Event Coreference Resolution Based On Heterogeneous Information Network

Tao Sun

School of information, Qilu University of Technology
Jinan, Shandong 250353, China
suntao0906@163.com

Abstract

Event coreference resolution is an important step for business intelligence. In this paper, we present a method for event coreference resolution based on heterogeneous information network. We adopt holistic, hierarchical clustering algorithm to solve this problem. In terms of similarity measure, the proposed method utilizes the relationship between entities, events, documents and data source, calculates event comprehensive similarity through the event properties similarity. Experiment on business event sets, characters event sets and product event sets, the experiment results show that the algorithm of this paper can effectively complete event coreference resolution, and has good recall and precision.

Keywords: *event coreference resolution; heterogeneous information network; entity coreference resolution; event fusion; aggregation entity recognition*

1. Introduction

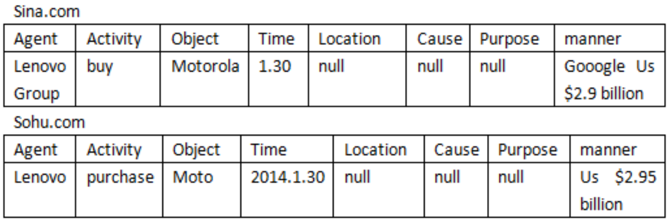
It is very common that the same event is described by many websites. However, the expression style and detail level is very different. It brings much redundant and inconsistent data, but on the other hand, one event mention rarely can describe the whole thing, and different event mention can complement each other, making it possible to fully describe the event. In order to accurate and comprehensive describe the event, it need to identify the different mention of the same event, this problem is called event coreference resolution [1]. The research of event coreference resolution far less than entity coreference resolution, but it can provide high-quality data for analysis application, so attract more and more scholars [2-5].

The most intuitive solution of event coreference resolution is pair-wise style, compare the event attribute literal similarity to get the event mention similarity. However, it is often fail to use literal similarity for event coreference resolution, because of different expression style and absence of event attribute.

For example, the event of “Lenovo acquire Motorola from Google with $ 2.9 billion in January 30, 2014” extracted from two different Chinese website, as shown in Figure 1, it cannot use the literal attribute similarity for event coreference resolution.

Event often appears in web page or document, the publish time, URL and authors are important to determine the relationship of the event mentions. Entities, events, documents and data source constitute a complex network. Homogeneous information network cannot cope with this problem, and the heterogeneous network is suitable to model this complex network. The node and association of heterogeneous information network are diversity.

ISSN: 2005-4262 IJGDC



Sina.com

Agent	Activity	Object	Time	Location	Cause	Purpose	manner
Lenovo Group	buy	Motorola	1.30	null	null	null	Gooogle Us $2.9 billion

Sohu.com

Agent	Activity	Object	Time	Location	Cause	Purpose	manner
Lenovo	purchase	Moto	2014.1.30	null	null	null	Us $2.95 billion

Figure 1. The Event of "Lenovo acquire Motorola" Extracted from two Chinese Website

The aim of event coreference resolution based on heterogeneous information network is that identify different event mention by identify coreference entities as much as possible, because of coreference event mentions have a lot of coreference entities.

This paper presents event coreference resolution based on heterogeneous information network. The main contribution of this paper is in the following aspects:

(1)This paper presents a two-stage event coreference resolution model, the two-stage model to fully consider the entity, event, data source and the relationships between them, to be more accurate identification of the correlation between events.

(2)Because of the deficiency in event attribute literal similarity, This paper try to reduce the literal similarity comparison, using entity recognition technology, including event participants, time, place and event type recognition, to synthetically identify event coreference resolution.

(3)The experiment shows that the model and key technologies proposed by this paper are feasibility and effectiveness, and with better recall and precision.

Section 2 of this paper gives the problem definition; Section 3 presents a two-stage event coreference resolution model; Section 4 presents the similarity measure method of event main elements; Section 5 is the experimental section, the model and the proposed algorithm testing; Section 6 gives related research work on event coreference resolution; summarized in Section 7.

2. Related Concepts and Definition

This section gives the related concepts and formal definitions of event coreference resolution.

Definition 1. Event coreference resolution. The event mention set E={e1,e2⋯ei...en} which extracts from different data source. N event mentions describe k events, and find the event mentions describe the same event in real world, ck={ej,ej+1,⋯em} (1≤ k≤ n,1 ≤ m≤ n,1≤ m-j≤ n).

Definition 2. The mention association graph. This graph represents the relationship between entities, event mention, documents and data source, as(V,L,DW,IW). Among them, V represents the node set which contains data source, documents, events and entity mention; L represents the edges set, real and virtual edges between the nodes represent direct and indirect association; DW and IW denote direct and indirect association weight(DW=0~1,IW=0~1).

Definition 3. Event coreference resolution based on heterogeneous network. For a given mention association graph, the goal of event coreference resolution is to find aggregation entity set which describes some event. The requirement is (1) the entity that describes the same event should belong to the same aggregation entity set; (2) the same aggregation entity sets that describes the same event.

This paper proposes a two-stage event coreference resolution model. The first stage is mining possible aggregation entity set C={C[1],C[2],…C[n]}. Under its guidance, take some entity set C[i] for event coreference resolution in mention association graph, forming coreference clusters in second stage. C[i][j] is the event mention which points to the same event C[i].

3. Two-Stage Event Coreference Resolution Model

Traditional event coreference resolution method uses event pairwise matching and transitive closure. This method requires event mention which will be recognized for pairwise comparison to make decisions. If the similarity is greater than a certain threshold, we consider it is coreference. This method requires decision must be made in a comparison. As described in the introduction, it is difficult to obtain sufficient "evidence" for coreference in many cases.

Comprehensive the above analysis, this paper puts forward a two-stage event coreference resolution model, as shown in figure 2. This model divides the whole process of event coreference resolution into the discovery of aggregation entity sets and coreference mention cluster.

(1)The first stage is mining aggregation entity sets according to the extracted event mention, that is a series of co-occurrence associated entity class C[i](i=1~n).

(2)The second stage is event coreference resolution based on aggregation entity set C[i] in mention association graph, to form event coreference set C[i][j] (j=1~m).

3.1. The First Stage

From the view of natural language processing, the event can be identified if the event participants, event type, time and place can be recognized. Therefore, the event participants (person or organization), time and location have a better identification and distinguish for event.

This paper uses entity recognition technology to identify participants of event, time and location. Since the form of date entity is more regular, use the rule entity recognition techniques which are similar with the regular expression string matching to complete it. The location, event participants and others are identified by adopting open source ICTCLAS system (http://www.ictclas.org/) which is developed by Institute of Computing Technology, Chinese Academy of sciences.

In order to improve the efficiency of event coreference resolution, this paper adds time window for event mention. Time window information exits in memory, and others are stored on the hard disk. We set the window size is 500, which 500 events can be stored in memory. In initial state, only a small portion of event is stored in time window. With the new event is added, first match the events in time window, if not matched, then search the event sets stored in outside the window. If new events cannot matched the events in time window, meanwhile the window is not full, then add new events to the window. Otherwise, remove the event which timestamp is the earliest.

3.2. The Second Stage

If multiple coreference entity exist in event mentions, and with the help of event type, then the event mentions would be coreference events.

Event participants (agent、object), time and location represent as {agent, object, time, location}.According to the description characteristic of event, if different event mentions has same time, same location, same participant, and do the same type event, so they are coreference and added to the same event mention set.

Table 1. Alogrithm of Holistic Event Coreference Resolution

```
Input: mention association graph r, similarity threshold $\tau$ and time widow;
Output: cluster of event C;

1. initialization: each event mention represents a class event in cluster set $C$; time window $windows = \phi$; extract k（k<=500）clustering entity sets which have the highest weight to set time window.
2. while $windows != \phi$ and r.article=unvisit do
3.     extract $< e_i, e_j, sim(e_i, e_j), e_i.type, e_j.type, >$ from $windows$; $sim(e_i, e_j)$ is the similarity of event mention $e_i, e_j$.
4.   if $sim(e_i, e_j) > \tau$ and $e_i$.type= $e_j$.type
5.      $e_{ij} = e_i \cup e_j$;
6.      $C = C \cup \{e_{ij}\} - \{e_i\} - \{e_j\}$;
7.      insert $e_{ij}$ into $windows$;
8.      remove $e_i, e_j$ from $windows$;
9.   end if
10.  if $sim(e_i, e_j) < \tau$ and r.article=unvisit
11.      if $windows$>=500 do
12.         remove max($e_i$.time) from $windows$;
13.         insert $e_k$ in r.article=unvisit into $windows$;
14.      end if
15.      else if all r.paper=visited then stop;
16.  end if
17. end while
18. return C;
```

The input of algorithm is mention association graph r, similarity threshold and time widow. The output is event cluster, each cluster represents an event in the real world, and the elements of the cluster represent event mention of this event. Algorithm 1 first initializes set, each event mention corresponds to a cluster, initialize time window to deposit entity sets (participants, time, location and event information), and extract entity which have the highest weight in first document of mention association graph r. Then start from 2 row entity matching, and calculate the similarity of event types.

As shown in line 4, if the similarity is greater than threshold, and the description is the same type of event, so merge the two events and put the event mention into cluster C, and remove the event ei and ej. Similarly in line 7-8, insert eij as the same time remove ei and ej. In line 10, if the comparison of event mentions has been completed in time window, and the documents which were not visited still exit in r, then if the time window is not full, then put the event mention ek of another document into time window. In line 15, if all documents have been visited, then end the process of event coreference resolution. Finally, in line 18, the algorithm returns the event cluster C, each cluster represents an event in the real world, and the elements of the cluster represent event mention of this event.

The efficiency is the weakness of this algorithm. Because the algorithm requires any two named entity for similarity comparison which has enormous comparison space, its time complex is (represents number of event mention). In addition, the merge of event mention will affect the relationship of entities, and vice versa. The re-calculate the similarity of event mention increases the complexity of the algorithm. Therefore, in order to improve the efficiency of the algorithm, we reduce the comparison time.

For all event mentions need pairwise comparisons in algorithm 1. If we find participant is different, so these event mentions cannot be coreference. Similarly, if event types of two event mentions are different, they also cannot be coreference. Thus, according to the Bootstrap idea, event mention can be classified by participants, then according or event type, so that can reduce the comparison range. Different range of event mention cannot be coreference, so as to improve the efficiency of the algorithm.

4. Similarity Measure

As the deficiency of literal similarity comparison in pair-wise fashion, this paper adopts entity recognition to compare the similarity of event mention. Event similarity refers to amount of information which two different events share. If the two event elements that correspond to two events are similar, the two events are similar, wherein the activity element is the core element of the event. Thus, the similarity measure of the event can be transformed into a similarity measure of event element, and then calculate the weighted comprehensive value of event element similarity. These factors mainly include the time, location and participants. In this section, we will introduce the similarity measurement of these three event elements.

4.1. Similarity Measurement of Participants

Participants involve the agent and object of event element, mainly are named entity of people and organizations, so the key of Similarity measurement for participants is the recognition and expression of named entity. In a document, participant may have more than one, when they appear in the same document; this constitutes the co-occurrence relationship, so the co-occurrence of named entity identifies the close degree of participants. Therefore, this paper in similarity computation of participants, we use the similarity calculation method based on co-occurrence relationship.

For participants (named entity) in the document, this paper uses the participants' co-occurrence vector $Lp(lp_1 : w_1, lp_2 : w_2, ..., lp_n : w_n)$, lp_i is named entity, and w_i is weight. The more the number of co-occurrence, then the relationship between these two named entities is closer. The weight is calculated as:

$$w(lp_i, lp_j) = \frac{\sum(lp_i, lp_j)}{\sum lp_i} \qquad （1）$$

In the formula (1), $w(lp_i, lp_j)$ represents co-occurrence weights of two named entities lp_i, lp_j, $\sum(lp_i, lp_j)$ represents the co-occurrence number of lp_i, lp_j, $\sum lp_i$ is the number of lp_i. Similarity calculation of participants is transformed into the between co-occurrence vectors' calculation, formula is:

$$sim(lp_i, lp_j) = \cos ine(LP_i, LP_j) \qquad （2）$$

In an event mention, participant may have more than one, so in the whole event mention, the similarity of participant is calculated as:

$$sim_{name}(lp_i, lp_j) = \frac{\sum_{i=1}^{m} \max(sim(l_i, P)) + \sum_{j=1}^{n} \max(sim(p_i, L))}{m+n} \qquad (3)$$

In formula (3), L and P are participant vector of event representation e1 and e2, $L=\{l_1:w_1,l_2:w_2...,l_m:w_m\}$, $P=\{p_1:v_1,p_2:v_2...,p_n:v_n\}$.

4.2. Similarity Measurement of Time

The specific time of events is usually included in the content of the document, and there is no standard consistent form, often used such as: on the morning of September 6th, last Saturday, three years ago, today, one o'clock on yesterday afternoon etc. Therefore, it's need to standardized time format. Time specification strategy adopted in this paper is: for the previous time expression has not years or months, adopts the same year or the same month with the document, like " yesterday, last month, last year and last week", get it by year, month that correspond to document published time minus certain date. For example, the specification of time to express the format is "at 21:50 on March 15, 2014".

After get a standardized time presentation format, we use formula 4.5 to calculate the similarity of time entity:

$$sim(t_i,t_j)=\begin{cases}\dfrac{|t_i\cap t_j|}{|t_i|\cup|t_j|}, & t_i\cap t_j\neq\varnothing\\ 0, \text{ot her s} & \end{cases} \qquad (4)$$

In formula (4), t_i is the time vector of event mention e_i, t_j is the time vector of event mention e_j, $t_i=\{t_{i0},t_{i1}\}$, $t_j=\{t_{j0},t_{j1}\}$. t_{i0}, t_{j0} is the start time of event, t_{i1}, t_{j1} is the end time of event.

4.3. Similarity Measurement of Location

Document in Web describes the location mainly uses the continent, state, province, city, county, district, township, village, street etc. Usually these symbols are divided into different levels, such as continent is the first level, nation is the second level and province is the third level. This paper calculates similarity of location by establishing a location ontology tree.

Location ontology tree's construction is continent -nation - province (state or district) - city - county - the town (village). Location similarity of two event representations is calculated as the following formula:

$$sim(s_i,s_j)=\frac{\alpha\times|s_i\cap s_j|}{|s_i|\cup|s_j|} \qquad (5)$$

In formula (5), s_i,s_j is the location vector of event e_i,e_j; $|s_i|,|s_j|$ is the path length which from the s_i,s_j to the ontology tree root node, $|s_i|\cap|s_j|$ is the first overlapping path length which from s_i,s_j to root node, α is parameter, location which appears in the first paragraph is usually the actual location, other place in document may not the true which only for a reference, so the value of α as follows:

$$\alpha=\begin{cases}1, \text{if } s_\text{i} \text{ and } s_\text{j} \text{are taken from the first paragraph}\\ 0.7, \text{if only one of } s_\text{i} \text{ and } s_\text{j} \text{ is taken from the first paragraph}\\ 0.5, \text{if } s_\text{i} \text{ and } s_\text{j}\text{are not taken from the first paragraph}\end{cases}$$

4.4. Similarity Measurement of Event Entity Factor

After calculating the similarity of participant, time and location in event mention, we combine them to constitute similarity measurement based on event entity factor, which is calculated as follows:

$$sim(e_i, e_j) = \beta \times sim(lp_i, lp_j) + \delta sim(t_i, t_j) + \varepsilon sim(s_i, s_j) \qquad (6)$$

$\beta, \delta \text{ and } \varepsilon$ is the parameter of participant, time and location, respectively. Specific weight is set by the user according to specific environment and the field. In the experiment, we set $\beta = 0.6, \delta = 0.2, \varepsilon = 0.2$.

4.5. Event Type Recognition

Because the event type is very important, we recognize event type by activity attribute in event mention. To do this, we can construct a verb form, each one marks verb and its corresponding event type, if event mention includes the verb in table, the type of event mention is the type which corresponds to the verb. The recall rate of this method is higher in large scale of verb table. But the exact rate is not good, because the natural language would often be polysemy. Therefore, this paper uses a method which is similar to TFIDF, to calculate a score for each verb, and filter out interferential verb by score level. The formula is as follows:

$$Score = etf * eidf \qquad (7)$$

$$etf = \frac{|T_{tw}|}{|T|} \qquad (8)$$

$$eidf = \log_2 \left(\frac{n}{edf} \right) \qquad (9)$$

In formula (7), etf is the proportion which total number of some event produced by verb in this event, That is, the contribution degree of verb for this event. $|T_{tw}|$ is the total number of some event which contains verb tw in the corpus. In In formula (8), $|T|$ is the total number of this event, $eidf$ is the frequency degree of verb tw . n is the total number of event in the corpus. edf is the total number of event which contains this verb. $Score$ of some verb is more bigger, the ability that represents some event is more stronger. Therefore, it can filter out the interference verb which the value is less than this threshold by assigning threshold. This method can improve the accuracy, but abandon the recall rate.

This paper summarizes the type of event recognition and machine learning verbs extensions pros and cons of both approaches, combining the two methods, a method for event type recognition as following:

Calculate the $Score$ of each verb in verb table.

According to the calculation of $Score$, set a high threshold, which is greater than $Score$ of the verb can be directly through the verbs to judge the type of event

For less than the threshold, use ME and SVM for event type recognition, respectively.

For the verb that is greater than the threshold, which is usually a single meaning of the word and appear less frequently, as long as it appears this method can determine the type of event. If this type of event use machine learning, because less example lead to recognition error and reduce recall rate.

For the verb that is less than the threshold, which is usually polysemy word, they are less accurate, when using the method of verb table lookup. Because more examples, therefore it is more suitable for the event type recognition by machine learning.

5. Experimental Results and Analysis

5.1. Data Set and Evaluation Criteria

This section uses the following real data set to assess the proposed method, in order to guarantee the quality of main data, the ratio of positive examples and counter-examples of data is 1:4.

- Business event dataset

The data set use well-known domestic and foreign enterprises for study, including Google, Apple, IBM, Lenovo and other 200 companies. Choose 100 statements the companies involved in the event from Wikipedia Introduction and related links, total 20,000. This part makes up corporate events positive examples data set, then randomly choose 400 statements the companies not involved in the event to constitute a counter-example data set, total 80,000.Then business event dataset has 100,000 event representation in total.

- Character event dataset

The data sets selected famous people's event from different areas for the study, including entrepreneurs, scientists, entertainers, politicians and other 100. Use the obtain method which similar with corporate event data sets, extracted 10,000 Positive examples and 40,000 counter-example data set from Wikipedia dataset, in total, product data set has 50,000 event representations.

- Product event dataset

The data sets selected event containing the product (mainly in information products) from different areas for the study, includes IPhone, Ipad, widows, office, DB2, Thinkpad and other 50 kinds. Use the obtain method which similar with first two data sets, extracted 5,000 Positive examples and 20,000 counter-example data set from Wikipedia dataset, in total, product data set has 25,000 event representations.

This paper adopts common web information retrieval standards to evaluate the experimental results through recall, precision and F1 measure.

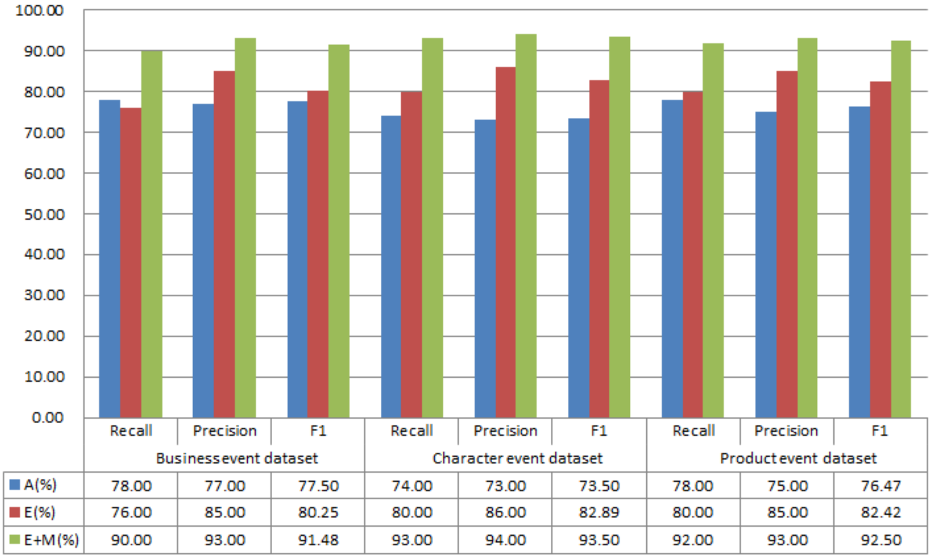
5.2. Experimental Results and Analysis

In order to verify the effect of the proposed method in this chapter, we make three experiments respectively on business event data sets, characters events data sets, and product event data sets.

In order to distinguish between different similarity metrics methods, (A) represents the method which implements event coreference resolution with only literal attribute similarity, (A+E) represents the method which implements event coreference resolution with attribute literal similarity and entity similarity(includes event participants, time, place), (A+E+M)represents the method which implements event coreference resolution with attribute literal similarity, entity similarity and event type. By constantly adding new similarity measure on the basis of the attribute literal similarity methods, we can verify the effect of newly introduced events similarity metrics method. Table 1 shows the three experiments' performance results on the recall, precision and F1 measure.

As we can see from the results given in Table 1, overall performance can be improved significantly through integrated use multiple events' similarity measure method. For example, in the characters event data sets and business event data sets, when we integrated use three events similarity metrics method, compare with use only literal attribute similarity method, its F1 measure respectively enhance 3.6% and 3.9%.

Meanwhile, not all properties can be enhanced after using similarity metrics method. For example, in business event data set, compared to method (A+E)and method(A), its recall rate did not increase, but decreased 3.2%, because when we consider more event similarity features, the accuracy of the matching has been guaranteed, but it is difficult to match decision when the evidence is insufficient. Experiment shows that compare with single event similarity measure method, the measure method which integrated use multiple events' similarity can improve the overall performance of event coreference resolution effectively.



	Business event dataset			Character event dataset			Product event dataset		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
🟦 A(%)	78.00	77.00	77.50	74.00	73.00	73.50	78.00	75.00	76.47
🟥 E(%)	76.00	85.00	80.25	80.00	86.00	82.89	80.00	85.00	82.42
🟩 E+M(%)	90.00	93.00	91.48	93.00	94.00	93.50	92.00	93.00	92.50

Figure 2. Comparison of Different Similarity

6. Related Work

The research of event coreference resolution is less than entity coreference resolution， using pair-wise manner is the most direct way to coreference resolution， determine whether it points to the same event in the real world through comparing the attributes of two event mentions [3, 5, 1]. The researches have noticed that the method to implement event coreference resolution with machine learning is important. The supervised and unsupervised methods adopted by CA Bejan, H Lee [2, 4] implements event coreference resolution with syntax, semantics, properties and other characteristics.

Existing methods mainly use attribute information of event co-occurrence to implement event coreference resolution. There is a large number of entity information in event mention, including the event's participants, time, location, etc. If there are a large number of entities in two events mention, the possibility of two event mentions are coreference. Currently, research on event coreference resolution with coreference entities information only has been seen in [4, 6]. Moreover, In addition to entity information, information such as document, the data source, the type of event and relationship among them are able to facilitate event coreference resolution. The research in this area has not yet been seen.

The preferred organization of objects such as entities, events, documents and a data source is heterogeneous information network. As the trend of heterogeneous information network develops, research of heterogeneous information network has become a new hotspot, influential study includes RankClus [7] and NetClus [8]. Unlike traditional sorting method of information network analysis, RankClus extract the characteristics of heterogeneous information networks, and implement sorting by analyzing the relationship between different objects and class association. NetClus expands the components number of heterogeneous information networks, implement data sorting in different categories with clustering. Currently, the research to implement event coreference resolution with heterogeneous information network has not seen.

Based on the above analysis, we organize objects such as entities, events, documents and a data source with heterogeneous information network, compare the similarity of a variety of different elements of events, measure through comprehensively utilize similarities such as literal properties, events, participants, time, location and event type. It can make up the lack of simply using the attribute similarity measure solutions to solve event coreference resolution and improve the overall performance in some degree.

7. Summary

In order to adapt to the needs of analysis application, we put forward a integral method of event coreference resolution based on heterogeneous information network. The method uses a hierarchical clustering integral event coreference resolution algorithm, performs iterative event coreference resolution with the interaction between decision-match. In terms of similarity measure, the method presented in this paper comprehensively utilize the relationships among entities, events, documents and data sources, measure events similarity with characteristics of different angles, and get a reasonable comprehensive similarity of event mention. The experiments base on the business event data sets, characters events data sets and product event data sets show that the algorithm can complete the task of event coreference resolution effectively, it also has a better recall and precision.

References

[1] Z. Chen, H. Ji and R. Haralick, “A pairwise event coreference model, feature impact and evaluation for event coreference resolution”, Proceedings of the Workshop on Events in Emerging Text Types, Association for Computational Linguistics, **(2009)**, pp. 17-22.

[2] H. Lee, M. Recasens, A. X. Chang, M. Surdeanu and D. Jurafsky, “Joint Entity and Event Coreference Resolution across Documents”, EMNLP-CoNLL **(2012)**, pp. 489-500.

[3] B. Chen, J. Su, S. Jialin Pan and C. Lim Tan, “A Unified Event Coreference Resolution by Integrating Multiple Resolvers”, IJCNLP, **(2011)**, pp. 102-110.

[4] C. Adrian Bejan and S. M. Harabagiu, “Unsupervised Event Coreference Resolution with Rich Linguistic Features”, ACL, **(2010)**, pp. 1412-1422.

[5] Z. Chen and H. Ji, “Graph-based Event Coreference Resolution”, Graph-based Methods for Natural Language Processing, **(2009)**, pp. 54-57.

[6] T. Y. He, “Coreference resolution on entities and events for hospital discharge summaries”, Massachusetts Institute of Technology, **(2007)**.

[7] Y. Sun, J. Han and P. Zhao, “Rankclus: integrating clustering with ranking for heterogeneous information network analysis”, Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. ACM, **(2009)**, pp. 565-576.

[8] Y. Sun, Y. Yu and J. Han, “Ranking-based clustering of heterogeneous information networks with star network schema”, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, **(2009)**, pp. 797-806.

Author

Sun Tao, he is a PhD, his current research interests focus on big data, data integration and cloud computing.