International Journal of Innovative Computing, Information and Control
Volume **11**, Number **2**, April **2015**

ICIC International ©2015 ISSN 1349-4198
pp. 407–415

A METHOD FOR GENERATING VIETNAMESE TEXT SENTENCE REDUCTION BASED ON BAYESIAN NETWORK

Ha Nguyen Thi Thu[1] and An Nguyen Nhat[2]

[1]Information Technology Faculty
Vietnam Electric Power University
235 Hoang Quoc Viet, Hanoi, Vietnam
hantt@epu.edu.vn

[2]Military Information Technology Institute
17 Hoang Sam, Nghia Do, Cau Giay, Hanoi, Vietnam
nguyennhatan@gmail.com

Received March 2014; revised August 2014

Abstract. *Sentence reduction is one of approaches for text summarization that has attracted many researchers and scholars of natural language processing field. In this paper, we present a method that generates sentence reduction and applying in Vietnamese text summarization based on Bayesian Network model. Use Bayesian network model to find the best likelihood short sentence through comparing difference of probability. Experimental results with 980 Vietnamese sentences, show that our method really is effective in generating sentence reduction that is understandable, readable and exactly grammatical.*

Keywords: Sentence reduction, Natural language processing, Text summarization, Bayesian network, Probability

1. Introduction.

The development of the information brings useful things to the people. When we look up any information on the Internet such as Google, Bing, it will return several hundred million results in a few seconds. However, it is usually very long texts, very hard to select important information (that users need) from these texts. So that, text summarization can bring to user the most significant information from original text, when the amount of information is huger and huger.

Text summarization can be classified into many different types, but in the final, there are two types: extraction and abstraction. Most text summarization systems are based on extraction approach. These systems extract sentences to generate a summary. With this approach, the weight of sentence is calculated based on some features that we think it is important like term frequency, sentence position, sentence length. And then, sentences will be sorted by its weight and extracted based on the rate (extraction rate). A text summary includes sentences that have maximum weight from the original text. Figure 1 depicts the general summary systems based on text extraction approach.

With this approach, text summary will be the synthesis of the discrete sentences from original text, and it can be:

- Text summarization is seamless because sentences are not linked by content in the text, specially, when the extraction rate is smaller, it will be greatly discrete.
- Text summarization is confusing sometimes since it may lose important information in original text by some sentences that have not been extracted.
- With the development of social networks like Facebook, Twitter, Twoo, . . . the summary of information and extracting the gist sentences from them are unreasonable because the information from these social networking sites is discrete.

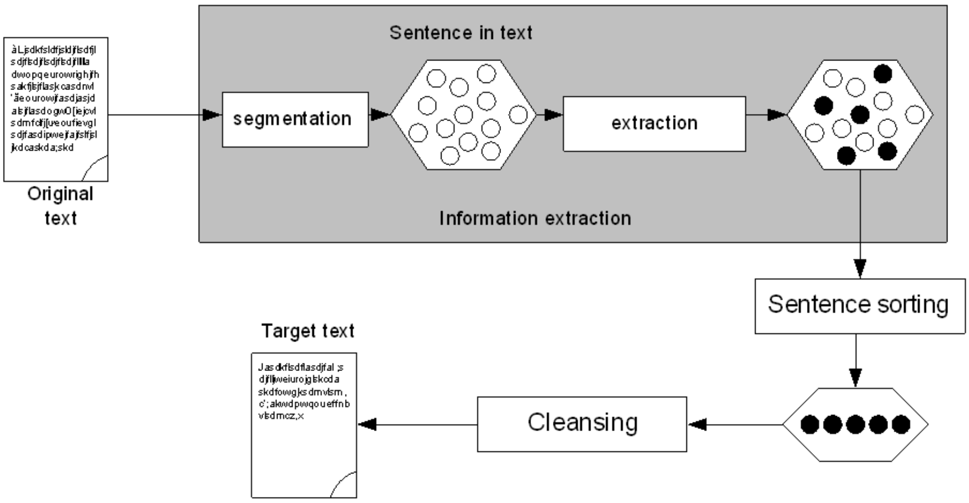


Figure 1. Text summarization based on extraction method

Therefore, we have chosen the abstraction approach for text summarization. In this approach, sentences can be generated by removing some redundant words in it, also known as "sentence reduction" or "sentence compression". The sentence reduction approach for processing in sentence level, removes not important words in the sentence and generates new sentence and creates summary. Text summarization will overcome some disadvantages that have been analyzed above.

Sentence reduction problem is described by word removing problem. Give an original sentence x including n words. $x = x_1, x_2, \ldots, x_n$ and sentence $y = y_1, y_2, \ldots, y_m$ is reduced from sentence x where $m < n$ and set of word in y is subset of words set in x. Reduced sentence remains gist information from original sentence, shorter and good grammar.

This approach is proposed by many researchers as: H. Jing et al. [9-14], D. M. Zajic et al. in 2007 [31], or sentence fusion by Barzilay and McKeown in 2005 [27].

In this paper, we present a sentence reduction method based on Bayesian network, that each word in original text is considered as a node of the network. Reduced sentence is generated by finding a path that is the shortest and greatest score; we called it *"the best likelihood path"*.

The next structure of this paper is as follows. In Section 2 we present an overview of related work. Section 3 presents the methodology of sentence reduction based on Bayesian network method. Section 4 shows the experimental results and finally is the conclusion.

2. **Related Works.** The researchers said that it is very difficult to use abstract approach for text summarization because it needs to use deep language. However, with this approach, they also said that the target text will be more concise, seamless and its quality is higher than extraction approach.

Almost related works focus on building lexical rules model or syntax parser tree. First Aho and Ullman used synchronous context free grammars (SCFGs) to be generalization of context free grammar formalism to simultaneously produce strings in two languages [29]. D. Wu in 1997 has proposed a method that included inversion transduction grammar [30] and some other related research with CFG like head transducers have been proposed by H. Alshawi, S. Douglas and S. Bangalore in 2000 [32].

Knight and Marcu in 2002 proposed a noisy channel formulation of sentence compression [17]. Their model consists of two components: a language model $P(y)$ whose role is to guarantee that the compression output is grammatical and a channel model $P(x|y)$

capturing the probability that the source sentence x is an expansion of the target compression y. Their decoding algorithm searches for the compression y which maximizes $P(y)P(x|y)$. The channel model is a stochastic SCFG, the rules of which are extracted from a parsed parallel corpus and their weights are estimated using maximum likelihood [17].

With tree model, D. Vickrey and D. Koller in 2008 proposed a sentence reduction method that used syntax into small and applied a series of hand-written transformation rules corresponding to basic syntactic patterns [6]. An unsupervised method for sentence reduction which relies on a dependency tree representation and shortens sentences by removing subtrees has been proposed by K. Filippova and M. Strube [16], and related work is proposed by M. Gagnon and L. Da Sylva in 2005 pruning dependency trees by removing prepositional complements of the verb, subordinate clauses and noun appositions [8]. T. Nomoto and Y. Matsumoto also used tree model in his research [33], broadly in line with prior work H. Jing and K. R. McKeown, 2000 [9]; Dorr et al. [3], 2003; S. Riezler et al., 2003 [28]; A. Arbor et al., 2006 [1].

And some other related works: generative models aim to model the probability of a target compression given the source sentence either directly by M. Galley and K. McKeown in 2007 [23]. J. Turner and E. Charniak [15] whereas proposed discriminative formulations that attempt to minimize error rate on a training set. S. Riezler et al. in 2003 used maximum entropy for generating sentence compression [28]. L. M. Nguyen and S. Horiguchi in 2003 applied syntax tree [24].

With Vietnamese sentence reduction approach, most of the methods are applied from English method. However, the performance of this method is not high when applied to Vietnamese language. Because of single syllable language, in Vietnamese, words cannot be determined based on space. So that, they often use extraction approach for building Vietnamese text summarization systems, and there are some methods that use reduction approach but they are not effective.

3. Methodology of Sentence Reduction Based on Bayesian Network.

3.1. Bayesian network.

Bayesian network, also called as belief networks is one of probabilistic graphical models. These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods. Hence, Bayesian network combines principles from graph theory, probability theory, computer science, and statistics.

A Bayesian network B is an annotated acyclic graph that represents a joint probability distribution over a set of random variables V. The network is defined by a pair $B = < G, \Theta >$, where G is directed acyclic graph whose nodes $X_1, X_2, \ldots, X_n$ represent random variables, and whose adages represent the direct dependencies between these variables. The graph G encodes independence assumptions, by which each variable X_i is independent of its non=descendants given its parents in G, denoted generically as Π_i. The second component, Θ, denotes the set of parameters of the network. This set contains the parameter $\theta_{x_i|\Pi_i} = P_B(x_i|\Pi_i)$ for each realization of x_i of X_i conditioned on Π_i, the set of parents of X_i in G. Accordingly, B defines a unique joint probability distribution over V, namely

$$P_B(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i|\pi_i) = \prod_{i=1}^{n} \theta_{X_i|\pi_i} \tag{1}$$

3.2. **Methodology of sentence reduction based on Bayesian network.** In text processing, there are three levels:

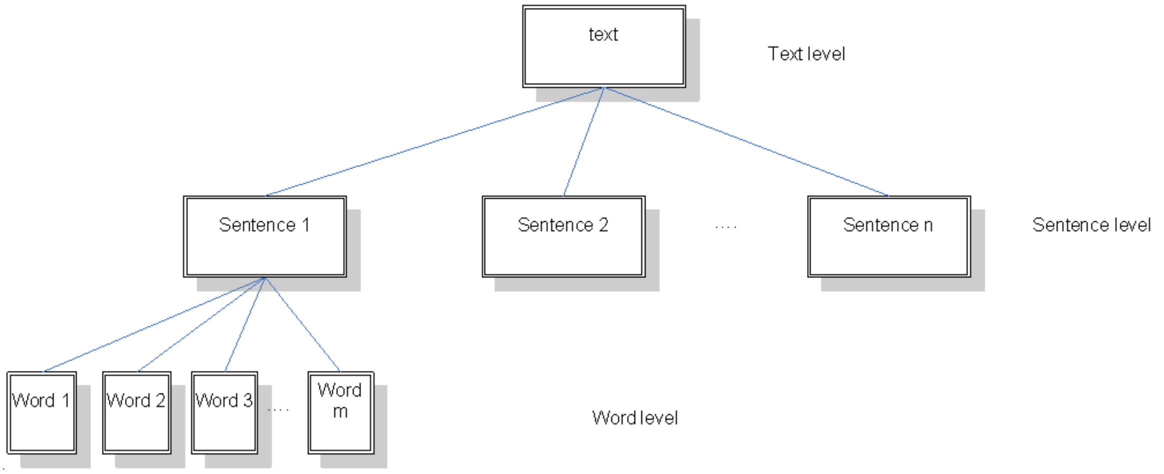


FIGURE 2. Text summarization based on extraction method

In our method we process level of sentence. Bayesian network is a cause-results network. Supposed that, each word w_k was generated by w_{k-1} or can be generated by w_{k-2} w_{k-n}. Then, we can build an improving Bayesian network that can find a reduced sentence based on probability of n-grams between word w_k and word w_{k-n} with $n = \overline{1,(k-1)}$.

Example 3.1. *Suppose sentence S that includes six words can be described as:*

$$S = w_1 w_2 w_3 w_4 w_5 w_6.$$

Sentence S is segmented into six words: w_1, w_2, w_3, w_4, w_5, w_6 and are matched words together to become reduced sentence allowing a Bayesian network model. First we need to find an initial state over all possible states like Figure 3.

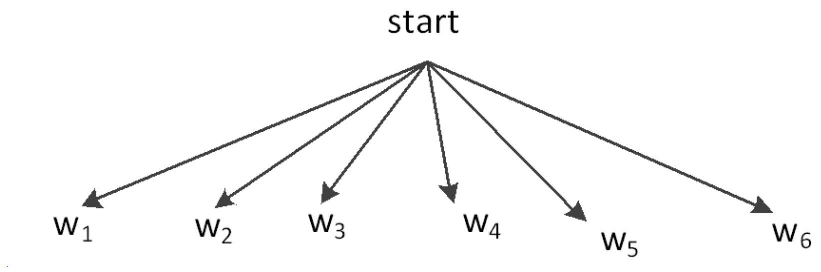


FIGURE 3. Initial state

Supposed that, the set of initial states with probability:

- Start -> $w_1 = 0.6$
- Start -> $w_2 = 0.32$
- Start -> $w_3 = 0.47$
- Start -> $w_4 = 0.56$
- Start -> $w_5 = 0.2$
- Start -> $w_6 = 0.11$

So that, we choose w_1 to be the initial state, so reduced sentence will be started by w_1.

Figure 4 below illustrates the structure of Bayesian network with six words in sentence S. Word w_1 is considered the first likely node to the next words in sentence S. And word w_2 can be created a path to w_3, w_4, w_5, w_6.

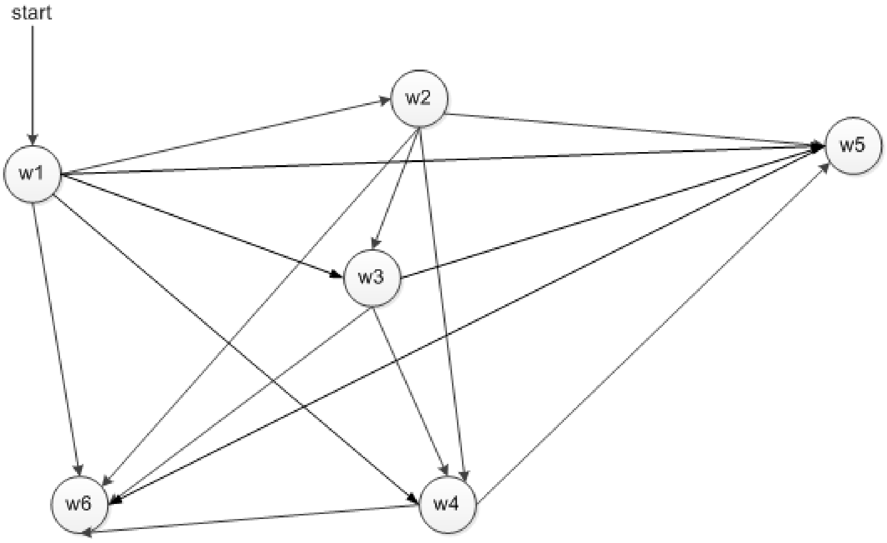


FIGURE 4. Bayesian network model with 6 words

To overcome the computational complexity, in this proposed method we use dynamic programming and do not need to compute n-grams probability over all nodes. For example, have probability as

- w_1 -> $w_2 = 0.3$
- w_1 -> $w_3 = 0.6$
- w_1 -> $w_4 = 0.042$
- w_1 -> $w_5 = 0.002$
- w_1 -> $w_6 = 0$

Choose one state that has maximum probably. So that, w_3 will be chosen and use a path that contains w_3.

For ease of visualization, Bayesian network is described as a tree as Figure 5.

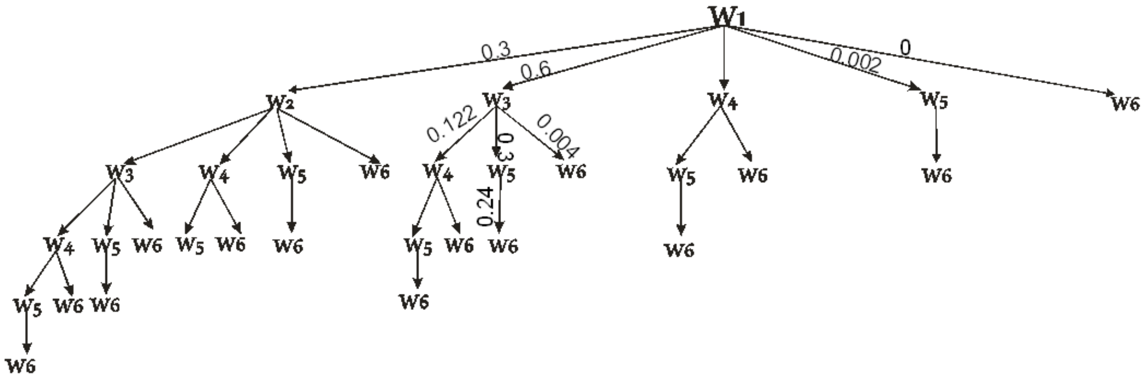


FIGURE 5. Bayesian network with probability

In this Bayesian network. Probability ability on the branch is calculated by n-grams. In this example, path from word w_1 to word w_2 has probability to be 0.3.

The first point of the sentence S is w_1. From word w_1 sentence is reduced by:

- From w_1 we have some likely paths to be w_2, w_3, w_4, w_5, w_6.

- Select the path which is weighted with the highest probability. For example in Figure 5 is w_1 -> $w_3 = 0.6$.
- Save point with the highest path. For example w_3 will be stored.
- From this point of highest path find some paths to the other words, choosing the most likelihood path. For example w_3 -> w_4.
- Continue to repeat the last word of the sentence S.

Finally, we have a sentence that has been reduced. In Figure 6 there is the most likelihood path (bold path) and reduced sentence includes four words w_1, w_3, w_5, w_6.

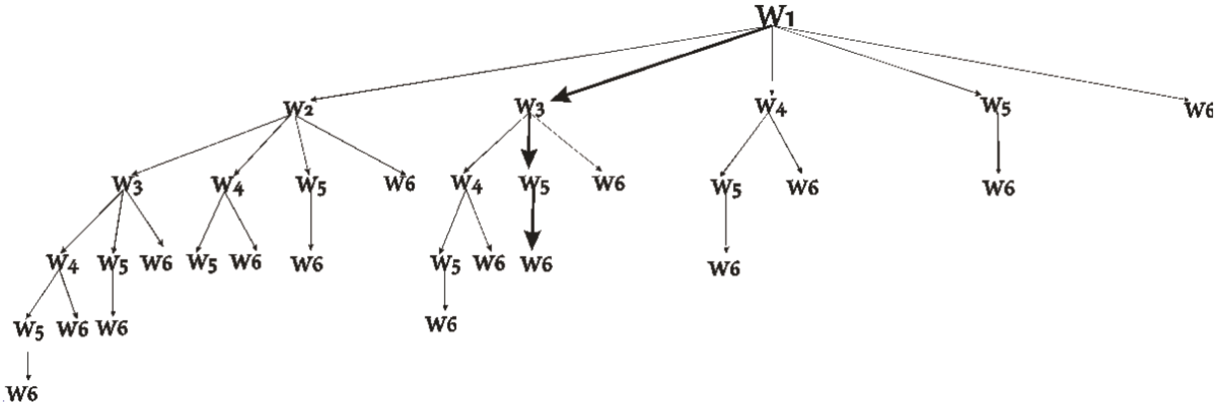


Figure 6. Sentence reduction

```
Sentence Reduction Based on Bayesian Network Algorithm
Input:
        $S$: original sentence;
Output:
        $S'$: reduced sentence;
1. Initialization
            $T = \phi$; $T' = \phi$; $N = \phi$; $i = 1$; $j = 0$;
2. Separate words from S
        For $i = 1$ to Length($S$)
          T(i)←Separate($S$);
3. Selecting word from original sentence
        While $i <$ Length($S$) do
        begin
            for $j = i + 1$ to Length($S$) do
                begin
                    $N(j) \leftarrow$ N-grams($w_j, w_i$);
                    point = argmax($N(j)$);
                    $T' = T' \cup w_j$
                end;
                $i = j$;
                $N = \phi$;
        end;
4. Generating sentence reduction
        $S'$ = Order($T'$)
```

Figure 7. Sentence reduction based on Bayesian network algorithm

In Figure 7, we simulated an algorithm which is called SRBBN (Sentence Reduction Based on Bayesian Network).

In this algorithm, we use some functions. ***Separate***() is used to separate words in sentence. ***Length***() returns length of sentence. N-grams is used for calculating n-grams of word w_i and word w_j that learn from training data. ***argmax***() takes the largest value in the set N. ***Order***() is used for sorting words in the original sentences to generate reduced sentence.

4. Experiments.

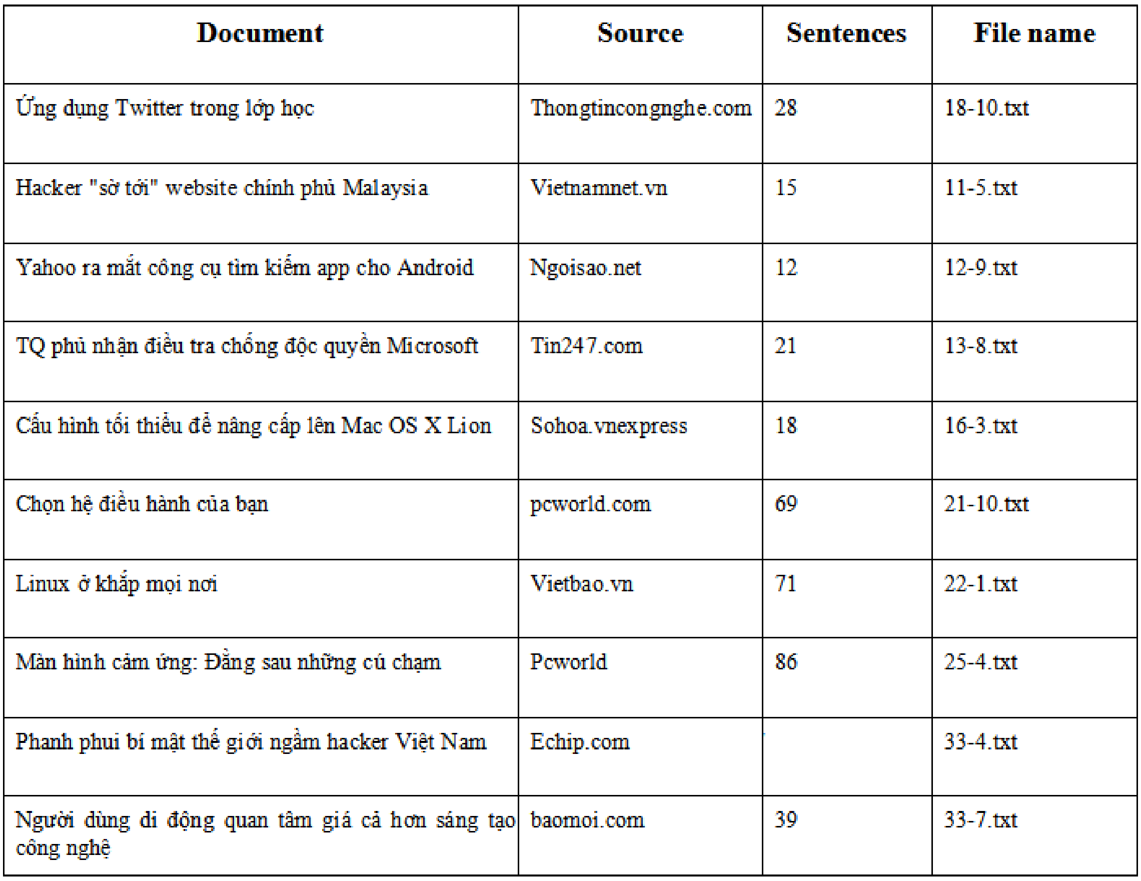
4.1. **Corpus.** There is no standard corpus for Vietnamese text summarization now. So that, in our experiment, we built corpus by manual. Documents in this corpus have been downloaded from website's news as: http://thongtincongnghe.com, http://echip.com, http://vnexpress.net, http://vietnamnet.vn, http://tin247.com. Corpus's title is "information" and "technology". There are over 300 documents in it. We segmented words from 300 documents into 16,117 sentences.

All file downloaded from website will be saved in corpus by *.txt and preprocessed. Table 1 illustrated a file in corpus that is preprocessed.

4.2. **Word segmentation.** We use a Vietnamese text segmentation tool that is used for word segmentation. We selected 814 sentences in the corpus for signed label.

4.3. **Results.** It is difficult to compare our method with previous ones, because there were no widely accepted benchmarks for Vietnamese text reduction sentence. Therefore, we compare our proposed method with manual sentence reduction generated by humans,

TABLE 1. Some documents in corpus



Document	Source	Sentences	File name
Ứng dụng Twitter trong lớp học	Thongtincongnghe.com	28	18-10.txt
Hacker "sờ tới" website chính phủ Malaysia	Vietnamnet.vn	15	11-5.txt
Yahoo ra mắt công cụ tìm kiếm app cho Android	Ngoisao.net	12	12-9.txt
TQ phủ nhận điều tra chống độc quyền Microsoft	Tin247.com	21	13-8.txt
Cấu hình tối thiểu để nâng cấp lên Mac OS X Lion	Sohoa.vnexpress	18	16-3.txt
Chọn hệ điều hành của bạn	pcworld.com	69	21-10.txt
Linux ở khắp mọi nơi	Vietbao.vn	71	22-1.txt
Màn hình cảm ứng: Đằng sau những cú chạm	Pcworld	86	25-4.txt
Phanh phui bí mật thế giới ngầm hacker Việt Nam	Echip.com		33-4.txt
Người dùng di động quan tâm giá cả hơn sáng tạo công nghệ	baomoi.com	39	33-7.txt

TABLE 2. Experimental result for sentence reduction

Method	**Compression**	**Grammaticality**	**Word significance weight**
Baseline	X	X	X
SRBBN	65.82	84.2	78.4
Human	61.2209	83.33333	63.5
Syn.con	67	65.7	6.11

called Human, a sentence reduction method using syntax control, called Syn.con and one of our sentence reduction method called SRBBN.

In this experiment, we use the evaluation way as K. Knight and D. Marcu [17]. Table 2 shows the sentence reduction results that are carried out by SRBBN method, Human and Syn.con for Vietnamese text.

5. **Conclusion.** In this paper, we have proposed a novel method, called the SRBBN for reducing Vietnamese sentence. This method is based on Bayesian Network. Our experimental results on a corpus of 16,117 sentences of Vietnamese text show that the proposed sentence reduction method achieved acceptable results when compared to human manual. Sentence after reduced satisfies user requirement, is readable, understandable and has good grammar.

Acknowledgements. We would like to thank the experts of University of Engineering and Technology, Vietnam of University, and Japan Advanced Institute of Science and Technology, Dr Nguyen Le Minh, Dr Nguyen Van Vinh, Dr Nguyen Phuong Thai, Dr Le Anh Cuong, Dr Pham Bao Son for their great help in building the experimental summarizing application.

REFERENCES

[1] A. Arbor, J. J. Clarke and M. Lapata, Constraint-based sentence compression: An integer programming 306 approach, *Proc. of the COLING/ACL*, pp.144-151, 2006.

[2] R. Blanco and C. Lioma, Graph-based term weighting for information retrieval, *Information Retrieval*, pp.54-92, 2012.

[3] B. Dorr, D. Zajic and R. Schwartz, Hedge trimmer: A parse-and-trim approach to head-line generataion, *Proc. of the HLT-NAACL.Text Summarization Workshop and Document Under-standing Conderence*, Edmon-ton, Canada, pp.1-8, 2003.

[4] C.-Y. Lin and E. Hovy, The potential and limitations of automatic sentence extraction for summarization, *Proc. of the HLT-NAACL 2003 Workshop on Automatic Summarization*, Edmonton, Canada, 2003.

[5] C. Napoles, C. Callison-Burch, J. Ganitkevitch and B. Van Durme, Paraphrastic sentence compression with a character-based metric: Tightening without deletion, *Workshop on Monolingual Text-To-Text Generation, Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, pp.84-90, 2011.

[6] D. Vickrey and D. Koller, Sentence simplification for semantic role labeling, *Proc. of ACL-08: HLT*, Columbus, Ohio, USA, pp.344-352, 2008.

[7] Y. Feng and M. Lapata, Automatic image annotation using auxiliary text information, *Proc. of ACL-08: HLT*, pp.272-280, 2008.

[8] M. Gagnon and L. Da Sylva, Text summarization by sentence extraction and syn-tactic pruning, *Proc. of Computational Linguistics in the North East*, Gatineau, Quebec, Canada, 2005.

[9] H. Jing and K. R. McKeown, Cut and paste based text summarization, *Proc. of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pp.178-185, 2000.

[10] H. Jing and K. McKeown, The decomposition of human written summary sentences, *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.129-136, 1999.

[11] H. Jing, Sentence reduction for automatic text summarization, *Proc. of the Conference on Applied Natural Language Processing*, pp.310-315, 2000.

[12] H. Jing, Using hidden Markov modeling to decompose human written summaries, *Computational Linguistics*, vol.28, no.4, pp.527-543, 2002.

[13] H. Jing and K. R. McKeown, Cut and paste based text summarization, *Proc. of the North American Chapter of the Association for Computational Linguistics Conference*, pp.178-185, 2000.

[14] H. Jing, *Cut-and-Paste Text Summarization*, Ph.D. Thesis, Columbia University, 2001.

[15] J. Turner and E. Charniak, Supervised and unsupervised learning for sentence compression, *Proc. of the 43rd Annual Meeting of the ACL*, pp.290-297, 2005.

[16] K. Filippova and M. Strube, *Dependency Tree Based Sentence Compression*, pp.25-32, 2008.

[17] K. Knight and D. Marcu, Summarization beyond sentence extraction: A probabilistic approach to sentence compression, *Artif. Intell.*, vol.139, no.1, pp.91-107, 2002.

[18] E. Lloret, A. Balahur, M. Palomar and A. Montoyo, Towards building a competitive opinion summarization system: Challenges and keys, *Proc. of the NAACL. Student Research Workshop and Doctoral Consortium*, pp.72-77, 2009.

[19] E. Lloret, H. Saggion and M. Palomar, Experiments on summary-based opinion classification, *Proc. of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp.107-115, 2010.

[20] E. Lloret and M. Palomar, *Text Summarisation in Progress: A Literature Review*, Springer Science & Business Media, pp.1-41, 2012.

[21] I. Mani, *Automatic Summarization*, John Benjamins Publishing Co. Amsterdam, Philadelphia, USA, 2001.

[22] J. Mei and L. Chen, SumCR: A new subtopic based extractive approach for text summarization, *Knowledge and Information Systems*, pp.527-545, 2012.

[23] M. Galley and K. McKeown, Lexicalized Markov grammars for sentence compression, *Proc. of the HLT-NAACL*, pp.180-187, 2007.

[24] M. L. Nguyen and S. Horiguchi, A sentence reduction using syntax control, *Proc. of the 6th Information Retrieval with Asian Language*, pp.139-146, 2003.

[25] M. L. Nguyen, A. Shimazu, S. Horiguchi, B. T. Ho and M. Fukushi, Probabilistic sentence reduction using support vector machines, *Proc. of the 20th International Conference on Computational Linguistics*, 2004.

[26] M. Johnson and E. Charniak, A TAG-based noisy-channel model of speech repairs, *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pp.33-39, 2004.

[27] R. Barzilay and K. R. McKeown, Sentence fusion for multidocument news summarization, *Computational Linguistics*, vol.31, no.3, pp.297-328, 2005.

[28] S. Riezler, T. H. King, R. Crouch and A. Zaenen, Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical functional grammar, *Proc. of HLT-NAACL*, pp.118-125, 2003.

[29] V. Aho and J. D. Ullman, Properties of syntax directed translations, *Journal of Computer and System Sciences*, vol.3, pp.319-334, 1969.

[30] D. Wu, Stochastic inversion transduction grammars and bilingual parsing of parallel corpora, *Computational Linguistics*, vol.23, no.3, pp.377-404, 1997.

[31] D. M. Zajic, B. J. Dorr, J. Lin and R. Schwartz, Multi-candidate reduction: Sentence compression as a tool for document summarization tasks, *Information Processing and Management Special Issue on Summarization*, 2007.

[32] H. Alshawi, S. Douglas and S. Bangalore, Learning dependency translation models as collections of finite-state head transducers, *Computational Linguistic*, vol.26, no.1, pp.45-60, 2000.

[33] T. Nomoto and Y. Matsumoto, Discourse parsing: A decision tree approach, *Proc. of the 6th Workshop on Very Large Corpora*, pp.216-224, 1998.