J Inf Process Syst, Vol.15, No.6, pp.1365~1377, December 2019
https://doi.org/10.3745/JIPS.02.0126

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Summarizing the Differences in Chinese-Vietnamese Bilingual News

Jinjuan Wu*, Zhengtao Yu*, Shulong Liu*, Yafei Zhang*, and Shengxiang Gao*

Abstract

Summarizing the differences in Chinese-Vietnamese bilingual news plays an important supporting role in the comparative analysis of news views between China and Vietnam. Aiming at cross-language problems in the analysis of the differences between Chinese and Vietnamese bilingual news, we propose a new method of summarizing the differences based on an undirected graph model. The method extracts elements to represent the sentences, and builds a bridge between different languages based on Wikipedia's multilingual concept description page. Firstly, we calculate the similarity between Chinese and Vietnamese news sentences, and filter the bilingual sentences accordingly. Then we use the filtered sentences as nodes and the similarity grade as the weight of the edge to construct an undirected graph model. Finally, combining the random walk algorithm, the weight of the node is calculated according to the weight of the edge, and sentences with highest weight can be extracted as the difference summary. The experiment results show that our proposed approach achieved the highest score of 0.1837 on the annotated test set, which outperforms the state-of-the-art summarization models.

Keywords

Bilingual News, Chinese-Vietnamese, Sentence Similarity, Summarizing the Difference, Undirected Graph

1. Introduction

In the Internet Age, information spreads rapidly regardless of borders. The media in different countries will report on the same event and express different opinions because of different positions. For example, on the theme of "One Belt, One Road", news reports in both Chinese and Vietnamese describe the content of the cooperation project agreement. However, Chinese news tend to emphasize the promotion of trade cooperation and cultural exchange, while Vietnamese articles tend to describe improvements in infrastructure construction and industrial development. This paper aims to summarize the differences in reporting between different languages, and generate a difference summary to help people understand events more comprehensively and accurately.

Within the field of summarizing the difference in bilingual news, cross-language analysis is a difficult issue. This problem generally can be addressed by the bilingual dictionary approach [1], parallel corpus approach [2,3] and machine translation approach [4,5]. The dictionary approach first builds a bilingual

Manuscript received November 6, 2017; first revision May 17, 2018; second revision July 24, 2018; accepted July 29, 2018.
Corresponding Author: Zhengtao Yu (ztyu@hotmail.com)
* Dept. of Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, China
(2698925796@qq.com, ztyu@hotmail.com, {1217194209, 76326474}@qq.com, gaoshengxiang.yn@foxmail.com)

alignment dictionary, and aligns the key words (e.g., emotional words or entities) as required. This method assumes that bridges can be built between different languages through clearly aligned keywords. For example, Mihalcea et al. [1] use both an English-Romanian and general dictionary to construct a bilingual-aligned subjective dictionary. The parallel corpus approach mainly uses the alignment relationship of a parallel corpus which is composed of the source language and its translation into other languages. The parallel corpus includes word-level, sentence-level and chapter-level alignment, but the difficulty lies in the fact that the corpus is not easy to obtain. For example, Banea et al. [3] propose to use subjective and objective classifiers of source language and parallel corpus to classify objective language. In recent years, the results of machine translation have improved, and it is gradually becoming an effective means of cross-language analysis. For example, Banea et al. [4] propose two different cross-language methods using machine translation: source language translation to target language, and target language translation to source language. Research into news summary extraction can be divided into topic representation approaches and indicator representation approaches [6]. Topic representation approaches first convert the text into a series of topics, then calculate the importance of the sentences according to the topic, and finally select the important sentences as the summary. Gillick et al. [7] propose the use of higher frequency words as topic representations, and these higher frequency words tended to be domain specific. Celikyilmaz and Hakkani-Tur [8] suggest using the hLDA model to calculate important topics in multi-document news, and then generate a summary. The authors [9] propose to use cosine distance to compute sentence similarity, and cluster sentences to extract the topic. Indicator representation approaches directly express the sentence into the feature vector and then calculate the importance of the sentence. For example, graph models [10,11] are used to calculate the importance of sentences, where graph vertices represent sentences, edges represent cosine similarity between sentences, the random walk algorithm is used to calculate the weight of the vertices, and the high weight would be used to select the most important sentence as a summary. Wan and Zhang [12] propose a novel system to incorporate the new factor of information certainty into the summarization task, which produce better content quality. The rise in the study of deep learning has also contributed to the extractive summarization task. Some methods use neural networks in the single document summarization framework [13-15]. They formulate sentence ranking as a hierarchical regression process. Given sentences with labeled importance scores [13], or the symbol [14,15] of 0 or 1, which indicates whether to extract the sentence into the summary or not. Unfortunately, the application of neural networks methods to bilingual multi-document summarization is difficult. Not only encoding and decoding for a long sequence of multiple sentences still lack satisfactory solutions [16], but it also lacks a large-scale corpus for training.

The existing approaches to summary extraction mainly involves single language documents, which aim to extract the important content of news and eliminate redundant information. In this paper, we analyze the multilingual news and extract different information. Singh et al. [17] propose to use a restricted Boltzmann machine to generate a summary retaining its important information. In recent years, graph-based ranking algorithm has been widely used for this task, such as the research conducted by Wan et al. [18], who propose a ranking method based on a graph to score the importance and differences in Chinese-English documents and then select sentences with high scores to generate a summary. The current article focuses on Chinese and Vietnamese news documents, with the research methods divided into two steps. First, the similarity information is filtered according to the cosine similarity between Chinese and Vietnamese news sentences. Second, the graph model is constructed, and the random walk algorithm is used to extract the representative sentences to generate the summary.

2. A Summary Method of News Difference Based on a Graph Model

To reflect the difference between Chinese and Vietnamese news. First, we extract the elements contained in the news documents to characterize the sentences. Second, we calculate the similarity between cross-language news to filter out the highly similar sentences. Third, the sentences that had not been filtered out as the vertices to construct the graph model. Finally, we use the random walk algorithm to obtain the weight of the vertices, that is the importance of the sentence, with the most important (n) selected as the summary.

The method of implementation is shown in Fig. 1.

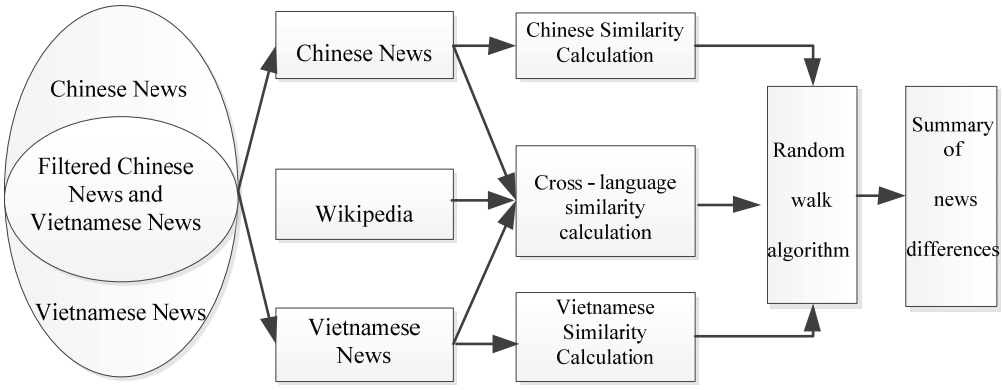


Fig. 1. A summary method of news difference based on a graph model.

2.1 The Extraction of Bilingual News Elements

The elements [19] contain important information such as the time, place, participant, and institution in the news events.

This paper aims to extract elements contained in Chinese and Vietnamese sentences, and use them to characterize the sentences. The extraction of Chinese elements use the LTP cloud platform [20]. We set named entities as news elements then obtain the collection of Chinese elements $E_{cn} = \{e_{c1}, e_{c2}, ..., e_{cm}\}$. Due to the lack of Vietnamese named entity recognition tools in the process, word segmentation tool [21] can be used to segment sentences, and part-of-speech tagging. We then manually extract the elements according to the processing results to obtain the collection of Vietnamese elements $E_{ve} = \{e_{v1}, e_{v2}, ..., e_{vn}\}$. Chinese and Vietnamese sentences are characterized by elements, for example $S_k = \{e_1, e_2, ..., e_k\}$.

2.2 Filter Similar News Sentences

Chinese and Vietnamese sentences with high similarity will not reflect differences. Based on this consideration, initial filtering is carried out according to the similarity before the sentence is analyzed. As Chinese-Vietnamese machine translation technology is not mature, we cannot simply translate bilingual news into one language. We therefore seek help from the multi-language concept description pages on Wikipedia [22]. The translation between concepts corresponded, so this is used in the calculation of Chinese/Vietnamese semantic similarity to realize the analysis of sentence relations.

There are many language options in Wikipedia, in which Chinese and Vietnamese concepts are the basis for similarity calculation between Chinese and Vietnamese words. Using this method [22], we first extract the Chinese/Vietnamese concept set with correspondences in Wikipedia, constructing a bilingual concept feature space. Then, words are represented as vectors by the mapping of feature spaces. Finally, the similarity between two vectors is calculated by the cosine. In our proposed approach, the input are the Chinese word e^{cn} and the Vietnamese word e^{ve} , let the two vectors are represented by $\vec{e}^{cn} = \left\{e_1^{cn}, e_2^{cn}, ..., e_n^{cn}\right\}$ and $\vec{e}^{ve} = \left\{e_1^{ve}, e_2^{ve}, ..., e_n^{ve}\right\}$, respectively. The formula for semantic similarity of Chinese and Vietnamese words is as follows:

$$Sim(e^{cn}, e^{ve}) = \frac{\sum_{i=1}^{n}(e_i^{cn}, e_i^{ve})}{\sqrt{\sum_{i=1}^{n}(e_i^{cn})^2}\sqrt{\sum_{i=1}^{n}(e_i^{ve})^2}} \tag{1}$$

Each news sentence is characterized by one or more elements, so similarity of the sentences can be computed by the similarity of the elements it contains. Assuming two sentences s_i and s_j contain the elements $e_1, e_2, \cdots e_m$ and $e_1, e_2, \cdots e_n$ after the word segmentation and part-of-speech tagging, that is, s_i is composed of m words and s_j is composed of n words.

The sentence similarity calculation method is based on the set of extracted elements. Words are selected one by one from the set of elements in a sentence to calculate the similarity with words in the element set of the same language documents. The word pair that obtains maximum similarity will be selected until the sentence element collection is void. Then the similarity of these word pairs will be added, and divided by the number of words contained in the sentence element set to determine similarity of the two sentences. The formula is as follows:

$$w_{ij} = \sum_{u=1}^{m} \max Sim(e_i, e_j) / m \tag{2}$$

where w_{ij} represents the similarity between the sentence s_i and s_j in the same language document, and $sim(e_i, e_j)$ means the similarity between the elements e_i and e_j. Assuming that $S_{cn} = \left\{s_1^{cn}, s_2^{cn}, ..., s_m^{cn}\right\}$ contains m sets of Chinese sentences, $S_{ve} = \left\{s_1^{ve}, s_2^{ve}, ..., s_n^{ve}\right\}$ contains n sets of Vietnamese sentences, and $W_{ij}, i \in [1, m], j \in [1, n]$ represents the similarity matrix between Chinese and Vietnamese sentences, which can be shown as:

$$W_{ij} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n-1} & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n-1} & w_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{m-11} & w_{m-12} & \cdots & w_{m-1n-1} & w_{m-1n} \\ w_{m1} & w_{m2} & \cdots & w_{mn-1} & w_{mn} \end{bmatrix} \tag{3}$$

After obtaining the similarity between Chinese and Vietnamese sentences, it is obviously unreasonable to filter sentences directly according to the similarity. For example, assume that the threshold is α, and

$w_{24} \geq \alpha$, which satisfy the condition that the similarity is greater than the threshold. If the sentences s_2^{cn} and s_4^{ve} are directly filtered only because of high similarity between them, the accuracy of the summary will be affected. $w_{24} \geq \alpha$ can only indicate that there is little difference between sentence s_2^{cn} and s_4^{ve} ,but sentence s_2^{cn} may still be different from Vietnamese sentences with the exception of s_4^{ve}. Sentence s_4^{ve} may still be different from Chinese sentences except for s_2^{cn} .

Based on the above considerations, the following method is adopted. First, global similarity is calculated for each sentence. Second, the sentences are filtered according to whether the global similarity of the sentences satisfy the threshold condition.

$$sim(s_i^{cn}) = \frac{1}{n}\sum_{j=1}^{n} w_{ij}, i = 1,2,...,m \tag{4}$$

$$sim(s_j^{ve}) = \frac{1}{m}\sum_{i=1}^{m} w_{ij}, j = 1,2,...,n \tag{5}$$

where $sim\left(s_i^{cn}\right)$ and $sim\left(s_j^{ve}\right)$ represent the global similarity of Chinese sentence s_i^{cn} and Vietnamese sentence s_j^{ve} , respectively. To be specific, $sim\left(s_i^{cn}\right)$ measures the similarity between a Chinese sentence and the Vietnamese full text articles. If the global similarity is higher than the threshold, it means that the difference of the Chinese sentence and Vietnamese full text is small and the sentence should be filtered out. The Vietnamese news sentences are handled in a similar way. We set the global similarity threshold to 0.2 during the experiment.

2.3 Graph Model Construction

After initial filtering of the sentence, the Chinese sentence $S_{cn} = \left\{s_1^{cn}, s_2^{cn}, ..., s_m^{cn}\right\}$ and the Vietnamese sentence $S_{ve} = \left\{s_1^{ve}, s_2^{ve}, ..., s_n^{ve}\right\}$ are obtained, where *m* and *n* are used to indicate the quantity of the remaining Chinese and Vietnamese sentences, respectively.

The remaining sentences can, in some cases, reflect differences between different languages news. The purpose of this paper is to summarize the differences between Chinese and Vietnamese news. To achieve this goal, we need to meet two conditions. First of all, the extracted news sentences should reflect that different language sentences contain different information. Second, the extracted sentences should reflect the nature of the summary, that is, they should be representative or important. In the first step, the filtering based on the global similarity of the sentence satisfied the first condition, so the Chinese and Vietnamese sentences remaining after filtering need to be processed into a summary. To achieve this goal, we calculate the scores of the different language sentences separately, and extract high value (n) scores as a summary of news differences.

To evaluate the importance of a sentence, we can consider the following features: the similarity of sentences in the same language documents, the difference of sentences in the different language documents. The higher the similarity of the same language documents, the more the sentence can reflect the news document content. The higher the difference of the different language document sets, the more the difference can be reflected in the Chinese and Vietnamese news expressions. Using this analysis, we construct the undirected graph model shown in Fig. 2.

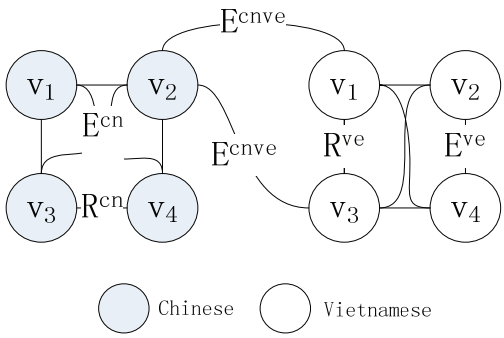


Fig. 2. Undirected graph model.

The vertices in the Fig. 2, indicate Chinese or Vietnamese sentences. E^{cn} represents the similarity between Chinese sentences; E^{ve} represents the similarity between Vietnamese sentences, and E^{cnve} represents the difference between Chinese and Vietnamese sentences. The similarity between sentences in the same language is calculated by cosine similarity. We select unigram + bigram as the feature. The sentence is represented as a vector by using the vector space model (VSM) model, and the similarity is calculated according to the cosine distance of the vector.

Bilingual sentence similarity calculation is based on Wikipedia, and the similarity of every sentence is obtained by calculating the Euclidean distance between the element vectors. Chinese word vector $\vec{e}^{cn} = \left\{e_1^{cn}, e_2^{cn}, ..., e_n^{cn}\right\}$, and Vietnamese word vector $\vec{e}^{ve} = \left\{e_1^{ve}, e_2^{ve}, ..., e_n^{ve}\right\}$.

The formula of similarity for Chinese and Vietnamese words is as follows:

$$Dis(e_i^{cn}, e_j^{ve}) = \frac{\left\|\vec{e}_i^{\,cn} - \vec{e}_j^{\,ve}\right\|}{\left\|\vec{e}_i^{\,cn}\right\| + \left\|\vec{e}_j^{\,ve}\right\|} \tag{6}$$

Similarity between Chinese-Vietnamese news sentences is as follows:

$$w_{ij} = \sum_{u=1}^{m} \max Dis(e_i^{\,cn}, e_j^{\,ve}) / m \tag{7}$$

where w_{ij} represents the similarity between sentences s_i and s_j in different language document sets, and $Dis(e_i, e_j)$ represents the similarity between elements e_i and e_j .

We construct the similarity matrix between Chinese and Vietnamese $W_{ij}^{cnve}, i \in [1,m]; j \in [1,n]$, and let $\left(W_{ij}^{cnve}\right)^T = W_{ij}^{vecn}$.

2.4 Graph Model Solving

The matrices w_{ij}^{cn} , w_{ij}^{ve} , w_{ij}^{cnve} represent the similarity between Chinese sentences, the similarity between Vietnamese sentences and the similarity between Chinese and Vietnamese sentences. The element of the matrix is equivalent to the weight of the edge in the graph model. The weight of the vertex can be calculated by the weight of the edge [18], letting $u(s_i^{cn})_{m\times 1}$ and $v(s_j^{ve})_{n\times 1}$ represent the scores of the Chinese sentence and the Vietnamese sentence. To achieve this goal, each matrix is first normalized to

get $\tilde{w}_{ij}^{cn}$, $\tilde{w}_{ij}^{ve}$, $\tilde{w}_{ij}^{cnve}$, ensuring the sum of the elements in each row of the matrix is 1.

$$u(s_i^{cn}) = \alpha \sum_j \tilde{w}_{ij}^{cn} u(s_j^{cn}) + \beta \sum_j \tilde{w}_{ij}^{vecn} u(s_j^{ve}) \tag{8}$$

$$u(s_j^{ve}) = \alpha \sum_i \tilde{w}_{ij}^{ve} u(s_i^{ve}) + \beta \sum_i \tilde{w}_{ij}^{cnve} u(s_i^{cn}) \tag{9}$$

where α and β indicate the effect of the same and different language similarity. Based on these assumptions we can see $\alpha > 0$, $\beta > 0$, let $\alpha + \beta = 1$. The above formulae are iteratively solved. To make the solution converge, $u(s_i^{cn})_{m\times 1}$ and $v(s_j^{ve})_{n\times 1}$ are normalized after each iteration. When the difference between the results of the two iterations is less than the threshold, it is assumed that the iteration ends. The scores of Chinese sentences $u(s_i^{cn})_{m\times 1}$ and Vietnamese sentences $v(s_j^{ve})_{n\times 1}$ are obtained by this method.

To further filter redundant information, we choose the greedy algorithm [23] to deal with the current score, and get the final sentence score. The algorithm for dealing with Chinese sentences was as follows:

(1) Initializes the two collections: $A = \varphi$, $B = \{s_i, i = 1,2,...,m\}$, where set B represents the Chinese sentence set.

(2) The elements in set B are sorted in reverse order with the original score $u(s_i^{cn})_{m\times 1}$.

(3) Assuming that s_i is ranked first, it is moved from set B to set A, and then the sentence score recalculated for similarity with s_i in set B. s_j is used to represent the sentence with the similarity to s_i . The score is calculated as follows: $score(s_j) = u(s_j^{cn}) - \varphi \cdot w_{ij}^{cn} \cdot u(s_j^{cn})$, where $u(s_j^{cn})$ represents the original score of sentence s_j , φ represents penalty factor, and w_{ij}^{cn} represents the similarity of s_i and s_j . When the penalty factor φ is 0 there is no penalty, and the sentence s_j score is unchanged. We use an experimental selection penalty factor of 0.5.

(4) The score calculated in the previous step was used to reverse the order of the elements in set B and then return to the third step until the number of elements in set B was zero.

The algorithm for Vietnamese news processing is consistent with the above. It needs to replace the input into Vietnamese sentence sets, so that the original score matrix is replaced by $v(s_j^{ve})_{n\times 1}$, and the similarity matrix is replaced by w_{ij}^{ve} . Using this method to calculate the final score of Chinese and Vietnamese sentences, and sort the sentences according to the final scores for each language, the top (n) sentences are extracted as summaries of news differences.

3. Experiments and Result

3.1 Data Set

The experimental data set contains Chinese and Vietnamese news of three topics. We searched http://google.com.hk/ to obtain the news documents related to the topics. Some documents were collected manually as data sets for the experiments. The specific information is shown in Table 1.

Table 1. Specific data for the experiment

Topic	Language	Number of sentences	Average length
Nguyen Phu Trong's visit to China	Chinese	421	34
	Vietnamese	394	46
Releasing water into the Mekong river	Chinese	405	29
	Vietnamese	399	31
Defense Minister meeting	Chinese	388	38
	Vietnamese	149	42

To evaluate the results, we read the Chinese and Vietnamese sentences on each of the three topics. Based on our full understanding of these news items, we chose 5 sentences from each language to form a summary of the differences in the news.

3.2 Evaluation Metrics

The top 5 sentences from each language were extracted as different sentences for the experiment. To evaluate the effect of the algorithm, we used the n-gram co-occurrence measure proposed by Lin and Hovy [24]. This method evaluates the model by calculating the degree of n-gram co-occurrence between the model summary and the manual summary. The higher the co-occurrence, the better the effect of the model. The calculation method is as follows:

$$C_{\mathrm{n}}=\frac{\sum_{C\in\{M\,\mathrm{od}el\}}\sum_{n-gram\in C}Count_{match}\left(n-gram\right)}{\sum_{C\in\{M\,\mathrm{od}el\}}\sum_{n-gram\in C}Count\ \left(n-gram\right)} \tag{10}$$

where $Count_{match}\left(n-gram\right)$ represents the number of n-gram co-occurrences between the model summary and the manual summary, $Count\left(n-gram\right)$ represents the number of n-gram in the model summary.

$$\mathrm{N}gram\left(i,j\right)=\exp\left(\sum_{i}^{j}w_n\log C_n\right)i\le j;i,j\in\left[1,4\right] \tag{11}$$

where W_n is the normalization factor and $w_n=\frac{1}{j-i+1}$, when $i=j=1$, $Ngram(1,1)$ represents the degree of unigram co-occurrence, $Ngram(1,2)$ represents the degree of unigram+bigram co-occurrence.

3.3 Evaluation Results

This paper selected the following three baselines methods to show the effectiveness of our proposed approaches.

Centroid [25]: A centroid-based method is used to calculate the saliency scores of sentences in the different languages. First, we calculate three scores: the centroid value, the position value and the overlapping value of the first sentence. Second, the three values are linearly summed to get the sentence score. Finally, the redundant information is removed to obtain the summary sentence. It is worth noting that this method does not use cross-language information.

Centroid++: This is an improved method based on the centroid method, which integrates cross-language information. The final score of the sentence comes from subtracting cross language similarity from scores calculated by the centroid method, and further reflects the differences between the different languages.

PBES [26]: Phrase-based extractive summarization [26] uses phrase-based scoring to represent saliency scores of sentences. We can assign phrase-based scores to sentences from the translated documents for summarization purposes. The model can operate on lexical entries with more than one word in the source and target languages. This works well with cross-language document summarization.

In the initial filtering of bilingual sentences according to global similarity, the global similarity threshold was set to 0.2, i.e. when we take 0.2, about 30% of the sentences are filtered out. In addition, the purpose of this paper was to extract the difference summary, not only concerned with the difference between the languages, but also the importance of sentences in the same language. We set $\alpha = 0.5$, $\beta = 0.5$ in the random walk algorithm, which means the similarity between different languages and between the same languages contribute equally to the final score of the sentence. We used the settings to implement the methods given in this article, and to achieve three baseline methods. Based on the n-gram co-occurrence measure, the $Ngram(1,1)$ and $Ngram(1,2)$ of the three different methods were calculated. Table 2 shows the experimental results of the Chinese difference summary. Table 3 shows the experimental results of the Vietnamese difference summary.

Table 2. Chinese difference summaries results

	$Ngram(1,1)$	$Ngram(1,2)$
Centroid	0.1056	0.0837
Centroid++	0.1301	0.1194
PBES	0.1536	0.1372
Our method	0.1837	0.1481

Table 3. Vietnamese difference summaries results

	$Ngram(1,1)$	$Ngram(1,2)$
Centroid	0.0949	0.0643
Centroid++	0.1403	0.1125
PBES	0.1447	0.1292
Our method	0.1821	0.1462

We compared the output of the model to other summary systems. The first two methods pay more attention to the location characteristics of sentences when extracted. PBES analyzes the relation between bilingual sentences by machine translation. This paper studies the problem of cross-language document summarization in Chinese and Vietnamese. Vietnamese is a minority language, and the results of machine translation are not optimal. In response, our method builds a bridge between different languages based on Wikipedia's multilingual concept description page, extracting elements to represent the sentences. It can be seen from Tables 2 and 3 that our method is superior to Centroid, Centroid++ and PBES under the same evaluation method, whether Chinese or Vietnamese news data is examined.

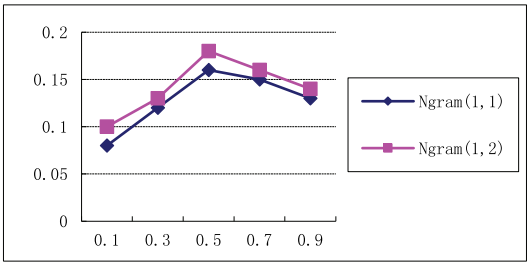


Fig. 3. The influence of α value on Chinese summary.

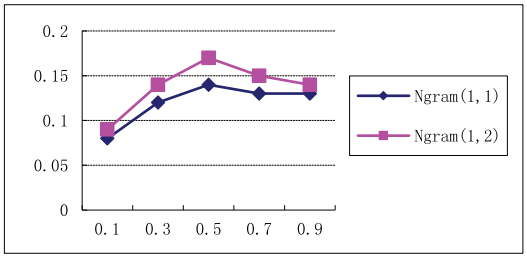


Fig. 4. The influence of α value on Vietnamese summary.

The effect of α on the experimental results can be observed in Figs. 3 and 4, which indicate the influence on Chinese and Vietnamese respectively. It can be seen that the experimental results gradually improve with the increase of α value, peaking at about 0.5, then gradually decreasing with the increase of α value.

Table 4. Summary of news differences

Chinese difference summary	Vietnamese difference summary
很多人为中国慷慨激昂的大国风范点赞；越南干旱的问题显然不能怨我们, 那到底该怨谁呢？ 越南正在遭遇近一个世纪以来的最严重干旱， 湄公河三角洲地区农业受到严重打击. 而且由于当地种植水稻, 需水量也进一步加大, 使得旱情显得越发严重. 这次是当地90年来最严重旱灾, 近百万人缺乏日常用水, 近16万公顷稻田受灾. (A lot of people praise China for our manners as a big country. The drought in Vietnam obviously cannot be blamed on us, but who should take the blame? Vietnam is now suffering its worst drought in nearly a century, and the agriculture of the Mekong Delta region has been hit hard. The drought becomes even more severe because of the increasing demand for water as a result of local planting of rice. This is the most serious drought in 90 years, which causes nearly a million people to lose access to water for everyday needs and nearly 160,000 hectares of rice fields are affected).	Trong khi đó, mùa này, nước thượng nguồn sông Mê Kông lại đổ về rất ít do bị ngăn cản bởi hàng loạt các công trình thủy điện của Trung Quốc, Lào, Campuchia. Vậy thì làm sao đập Cảnh Hồng có thể xả cho chúng ta trong nhiều đợt khi không đủ nước? Dòng chảy sẽ chảy qua các nước phía trên, trong khi, Thái Lan, Lào, Campuchia cũng đang bị hạn rất nặng nề. Hiện nay, các nước thuộc hệ thống sông Mekong có một cơ chế hợp tác quan trọng thông qua Hiệp hội sông Mekong. Việt Nam đề nghị Trung Quốc xả lũ cứu hạn đồng bằng sông Cửu Long. (Meanwhile, the upstream Mekong River is falling back very little in this season because it is blocked by a series of hydropower projects in China, Laos and Cambodia. Why are we still lacking water when Jinghong dam released water? The water will flow through Thailand, Laos, Cambodia and other countries which also suffer severe drought. The Mekong River Basin countries have important cooperation mechanisms through the Mekong River Commission at present. Vietnam calls on China to increase its discharge flow in the Mekong Delta).

Finally, we selected the topic "Mekong River", and used this method to summarize the differences in Chinese-Vietnamese bilingual news as shown in Table 4. Here the proposed method extracts different viewpoints from the Chinese and Vietnamese news on the Mekong River topic. The Chinese summary paid attention to Vietnam's severe drought and provides an objective analysis of the shortage of water resources. The Vietnamese summary emphasized the limited flow of the Mekong to particular areas and the need of the China Hydropower Station to discharge water. To a certain extent, the differences between Chinese and Vietnamese news are reflected here.

4. Conclusions

In this paper, we have proposed a method based on a graph model to summarize the differences between Chinese and Vietnamese bilingual news. In the proposed method, multilingual conceptual description pages on Wikipedia were used to analyze sentences similarity, which contribute to solving the graph model and further complete the summary task. The experiments are giving to show the effectiveness of our proposed approach.

Acknowledgement

This work was supported by National key research and development plan project (No. 2018YFC0830105, 2018YFC0830100), National Nature Science Foundation (No. 61732005, 61672271, 61761026, 61662041, 61762056), High-tech Industry Development Project of Yunnan Province (No. 201606), and Natural Science Foundation of Yunnan Province (No. 2018FB104).

References

[1] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 2007, pp. 976-983.

[2] M. S. Almeida, C. Pinto, H. Figueira, P. Mendes, and A. F. Martins, "Aligning opinions: cross-lingual opinion mining with dependencies," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, 2015, pp. 408-418.

[3] C. Banea, R. Mihalcea, and J. Wiebe, "Porting multilingual subjectivity resources across languages," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 211-225, 2013.

[4] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: are more languages better?," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010, pp. 28-36.

[5] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan, "Multilingual subjectivity analysis using machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, HI, 2008, pp. 127-135.

[6] A. Nenkova and K. McKeown, "A survey of text summarization techniques," in *Mining Text Data*. Boston, MA: Springer, 2012, pp. 43-76.

[7] D. Gillick, B. Favre, and D. Hakkani-Tur, "The ICSI Summarization System at TAC 2008," 2008; https://pageperso.lis-lab.fr/benoit.favre/papers/favre_tac2008.pdf

[8] A. Celikyilmaz and D. Hakkani-Tur, "A hybrid hierarchical model for multi-document summarization," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 815-824.

[9] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization," *Information Processing & Management*, vol. 33, no. 2, pp. 193-207, 1997.

[10] Y. Li and S. Li, "Query-focused multi-document summarization: combining a topic model with graph-based semi-supervised learning," in *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, 2014, pp. 1197-1207.

[11] D. Parveen and M. Strube, "Integrating importance, non-redundancy and coherence in graph-based extractive summarization," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp. 1298-1304.

[12] X. Wan and J. Zhang, "CTSUM: extracting more certain summaries for news articles," in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, Gold Coast, Australia, 2014, pp. 787-796.

[13] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, "Ranking with recursive neural networks and its application to multi-document summarization," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX, 2015, pp. 2153-2159.

[14] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," 2016; https://arxiv.org/abs/1603.07252.

[15] S. Narayan, N. Papasarantopoulos, S. B. Cohen, and M. Lapata, "Neural extractive summarization with side information," 2017; https://arxiv.org/abs/1704.04530.

[16] J. G. Yao, X. Wan, and J. Xiao, "Recent advances in document summarization," *Knowledge and Information Systems*, vol. 53, no. 2, pp. 297-336, 2017.

[17] S. P. Singh, A. Kumar, A. Mangal, and S. Singhal, "Bilingual automatic text summarization using unsupervised deep learning," in *Proceedings of 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, India, 2016, pp. 1195-1200.

[18] X. Wan, H. Jia, S. Huang, and J. Xiao, "Summarizing the differences in multilingual news," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing, China, 2011, pp. 735-744.

[19] Linguistic Data Consortium, "ACE 2005 – Chinese entities V5.5," 2005; https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications.

[20] W. Che, Z. Li, and T. Liu, "LTP: a Chinese language technology platform," in *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, Beijing, China, 2010, pp. 13-16.

[21] SourceForge.Net, "JVnTextPro: A Java-based Vietnamese Text Processing Tool," 2010; http://jvntextpro.sourceforge.net/.

[22] Q. Yang, Z. Yu, X. Hong, S. Gao, and Z. Tang, "Chinese-Vietnamese word similarity computation based on Wikipedia," *Journal of Nanjing University of Science and Technology*, vol. 40, no. 4, pp. 461-466, 2016.

[23] X. Wan, J. Yang, and J. Xiao, "Manifold-ranking based topic-focused multi-document summarization," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 2007, pp. 2903-2908).

[24] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada, 2003, pp. 150-157.

[25] D. R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919-938, 2004.

[26] J. G. Yao, X. Wan, and J. Xiao, "Phrase-based compressive cross-language summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 118-127.

Jinjuan Wu https://orcid.org/0000-0003-1577-6445

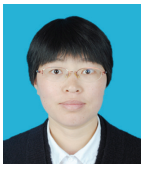
She is current a postgraduate in the Kunming University of Science and Technology, Kunming, China. She focus on nature language processing and information retrieval.

Zhengtao Yu https://orcid.org/0000-0002-4012-461X

He is currently a professor and Ph.D. supervisor at School of Information Engineering and Automation, and the chairman of Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, China. He received the Ph.D. degree in Computer Application Technology from Beijing Institute of Technology, Beijing, China, in 2005. His main research interests include natural language processing, machine translation and information retrieval.

Shulong Liu https://orcid.org/0000-0003-3063-8454

He is current a postgraduate in the Kunming University of Science and Technology, Kunming, China. He focus on nature language processing and information retrieval.



Yafei Zhang https://orcid.org/0000-0003-2347-5642

She is currently a lecturer and master's supervisor at College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. She received the Ph.D. degree in Signal and information processing from Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2008. Her main research interests include image processing and natural language processing.

Shengxiang Gao https://orcid.org/0000-0002-2980-8420

She is lecturer at Kunming University of Science and Technology, Kunming, China. She is also a CCF member since 2013. She received the bachelor's degree in industrial automation, the M.S. degree in pattern recognition and intelligent system and the Ph.D. degree from Kunming University of Science and Technology in 2000, 2005, and 2016, respectively. Her research interests include nature language processing, machine translation, and information retrieval.