An Event-oriented Multi-Pass Sieve Module for Coreference Resolution

Qiang Li, Zongtian Liu, Lei Chen, Xianchuan Wang
School of Computer Engineering and Science
Shanghai University
Shanghai, China
e-mail: 603265406@qq.com

***Abstract*—Coreference resolution is one of the key issues in the natural language processing, it can eliminate uncertain problems of event in the event-oriented natural language processing, and that is important for the upper application of event. This paper builds an event-oriented multi-pass sieve module for coreference resolution, and combined with the characteristics of the event, we add the constraint conditions to each sieve to improve the accuracy of each sieve. We use this module to carry on experiment for object elements of the event. Compared with machine learning method based on C4.5 decision tree, it has a very big enhancement on the performance.**

Keywords-event-oriented; object element; multi-pass sieve ; coreference resolution

I. INTRODUCTION

Coreference is a common language phenomenon in natural language, it links a concise language unit to a complex language unit in the context. That makes the language expression concise and coherent, and has a distinct level. But various kinds of coreference increase the difficulty that computers understand natural language. Therefore, it is necessary to find the different expression of the same entity to eliminate this obstacle. That is the purpose of coreference resolution.

In the event-oriented natural language processing, as a knowledge representation unit, events include many event elements such as object, time and environment and so on. There are a lot of coreference for events and event elements. Coreference in the event, for example, an agent object is pronoun in the event, but it can't give a concrete object information for event-based applications such as reasoning and automatic abstracting because of the abstractness of pronouns. That leads to many difficulties when computer understands the whole article and analysis the relationship between events. The research of event-based coreference resolution contributes to solve these problems and improve the performance of event-based applications.

This paper mainly focuses on the research of object element coreference resolution in the event. Coreference has a variety of expression forms. In the event-oriented text, there are four types for the coreference of object elements: the same description coreference, the different description coreference, the contraction coreference and the pronoun coreference. The same description coreference is the exact string match; The different description coreference needs to determine whether has a coreference relation according to the context and the semantic relation between mentions; The contraction coreference refers to two elements which have coreference relation have a part of same string, but not all string; The pronoun coreference refers to the type which the anaphor element is pronoun and the antecedent element is the concrete object. According to these four types of coreference, this paper builds the multi-pass sieve module.

II. RELATED WORK

Through the development of nearly 50 years, it has made some achievement for coreference resolution at home and abroad, especially with the development of international conferences about coreference resolution, like MUC, ACE, ARE, it obtained the fast development. Throughout the researchs of these years, the method of coreference resolution is basically divided into two broad categories: the method based on heuristic rules of linguistics and the method based on data driven [1]. If it can get feature informations of high quality, it can obtain good effect regardless of what kind of method.

Foreign researchers study coreference resolution in the early days, and put forward many classical algorithms. Hobbs put forward a coreference resolution algorithm about English personal pronouns [2], combining with grammar rules to address coreference resolution in the syntax analysis tree; Lappin et al. propose a RAP algorithm [3], it obtains grammatical structure of documents through slot grammar which is proposed by McCord, then implements resolution of the third person pronoun and the reflexive pronoun in inner-sentence or between sentences through calculating salience of candidate antecedent and determine antecedent by filter rule; Soon provides a complete implementation steps of coreference resolution system based on classification algorithm of the decision tree in 2001 [4], and obtains good effect; Raghunathan et al. propose a simple model of coreference resolution based on multi-pass sieve frame in 2010 [5], the result in standard test set is better than the method of machine learning; Lee et al. extend Raghunathan's model [6], and obtain highest precision in CoNLL-2011 shared task.

Domestic researchers study coreference resolution is later than foreign researchers, but also obtain a lot of achievements. According to the characteristics of Chinese, Houfeng Wang et al. adopt the method of weakening the language knowledge which is proposed by Mitkov to address

978-1-4673-9323-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISKE.2015.56

coreference resolution of personal pronouns [7]; Junsheng Zhou uses unsupervised clustering algorithm through introducing a weighted graph to address coreference resolution of noun phrases [8]; Muyu Zhang et al. propose a competition model, fuse head constraints into instance matching algorithm [9], and improve the effect.

Above methods mostly only identify a type of coreference, this paper uses the method which is proposed by Lee et al., combines with the characteristics of the event, and builds a multi-pass sieve module to address four types of coreference mentioned above. Sieves in this module are sorted from highest to lowest precision, each tier builds on the entity clusters constructed by previous models in the sieve. And combined with the characteristics of the event, we add the constraint conditions to each layer to improve the accuracy in each layer. In addition, we builds a classification method based on C4.5 decision tree, and use it as comparison. Through testing on CEC corpus, we find that the performance of the former is better than the latter.

III. RELATED DEFINITION

Definition 1 (Event) [10] We define event as a thing happens in a certain time and environment, which some actors take part in and show some action features. Event e can be defined as a 6-tuple formally:

e :: =（A，O，T，V，P，L）

We call elements in 6-tuple event factors. A means an action set happen in an event. It describes the process of event happens. These actions executed in sequence or concurrently while event happens. O means objects take part in the event, including all ac-tors and entities involved in the event. T means the period of time that event lasting. The time period includes absolute time and relative time. V means environment of event, such as location of event. P means assertions on the procedure of actions execution in an event. Assertions include pre-condition, post-condition and intermediate assertions. L means language expressions, including Core Word Set, Core Words Expressions and Core Words Collocations.

Definition 2 (Idea Event) An idea event is an event including idea language which is generated in one's mind. The expression way of idea language is verbal expression, or describes in words, or leaves in heart by self-knowledge.

Idea language: the content that actor expresses ideas, views, attitude and the describing facts. It can be expressed as: idea language = {{narration} {idea event}}. Idea language is make up of narration or idea event.

Narration: it can be translated into first-order predicate which describes event contents.

Definition 3 (event trigger) It also called event pointer word or event core word. It is a word which can clearly express the happened event in the text. In general case, it is a main verb in the sentence (may also be a noun), and it describes the event directly.

Definition 4 (antecedent element and anaphor element) If coreference relations between two elements is existed in event-oriented text, antecedent element is a specific element, and anaphor element is an abstract element.

Definition 5 (event-oriented coreference resolution) It is a process which can look for the relation between antecedent element and anaphor element in event-oriented text, and giving the antecedent element which is pointed by anaphor element.

IV. THE FRAMEWORK OF COREFERENCE RESOLUTION

This system is an event-oriented Chinese coreference resolution platform, the corpus adopts CEC(Chinese Event Corpus) which is annotated coreference relations. This system consists of three main modules: the first module is mention detection with the purpose of identifying candidate antecedent elements. There are two semantic types for object element in the event: participant and object, the former is associated with people, and the latter is associated with objects or people, they should be handled separately; The second module is coreference resolution, it consists of five sieves, and sieves in this module are sorted from highest to lowest precision, each tier builds on the entity clusters constructed by previous models in the sieve. And combined with the characteristics of the event, we add the constraint conditions to each layer to improve the accuracy in each layer; The third module is post-processing, it is performed to adjust our output to the form of annotated coreference relations in CEC.

The whole system's framework as show in the figure 1 below.

A. Mention Detection Module

This module's purpose is identifying candidate antecedent element, its accurate identification can improve the performance of the whole system. The importance of this paper is researching and investigating coreference resolution module. However, this module can bring into inaccurate candidate antecedent element or miss parts of candidate antecedent element. To some extent that will influence the building of the second module, and is not conducive to the full study of each layer. In order to ignore the influence of this module, we use standard corpus which is annotated by manual as test corpus.

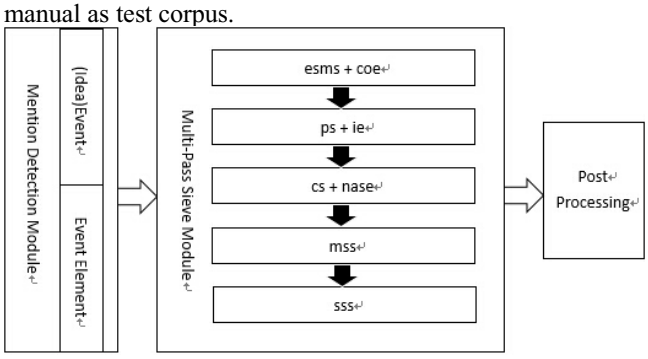

Figure 1. The framework of event-oriented multi-pass sieve for coreference resolution.

B. Coreference Resolution Module

This module consists of five layers: exact string match sieve, pronoun sieve, contraction sieve, morpheme similarity

sieve and semantic similarity sieve. Sieves in this module are sorted from highest to lowest precision, each tier builds on the entity clusters constructed by previous models in the sieve. And combined with the characteristics of the event, we add the constraint conditions to each layer to improve the accuracy in each layer. Each layer of this module addresses different coreference type. The first layer addresses the same description coreference, the second layer addresses the pronoun coreference, the third layer addresses the contraction coreference, the last two layers addresses the different description coreference. The following will be introduced respectively.

1) Exact String Match Sieve (esms)

This model links two elements only if they are the exact string match e.g., [中国地震局新闻发言人张宏卫]←[中国地震局新闻发言人张宏卫]([China Seismological Bureau news spokesman Hongwei Zhang]←[China Seismological Bureau news spokesman Hongwei Zhang]), but except for pronoun elements, such as 他(he), 这些(these) etc.. The identified method of this coreference is relatively simple. When antecedent element and anaphor element are the exact string match, they are coreference relation.

2) Pronoun Sieve (ps)

This layer addresses the pronoun coreference, mainly including personal pronoun, demonstrative pronoun etc.. The identified instance, e.g., [被撞伤男孩父亲]←[他] ([the father of injured boy] ←[he]). The identified method is that when anaphor element is pronoun, the closest element above which is not pronoun is its antecedent element.

3) Contraction Sieve (cs)

Two elements have coreference relation when they have a part of same string, but not all, e.g., [四川省地震局]←[省地震局]([Seismological Bureau of Sichuan province] ← [Provincial Seismological Bureau]). Identifying this coreference needs two kinds of rules, the first rule is AB←B, e.g., [四川省地震局]←[省地震局]([Seismological Bureau of Sichuan province] ← [Provincial Seismological Bureau]); The second rule is AEB←AB, e.g., [哥斯达黎加红十字会]←[哥红十字会]([The Red Cross in Costa Rica]←[the CR Red Cross]). Although this two kinds of rules can't cover all, they cover most of this kind of coreference. We find that introducing other rules can reduce the accuracy in the experiment.

4) Morpheme Similarity Sieve (mss)

This layer addresses the different description coreference, this characteristic only exists in the event. If a trigger word in one event and the core word of object element in another event is semantic similarity, object elements in these two event have coreference relations. The core word of object element is noun before the verb in the element, e.g., "困"(trap) in "被困人员"(trapped person). The identified instance, e.g., [修理巷道的 20 名矿工]←[被困人员] ([20 miners of repairing roadways]←[trapped person]), the trigger word in the event which including the object element "修理巷道的 20 名矿工"(20 miners of repairing roadways) is "困"(trap), and it is semantic similarity with "困"(trap) which is in object element "被困人员"(trapped person).

5) Semantic Similarity Sieve (sss)

This layer identifies coreference relations by HowNet, e.g., [西藏]←[自治区]([Xizang]←[municipality]). Because object elements have many modifiers, it need to extract the core noun. The core noun is the last noun in the word sequence after word segmentation.

C. Post Processing Module

This module is performed to adjust our output to the form of annotated coreference relations in CEC. The system make object elements which points to the same entity put into a coreference chain, e.g., A←B←C, but in CEC it is separate, e.g., A←B and B←C. So we should make their forms consistent. Detaching the coreference chain which system provides to adjust to the form of annotated coreference relations in CEC.

D. Constraint Conditions

In the second module, although each layer can identify most of coreference relations, the limitations of the characteristics also lead to many inaccurate identification and unidentified coreference for the first three layers. So we add the constraint conditions to each layer to improve the accuracy in each layer.

1) The Correlation of Elements (coe)

This condition aim at exact string match sieve. For elements which represented as quantifiers and nouns, most of them are not coreference relations in this sieve, e.g., "11 人"(eleven persons), "一个男子"(a man). The effect of string match is very accurate for object elements which expresses specifically. But above type of object elements is abstract, and can't express the concrete information of object. If we only use string match rules, this sieve will identify them as coreference relation, so we should add constraint conditions to eliminate it.

This constraint condition works by comparing correlations of trigger words and other existent elements in the event which including this type of object element. Firstly identifying this type of object elements by use tokenizer called Nlpir to obtain part of speech of each word. Except for the limitation of quantifiers and nouns, ensuring that number of words can't be more than four words, that can ensure abstraction of object elements. For this type of coreference, if the trigger word, time elements and location elements are semantic similarity respectively in the event including this type of object element, they have coreference relations. In fact, time elements and location elements may be elliptical, so comparing the semantic similarity of the trigger word is very general in most cases. As show in Fig. 1, the object element in the event labeled e8 and the object element in the event labeled e9 are exact string match, but they don't point to the same entity obviously. But they can eliminate by above condition. We use a dictionary called tongyici cilin to compare semantic similarity.

```
<Event eid="e8">
    6名病危人员已有(In the six critically ill persons)
    <Participant sid="s8">3人(three persons)</Participant>
    <Denoter type="stateChange" did="d8">脱离危险(are out of danger)</Denoter>
</Event>:
<Event eid="e9">
    5名病重人员已有(In the five critically ill persons)
    <Participant sid="s9">3人(three persons)</Participant>
    明显(are obviously)
    <Denoter type="stateChange" did="d9">好转(take a turn for the better)</Denoter>
</Event>。
```

Figure 2. The correlation of elements.

2) Idea Event (ie)

This condition aim at pronoun sieve. This sieve use recent principles which can only identify previous element which is closest to pronoun element and is not pronoun. But this principle can bring into many inaccurate coreference, and this rule can only identify previous element which is closest to pronoun element. If antecedent element is behind pronoun element or is the previous element which is not closest to pronoun element, this rule can't identify that. Therefore, we introduce idea events to identify these particular cases.

Idea events draw forth a idea language, idea languages also consist of many event, and the pronoun elements in these event tend to the sponsor of the idea event, such as "我"(I)、"他"(he). For example, <Event>小明说(Xiao Ming said)</Event>："<Event>我想去打羽毛球(I want to play badminton)</Event>。" The previous event is an idea event, the latter is an idea language which is drew forth by Xiao Ming, so the coreference relation is [小明]←[我]([Xiao Ming]←[I]).

As show in Fig. 2, the case that antecedent element is behind pronoun element can't be identified by recent rules. After joining idea events, the coreference relation, like [居民徐先生]←[我]([the resident Mr. Xu] ← [I]) can be identified.

The case that the antecedent element is the previous element which is not closest to pronoun element also can be solved by idea events. As show in Fig. 3, [三年级中毒学生小芳]←[她]([the toxic student in grade 3 called Xiao Fang] ← [she]) can be identified by joining idea events.

```
<Event type="thoughtevent" eid="e13">"
    <Time type="relTime" tid="t13">7点不到(less than 7 o'clock)</Time>
    <Participant sid="s13">我(I)</Participant>
    <Denoter type="perception" did="d13">听到(heard)</Denoter>
</Event>
.........
<Event eid="e19">
    <Location lid="l19">纸厂(paper mill)</Location>
    <Denoter type="stateChange" did="d19">起火(was on fire)</Denoter>了
</Event>。"
<Event type="thoughtevent" eid="e20">
    <Location lid="l20">起火纸厂附近(nearby paper mill which was on fire)</Location>
    <Participant sid="s20">居民徐先生(Mr.Xu)</Participant>
    <Denoter type="statement" did="d20">说(said)</Denoter>
</Event>,
```

Figure 3. The case that antecedent element is behind pronoun element.

```
<Event type="thoughtevent" eid="e5">
    <Participant sid="s5">三年级中毒学生小芳(the toxic student in grade 3 called Xiao Fang)</Participant>
    <Denoter type="statement" did="d5">说(said)</Denoter>
</Event>
.........
<Event eid="e7">将(when)
    <Object oid="o7">饭菜(meals)</Object>
    <Denoter did="d7" type="operation">收回(were taken back)</Denoter>时
</Event>
<Event eid="e8">
    <Participant sid="s8">她(she)</Participant>已经(has already)
    <Denoter type="action" did="d8">吃了(eat)</Denoter>
    <Object oid="o8">半碗米饭(Half a bowl of rice)</Object>
</Event>
```

Figure 4. The case that the antecedent element is the previous element which is not closest to pronoun element.

3) Number and Special Events (nase)

This condition aim at the contraction sieve. We add two constraint conditions for the type of AB←B: number and special events. Number is that the numeric value of antecedent elements and anaphor elements should be consistent. Special events consist of events that the object element is "人员"(personnel) an the trigger word is "伤亡"(injuries and deaths). This type of event has no other elements except above object element. This sieve will eliminate this type of event, because this type of event brings into many coreferences, and they are all wrong.

We only add number consistent for the type of AEB←AB.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In the experiment, the corpus adopts CEC (Chinese Event Corpus) which is annotated coreference relations. The number of corpus is 100. Among this corpus, the number of event is 1778, the number of the trigger word is 1778, the number of the object element is 1629, and the number of the coreference of object elements is 287. There are many common corpuses of coreference resolution on the international, e.g., ACE, OntoNotes etc.. Although they also contain Chinese corpus, events annotated in the text only contain verb, event coverage is narrow, and idea events and event elements are not annotated. So test corpus selected is CEC annotated by Semantic intelligence laboratory of Shanghai University, it solves the above problems.

A. Single Result

Experimental results of each layer are shown in table 1.

Each layer is respectively exact string match sieve, pronoun sieve, contraction sieve, morpheme similarity sieve and semantic similarity sieve from top to bottom. As show in table 1, the first two layers have a higher contribution degree for the module, and the last two layers have a relatively lower contribution degree. And from the recall, we can find the ratio of four type of coreference in corpus and the ratio of the kind of the different description coreference is lowest. By comparing with the accuracy of each layer before and after adding constraint conditions, we can find that constraint conditions enhance the accuracy in the corresponding layer to a great extent.

B. Cumulative Result

The table 2 show the cumulative result as each layer are added to the module from highest to lowest precision.

TABLE I. SINGLE RESULT

	no constraints			add constraints		
	Precision	Recall	F1	Precision	Recall	F1
1	93.5%	44.9%	60.7%	97.7%	44.9%	61.6%
2	82.2%	12.9%	22.3%	88.9%	13.9%	24.1%
3	70.2%	29.6%	41.7%	83.8%	28.9%	43.0%
4	66.7%	1.4%	2.7%	--	--	--
5	60.0%	1.0%	2.0%	--	--	--

TABLE II. CUMULATIVE RESULTS

	Precision	Recall	F1
1	97.7%	44.9%	61.6%
1,2	88.7%	54.7%	67.7%
1,2,3	84.9%	74.6%	79.4%
1,2,3,4	84.8%	76.0%	80.1%
1,2,3,4,5	84.3%	77.0%	80.5%

As shown in table 2, as each layer adds in turn, precision is declining, recall is going up and F1 is also going up. The reason of the decline of the precision is the accession of the sieve of the lower precision. The rise of F1 illustrates that each sieve is useful to the system.

C. *Compared with The Decision Tree Method*

According to the reference [11], this paper builds a coreference resolution system based on C4.5 decision tree to compare with the multi-pass sieve module. This method adopts six attributive characters: distance, string match, sex, pronoun, semantic category and number consistency. The experiment result is similar to the result of this reference.

The multi-pass sieve module is the method base on rules, and decision tree is the classic supervised machine learning method in the data driven method. The comparison between this two methods can show that if we can get enough knowledge to express information, the method base on rules will obtain the good effect. As shown in table 3, we can clearly see the result of two types of method.

Through the above comparison, we find that the precision, recall and F1 of the multi-pass sieve module is better than the decision tree. The reason mainly has the following points: (1) Machine learning method relies on the corpus, and needs to learn knowledge in the corpus, so it is important for the number of corpus. Plenty of corpus can help machine learning method learn more knowledge, but the number of the corpus we used is not enough. (2) In the decision tree method, the number of counterexample is far greater than the number of positive example. That also influences the result. (3)Decision tree models adopt mention-pair model, so judging coreference relations only extract information from two words. And in the multi-pass sieve model, each tier builds on the entity clusters constructed by previous models in the sieve. So it adopts entity-mention model [12], the information content obtained is greater than the former. (4) Decision tree models judge various kinds of coreference together, and that will influence the judgment of each type of coreference. And multi-pass sieve model solves this problem, each layer only address one type. (5) According to the characteristics of the event, multi-pass sieve model adds constraint conditions which is associated with events. That improves the performance of the system.

TABLE III. THE COMPARISON OF EXPERIMENTAL RESULT

	Precision	Recall	F1
C4.5 decision tree	73.7%	55.7%	63.5
Multi-pass sieve	84.3%	77.0%	80.5%

But because ignoring the mention detection module, only when the F1 value of the first module is greater than 80%, adopting the method of this paper have the above advantages.

VI. CONCLUSION

According to four types of coreference, this paper builds the multi-pass sieve module. And combined with the characteristics of the event, we add the constraint conditions to each sieve to improve the accuracy in each sieve. Compared with machine learning method based on C4.5 decision tree, it has a very big enhancement on the performance. This paper introduces the single result of each layer and cumulative results of each layer. That proves the feasibility of the module. But it remains to be improved about the realization of each layer in this paper, especially contraction sieve and semantic similarity sieve. We will continue to improve this module in the future work. And we will aim at the characteristics of Chinese events, find out more effective features, and dig out more discourse knowledge for event-oriented coreference resolution.

ACKNOWLEDGMENT

This paper is supported by the Natural Science Foundation of China, No.61305053 and No.61273328.

REFERENCES

[1] Xuanyu Zhou, Juan Liu, Xiao Lu. Intra-Document Anaphora Resolution: A Survey[J]. Journal of Wuhan University(Natural Science Edition), 2014, 01: 24-36.

[2] Hobbs, J.R., Resolving pronoun references. Lingua, 1978. 44(4): p. 311-338.

[3] Lappin, S. and H.J. Leass, An algorithm for pronominal anaphora resolution. Computational linguistics, 1994. 20(4): p. 535-561.

[4] Soon, W.M., H.T. Ng, and D.C.Y. Lim, A machine learning approach to coreference resolution of noun phrases. Computational linguistics, 2001. 27(4): p. 521-544.

[5] Raghunathan, K., et al. A multi-pass sieve for coreference resolution. in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010. Association for Computational Linguistics.

[6] Lee, H., et al. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. in Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. 2011. Association for Computational Linguistics.

[7] Houfeng Wang, Zheng Mei. Robust Pronominal Resolution within Chinese Text[J]. Journal of Software, 2005, 05: 700-707.

[8] Junsheng Zhou, Shujian Huang, Jiajun Chen, et al. A New Graph Clustering Algorithm for Chinese Noun Phrase Coreference Resolution[J]. Journal of Chinese Information Processing, 2007, 02:77-82.

[9] Muyu Zhang, Yaobing Li, Bing Qin et al. Coreference Resolution Based on Head Match[J]. Journal of Chinese Information Processing, 2011, 03:3-8.

[10] Zongtian Liu, Meili Huang, Wen Zhou et al. Research on Event-oriented Ontology Model[J]. Computer Science, 2009, 11:189-192.

[11] Lihong Wei. Chinese Coreference Resolution Based on Decision Tree[J]. Software Guide, 2014, 03:31-33.

[12] Yang Song, Houfeng Wang. A Survey of Coreference Resolution Research Methods[J]. Journal of Chinese Information Processing, 2015, 01:1-12.