The Chinese Vietnamese Bilingual News Event Ranking Method Based on Attribute Association Graph

Mingwei Zhu[1,2],Zhengtao Yu[1,2], Guangshun Qin[1,2], Hua Lai[1,2], Shengxiang Gao[1,2]

1(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;)

2(Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Correspongding author:Hua Lai, Email:405904235@qq.com

***Abstract*—Facing with bilingual news event's attributes association about Chinese and Vietnamese, we propose a bilingual news event sequencing method based on graph of attributes association. This method built the graph of attributes association based on the relations between event properties and it sequences the Chinese and Vietnamese bilingual news event using this graph. This method first extracts the titles, elements, topic sentences which characterize bilingual news event properties. And it uses the extractions to describe events in the document. Then we translate these title, elements and topic sentences by using the bilingual dictionary and Vietnamese-Chinese Alignment Corpora. And we build the graph of attributes association. We calculate the weights of edges by using word2vec and elements co-occurrence strength. In the end, it sequences the bilingual news using similarity among event node, query keyword and connection between the graph's nodes and event. The experimental results show that the proposed method can effectively improve the performance of the ranking of the news events in the Chinese and Vietnamese, and the attribute relevance has a good effect on the ranking of the Chinese and Vietnamese news events.**

Keywords—attribute association graph, news events, ranking method

I. INTRODUCTION

Under the situation of frequent exchange between China and Vietnam, bilingual news event capture becomes more important. Bilingual news sequencing is key to bilingual news information retrieval and very valuable in bilingual information capture. Presently the domestic and overseas scholars have done a lot of research on entity ranking. Zhaoman Zhong et al. [1] studied a sorting method based on factors combination of news event. Using event retrieval formal rules based on BNF, they proposed the algorithm calculating the similarity between document and event elements. This algorithm integrates Portfolio Analysis for Web news features, retrieval based on multi-elements of event, distance between constrain item and event entry. Qin et al. [2] proposed a kind of news event list sorting method which combines the attributes and relationships. In this method, they fuse the event attribute and the relation feature to the ListNet ranking model. By that, a news event list scheduling model can be built. Chen et al [3] proposed the expert ranking method based on ListNet with multiple features. Firstly, multiple features was selected through the analysis of expert pages; secondly, in order to learn parameters through gradient descent and construct expert ranking model, all features were integrated into ListNet ranking model; finally, expert ranking contrast experiment will be performed using the trained model. Xianhui Li et al. [4] use depth learning algorithm to input expert document corpus Restricted Boltzmann machine for training. Based on the weight of vector cosine value, a depth based learning expert ranking model is constructed. In the literature, the paper[5] proposes a web ranking method based on social tagging, using a large number of user annotation data to construct the theme model of user pages, and combines the query correlation and the enhancement of the interaction to sort the pages. Wang Yang et al, [6] who use machine learning ideas, query - Document on the construction of a sort model, so as to achieve the sort of documents and web pages. Yao Xu et al [7] proposed Entity ordering algorithm based on graph structure. They got the relevant documentation using Random Walking Algorithm and sequence the entity which has been extracted from relevant documentation.

The methods above are worked in single language environment. Because of computational difficulty of bilingual similarity computation, bilingual news Event sequencing gets harder. Event property is key to bilingual news Event sequencing. See event as nodes and Event Correlation as edges. Integrating the thought of Graph-Ranking overall thought and build the graph of attributes association based on event correlation Finish the Event sequencing by using this graph.

II. THE CHINESE VIETNAMESE BILINGUAL NEWS EVENT RANKING METHOD OF ATTRIBUTE ASSOCIATION GRAPH

The flow chart of the scheduling algorithm of Chinese and Vietnamese bilingual news events based on attribute association graph mentioned in this paper is shown in Figure 1.

As shown in figure 1, firstly, we extract the event tuples (title, event element and topic sentence) in Chinese and Vietnamese bilingual document. Secondly, we extract the content of the translation through the Chinese Vietnamese bilingual dictionary and the Chinese Vietnamese entity alignment corpus to research the realization in the same

978-1-5386-1981-0/17/$31.00 ©2017 IEEE

language environment. Thirdly, constructing the association graph of event and calculating the edge weights of the graph. In the end, the sort of news event in Chinese and Vietnamese will be realized.

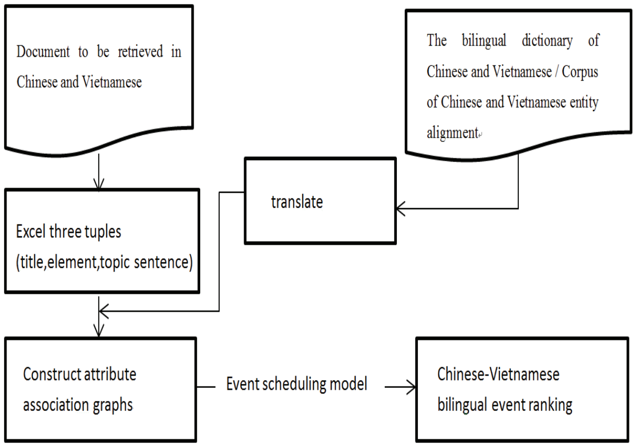
Document to be retrieved in Chinese and Vietnamese
The bilingual dictionary of Chinese and Vietnamese / Corpus of Chinese and Vietnamese entity alignment
translate
Excel three tuples (title,element,topic sentence)
Construct attribute association graphs
Event scheduling model
Chinese-Vietnamese bilingual event ranking

Fig. 1. Flow chart of news event sorting algorithm

2. 1 Three tuples analysis and extraction

We think the most important parts of an event-oriented news document are title, event elements and event topic sentences. So these three elements are used as tuple of a document in this paper.

1)As the essence of the news, title plays the role of revealing, clarifying and evaluating the news content, and is an important feature of sorting news pages. We extract the titles by using the Web crawler technology.

2) The event elements are trigger words, time, place and object. News event element extraction uses NER (Named Entity Recognition, NER) [8] and POS (part-of-speech) tagging technology [9], etc.

3) In a news document, we call the sentence which include a core element "object ,event trigger words" and a secondary element "time or place" as event topic sentence.

2. 2 The corresponding relationship between Chinese and Vietnamese

Corpus resources. The premise of the bilingual news event ranking is the correspondence of the bilingual words. We have constructed a large scale "Chinese and Vietnamese bilingual dictionary"and now, the dictionary contains about 260 thousand bilingual entries and they are very practical for our research. The event attributes contain many entities, such as the names of people, place names and organization names. The accuracy of these entities translated by machine or dictionary is not high. We extracted the bilingual entity pair in Chinese and Vietnamese from Wikipedia to construct "Chinese and Vietnamese entity alignment corpus". It contains more than five thousand alignment entities, and it's a great help to translate the element entity in Chinese and Vietnamese events.

Table 1. The bilingual dictionary of Chinese and Vietnamese

No.	Vietnam words	Chinese words
1	Một	阿(Ah)
2	Ah Q	阿 Q(Ah Q)
3	Một q	阿 q(Ah q)
4	Abadan	阿巴丹(Abadan)
5	Thành phốAbidjan	阿比让市(Abidjan)

Translation of words. This paper is mainly about the translation of word-level. For different types of words and expressions, different translation strategies are adopted. The translation of words are divided into three categories: time words, entity words and general words. For time words, the method of translation is by language features of Chinese and Vietnamese. Through the analysis, we can know that the order of the specific date (year, month, date) in the Vietnamese language is a mirror image of the Chinese language. That is to say the time words in Vietnamese in accordance with the order date, month and year, and the number should be in the back. For example, the time words "2015 年 12 月 12 日" are translated as "ngày 12 tháng 12 năm 2015" in Vietnamese. " ngày, tháng, năm" correspond to "日,月,年" respectively. For entity words, using "Chinese and Vietnamese entity alignment corpus" to translate them. If the translation is inaccurate, we can use "Chinese and Vietnamese bilingual dictionary" and online translation tools such as Google translation to correct it. For general words, we can use the "Chinese and Vietnamese bilingual dictionary" to translate them.

Table 2. The bilingual dictionary of Chinese and Vietnamese

	Chinese entity	Vietnamese entity
1	越南(Vietnam)	Việt Nam
2	河内市(Hanoi)	Thành phố Hồ nội
3	胡志明市(Ho Chi Ming City)	Thành phố Hồ Chí Minh
4	张晋创(Truong Tan Sang)	Trương Tấn Sang
5	外戚(Dukes)	Minh Thành Tổ

2. 3 Construction of event attribute association graph

We take the event extracted from the news document as the node of the event attribute association graph. When the same elements occur together between the events, they will have an association. In this case, the node in the figure will have connected edges. Based on this idea, we construct the event attribute Association graph. Fig. 2 is a graph consisting of eight events, and node for each event contains three elements: title, event element and event topic sentence. If the two event nodes are connected in the event attribute association graph, it shows that the two events have certain relevance. So, the events can be seen as the three tuples<title, event elements, event topic sentences>, then the association between events can be seen as the correlation of these three elements. Because the title can also be seen as a sentence, we can use word2vec [10] to calculate the similarity between the title and the topic sentence in two events.

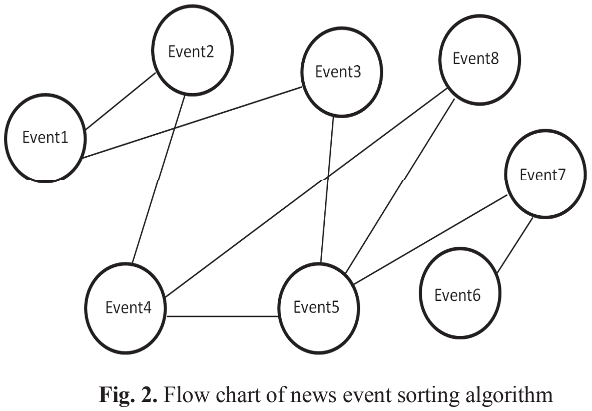

Fig. 2. Flow chart of news event sorting algorithm

Then we can get the association of event elements by co-occurrence of event element. Based on the analysis above, the edge weights of event and are calculated as follows:

$$Weight(E_i,E_j)=\alpha\times Sim(T_i,T_j)+\beta\times Sim(S_i,S_j)+\lambda\times Relation(E_i,E_j) \quad (1)$$

Among them, α,β,λ are the parameters between the interval [0, 1] and the sum of them is 1. Here α and β are both set to 0. 3, and λ is set to 0. 4. $Sim(T_i,T_j)$ and $Sim(S_i,S_j)$ represent the similarity between the two event titles and topic sentences. Using word2vec to construct a word vector space, then the words in the two sentences are mapped to the word space to get two vectors. We use vector cosine algorithm to calculate the similarity of sentences:

$$Sim(S_i,S_j)=\cos\theta=\frac{\sum_{k=1}^{m} w_{ik}\times w_{jk}}{\sqrt{(\sum_{k=1}^{m} w_{ik}^2)(\sum_{k=1}^{m} w_{jk}^2)}} \quad (2)$$

W_{ik} and W_{jk} represent values of the vector of sentence and in dimension,and represents the dimension of vector space. $\mathrm{Re}lation(E_i,E_j)$ represents the co-occurrence strength between the two event elements. Intensity value of co-occurrence are shown in table 3:

Table 3. Co-occurrence strength of event

Number	Element co-occurrence description	Co-occurrence strength value
1	the co-occurrence of an element between the events	0. 25
2	the co-occurrence of two elements between the events	0. 5
3	the co-occurrence of three elements between the events	0. 75
4	the co-occurrence of four elements between the events	1

2. 4 Event ordering model based on attribute association graph

According to the global characteristics of the graph ordering, and the basic idea of event sorting algorithm based on attribute association graph is that the score of an event is not only related to the similarity between the event and the query, but also need considering the relationship between the event and the other events. If the event e_i is associated with the event e_j in the graph, then the scores of similarity between these two events and query should be relatively close to each other. If event sets $E=\{e_i \mid i=1,2,\ldots,N\}$, e_i represents event i in E, N represents the number of the event. Based on the above considerations, for the event in event set , the scores in a given query q can be expressed as:

$$Score(e\mid q)=d\times\sum_{v\in E}\frac{Weight(e,v)}{\sum_{z\in E}Weight(z,v)}Score(v\mid q)+(1-d)\times\frac{Sim(e,q)}{\sum_{z\in E}Sim(z,q)} \quad (3)$$

In the formula, d is the damping coefficient, and it is an adjustable parameter. Here we assign it as 0. 85 and it is a commonly used value in classic sorting algorithm PageRank [11]. $Weight(e,v)$ represents the similarity between two events e and v ,that is the edge weights between nodes in the graph. $Sim(e,q)$ represents the similarity between event e and query q , and the calculation of their similarity is based on the relationship between event title and topic sentences and queries:

$$Sim(e,q)=\alpha\times Sim(t,q)+\beta\times Sim(s,q) \quad (4)$$

Here,A is 0. 3 and B is 0. 7, $Sim(t,q)$ and $Sim(s,q)$ show the similarity of the title and topic sentence of the event and query. Using word2vec to calculate it and the method is shown in formula (2). In practical operation, we need to set the initial value for the $Score(e\mid q)$ and we set k as the initial value of Score(e|q), then perform iterative calculation until the scores of all events converge. We use the formula (3) to calculate the ranking score of query and the candidate events,then put the scores in descending order. Thus, we can obtain a list of candidate event sets. Event ordering process based on attribute association graph is shown in algorithm 1.

Algorithm 1 Event ordering algorithm based on attribute association graph

```
Input: Candidate news event document set $D$ ,Query $q$
Output: Event set sort list $E_{ranking}$
BEGIN
1) Extract event set E from document set D
2) Construct graph $G(V,E)$ by formula (1)
3) Set the initial value of $Score(e\mid q)$ of the vertices of each event in Gas $k$
4) Calculate the event vertices $Score(e\mid q)$ by formula (3)
5) Implement step 4,until $Score(e\mid q)$ converges
6) Get $E_{ranking}$ by descending order according tont in event set $E$
END
```

III. Experiment and result analysis

3. 1 Experimental data show

The corpus of the experimental data in this paper are collected from the 3000 event oriented bilingual news documents, including military, finance, sports, science and technology and education. Each news document includes title and text, Word segmentation, tagging part of speech, and stopping words, then we can extract the event elements which appear in the documents, and the statistical information of the corpus is shown in Table 4.

Table 4. Corpus statistics

	Military	Finance	Sports	Technology	Education	Total
Chinese	316	323	348	331	312	1630
Vietnamese	254	263	287	288	278	1370

Table 5. Query topic information

Number	Chinese	Vietnamese
1	汶 川 (Wenchuan) ， 地 震 (earthquake)	V•n Xuyên, •ng •t
2	香港(Hong Kong),	H•ng Kông,trong k• toán
3	占中(Occupy Central)	Th••ng H•i, B•n Th••ng H•i
4	失联(Out of contact)	Ngân Xuyên xe buýt, •t nhà

Query terms are set in a way that is consistent with the user's search engine. We define eight event oriented query topics, respectively in Chinese and Vietnamese. For each query topic, we use MAP and p@5 as evaluation criteria. MAP represents the average value of the accuracy of each query in the query set, which can well evaluate the retrieval performance. P@n is the accurate rate of return to the previous n results of the sorting system. Sometimes the users may be interested in the previous n results returned by the retrieval system, so p@n is a measure of the performance of the sort from this point of view. As shown in Table 5.

3. 2 Experimental design and results analysis

In order to better verify the effectiveness of the method, two contrast experiments were designed in the paper.

Contrast experiment of the ranking results monolingual environment.

In the case of the same data set, we use news event list sorting method based on ListNet to do single language event scheduling, and take the results as a benchmark to compare with the methods mentioned in this paper. MAP is used as evaluation index in the experiment. And the experimental results are shown in Table 6 and 7.

Table 6. Chinese Query results

	MAP	% of Monolingual
Monolingual	0. 32412	
Bilingual	0. 27924	86. 2

Table 7. Vietnamese query results

	MAP	% of Monolingual
Monolingual	0. 29675	
Bilingual	0. 25086	84. 5

Seen from the above two tables, the method of sorting the Chinese Vietnamese bilingual events is proposed in the paper has achieved a good cross language sorting effect on the MAP index. Among them, compared with the Chinese query, query the Vietnamese ranking results have decreased significantly. The main reason is that in the document preprocessing, Vietnamese word segmentation and POS tagging accuracy is lower.

Comparison of different methods of ranking results.

In order to verify the effectiveness of the proposed sorting method, we designed comparative experiments with the current commonly used cross-lingual information retrieval method for comparison.

In the experiment, we put the document translation method based on the Machine Translation system as M1, and using Google translation as the Machine Translation system; we put the query translation method based on bilingual dictionary as M2. We put the method mentioned in this paper as M3. And evaluating indicator in the experiments is MAP. The experimental results are shown in Figure 3:

Figure 3 shows that under the same conditions, the ranking performance of the Chinese and Vietnamese bilingual news event ranking method based on the attribute association graph proposed in this paper is better than that of the other two methods. In indicator MAP, the average value reached 26.08%, which increased by one percentage point. In indicator P@5, the average value reached 64.26%, which increased by five percentage point. The news event ranking method takes into account the event attribute which is the key to the event scheduling task. Therefore, the translation of bilingual texts can be translated into the translation of event attributes, which can reduce the difficulty of translation. At the same time, using the characteristics of the inter event correlation, combined with the overall thinking of the graph sort and considering the interaction between multiple events in the sorting process, all this can achieve good sorting effect.

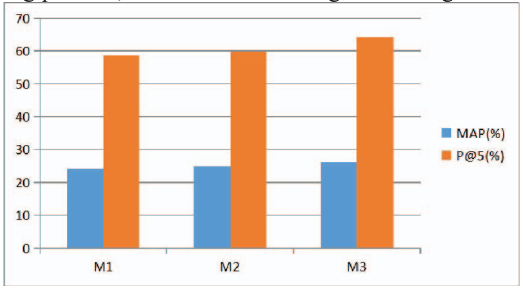

Fig. 3. Experimental results of different sorting methods

IV. Conclusions

In the paper, we use the characteristics of event correlation to construct ranking method based on attribute association graph. In further research, we can also sort the news events from the level of Chinese and Vietnamese bilingual texts. At the same time, the event attribute can be used as the evaluation index of the ranking to improve the efficiency.

V. Acknowledgement

This work was supported by National Nature Science Foundation (Grant Nos.61472168,61672271,61732005), Emerging Industry Technology Development Project of Yunnan Province (Grant Nos.201606), Science and Technology Innovation Talents Fund Projects of Ministry of

Science and Technology (Grant No.2014HE001), Key Province, Innovation Team Projects of Yunnan Province（Grant No.2014HC012）.The corresponding author is Hua Lai. His email is 405904235@qq.com.

REFERENCES

[1] ZhongZhaoman,Li Cunhua,Liu Zongtian,Dai Hongwei. Web News Oriented Event Multi-Elements Retrieval[J]. Journal of Softwar,2013, 24(10):2366−2378.

[2] Qin Guangshun, Yu Zhengtao, Hong Xudong, Liu Xiaohui, Guo Jianyi. A News Event Ranking Method Based on List with Attributes and Relationship[J]. Journal of Information and Computational Science, 2015, 12(11): 4487-4494.

[3] Chen Fangqiong, Yu Zhengtao, Wu Zejian, et al. Expert ranking method based on ListNet with multiple features[J]. Journal of Beijing Institute of Technology, 2014(2):240-247.

[4] LIXian-Hui,YUZheng-Tao,WEISi-Chao,GAOSheng-Xiang, WANGLi-Ren. Deep Learning Expert Ranking Method Based on Listwise[J]. Pattem Recognition and Aitificial Intelligence,2015, 28(11):976-982.

[5] Liu Kaipeng,Fang Binxing. A Novel Page Ranking Algorithm Based on Social Annotations[J]. Chinese Journal of Computers,2010, 33(6):1014-1022.

Laboratory in Regional University Projects of Yunnan

[6] Wang Yang,Huang Yalou,Xie Maoqiang,Liu Jie,Lu Min,Liao Zhen. A Multiple Query Dependent Banking SVM Aggregation Algorithm[J]. Journal of Computer Research and Development,2011, 48(4):558-566.

[7] XU Yao, ZHAO Zhengwen, Chen Qun, Liu Hailong, DU Jing, Hu Jiaqi, LI Zhanhuai. Graph-based Web Entity Ranking Method[J]. Computer Science,2014, 41(5):219−222.

[8] Zhao Jun. A Survey on Named Entity Recognition,Disambiguation and Cross-Lingual Coreference Resolution[J]. Journal of Chinese Information Processing,2009, 23(02):3-17.

[9] Yang Er-hong,Fang Ying,Liu Dongming,Qiao Yu. The Evaluation of Chinese Word Segmentation and POS Tagging[J]. Journal of Chinese Information Processing,2006, 20(01):44-49.

[10] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer networks and ISDN systems, 1998, 30(98): 107-117.

[11] Wang Wei,Zhao Dongyan,Zhao Wei. Identification of Topic Sentence about Key Event in Chinese News[J]. Acta Scientiarum Naturalium Universitatis Pekinensis,2011,47(5):789−796.