Topic Structure Identification of PClause Sequence Based on Generalized Topic Theory*

Yuru Jiang[1,3] and Rou Song[1,2]

[1] Computer School, Beijing University of Technology, Beijing, China
jiangyuru@bistu.edu.cn, songrou@blcu.edu.cn
[2] Information Science School, Beijing Language and Culture University, Beijing, China
[3] Computer School, Beijing Information and Science Technology University, Beijing, China

Abstract. To solve the problem of topic absence at the beginning of Chinese Punctuation Clause(abbreviated as PClause), this study, with due regard to the characteristics of topic structure and the stack model having been clearly explained by Generalized Topic Theory, proposes a scheme for identifying the topic structure of PClause sequence. The accuracy rate for open test is 15 percent higher than the baseline, which proves the effectiveness of employing Generalized Topic Theory in identifying the topic structure of PClause sequence.

Keywords: PClause sequence, generalized topic, topic structure, topic clause.

1 Introduction

The study of discourse structure plays a crucial role in language engineering, including but not limited to summarization, information extraction, essay analysis and scoring, sentiment analysis and opinion mining, text quality assessment, as well as machine translation[1]. A common practice adopted by present studies is to decompose the text into small units such as sentences, phrases and words, which are selected as features in statistical methods or machine learning approaches. However, the characteristics of the discourse structure are rarely exploited.

Chinese discourses are characterized with a high frequency of anaphora, especially zero anaphora[2], so that when a Chinese discourse is decomposed into sentences, some anaphoric components will be missing. This has been a big problem affecting the discourse-related NLP applications. Chinese linguists have done a lot of theoretical researches on the zero anaphora in Chinese from four aspects, namely, syntax, pragmatics, discourse analysis and cognitive linguistics. But the characteristics and distribution rules of zero anaphora having been found are hard to formalize and hence inapplicable in computerization. On the other hand, many insightful statistics-based and rule-based studies on anaphora resolution in Chinese by the NLP researchers

* This study is supported by National Natural Science Foundation of China, subject No. 60872121, 61171129 and 61070119.

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 85–96, 2012.

exploiting linguistics knowledge are largely focused on the resolution of pronouns and nouns, with quite little research on the resolution of zero anaphora[3].

With regard to the characteristics of Chinese discourses, Generalized Topic Theory[4] sets punctuation clauses (PClauses hereafter), which have clear boundaries, as basic units of Chinese discourse, and proposes the concepts of generalized topic and topic clause, so that such characteristics as the discourse structure and topic clauses are explicitly described. Within this framework, a stack model of the dynamic generation of topic clauses is devised, providing theoretical basis and formal approach for Chinese discourse analysis.

Based on the Generalized Topic Theory, a study on identifying the topic clause of individual PClause has been done[5]. In that work, a topic clause corpus has been constructed and a scheme for constructing a candidate topic clause(CTC) set has been devised. Then semantic generalization and editing distance are employed to select the correct topic clause. Experiments yield an accuracy rate of 73.36% for open test, which has great significance for discourse related Chinese processing. On the basis of this work, this paper presents a study on identifying the topic structure of PClause sequences.

In the rest of the paper, section 2 briefly introduces the Generalized Topic Theory and its concepts relevant to this study; section 3 describes the scheme for identifying the topic structure of a PClause sequence; section 4 presents the corpus, baseline and evaluation criteria of the identification experiment; section 5 shows the experiment result and the analysis on it; and the last section provides a summary and future work.

2 Generalized Topic Theory

2.1 PClause Sequence

The basic unit of Chinese discourse is PClause, which is a string of words separated by punctuation marks of comma, semicolon, period, exclamation mark or question mark or direct quotation marks. [6]

E.g.1. (Fortress Besieged by Ch'ien Chung-shu)
这几个警察并不懂德文，居然传情达意，引得犹太女人格格地笑，比他们的外交官强多了。
(These policemen knew no German, but were trying to flirt, made the Jews women giggle, were much better than their diplomats.)[1]

In E.g.1, the discourse fragment consists of four PClauses, and its PClause sequence can be represented as below:

c_1. 这几个警察并不懂德文， (These policemen knew no German,)
c_2. 居然传情达意， (but were trying to flirt,)

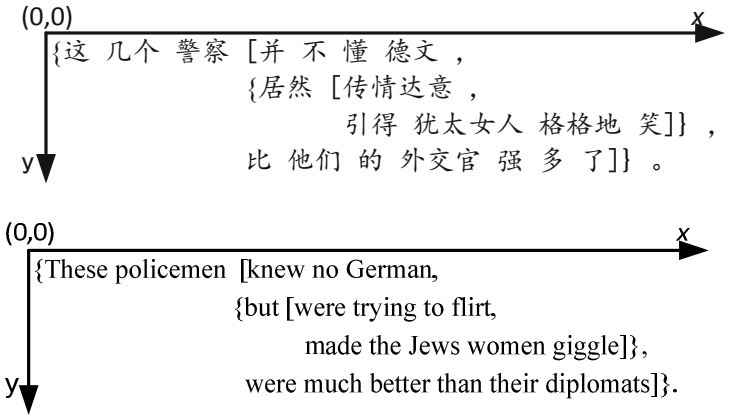
[1] Since word order differs significantly in Chinese from in English, which directly affects the topic clause recognition, the translation of examples in this paper will be direct, keeping as much syntactic information in Chinese as possible.

c_3. 引得犹太女人格格地笑, (made the Jews women giggle,)
c_4. 比他们的外交官强多了。 (were much better than their diplomats.)

Some PClauses can stand alone as a sentence, having a complete structure of topic--comment. For example, in c_1, the topic is "这几个警察(these policemen)", the comment is "并不懂德文(knew no German)". But some PClauses, which may miss sentence components, cannot stand alone. c_2, for example, is just a comment. Its topic "这几个警察(these policemen)" is in c_1.

2.2 Topic Structure

The topic structure of Chinese is the syntactic structure of PClause sequence, which is composed of a generalized topic and one or more comments. A comment itself can be a topic structure too, so that one topic structure can be nested in another topic structure. Such a structure can be represented by indented new-line representation[6]. For instance, E.g.1 can be represented as below.



In the above graph, what is quoted by the "[]"marks is comment, the left of which is the topic. And what is quoted by the "{}"marks is the topic structure. The left part of c_1 is at x=0, and the other PClauses are indented to the right edge of its topic.

In the indented new-line representation, if the x value of the topic of a topic clause is 0, then it is the outmost topic of the topic structure.

2.3 Topic Clause

If all the missing topic information, including the outmost topic, of each PClause is filled up, then the complete structure is termed as a topic clause.

Given a PClause c_i and its topic clause t_i, each PClause and its respective topic clause are:

c_1.这几个警察并不懂德文， t_1.这几个警察并不懂德文，
(These policemen knew no German,)
c_2.居然传情达意， t_2.这几个警察居然传情达意，
(These policemen but were trying to flirt,)
c_3.引得犹太女人格格地笑， t_3.这几个警察居然引得犹太女人格格地笑，
(These policemen but made the Jews women giggle,)
c_4.比他们的外交官强多了。 t_4.这几个警察比他们的外交官强多了。
(These policemen were much better than their diplomats.)

Below is the generating process of the above topic clauses.

t_1=c_1;

The topic of c_2 is"这几个警察(these policemen)" in t_1. Remove the right side of this topic in t_1 and concatenate the rest with c_2, and we will have t_2;

The topic of c_3 is "居然(but)" in t_2. Remove the right side of this topic in t_2 and concatenate the rest with c_3, and we will have t_3;

The topic of c_4 is"这几个警察(these policemen)" in t_3. Remove the right side of this topic in t_3 and concatenate the rest with c_4, and we will have t_4.

If we regard the beginning and the end of a topic clause respectively as the bottom and top of a stack, then the removing and concatenating actions in the generating process of topic clause are typical stack operations. Therefore, the generating process of topic clause can be formalized by a stack model.

From the above procedure, given a PClause c, its topic clause t can be c itself, but it can also be a string concatenating the former part of the preceding topic clause t_{pre} with c. Hence a topic clause can be defined as

$$t = tw_1^i \backsim c \tag{1}$$

Here tw_1^i is a string of i words in the former part of t_{pre}, $i \in [0,n]$, and $\backsim$ is an operator that concatenates the two strings together. When i=0, tw_1^i is an empty string. In the following part of the paper, i will be referred to as PClause depth.

3 Identification Scheme

Strategies employed in identifying the topic clause of individual PClause[5] include the following strategies: using stack model to generate CTCs, using Edit Distance to calculate similarity, using semantic generalization to solve the problem of data sparse, and using completeness of topic clause restriction to select the optimum CTC. This will be the basis for our work to devise the topic clause identification scheme of PClause sequence.

3.1 Identification Objective

For a PClause sequence, if its topic clause sequence can be identified, then the topic structure of the PClause sequence can be obtained. Therefore, the objective of this

paper is to identify the topic clause sequence for a given PClause sequence, viz., given a PClause sequence $c_1,\ldots,c_n$, where the topic clause of c_1 is c_1 itself, then the objective is to identify the topic clause of each PClause $t_1,\ldots,t_n$.

3.2 Identification Process

The process of identifying the topic clauses of a PClause sequence can be stored and represented as a tree. E.g.2 is the first four PClauses (in word-segmented form) from an article about "raja kenojei". The identification process of this PClause sequence is shown in Fig.1.

E.g.2. raja kenojei (from China Encyclopedia)
c_1. 斑鳐 是 鳐形目 鳐科 鳐属 的 1 种 。
c_2. 吻 中长 ，
c_3. 尖 突 。
c_4. 尾 细长 ，
(c_1. raja kenojei is rajiformes rajidae raja de one species,
c_2. snout medium-sized,
c_3. tip projecting.
c_4. tail slim and long.)

The topic clause of c_1 is t_1, same as c_1, which is the root node in Fig.1.

According to formula (1), for the given t_1 and c_2, since there are eight words in t_1, the value range of i in tw_1^i that generates t_2 is [0,8], and the following strings can be generated as the CTCs of c_2.

[1].吻 中长 ，
[2].斑鳐 吻 中长 ，
[3].斑鳐 是 吻 中长 ，
[4].斑鳐 是 鳐形目 吻 中长 ，
[5].斑鳐 是 鳐形目 鳐科 的 吻 中长 ，
[6].斑鳐 是 鳐形目 鳐科 鳐属 吻 中长 ，
[7].斑鳐 是 鳐形目 鳐科 鳐属 的 吻 中长 ，
[8].斑鳐 是 鳐形目 鳐科 鳐属 的 1 吻 中长 ，
[9].斑鳐 是 鳐形目 鳐科 鳐属 的 1 种 吻 中长 ，
([1].snout medium-sized,
[2].raja kenojei snout medium-sized,
[3].raja kenojei is snout medium-sized,
[4].raja kenojei is rajiformes snout medium-sized,
[5].raja kenojei is rajiformes rajidae snout medium-sized,
[6].raja kenojei is rajiformes rajidae raja snout medium-sized,
[7].raja kenojei is rajiformes rajidae raja de snout medium-sized,
[8].raja kenojei is rajiformes rajidae raja de one snout medium-sized,
[9].raja kenojei is rajiformes rajidae raja de one species nout medium-sized,)

They form the nodes on the second level in the tree shown in Fig.1. (For convenience, only 3 nodes are displayed.)

Similarly, the CTCs of c_3 can be generated by using c_3 and the CTC of c_2 (although t_2 still uncertain yet, it must be a node on the second level in the tree). For instance, given a CTC on the second level in the tree, such as "斑鳐 是 鳐形目 吻 中长 (raja kenojei is rajiformes snout medium-sized)", and c_3, we can have the CTCs shown in Fig.1 A part; given "斑鳐 是 吻 中长 (raja kenojei is snout medium-sized)", and c_3, we can have the CTCs shown in Fig.1 B part; given "斑鳐 吻 中长 (raja kenojei snout medium-sized)", and c_3, we can have the CTCs shown in Fig.1 C part. They form the nodes on the third level in the tree shown in Fig.1. By the same token, a tree of CTCs can be generated for the text 斑鳐(raja kenojei).

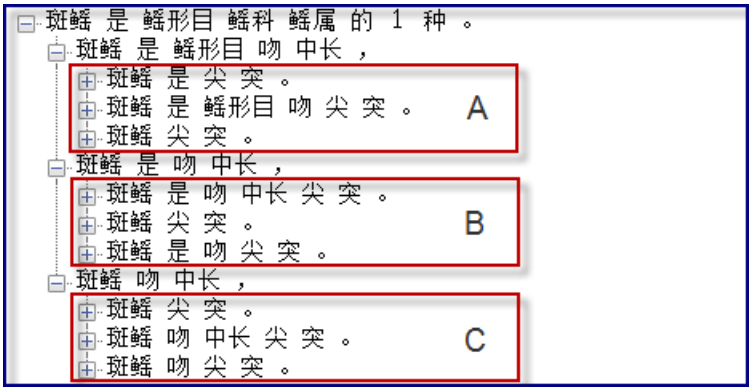


Fig. 1. Part of the CTC Tree of Example 2

Adopting proper strategies to calculate the value of each node in the CTC tree, we can then calculate the path value of each leaf node to the root node. The path with the largest path value can be found and the nodes on it from the root node to the leaf node are the topic clause sequence to be found.

In this way, the task of identifying the topic clause of a PClause sequence is converted to searching the maximum path in the tree.

3.3 Recognition Algorithm

Input: PClause Sequence $c_1,\ldots,c_n$
Output: topic clause sequence $t_1,\ldots,t_n$
1. Generate CTC tree of a PClause Sequence

As section 3.2 describes, given a PClause sequence $c_1,\ldots,c_n$, a CTC tree can be constructed correspondingly, which has n levels. On each level, there will be a number of nodes, and each node can be described with the 5-tuple $<k,ct,v,n_1,n_2>$, which corresponds to one CTC of c_k , where k is the level id, with $1 \le k \le n$, ct is the string sequence of the CTC, v is the path value from the root node to the current node, n_1 is

the sequence number of the node in level k, and n_2 is the sequence number of the father node in level k-1.

Level 1: [<1,ct_1,0,1,0>], $ct_1=c_1$.

Given level k-1 is [<k-1,ct_1,v_1,1,n_1>, <k-1,ct_2,v_2,2,n_2>,…, <k-1,ct_m,v_m,m,n_m>], the k^{th} level will be generated by three steps.

Step 1: Crude Generation – to construct all possible CTC nodes from every node on level k-1 and c_k. Supposing p nodes have been generated on level k from the i-1 nodes on level k-1, and the word sequence ct_i, of the CTC on the i^{th} node on level k-1 is [tw_1,…,tw_s], then the j^{th} node on level k generated from node <k-1,ct_i,v_i,i,n_i> is

<k,$tw_1^{j-1} \frown c_k$,v_i+$score(tw_1^{j-1} \frown c_k)$,$p$+$j$,$i$>

where $1 \leq j \leq s+1$, tw_1^0=nil, tw_1^j=[tw_1,…,tw_j], and score is a scoring function for CTC, the definition of which please refer to section 3.4.

Step 2: Pruning – to delete the nodes of low score by some strategies.

Two pruning strategies are adopted.

a. single-node-based pruning: For all the CTC nodes constructed from a node on level k-1 and c_k, if the number of nodes is greater than 3, only the top 3 nodes with the largest node value are kept, the rest being pruned. If two CTCs have the same value, then the shorter one is given priority.

b. level-based pruning: If the number of nodes on level k-1 is greater than 50, then the top 50 nodes with largest path value are kept and the rest are deleted.

Step 3: Sorting – After the above steps, the nodes on level k are sorted in descending order by v value. The sorted nodes on this level are then numbered starting from one.

2. Generating CTC Sequence

Given that the set of CTC nodes on level i in the tree is marked as $tcs(i)$, then $t_n=ct_i$, where <n,ct_j,v_j,j,n_j> $\in$ $tcs(n)$, and $v_j=max\{v_p | <n,ct_p,v_p,p,n_p> \in tcs(n)\}$.

For a known t_k ($2 \leq k \leq n$), then $t_{k-1}=ct_r$, where <k,t_k,v,j,n_j> $\in tcs(k)$, <k-1,ct_r,v_r,r,n_r> $\in tcs(k-1)$, $r=n_j$.

So that $t_1=c_1$.

3.4 CTCs Scoring Function

In the previous work[5], a Topic Clause Corpus (Tcorpus) is used and two approaches of similarity calculation and semantic generalization are adopted to select the optimum topic clause from the CTCs.

Given a CTC d of PClause c, a topic clause most similar to d is found from the corpus, whose similarity is marked as $sim\_CT(d)$. For any two strings x and y, given that their similarity is $sim(x,y)$. $sim\_CT(d)$ is defined as

$$sim\_CT(d) = \max_{t \in Tcorpus} sim(d,t) \tag{2}$$

From the CTC generating process it can be seen that the topic clause of a PClause is related to the PClause itself and the topic clause of precedent PClause which are the context of the topic clause. Therefore, when identifying the topic clause of a PClause, the context in which the topic clause is generated must be taken into account.

Given a CTC *d*, the PClause *d_c* from which *d* is generated, and the topic clause of the PClause preceding *d_c* is d_t_{pre}, the context similarity of *d* is defined as

$$ctxSim\_CT(d) = \max_{t \in Tcorpus} (\lambda_1 sim(d,t) + \lambda_2 sim(d\_c, t\_c) + \lambda_3 sim(d\_tc_{pre}, t\_tc_{pre})) \quad (3)$$

where *t* is the topic clause in Tcorpus, *t_c* is the PClause from which *t* is generated, and t_t_{pre} is the topic clause of the PClause preceding *t_c*.(Please note that in Tcorpus, for the sake of calculating the context similarity of *d*, *t_c* and t_t_{pre} are kept as well as the information of topic clause *t*.)

Experiments have been done which separately adopted *sim_CT(d)* and *ctxSim_CT(d)* as the scoring functions, and the result shows that the accuracy rate using the latter is 11.3% higher (see section 4). For E.g.3:

E.g.3.

d_tc_{pre}： A 一般 均 具 H 或 H C ,
(A usually is equipped with both H or H C ,)

d_c： 用以 引诱 食饵 。(for baiting.)

t_1： A 一般 均 具 **H** 用以 引诱 食饵 。
(A usually all is equipped with H for baiting.)

st_1： A C 一般 具 **H** ， (usually A C is equipped with H,)

t_2： A 一般 均 具 H 或 H C 用以 引诱 食饵 。
(A is usually be equipped with H or H C for baiting.)

The above is the generalized form of some sentences, which is used in similarity calculation. Alphabet symbols are the generalized mark for words, e.g. A stands for fish terms, H for appearance and C for body parts. The none-alphabetical words are not generalized in current stage. t_1 is the topic clause of *d_c* identified by *sim_CT*, but is not the correct one. The topic clause most similar to t_1 is st_1 in Tcorpus, but the similar parts of them (the shaded words) has nothing to do with *d_c*.

The correct topic clause is t_2, which is identified by *ctxSim_CT*. The topic clause in Tcorpus most similar to t_2 in E.g.3 and its context are as below:

t_tc_{pre}： A 有些 B C 具 C， (A has some B C is equipped with C,)

t_c： 以 引诱 食饵， (for baiting.)

t： A 有些 B C 具 C 以 引诱 食饵 ,
(A has some B C is equipped with C for baiting,)

It can be seen that *d_c* is very close to *t_c*, t_2 and *t* share many similar components, and that d_t_{pre} and t_t_{pre} also have some components in common (similar components are underlined). Therefore the topic clauses in Tcorpus identified by *ctxSim_CT* can more objectively test whether the CTC is the right one for the PClause at hand. In other words, *ctxSim_CT* is better than *sim_CT* in CTC evaluation.

4 The Experiment Process

4.1 Corpus

This study exploits 202 texts about fish in the Biology volume of China Encyclopedia, which consists of 9508 PClauses whose topic clauses are manually tagged. A modern Chinese general-purpose word segmentation system[7] developed by Beijing Language University is used for word segmentation and generalization. To ensure the accuracy rate for word segmentation, the original GPWS vocabulary bank, ambiguous word bank and the semantic tag bank are extended.

15 texts are used for test in the experiment. When identifying the topic structure of one text, the topic clauses in the rest 201 texts constitute the training corpus Tcorpus. There are 717 PClauses and 46 topic structures in these 15 texts. On average, each topic structure consists of 15.59 PClauses. In the 717 PClauses, 452 share the component of fish names, a proportion of 63.04%.

4.2 Baseline

In the texts about fish in the encyclopedia, the topic clause of a PClause is mostly obtained by simply concatenating the fish name (the title of the text) to the beginning of the clause. Therefore, the baseline is defined as

baseline=number of PClauses whose topic is the text title/total number of PClauses (4)

According to the statistics on the topic clauses of the 9508 PClauses in the 202 texts about fish, the number of PClauses whose topic clause is the PClause concatenated with the text title is 5786. Therefore, the baseline is 5786/9508=0.6085.

4.3 Evaluation Criteria

For N PClauses, if the number of PClauses whose topic clauses are correctly identified is hitN, then the identification accuracy rate is hitN/N.

5 Experiment Result and Analysis

5.1 Experiment Result

The result of open test on 15 texts is shown in Fig.2. If *sim_CT(d)* is used as the scoring function, the accuracy rate is 64.99%, 4.14 percent points higher than the baseline. But if *ctxSim_CT(d)* is used as the scoring function, when λ_1=0.5, λ_2=0.4, λ_3=0.1, the accuracy rate reaches 76.25%, 15.44 percent points higher than the baseline.

5.2 Analysis

(1) The reason for the low accuracy rate for texts about barbinae

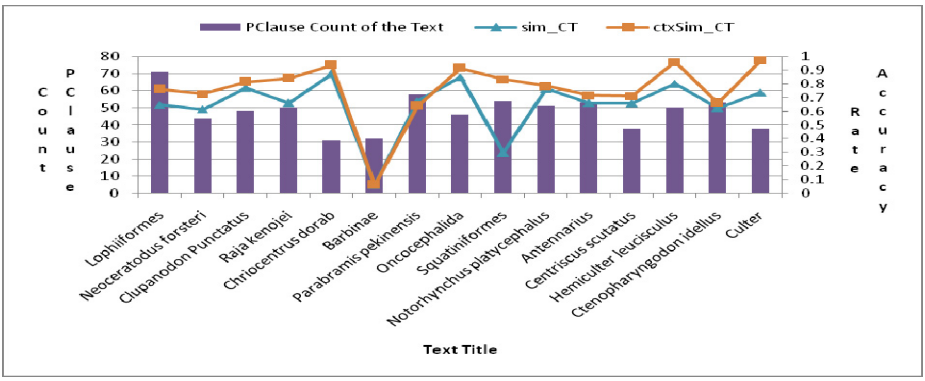


Fig. 2. PClause Count and Accuracy Rate for Topic Clause Identification about 15 texts

Most PClauses have only one topic clause, so that in the experiment there is only one correct answer for each PClause. However, some PClauses may have more than one CTCs that can be regarded as the correct answer. For example:

E.g.4.

tc$_{pre}$. 本亚科 鱼类 通称 鲃 。
(Fish of this subfamily are generally called barbels.)

c. 体 近 纺锤形， (have a spindle-shaped body,)

ctc$_1$. 本亚科 体 近 纺锤形， (this subfamily have a spindle-shaped body,)

ctc$_2$. 本亚科 鱼类 体 近 纺锤形，
(Fish of this subfamily have a spindle-shaped body,)

In E.g.4, *ctc*$_1$ and *ctc*$_2$ are two correct answers. However, *ctc*$_2$ is the specified right answer while *ctc*$_1$ is the one selected by program. In text about barbinae, there are 23 such "mistakes" have taken place. Taking this into consideration, the number of correctly identified topic clauses should be 25, reaching an accuracy rate of 78.13%. Other texts also have similar issues.

(2) On some levels in the CTC tree, there may be nodes with the same CTC string. In extreme cases, all the nodes in one level have the same CTC string.

For example, on the 3rd level in Fig.1, the node "斑鳐 尖 突 (raja kenojei tip projecting)" appears three times. In some texts for testing, there are cases that in the CTC tree, some levels may have nodes that contain identical CTC information. If the CTC is not the correct one, and the topic information for the subsequent PClauses is absent, a chain of errors will be caused.

Therefore, to ensure the heterogeneity of the nodes on each level is an issue to be considered in future work. A plausible approach could be that when constructing the nodes on each level in the CTC tree, same brother nodes, if any, will be merged into one node while keeping the total number of nodes on each level unchanged. The CTC

tree will thus be transformed into a CTC graph, which preserves more path information with the space complexity unchanged.

(3) The relation between accuracy rate and the PClause position(Sequence number of the PClause in the text).

From the PClause identification process, esp. the construction process of CTC tree, it can be seen that the accuracy rate on an upper level may affect that on the lower levels. It appears that the farther a PClause is from the beginning, the lower the accuracy rate for its topic clause identification. But as a matter of fact, there is no positive correlation between the PClause position and the accuracy rate for topic clause identification.

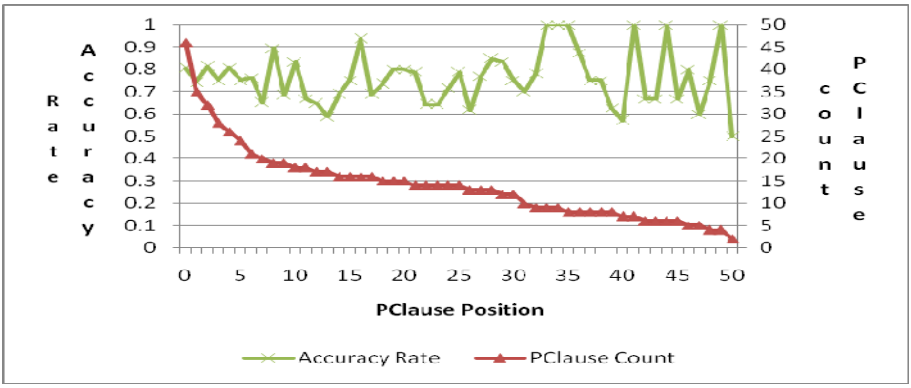


Fig. 3. PClause Position, PClause Count and the Accuracy Rate for Topic Clause Identification

(4) The relation between the accuracy rate and the PClause depth

Fig. 4 shows that the PClause depth may contribute to the decline of accuracy rate. There are as many as 139 PClauses with depth of 2, the accuracy rate for their topic clause identification being as low as 53.96%.

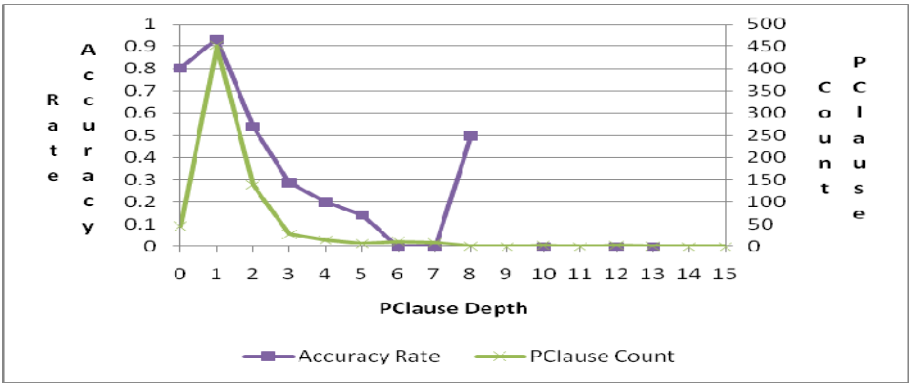


Fig. 4. PClause Depth, PClause Count and the Accuracy Rate for Topic Clause Identification

6 Summary and Future Work

This paper briefly describes Generalized Topic Theory, on the basis of which it proposes a research scheme for identifying the topic structure of PClause sequence. Correspondingly, experiments are devised and completed. In the study, tree structure is adopted to store the CTC sequence generation process. Edit Distance and semantic generalization are employed as the basis for context-based similarity calculation to evaluate the CTC nodes in the tree. Finally, by building the scoring function for the nodes and with proper pruning strategies, the topic structure of PClause sequence is identified with satisfying experiment results.

However, there are some aspects where this work needs to be improved. First, the values of λ_1, λ_2 and λ_3 are given empirically in the context similarity calculation. More scientific methods should be found to calculate the values reasonably. Second, it is a question how to keep heterogeneity of the nodes on each level in the tree. In addition, the achievement of topic clause identification in encyclopedia texts about fish can be extended to other encyclopedia corpora. Further efforts should be made to probe into the application of this experiment scheme to more fields.

References

1. Webber, B., Egg, M., Kordoni, V.: Discourse Structures and Language Technology. Journal of Natural Language Engineering 1, 1–40 (2012)
2. Chen, P.: Discourse Analysis of Zero Anaphora in Chinese. Zhongguo Yuwen 5, 363–378 (1987)
3. Huang, X., Zhang, K.: Zero Anaphora in Chinese—the State of Art. Journal of Chinese Information Processing 23(4), 10–15 (2009)
4. Song, R., Jiang, Y., Wang, J.: On Generalized-Topic-Based Chinese Discourse Structure. In: 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp. 23–33. Tsinghua University Press, Beijing (2010)
5. Jiang, Y., Song, R.: Topic Clause Identification Based On Generalized Topic Theory. Journal of Chinese Information Processing 26(5) (2012)
6. Song, R.: Research on Properties of Syntactic Relation Between PClauses in Modern Chinese. Chinese Teaching in the World 2, 26–44 (2008)
7. A modern Chinese general-purpose word segmentation system v3.5, `http://democlip.blcu.edu.cn:8081/gpws/`