J Ambient Intell Human Comput (2012) 3:305–313
DOI 10.1007/s12652-012-0143-x

ORIGINAL RESEARCH

TSGVi: a graph-based summarization system for Vietnamese documents

Tu-Anh Nguyen-Hoang · Khai Nguyen · Quang-Vinh Tran

Received: 30 June 2011 / Accepted: 5 June 2012 / Published online: 27 June 2012

Abstract This paper proposes an automatic method to generate an extractive summary of multiple Vietnamese documents which are related to a common topic by modeling text documents as weighted undirected graphs. It initially builds undirected graphs with vertices representing the sentences of documents and edges indicate the similarity between sentences. Then, by adopting PageRank algorithm, we can generate salient scores for sentences. Sentences are ranked according to their salient scores and selected based on maximal marginal relevance to form the summaries. These summaries are combined and applied the same process one more time to form the final extractive summary of the document set. A series of experiments are performed on Vietnamese news articles and English data of DUC 2002, 2003, 2007. The results demonstrate the effectiveness of the proposed technique over reference systems.

Keywords Graph model · Weighted PageRank · Sentence extraction · Multi-document summarization · Vietnamese

1 Introduction

The rapid growth of Internet has been accompanied by the explosion of online textual information. Thus, it becomes more difficult for a user to cope with all the text that is potentially of interest and it needs to develop more efficient and high quality summarization systems. The aim of text summarization is to extract content from an information source and present the most important content to the user in a condensed form and in a manner sensitive to the user's or an application's need (Mani and Bloedorn 1997). Text summarization techniques can be categorized along two categories: abstraction and extraction. An extract-summary consists of sentences extracted from the document while an abstract-summary may employ words and phrases that do not appear in the original document. Usually, abstractive summarization requires heavy machinery for language generation and is difficult to replicate or extend to broader domains. In contrast, simple extraction of sentences has produced satisfactory results in large-scale applications, especially in multi-document summarization.

Though many achievements are achieved for English documents, there are no excellent systems for Vietnamese text summarization yet because of the flexibility of the grammar in Vietnamese sentences. Primary study on single document summarization uses statistical-based sentence extraction approach (Ha et al. 2005; Nguyen et al. 2005). Vietnamese multi-document summarization is a new research focus and research work is being carried on recently (Phuc and Hung 2008).

In this paper we extend our initial results in Nguyen et al. (2010) and concentrate on the shallow approach of text summarization by using sentence extractions and propose a special method for Vietnamese multi-document summarization. This method mainly consists of three phases. Firstly, we add the structure to each document in the data set, which can then be viewed as an undirected weighted graph. These graphs are built based on title and sentences within the document. Secondly, the graph-based

T.-A. Nguyen-Hoang (✉) · K. Nguyen · Q.-V. Tran
Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh, Vietnam
e-mail: nhtanh@fit.hcmus.edu.vn

K. Nguyen
e-mail: nhkhai@fit.hcmus.edu.vn

Q.-V. Tran
e-mail: tqvinh@fit.hcmus.edu.vn

ranking algorithm weighted PageRank is performed on the graph for generating salient scores of each sentence in the document. Important sentences containing significant information of the text will get the higher scores or ranks. We extract several top-ranking sentences in the document to form the summary for this one. Then we merge all summaries into one single document. Finally, we apply the same process to this combined single document with a modification at the sentence extraction step. Redundancy is a problem in multi-document summarization due to the fact that sentences with similar meaning can come from different single-document. Therefore, instead of typically select top ranked sentences, we use the Maximal Marginal Relevance (MMR) algorithm to form the final extractive summary.

To the best of our knowledge this is the first time the graph model and ranking algorithm have been used for Vietnamese text summarization. Our proposed summarization method has several benefits. Firstly, because this method is an unsupervised learning approach, it requires no training data. Secondly, this method is domain-independent as well as language-independent, then we do not need to consider either domain-specific knowledge or deep linguistic analysis of texts. This method is considered appropriating to linguistic characteristics of Vietnamese and do not require more linguistic resources which are still limited in Vietnamese. It make ours easy to implement whereas still obtain acceptable and satisfactory result.

The rest of this paper is organized as follows. Section 2 briefly describes related work on text summarization. In Sect. 3, we deal with the graph based document representation model. Section 4 describes the method of Vietnamese multi-documents summarization using sentence extraction based on graph model. The results of experiments on the dataset of Vietnamese and English are discussed in Sect. 5. We conclude the paper in Sect. 6 with pointers to future research.

2 Related work

Automatic text summarization has attracted much attention since the original work by Luhn in the 50's (Luhn 1958). Automatic methods of summarization have used three main approaches: linguistic (McKeown et al. 2002; Mittal et al. 1999), statistical (Berger and Mittal 2000; Carbonell and Goldstein 1998; Nomoto and Matsumoto 2001) and combinations of the two approaches (Barzilay and Elhadad 1997; Goldstein et al. 1999; Schiffman et al. 2001). In this section, we review some work done in multi-document summarization, approaches relied on graph model, and methods have applied for Vietnamese text. Carbonell et al. create a multi-document summary by first finding passage similarity using MMR for multiple documents on the same topic (Carbonell and Goldstein 1998). MMR selects a sentence in such a way that it is both relevant to the query and has the least similarity to sentences selected previously. McKeown et al. introduces the Newsblaster summarizer (McKeown et al. 2002) which integrates machine learning and statistical techniques to identify similar sentences across the input articles.

Lin and Hovy introduces NeATS (Lin et al. 2002), which uses a six-step process for multi-document summarization. The system combines a number of techniques that had already been applied to single document summarization including sentence position, term frequency, topic signature, term clustering, MMR, word filters, and time stamps. The MEAD (Radev 2001) is a multi-document summarizer that uses the centroids of the clusters of related documents in order to extract sentences central to the topic and selects these sentences to form the summary. Sentences are scored based on a linear combination of their centroid score, text position value, and overlap with the title sentence. Salton et al were first to apply graph based degree centrality measure to extract important paragraphs from single document (Salton et al. 1997). The documents are modeled using undirected graphs with the vertices representing paragraphs, and edge weights representing the similarity between the paragraphs using cosine similarity measure.

Mani and Bloedorn propose a method to summarize similarities and differences in a pair of related documents using graph representation of text (Mani and Bloedorn 1997). They represent each document as a graph, where terms are nodes and edges correspond to semantic relationships between terms. Activated graphs of two documents are matched in order to find a graph corresponding to similarities and differences between the pairs. This graph is then used for synthesizing the summary.

Zha (2002) has used a bipartite graph representation of terms and sentences for generic text summarization. A spectral graph clustering algorithm is used to partition sentences of the documents into topical groups. Within each cluster the saliency scores for terms and sentences are calculated using mutual reinforcement principal which assigns high salient scores to the terms that appear in many sentences with high salient scores, and to the sentences that contain many terms with high salient score.

TextRank (Mihalcea and Tarau 2004) and LexRank (Erkan and Radev 2004), the graph-ranking based methods have been proposed for computing relative importance of sentences. They first build a directed or undirected graph, where individual sentences are modeled as nodes and edge is weighted to reflect the relationship between the two sentences it connects. LexRank (Erkan and Radev 2004) uses PageRank to determine sentence importance. TexRank (Mihalcea and Tarau 2004) examined several graph

ranking methods originally proposed to analyze webpage prestige, including PageRank and HITS for single-document summarization. They extended the algorithm (Mihalcea and Tarau 2004) for multiple documents. A meta-summary of documents was produced from a set of single-document summaries in an iterative manner (Mihalcea and Tarau 2005a).

Wei et al. (2010) integrate document-document and document-sentence relations into the graph-based models. These relations are used to adjust the weights of the sentence-level vertices and the strength of the sentence-level edges in the graph. They develop a graph-based sentence ranking algorithm, namely DsR (Document-Sensitive Ranking) to truly summarize multiple documents rather than a single combined document.

Previous works on Vietnamese text summarization have used statistical method (Ha et al. 2005), Self Organizing Map (SOM) (Phuc and Hung 2008), and machine learning approach (Nguyen et al. 2005) in developing automated text summarization system.

The authors of Ha et al. (2005) combine some statistical sentences extraction methods for single document summarization. They choose important sentences by assigning weights to each of them. The total weight of each sentence is calculated as liner combination of title weight, position weight, proper-noun weight, correlation weight, and TF-IDF weight. But the system is not generalized because these weights are mostly dependent on the type of the document.

Phuc and Hung (2008) use SOM with two dimension output layer for clustering documents representing by graphs. In this graph, nodes represent words, and there is an edge between two words if these words are adjacent somewhere in a document. A main idea of documents is the sentences containing as much as the words determined by the order of occurrence on the weighted graphs of SOM output layer. It is created based on the weighted graph representing a group of similar documents. The experimental results more concentrated on clustering performance than extraction main ideas.

The authors of Nguyen et al. (2005) propose a statistic machine learning approach, in which SVM ensemble is used to extract important sentences from single document. Because this is a supervised learning method so labeled data, indicate which sentence is important and which one is not, are needed to train the classifier. Then the outcome of the system is dependent on the training data and the topic of the document.

3 Graph based document representation model

In the graphic model, documents are transformed into a graph or set of graphs. The main benefit of graph-based techniques is to keep the inherent structural information of the original document. There are numerous methods for creating graphs from documents. Depending on the application, text units with various sizes and characteristics can be added to the graph as vertices, e.g. words, collocations, entire sentences, or others. In other words, it is the application that determines which type of relation is used to draw connections between two vertices, e.g. lexical or semantic relations, contextual overlap, etc.

For text summarization task, given a document d, let $G = (V, E)$ be an undirected graph represent the document d with the set of nodes V and set of edges E. Under this model, the nodes represent the sentences in d. Each edge e indicates the similarity between v_i and v_j. Two sentences are connected if and only if they are similar to each other and must satisfy a similarity threshold t. Each node in V is also labeled with their salient score. This score, computed by ranking algorithm, illustrates the amount of information that a sentence contains.

4 Vietnamese text summarization

We develop a multi-document summarization model for Vietnamese documents called TSGVi (*Text Summarization based on Graph for Vietnamese documents*). Figure 1 shows the overview of our summarization model. The input to the model is a set of related documents. Firstly, the set of documents is pre-processed. The undirected weighted graph is constructed for each document with sentences as nodes and similarities as edges. Thereafter, weighted ranking algorithm PageRank is performed on the graph to generate salient score for each sentence in the document. The sentences are ranked according to their salient scores. The top-ranking sentences are selected to form the summary for each document and MMR also is used to filter out redundant information. Secondly, all the single summary of each document are assembled into one document. Finally, the described above process is applied to this combining document to form the final extractive summary.

4.1 Pre-processing

Before constructing graph, the input set of related documents needs to be preprocessed. In the first step, input documents are parsed to extract all sentences. Those sentences, which are too short or almost contain no information, are eliminated. After that, these sentences are tokenized. While English is an inflexional language, Asian languages such as Chinese and Vietnamese are isolating languages. These languages have no explicit word boundary.

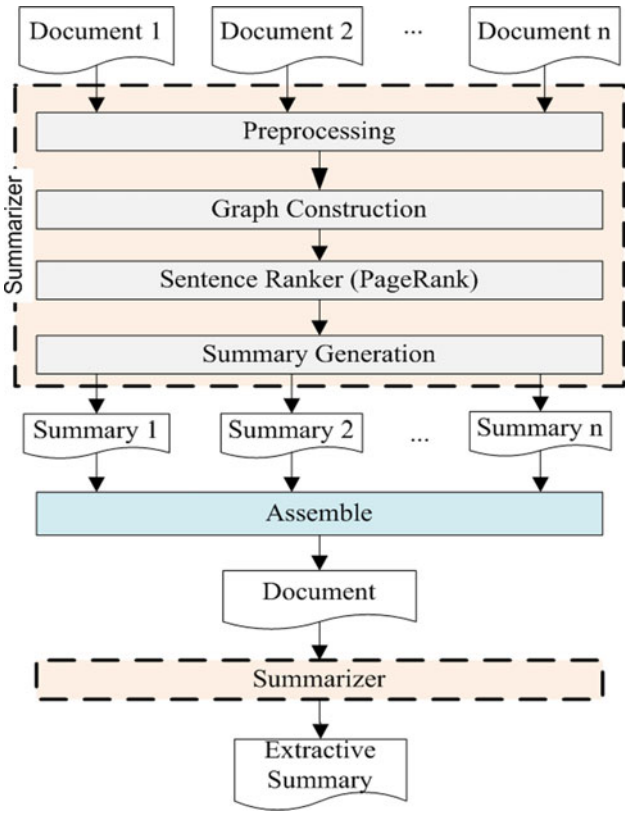

Fig. 1 The main process of our model TSGVi

Vietnamese has a special unit called *ting* which corresponds at the same time to a syllable in phonological respect, a morpheme in syntax respect, a semanteme in word structure respect, and a word in sentence constituent creation respect. There are three kinds of *ting*:

Firstly, *tings* with real meaning like *sng* (river), *ni* (mountain), *i* (go), *ng* (stand), *nh* (remember), *thng* (love tenderly) ..., which can stand alone as a sentence constituent and have all semantic and syntactic behaviour, are called typical words.

Secondly, *tings* like *nhng* (but), *m* (that), *tuy* (though), *nn* (so) ..., which cannot be a single sentence constituent but are used to compose sentence constituent and have syntactic meaning as typical words, are called tool words.

Finally *tings* that come from Chinese like *sn* (mountain), *thu* (water), *gia* (home), *bt* (not) ... or that have unclear meaning and usually composed with another syllable like *c* (*xe c*—vehicle), (*p*—beautiful), v(*vui v*—joyful) ... have role of creating word, and can be temporary used like word.

Among various definitions of Vietnamese word, the linguists reach the unanimous agreement that considers word like the smallest unit, which has fully specified meaning and stable structure and which is used to compose sentence constituents. Vietnamese lexicon contains:

Simple words or monosyllable words corresponding to *ting* of categories 1 and 2. Complex words having more than one syllable. There are principally three types of syllable combination: phonetic reduplication (e.g. *trng*/white—*trng trng*/whitish), semantic coordinated compound (e.g. *qun*/trousers, *o*/shirt—*qun o*/clothes) and semantic major/minor compound (e.g. *xe*/vehicle,/pedal—*xe p*/bicycle). We also notice the existence of some compound words whose syllable combination is no more recognizable (*b nng*/pelican).

Furthermore, idioms and locutions, which are generally considered as lexical units in sentence constituents.

Because of high compound word frequency, Vietnamese text tokenization task is rather complicated.

For the work of tokenizing Vietnamese sentence, we have developed a words segmentation tool based on Left-Right Maximum Matching algorithm. In order to improve the speed of our model, hash table is also used for indexing words in documents. After that, stop words which do not bring any information (e.g., *v*, *ca* and *l*) are removed. For this purpose, a list of stop words is prepared and used in the preprocessing phase as a stop list (about 900 words, collected manually).

4.2 Graph construction

This step transforms Vietnamese text documents into graph format. The undirected weighted graph $G = (V, E)$ represent each document is constructed as follow. Each sentence appearing in the document becomes a node in the graph representing that document. The edges of the graph represent similarity between the sentences. This similarity is computed by the TF-IDF function. Where *tf* is the term frequency in the document, and *idf* is the inverse document frequency. There are a lot of methods to define sentence similarity such as Jaccard, Word-Overlap, Dice can be applied for Vietnamese. We choose TF-IDF because this method considers the important of words based on its frequency when define the similarity between sentence. Note that, we do not implement semantic similarity methods because standard lexical database such as English WordNet is not yet available in Vietnamese. However, we are trying to build a small Vietnamese WordNet database to improve our work in the future. Formally, given two sentences S_x, S_y, the similarity between them can be defined as:

$$sim(S_x, S_y) = \frac{\sum_{w \in (S_x, S_y))} tf_{w,S_x} tf_{w,S_y} (idf_w)^2}{\sqrt{\sum_{w \in S_x} (tf_{w,S_x} idf_w)^2 . \sum_{w \in S_y} (tf_{w,S_y} idf_w)^2}} \quad (1)$$

Two sentences are linked if their similarity is greater than a predefined threshold t ($t = 0.5$ in the experiments). The result of this step is a highly connected graph. Each edge represents the relationship between the two sentences it connects. The edge weight reflects the strength of the

connection between the sentences in the document. This undirected weighted graph is input of the process in next section to calculate salient score for each sentence.

Table 1 shows the example of Vietnamese document written in a newspaper. The document has 10 sentences, identified from s1 to s10(the meaning of each is in the translation below). Thus, the graph is constructed with ten nodes. After the pre-processing step, the system begin to calculate the similarity between each pair of sentence. This similarity creates the weight of edges between nodes. Those, which are too small, is automatically eliminated. The result of the construction step is a high connection graph as in Fig. 2. In this figure, edges with higher weight are illustrated with stronger line.

4.3 Sentence ranker

Once document graph is built, the sentences in a document will be ranked through random walk on G. We compute a salient score for each node using *PageRank* algorithm (Brin and Page 1998). PageRank is one of the most popular link analysis algorithms and is used for web page ranking. It determines the importance of a node within a graph, based on information drawn from the graph structure. Originally, PageRank is applied to directed graph, but it can also work well on undirected graph. By this way, the output-degree and the input-degree for a vertex are the same. To integrate the weighted graph into the original

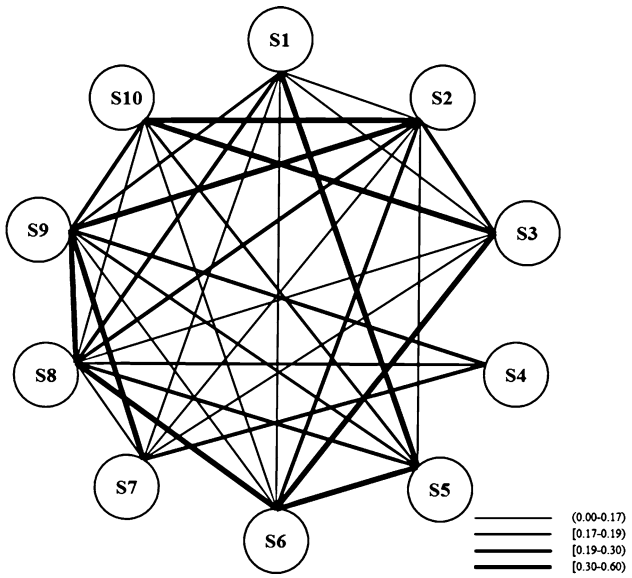

Fig. 2 The result of the graph construction

PageRank equation, we can compute a score *PR(u)* for a node u according to:

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in In(u)} \frac{w_{vu}}{\sum_{x \in Out(v)} w_{vx}} PR(v) \quad (2)$$

where N is the number of nodes in the graph, *In(u)* as the set of nodes that point to u, *Out(v)* is the set of nodes to which node v points, w_{vu} as the weight of the edge directing

Table 1 Example of Vietnamese document

ID	Text
s1	Hm th su, ngi i din ca tin o Giuseppe Rossi cho bit, Villarreal t chi ngh u tin t Barca
	On Friday, the representative of Giuseppe Rossi said that Villarreal has denied the first request from Barca
s2	T Goal nh gi chin thut tr gi ca Barca c thay i ng k sau khi Sandro Rosell c c chc ch tch
	The Goal imm the tactic of Barca has been changed significantly after Sandro Rosell become their president
s3	Hm th t, ch tch Rosell cng khai tuyn b gi tr ca Cesc Fabregas gim so vi n m ngo i
	On Wednesday, president Rosell stated that the value of Cesc Fabregas has been decreased compared to last year
s4	Do theo ng Arsenal khng th i ti 50 triu
	Thus, Arsenal cannot claim for 50 million, said the president
s5	B n cnh Rossi v Fabregas, Barca cng t ra h hng vi hai mc ti u Alexis Sanchez v Javier Pastore
	Along with Rossi and Fabregas, Barca also lost their interest in Alexis Sabchez and Javier Partore
s6	Ch tch Udinese, i s h u Sanchez, tuyn b Barca l ng c vin s mt trong cuc ua vi c c i gia bng Anh
	President of Udinese, Sanchez owner, said that Barca is prior in the race with the other clubs from Britain
s7	Tuy nhin, hm th hai tn hiu xu bt u xut hin khi c tin Man City sn sng tr 35 tri u
	However, on Monday Man City stated that they are willing to pay for 35 million
s8	Ch tch Palermo, i s hu Pastore, va ht gi 50 triu, tuy nhin p li t pha Barca l s im lng
	President of Palermo, Pastore owner, has claimed for 50 million, but all they got from Barca is a silence
s9	Hin Barca c khon n ngn hng ln n hng trm triu
	Barca still owe their creditor for hundreds of million
s10	Do ch tch Rosell t ra mt trong nhng mc tiu chnh trong nhim k l gim n v tng an ton ti chnh
	Thus, president Rosell stated that the main objective is to reduce the debt and increase the balance of financial

from node v to node u, and d is a constant *damping factor*, set at 0.85.

To calculate PR, an initial score of 1 is assigned to all nodes, and Eq. (2) is applied on a weighted graph G iteratively until the difference in scores between iterations falls below a threshold of 0.0001 for all nodes. The weights of the nodes are salient scores of the sentences. Sentences corresponding to nodes with higher scores are important, salient to the document, and have strong relationship with others sentences. After the ranking algorithm converges, the sentences are sorted according to their scores.

Algorithm 1 Pseudo code for sentence ranking

```
Require: G = (V, E)
  for all v ∈ V do
    PR(v) = 1
  end for
  repeat
    for all v ∈ V do
      find In(v), Out(v)
      compute PR^w(v) with (2)
    end for
  until converge
  rank V based on PR^w
  return the ranked list
```

4.4 Summary generation

After doing the ranking process, each sentence S_i has its salient score $PR(S_i)$. Simply, sentences with high ranking scores may be chosen as the final ones in the summary. However, there may be much redundancy among the top ranking sentences, since similar sentences tend to get similar ranking scores during the ranking process. Then, if we form the summary by select only top ranked-sentences, these similar sentences tend to be selected together and appear in the summary. This will cause the redundancy in the summary because too much similar sentences represent the same idea. Moreover, the other ideas of the documents, which contain the smaller group of similar sentences, may not be selected. Then the information of the documents can be lost.

The modified version of MMR (Carbonell and Goldstein 1998) is applied to re-rank and select sentences to add into summary. A sentence is added if it is high ranked and not too similar to any sentence existing in the summary. First, the sentence with highest rank is removed from ranked list and added to the summary. Then, the next sentence, which has the highest re-ranked score from Eq. (3), is chosen from the ranked list. This sentence is removed from the ranked list and added to the summary. This process is iterated until the summary reaches the predefined length.

$$MMR = argmax_{s_i \in R \setminus S}[\lambda.PR(s_i) - (1-\lambda).max_{s_j \in S}.sim(s_i, s_j)] \quad (3)$$

In this equation, R is the set of all sentences, S is the set of summary sentences and $PR(s)$ is the ranking score for sentences computed in previous section; λ is a tuning factor between a sentence's importance and its relevance to previously selected sentences. We choose the value $\lambda = 0.6$ for the best performance in the experiments. According to the way we construct the graph, the sentences that are similar to one or more other sentences, tend to have higher scores and thus higher ranks. These kinds of sentences are often selected to form the summary. In contrast, the sentences, which have less similar to the others, thus have less voting members, are hardly selected to the final summary. It also revealed that the use of MMR is necessary to reduce the redundancy issue.

5 Experimental evaluation

In this section, we conduct experiments to test our graph based summarization approach empirically. We used the ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin and Hovy 2003) automatic n-gram matching toolkit for evaluation, which was adopted by DUC (Document Understanding Conferences) for automatically summarization evaluation. It measures the summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. The higher the ROUGE score has, the better the system is.

5.1 Evaluation on Vietnamese dataset

The experiment corpus consists of 20 sets of related news documents which are in 6 topics, namely, politics, economics, society, sport, health and weather. All news documents are collected from various famous Vietnamese web pages such as VnExpress,[1] TuoiTre,[2] Thanhnien,[3] and Dantri.[4] Numbers of sentences contained in these news documents range from 5 to 60. Each set includes 8 to 13 documents with the same type. There are totally 207 documents in this data corpus. For each set of related documents, the experts manually constructed the 100-words summary. Details of this data corpus are given in Table 2.

In ours experiments, the proposed approach was compared with two state-of-the-art summarization systems:

[1] http://www.vnexpress.net.
[2] http://www.tuoitre.com.vn.
[3] http://www.thanhnien.com.vn.
[4] http://www.dantri.com.vn

Table 2 Detailed information on Vietnamese news documents corpus

Topic	No. set	No. doc	Avg. word
Economics	8	81	705
Society	4	40	714
Politics	3	31	564
Health	2	21	494
Weather	2	24	449
Sport	1	10	282
Total	20	207	618

Table 3 Comparison of summarization systems on Vietnamese

No	System	ROUGE-1	ROUGE-2
1	LEAD	0.5917	0.2036
2	LexRank	0.5816	0.2084
3	TextRank	0.6348	0.2869
4	TSGVi	0.6438	0.3096

Table 4 Rouge scores for each topic

ID	Topic	System	ROUGE-1	ROUGE-2
1	Economics	LEAD	0.54	0.149
		LexRank	0.535	0.167
		TextRank	0.561	0.195
		TSGVi	0.601	0.234
2	Society	LEAD	0.61	0.231
		LexRank	0.596	0.221
		TexRank	0.691	0.321
		TSGVi	0.655	0.303
3	Politics	LEAD	0.629	0.206
		LexRank	0.627	0.276
		TextRank	0.659	0.348
		TSGVi	0.75	0.545
4	Health	LEAD	0.62	0.219
		LexRank	0.631	0.233
		TextRank	0.679	0.224
		TSGVi	0.705	0.272
5	Weather	LEAD	0.685	0.322
		LexRank	0.63	0.254
		TextRank	0.631	0.292
		TSGVi	0.593	0.297
6	Sport	LEAD	0.629	0.285
		LexRank	0.635	0.348
		TextRank	0.698	0.399
		TSGVi	0.786	0.570

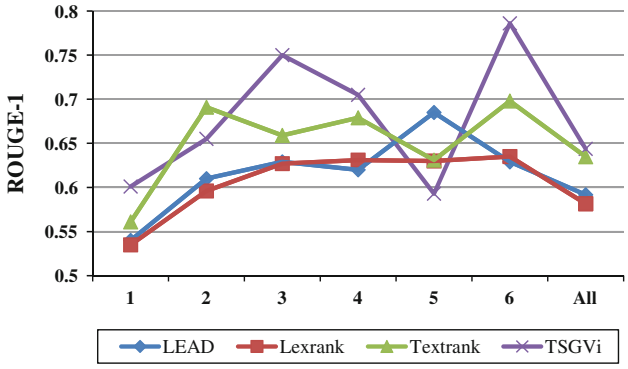

Fig. 3 Comparative ROUGE-1 scores for all topics

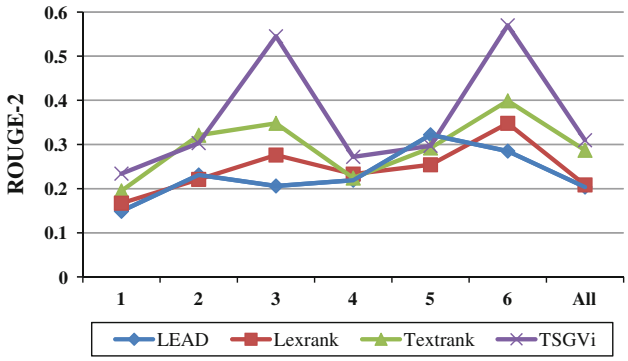

Fig. 4 Comparative ROUGE-2 scores for all topics

Table 5 Detailed information on DUC corpus

Dataset	No. doc	No. set	Avg. doc	Avg. word
DUC 2002	567	59	10	560
DUC 2003	298	30	10	543
DUC 2007	1,125	45	25	466

TextRank (Mihalcea and Tarau 2005a), LexRank (Erkan and Radev 2004), and LEAD baseline system. We have developed two summarization systems according to (Mihalcea and Tarau 2005a) and (Erkan and Radev 2004), respectively. The LEAD baseline takes the first sentences one by one from the first document to the last document in the collection, where documents are assumed to be ordered chronologically. Among the several options of Mihalcea's algorithm (Mihalcea and Tarau 2005a), the method based on the authority score of HITS on the directed backward graph in singe-document summarization phase and PageRank on undirected graph in meta-document summarization phase is the best. It is taken by us for comparison.

Table 6 Comparison of summarization systems on DUC 2002

System	TSGVi	Textrank	DUC 1	DUC 2	DUC 3	DUC 4	DUC 5
ROUGE-1	0.34826	0.3572	0.3047	0.3056	0.3264	0.3447	0.3578

Table 7 Comparison of summarization systems on DUC 2003

System	TSGVi	LexRank	DUC 1	DUC 2	DUC 3	DUC 4	DUC 5
ROUGE-1	0.36134	0.3646	0.3582	0.3607	0.366	0.3676	0.3798

Table 8 Comparison of summarization systems on DUC 2007

System	TSGVi	DUC 1	DUC 2	DUC 3	DUC 4	DUC 5	DUC 6	DUC 7
ROUGE-1	0.42102	0.453	0.44538	0.43489	0.43273	0.43226	0.4247	0.42232

For each set of documents, every system creates summary with 100 words length, same as the length of human-extracted summary. Because Vietnamese is written in Latin characters and so that we can use the ROUGE for comparing with the reference summary extracted by human. The 1-gram ROUGE score (a.k.a. ROUGE-1) has been found to correlate very well with human judgments at a confidence level of 95% based on various statistical metrics. Therefore, in these experiments we use the ROUGE-1 scores to evaluate the summary. Besides ROUGE-1, we use ROUGE-2 to improve the confidence of evaluation, especially when 2-grams words occur with highest frequency in Vietnamese.

Table 3 shows the ROUGE scores for each summarization system over all sets of documents. It can be seen from Table 3 that TSGVi gets better evaluation than TextRank, LexRank, and the LEAD baseline system for this data set. We found that the results generated by LexRank tend to have much redundancy while TSGVi and TextRank ranking system have little redundancy.

Table 4 shows ROUGE scores for each topic. In the Table 4 the proposed system TSGVi almost outperforms other systems over all topics except set of documents related to society and weather topics. We believe that it's due to the fact that the author of a news article usually summarizes the news at the beginning of the text, however, that's not the truth all the times. So that the backward graph is suitable with these topics and the LEAD baseline method sometimes gives higher result. However, our proposed system is better in all genres.

Figures 3 and 4 show comparison of ROUGE scores between TSGVi and others for each topic and overall dataset. As discussed, Vietnamese have compound words, which are used in both writing and speaking with a very high frequency. It means that the number of matched single words is higher than the number of matched compound words. Therefore, because based on 1-gram computation, the ROUGE-1 score in Vietnamese is generally higher than in English. It can be said that in evaluation for Vietnamese summary, the ROUGE-2 score is more appropriate than ROUGE-1 score.

For time consumption calculation, we test our system on Intel P8400 CPU with 3GB main memory. Our system generates summary at real-time 0.107s per document set.

5.2 Evaluation on DUC datasets

As stated above, the main purpose of this system is multi-document summary in Vietnamese, however, to test the stability and portability of the system when making a summary on other languages, we also conduct experiment on the datasets of DUC conferences over the years, namely DUC 2002, DUC 2003, DUC 2007, and Table 5 is the profile of these datasets. For the pre-processing step, mentioned in Sec 4.1, in English documents of DUC, we use the tool of SharpNLP,[5] which developed based on the maximum entropy models, to split sentence and tokenize word.

Thus, on DUC 2002 and DUC 2003, our system provides summaries of 100 words in length, and on DUC 2007 summary is limited to 250 words. We then evaluated based on scores from the tool ROUGE, concrete results for DUC 2002, DUC 2003, DUC 2007 are given in Tables 6, 7 and 8, respectively.

In this table, we list the ROUGE-1 of our system and others, including Textrank at DUC2002 (Mihalcea and Tarau 2005b), Lexrank at DUC 2003 (Erkan and Radev 2004) and some systems having the highest ranks at these conferences.

It can be seen from the table that TSGVi respectively ranked 4, 6, 8 on the data DUC 2002, DUC 2003, DUC 2007. Although the results TSGVi lower than some others,

[5] http://www.sharpnlp.codeplex.com/

but also fully demonstrates that TSGVi completely acceptable if this system is deployed for the English document summarization problem.

We believe, TSGVi can also make a summary on some other languages beside Vietnamese and English. In this case, we just only do a little modification in the implementation of pre-processing modules.

6 Conclusions

In this paper, we present an automatic method to generate an extractive summary of multiple Vietnamese documents based on graph model and ranking algorithm. We firstly create the graph to represent the structure of documents and then perform ranking propagation on this graph. We compare our proposed approach with other systems on Vietnamese news documents. All experiments results indicate that our method can work well in Vietnamese document without the deep knowledge of natural language processing. In addition, there is no need of training data. After that we continue conducting several experiments in the English documents from DUC 2002, 2003 and 2007. The result also shows that our approach is stable flexible for multi-lingual document summarization.

In next step, we are going to address the issue of how to improve the performance on Vietnamese text summarization by natural language processing technologies, and then how to apply related results to the fields of information extraction and recommendation also is the key point of our research work in the future. One problem of graph model is that if the document has more than one idea, the constructed graph can be scattered. Sentences, in the bigger scatters, will usually have the higher ranking score than sentences in the other scatters due to the number of connections. It can cause the loss of information during the summarization process. Therefore, finding the method to revise the ranking score is also one of the main points we target. Also, the database for Vietnamese documents are only 20 test sets and it is considered small compare with the DUC database for English. However, we are trying to build a larger database for the more confident result. Furthermore, we intend to utilize more evaluation methods to evaluate the proposed summarization system.

Acknowledgments The authors would like to thank Prof. Kiem Hoang from the University of Information Technology, VNU, HCM City for his invaluable and insightful comments. The authors also thank the anonymous reviewers for their helpful comments.

References

Barzilay R, Elhadad M (1997) Using lexical chains for text summarization. In: In Proceedings of the ACL workshop on intelligent scalable text summarization, pp 10–17

Berger AL, Mittal VO (2000) Ocelot: a system for summarizing web pages. In: SIGIR, pp 144–151

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw 30(1-7):107–117

Carbonell JG, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR, pp 335–336

Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22:457–479

Goldstein J, Kantrowitz M, Mittal VO, Carbonell JG (1999) Summarizing text documents: sentence selection and evaluation metrics. In: SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, August 15–19, 1999, Berkeley, CA, USA. ACM, New York, pp 121–128

Ha TL, Huynh TQ, Luong MC (2005) A primary studies on summarization of documents in Vietnamese. In: The First World Congress of the International Federation for Systems Research

Lin CY, Hovy EH (2003) Automatic evaluation of summaries using n-gram co-occurrence statistics. In: HLT-NAACL

Lin CY, Lin CY, Hovy E (2002) Automated multi-document summarization in neats. In: Proceedings of the human language technology conference (HLT2002), pp 23–27

Luhn HP (1958) The automatic creation of literature abstracts. IBM J Res Dev 2:159–165

Mani I, Bloedorn E (1997) Multi-document summarization by graph search and matching. In: AAAI/IAAI, pp 622–628

McKeown KR, Barzilay R, Evans D, Hatzivassiloglou V, Klavans JL, Nenkova A, Sable C, Schiffman B, Sigelman S, Summarization M (2002) Tracking and summarizing news on a daily basis with columbia's newsblaster

Mihalcea R, Tarau P (2004) Textrank: bringing order into text. In: EMNLP, pp 404–411

Mihalcea R, Tarau P (2005a) A language independent algorithm for single and multiple document summarization. In: Proceedings of IJCNLP'2005

Mihalcea R, Tarau P (2005b) Multi-document summarization with iterative graph-based algorithms. In: 1st International conference on intelligent analysis methods and tools (IA)

Mittal VO, Kantrowitz M, Goldstein J, Carbonell JG (1999) Selecting text spans for document summaries: heuristics and metrics. In: AAAI/IAAI, pp 467–473

Nguyen ML, Shimazu A, Phan XH, Ho TB, Horiguchi S (2005) Sentence extraction with support vector machine ensemble. In: The First World Congress of the international federation for systems research

Nguyen HTA, Nguyen HK, Tran QV (2010) An efficient Vietnamese text summarization approach based on graph model. In: RIVF

Nomoto T, Matsumoto Y (2001) A new approach to unsupervised text summarization. In: SIGIR, pp 26–34

Phuc D, Hung MX (2008) Using SOM based graph clustering for extracting main ideas from documents. In: RIVF, pp 209–214

Radev DR (2001) Experiments in single and multidocument summarization using mead. In: First document understanding conference

Salton G, Singhal A, Mitra M, Buckley C (1997) Automatic text structuring and summarization. Inf Process Manag 33(2): 193–207

Schiffman B, Mani I, Concepcion KJ (2001) Producing biographical summaries: combining linguistic knowledge with corpus statistics. In: ACL, pp 450–457

Wei F, Li W, Lu Q, He Y (2010) A document-sensitive graph model for multi-document summarization. Knowl Inf Syst 22(2): 245–259

Zha H (2002) Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In: SIGIR, pp 113–120