8344(+/-)0.007
AdaBoost	Dataset A	0.8509(+/-)0.006	0.8508(+/-)0.005	0.8398(+/-)0.005
	Dataset B	0.8508(+/-)0.006	0.8167(+/-)0.009	0.8298(+/-)0.007
Gradient Boost Regression Tree	Dataset A	0.8759(+/-)0.008	0.8701(+/-)0.006	0.8565(+/-)0.006
	Dataset B	0.8659(+/-)0.006	0.8322(+/-)0.010	0.8447(+/-)0.008

4 Our proposed model

Coreference resolution is a major task of discourse analysis. We designed a hybrid approach to coreference resolution combined with the sieves-based method and features-based method. Our approach employed a fairly standard architecture, performing mention detection prior to coreference resolution.

4.1 System architecture

Motivated by Chen et al. [Chen and Ng (2012)], our system framework is visualized in Fig. 5. This system takes an input discourse and outputs a confidence score for the mention-pairs. It primarily consists of the following three components: Mention Detection, Coreference Resolution and Post-Processing.

Mention Detection extracts mention objects and related features such as gender, single and complex information, etc., from the input text in preparation for the next join step via sieve-learning and feature-learning.

Coreference Resolution performs anaphora resolution for the mention objects using a joint learning mechanism based on heuristic filtering sieves and machine learning. Specifically, we use the open source implementation of Chen’s system [Chen and Ng (2012)] for the learning filtering sieves and features; however, our approach differs in that we add three new sieves related to the discourse micro-topic scheme to the filtering sieves used in the original system, Tab. 6 lists the heuristic filtering sieves that we use.

Post-Processing is performed mainly to adapt to the specific requirements of the task, e.g., to remove a single mention object.

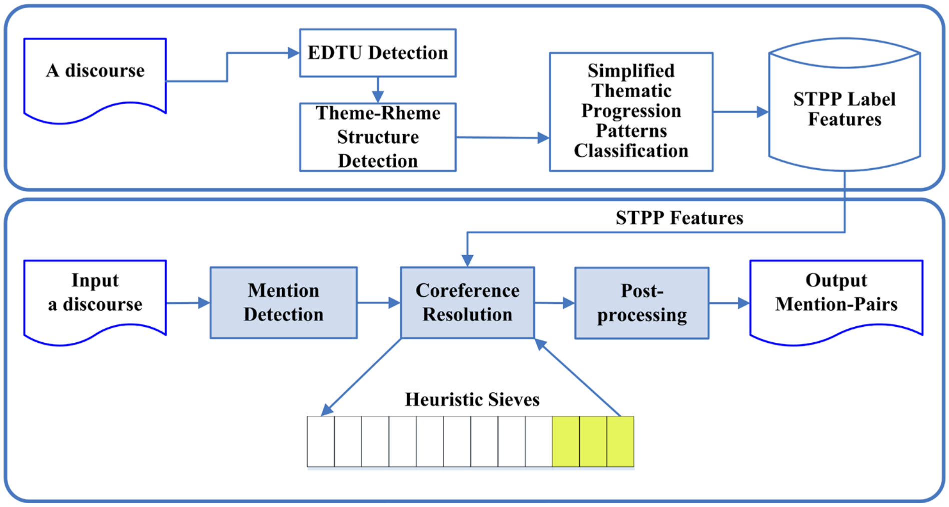


Figure 5: System architecture

4.2 Top-down strategy

Fig. 2(a) shows the micro-topic scheme representation of Example (b), which contains 7 static entities and constitutes 6 STPPs including T1←T2, T2←T3, T3←T4, T4←T5, T5←T6, T6←T7. We call this Relationship-A.

Fig. 2(b) shows the coreference resolution structure from Example (b). It contains the 9 mention objects. Among these, M1-M5 were derived from OntoNotes during the annotation of the mention object, and M6-M9 are annotated in our CDTC corpus. These 9 mention objects constitute the 6 coreference relationships and include M1←M2, M3←M4, M5←M6, M6←M7, and M7←M8, M8←M9. We call this Relationship-B.

The above M1-M9 mention objects are contained in static entities named (M1 in) T1, (M2 in) T2, (M3 in) R2, (M4 in) T3, (M5 in) T3, (M6 in) T4, (M7 in) T5, (M8 in) T6, and (M9 in) T7. According to Relationship-A, if we replace the static entities (Tx or Rx) with mention objects (Mx), then we obtain the new coreference relationships, including M1←M2, M2←M4, M4←M6, M6←M7, M7←M8, and M8←M9. We call this Relationship-C.

When comparing Relationship-B and Relationship-C, we found that both have high similarity. This sparked the motivation that if we already know the discourse micro-topic structure with STPP's feature at the discourse level (e.g., Relationship-A), we can infer the relationships between the lower anaphora resolutions (e.g., Relationship-C). We called this the top-down strategy.

4.3 The sieves

The 12 heuristic filtering sieves used in this paper are shown in Tab. 6. Among these, the first 9 rules are from [Chen and Ng (2012)]. As described in detail below, we introduce the last three new sieves, **GoldRoMTS Match**, **No GoldRoMTS Mismatch**, and **GoldRoMTS plus predicted RoCR**.

Table 6: Definition of the sieves

Order	Sieve Name
1	Chinese Head Match
2	Discourse Processing
3	Exact String Match
4	Precise Constructs
5	Strict Head Match A
6	Strict Head Match B
7	Strict Head Match C
8	Proper Head Match
9	Pronouns
10	**Gold RoMTS Match**
11	**No Gold RoMTS Mismatch**
12	**GoldRoMTS plus Predicted RoCR**

Definition 1: The coreference relationship between two mentions in the discourse is called a Relationship of Coreference Resolution (**RoCR**). Among these, a RoCR identified by the artificial tagging is called **Gold RoCR**, denoted by *RoCR#*, while a RoCR identified automatically by the machine is called **Predicted RoCR** and denoted by *RoCR*.

Definition 2: The micro-topic relationship between a theme and a rheme in the discourse is called a Relationship of Micro-Topic Structure (**RoMTS**). Among these, a **RoMTS** identified by the artificial tagging is called **Gold RoMTS**, denoted as *RoMTS#*, while a **RoMTS** identified automatically by the machine is called Predicted RoMTS and denoted by *RoMTS*.

Definition 3: The new sieve 01 (Gold RoMTS Match) If a Gold RoMTS between theme-rhemes from Mention A and Mention B is established, then the coreference relationship between Mention A and Mention B is established. The formal representation is

$$(\exists A \exists B)(RoMTS^{\#}(A,B)) \rightarrow RoCR(A,B) \tag{5}$$

Definition 4: The new sieve 02 (No Gold RoMTS Mismatch) If a Gold RoMTS between theme-rhemes from Mention A and Mention B is not established, then the coreference relationship between Mention A and Mention B is not established. The formal representation is

$$(\forall A \forall B)\neg(RoMTS^{\#}(A,B)) \rightarrow (\neg RoCR(A,B)) \tag{6}$$

Definition 5: The new sieve 03 (Gold RoMTS plus Predicted RoCR) If a Gold RoMTS between the theme-rhemes from Mention A and Mention B and a Predicted RoCR between Mention A and Mention B are established, then the coreference relationship between Mention A and Mention B is established. The formal representation is shown below:

$$(\forall A \forall B)(RoMTS^{\#}(A,B) \wedge RoCR^{*}(A,B)) \rightarrow RoCR(A,B) \tag{7}$$

5 Experiments

We conducted extensive experiments to investigate the performance of our approach, comparing with two baselines. Before presenting our experimental results, we first provide a brief description of each dataset used in our experiments, and then discuss the implementations and parameter settings of these experiments.

5.1 Datasets

We evaluated the system on the Chinese portion of the corpus provided by the CoNLL-2012 Shared Task and the CDTC corpus that we tagged. The CoNLL-2012 Shared Task corpus considers all pronouns (PRP, PRP$), noun phrases (NP) and heads of verb phrases (VP) as potential mentions. It contains 7 categories of documents (comprising over 2 K documents with 1.3 M words). We used the common part of the official train/dev/test datasets from the CoNLL-2012 Shared Task corpus and our CDTC corpus. To ensure the validity of the test data, we extracted the intersection data sets of the two corpuses, 320 annotated texts (chtb0001-chtb0320), as the experimental data set, named CoNLL-CDTC. Descriptive statistics for the experimental data set are shown in Tab. 7.

Table 7: CoNLL-CDTC dataset employed for this study

Sizes	Docs	Mentions	Chains	Static Entities	MTL
Train	256	25375	8941	12167	2632
Dev	32	3031	1048	1437	318
Test	32	3651	1293	1973	424

5.2 Evaluation metric

As metrics, we adopted MUC [Vilain, Burger, Aberdeen et al. (1995)], B-Cubed [Bagga and Baldwin (1998)] and CEAF [Luo (2005)]. These were the most commonly used metrics for the CoNLL-2011 and CoNLL-2012 Shared Tasks. Moreover, to ensure the validity of our results, we adopted the automatic evaluation program named Scorer (from CoNLL-2012) to calculate the experimental results[6].

5.3 Baseline systems

We applied CoreNLP[7] [Lee, Peirsman, Chang et al. (2011)] and SinoBerryPicker[8] [Chen and Ng (2012)] to the same data set to implement two types of baseline systems, named

[6] http://conll.cemantix.org/2012/

[7] CoreNLP http://stanfordnlp.github.io/CoreNLP/

[8] http://www.hlt.utdallas.edu/- yzcchen/coreference/

BaseX and BaseY, respectively. CoreNLP won the championship in the CoNLL-2011 Shared Task for English, while SinoBerryPicker achieved first place in the CoNLL-2012 Shared Task for Chinese. Tab. 8 shows a comparison between the results of these two baseline systems and the official results from the CoNLL-2012 Shared Task.

The first row in Tab. 8 indicates the official results of Chen's system [Chen and Ng (2012)] on the official Chinese datasets from the CoNLL-2012 Shared Task in Closed mode[9].

Table 8: Comparison of baseline system on gold datasets

Systems	Avg. of F1-Measure	
	Gold Mentions	Gold Mention Boundaries
Chen (Official)	77.77	68.56
BaseX	77.68	68.45
BaseY	70.36	60.33

The second row in Tab. 8 shows the results from our implementation of Chen's system [Chen and Ng (2012)] applied by us to the same CoNLL-2012 Shared Task Chinese datasets. A comparison of these two experimental results shows that the F1 mean values are very close, indicating that our implementation of the BaseX system is reliable and can be used as a baseline system.

Limited to the dataset of the CoNLL-2011 Shared Task, Lee's system [Lee, Peirsman, Chang et al. (2011)] had no experimental results for Chinese data sets. Therefore, we implemented this system and applied it to the CoNLL-2012 Shared Task Chinese datasets. The results are shown in the third row in Tab. 8. The results show that this system has a gap compared with its performance on the English dataset, revealing that the characteristics of the Chinese and English languages are somewhat different.

5.4 Results

Considering that the main purpose of this experiment was to verify the characteristics of the micro-topic scheme and the effect of the new sieves, we pay more attention to the second part of the model, coreference resolution.

To reflect the unique characteristics of the discourse micro-topic scheme, we used the standard Mention dataset from the official conference (Gold Mentions), thereby avoiding potential errors from our implementation of the processing module (Mention Detection) that may have affected its performance.

Gold Mention datasets are divided into two categories: Gold Mention Boundaries and Gold Mentions. Applied to the Gold Mention Boundaries datasets, the experimental results from the two baseline systems and our system on the CoNLL-CDTC are shown in Tab. 9. Then, applied to the Gold Mentions datasets, the experimental results from the two baseline systems and our system are shown in Tab. 10. Compared with the two baseline systems, the average F1 value of our system achieved a better performance.

[9] http://conll.cemantix.org/2012/

To evaluate the contribution of the 3 new sieves to our system's performance, we used different sieves to conduct experiments on the two types of data sets. The results are shown in Tabs. 11 and 12. Among the three new sieves, new sieve 1 substantially improves the recall (R), as shown in Tabs. 11 and 12. The results in the second row reveal the best achievement for sieve 1 according to the MUC valuation metric.

Table 9: Results of our system compared with two benchmarks via gold mention boundaries (supplementary) & closed in CoNLL-CDTC

Systems	**MUC**			**B-CUBED**			**CEAFe**			**Avg.**
	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	
BaseX	70.68	72.06	71.36	73.89	80.15	76.89	58.15	56.55	57.34	68.53
BaseY	65.56	61.86	63.66	67.78	70.34	69.04	45.30	49.67	47.38	60.03
OurSys	73.22	73.46	**73.34**	73.78	82.01	**77.68**	69.22	49.44	**57.68**	**69.57**

Table 10: Results of our system compared with two benchmarks via gold mentions (supplementary) & closed in CoNLL-CDTC

Systems	**MUC**			**B-CUBED**			**CEAFe**			**Avg.**
	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	
BaseX	76.32	92.12	83.48	72.89	91.36	81.09	83.28	57.92	68.32	77.63
BaseY	70.38	88.71	78.49	70.13	89.75	78.74	52.48	56.70	54.51	70.58
OurSys	77.38	92.71	**84.35**	77.63	93.75	**84.93**	77.48	64.70	**70.52**	**79.93**

5.5 Discussion

The overall results of the experiment show that the introduction of features from the discourse micro-topic scheme can help to improve the anaphora resolution performance. This result may occur because the addition of structural features such as text cohesion enrich the original feature space. Because of the lack of representation and corpus resources for discourse structure to facilitate access to the discourse structure, most traditional coreference resolution systems consider only word, syntactic and shallow semantic features; consequently, they are unable to achieve effective applications of the discourse feature.

Fortunately, our proposed discourse micro-topic scheme and construction of the initial corpus resources solve the lack of representation and resource scarcity problems, respectively. Our approach may provide future researchers with a generalized way to make use of discourse level features when conducting discourse analysis research. Based on the results from testing different sieve combinations, the three new sieves based on the micro-topic scheme provide different contributions to the P, R and F1 performances.

Table 11: Results of our system with different new sieves on the gold mention boundaries (supplementary) & closed in CoNLL-CDTC

Systems	MUC			B-CUBED			CEAFe			Avg.
	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	
New Sieve_01	74.50	68.29	71.26	72.98	76.38	77.64	59.78	57.65	58.70	68.20
New Sieve_02	70.58	72.86	71.70	69.88	81.29	75.15	58.99	56.83	57.89	68.25
New Sieve_03	70.89	73.02	71.94	73.39	81.06	77.03	62.32	55.76	58.86	69.28
New Sieve_01&02	74.10	69.81	71.89	72.69	80.67	76.47	59.69	52.44	55.83	68.06
New Sieve_01&03	72.62	73.35	72.98	73.89	81.33	77.43	65.74	50.38	57.04	69.15
New Sieve_02&03	69.94	73.68	71.76	70.68	83.08	76.38	66.45	54.32	59.78	69.31
New Sieve_All	73.22	73.46	**73.34**	73.78	82.01	**77.68**	69.22	49.44	57.68	**69.57**

Table 12: Results of our system with different new sieves via gold mentions (supplementary) & closed in CoNLL-CDTC

Systems	MUC			B-CUBED			CEAFe			Avg.
	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	
New Sieve_01	78.32	90.11	83.80	79.17	90.24	84.34	82.48	56.89	67.34	78.49
New Sieve_02	75.98	92.36	83.37	72.66	91.87	81.14	81.63	59.78	69.02	77.84
New Sieve_03	74.32	93.38	82.77	72.38	93.34	81.53	79.82	62.15	69.89	78.06
New Sieve_01&02	78.01	90.85	83.94	76.33	91.43	83.20	82.19	58.81	68.56	78.57
New Sieve_01&03	78.30	90.59	84.00	73.79	91.48	81.69	78.46	59.39	67.61	77.76
New Sieve_02&03	75.83	93.88	83.90	71.36	92.79	80.68	81.98	63.11	71.32	78.63
New Sieve_All	77.38	92.71	**84.35**	77.63	93.75	**84.93**	77.48	64.70	70.52	**79.93**

The primary reason may be that the coreference relationship is essentially a type of cohesion relationship that belongs to the micro-topic link category. Therefore, it is possible to directly improve the anaphora resolution performance using the sieves related to the micro-topic scheme. For instance, as shown in the 7th row of Tabs. 11 and 12, the new sieves 2 and 3 significantly improve the accuracy of the P value.

6 Conclusions and future work

The discourse topic structure of natural language and the relationships between discourses directly reflect the cohesion of the text, which is closely related to the anaphora resolution. However, the lack of representation of discourse topic structure and corresponding corpus resources make it difficult to provide good features related to discourse topic structure. This lacking information has affected the progress of discourse analysis research.

This paper provides three main contributions. First, to solve the feature representation problem of discourse topics, this paper proposes a formal representation for a discourse micro-topic scheme (MTS) based on theme-rheme theory and thematic progression theory, which is named *Simplified Thematic Progression Pattern* (STPP). Second, based on this proposed topic structure, some machine methods were adopted to automatically extract textual STPP features for a major task of discourse analysis, coreference resolution. Finally, based on previous works, a hybrid approach to Chinese coreference resolution is proposed. This approach not only capitalizes on the traditional shallow features, but also employs the discourse topic features via a top-down strategy and STPP. The results of experiments on the CDTC corpus and on Chinese datasets from the CoNLL-2012 Shared Task show the effectiveness of the proposed approach. Nevertheless, accurately predicting the thematic progression patterns remains a challenging task. To stimulate further research on this task, we will expand the CDTC corpus with crowdsourcing technology in terms of quantity and genre, and further improve the performance of predicting simplified thematic progression pattern with a generative adversarial network.

Acknowledgment: This research has been supported by the National Natural Science Foundation of China under grant 61728205, 61673290, 61672371, 61750110534, 61876217, Science & Technology Development Project of Suzhou under grant SYG201817.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Aone, C.; Bennett, S. W. (1995): Evaluating automated and manual acquisition of anophora resolution strategies. *Proceedings of the Annual Meeting Boston Association for Computational Linguistics*, pp. 122-129.

Bagga, A.; Baldwin, B. (1998): Algorithms for scoring coreference chains. *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pp. 563-566.

Bjorkelund, A.; Kuhn, J. (2014): Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 47-57.

Carlson, L.; Marcu, D.; Okurowski, M. E. (2003): Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Current and New Directions in Discourse and Dialogue*, pp. 85-112.

Chen, C.; Ng, V. (2012): Combining the best of two worlds: a hybrid approach to multilingual coreference resolution. *Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task*, pp. 56-63.

Chen, C.; Ng, V. (2015): Chinese common noun phrase resolution: an unsupervised probabilistic model rivaling supervised resolvers. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 763-774.

Clark, K.; Manning, C. D. (2016): Improving coreference resolution by learning entity-level distributed representations. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 643-653.

Danes, F. (1974): Functional sentence perspective and the organization of the text. *Papers on Functional Sentence Perspective*, pp. 106-128.

Du, L.; Buntine, W.; Johnson, M. (2013): Topic segmentation with a structured topic model. *Proceedings of the* 2013 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 190-200.

Du, L.; Pate, J. K.; Johnson, M. (2015): Topic segmentation with an ordering-based topic model. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2232-2238.

Fries, P. H. (1983): On the status of theme in English: arguments from discourse. *Micro and Macro Connexity of Texts*, pp. 116-152.

Halliday, M. A. K.; Matthiessen, C.; Halliday, M. (2004): *An Introduction to Functional Grammar*. Hodder Education, London.

Kundu, G.; Sil, A.; Florian, R.; Hamza, W. (2018): Neural cross-lingual coreference resolution and Its application to entity linking. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 395-400.

Lan, M.; Wang, J.; Wu, Y.; Niu, Z. Y.; Wang, H. (2017): Multi-task attention-based neural networks for implicit discourse relationship representation and identification. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1299-1308.

Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M. et al. (2011): Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 28-34.

Lee, K.; He, L.; Zettlemoyer, L. (2018): Higher-order coreference resolution with coarse-to-fine inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 687-692.

Lu, J.; Ng, V. (2017): Joint learning for event coreference resolution. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 90-101.

Luo, X. (2005): On coreference resolution performance metrics. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 25-32.

Meng, R.; Rice, S. G.; Wang, J.; Sun, X. (2018): A fusion steganographic algorithm based on faster R-CNN. *Computers, Materials & Continua*, vol. 55, no. 1, pp. 1-16.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L. et al. (2008): The Penn discourse treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp. 2961-2968.

Raghunathan, K.; Lee, H.; Rangarajan, S.; Chambers, N.; Surdeanu, M. et al. (2010): A multi-pass sieve for coreference resolution. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 492-501.

Rutherford, A.; Xue, N. (2015): Improving the inference of implicit discourse relations via classifying explicit discourse connectives. *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL*, pp. 799-808.

Salton, G.; Singhal, A.; Buckley, C.; Mitra, M. (1996): Automatic text decomposition using text segments and text themes. *Proceedings of the Seventh ACM Conference on Hypertext*, pp. 53-65.

She, X.; Jian, P.; Zhang, P.; Huang, H. (2018): Leveraging hierarchical deep semantics to classify implicit discourse relations via a mutual learning method. *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 3, pp. 21-36.

Song, R.; Jiang, Y.; Wang, J. (2010): On generalized-topic-based Chinese discourse structure. *Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp. 23-33.

Vilain, M.; Burger, J.; Aberdeen, J.; Connolly, D.; Hirschman, L. (1995): A model-theoretic coreference scoring scheme. *Proceedings of the 6th conference on Message understanding*, pp. 45-52.

Wang, L.; Li, S.; Lyu, Y.; Wang, H. (2017): Learning to rank semantic coherence for topic segmentation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1340-1344.

Weischedel, R.; Pradhan, S.; Ramshaw, L.; Palmer, M.; Xue, N. et al. (2011): OntoNotes Release 4.0. *Linguistic Data Consortium*, Philadelphia, PA.

Xi, X. F.; Zhou, G. D. (2017): Building a chinese discourse topic corpus with micro-topic scheme based on theme-rheme theory.

https://link.springer.com/article/10.1186%2Fs41044-017-0023-7.

Zhou, G. D.; Li, P. F. (2013): Improving syntactic parsing of Chinese with empty element recovery. *Journal of Computer Science and Technology*, vol. 28, no. 6, pp. 1106-1116.

Zhu, Y. S. (1995): Patterns of thematic progression and text analysis. *Foreign Language Teaching and Research*, vol. 3, pp. 6-12.