Extracting Phrases in Vietnamese Document for Summary Generation

Huong Thanh Le, Rathany Chan Sam
School of Information and Communication Technology
Hanoi University of Technology
Hanoi, Vietnam
huonglt-fit@mail.hut.edu.vn, rathany_cam@yahoo.com

Phuc Trong Nguyen
Thang Long Informatics Company
Hanoi, Vietnam
ntp3985@gmail.com

Abstract—**This paper describes an approach to Vietnamese text summarization, concentrated on the discourse structure of the text. Based on characteristics of Vietnamese, we propose rules for segmenting text into elementary discourse units (edus) and for recognizing discourse relations between textual spans. The score of an edu is computed based on the discourse tree. The edus with highest scores are chosen to put in the summary. Experiments show that this method can give promising results.**

Keywords - text summarization, discourse structure, rhetorical relation, Vietnamese

I. INTRODUCTION

Automatic text summarization is the technique which automatically creates an abstract or summary of a text. According to Mani and Maybury [8], three basic operations used in summarization are: (i) selecting more-salient or non-redundant information; (ii) aggregating information; and (iii) generalizing specific information with more general, abstract information.

There are several approaches to text summarization. These approaches are different in the method of evaluating and selecting salient textual spans, which are usually clauses or sentences. The simplest approach takes a shallow processing on the input text. It employs corpus-based, statistical techniques, surface linguistic analysis, and the use of large, public domain linguistic resources such as on-line text corpora and machine-readable lexicons. There are a lot of work concentrating on this approach (e.g., [4, 11]). Although this type of systems is quite efficient and robust, it lacks of a deep semantic processing.

The second approach is the entity-level one, which builds an internal representation for text, modelling text entities and their relationships. A typical representation of this approach is the text summarization system developed by Salton et al. [12]. This system automatically generates semantic hypertext links between paragraphs in the text. The more links a paragraph has, the more important it is. A number of the most important paragraphs are then included in the summary.

The most sophisticated approach is based on discourse analysis. It models the global structure of the text and its relation to communicative goals. The system developed by Marcu [9] follows this approach. Here, the discourse parsing algorithm is used to generate the semantic trees for the document. Then, scores are assigned to phrases on the tree. Basing on these scores, the phrases with highest scores are chosen as the summary.

Most works on Vietnamese text summarization (e.g., [1, 10]) are based on statistics. The disadvantage of this approach is that it always considers sentence as the smallest unit to be selected. As a result, the summary still has redundancy. This paper introduces our approach to Vietnamese text summarization using discourse structures. Since this approach considers clause as the elementary unit, it prevents redundant information in the summary. Based on previous works on discourse analysis of [6, 9] and the characteristics of Vietnamese language, we propose our method to construct discourse structure of Vietnamese text. The summary is then generated from the discourse structure of the text.

The rest of this paper is organized as follows. The rhetorical structure theory is introduced in Section II. Section III described our method to construct the discourse structure of Vietnamese text. Section IV represents our method of summarizing text using discourse structures. Experimental results are given in Section V. Finally, Section VI concludes the paper.

II. RHETORICAL STRUCTURE THEORY

Rhetorical Structure Theory (RST) is a method of representing the coherence of text. It models the rhetorical structure of a text by a hierarchical tree that labels discourse relations between spans. This hierarchical tree diagram is called a "rhetorical tree", "discourse tree", or "RST tree". The leaves of an RST tree correspond to edus, which are clauses or clause-like units with independent functional integrity, whereas the internal tree nodes correspond to larger spans.

Figure 1 represents the discourse tree of Example 1. Instead of displaying the full text of each tree node, we cite the first and last edus that contribute to it (e.g., "1.1-1.2", "1.1-1.3"). An internal tree node contains one or several names (e.g., ELABORATION, EXPLANATION) of the discourse relations that hold between adjacent, non-overlapping spans. The span that participates in a discourse relation is either a *nucleus* (N) or a *satellite* (S). The nucleus plays a more important role than the satellite in respect to the writer's intention. If both spans have equal roles, they are both considered as *nuclei* in the relation.

(1) [Bạn nên đến gặp Thành hôm nay$_{1.1}$] [sau khi xong việc$_{1.2}$]. [Ngày mai anh ấy sẽ đi Sài Gòn.$_{1.3}$]
[You should meet Thanh today$_{1.1}$] [after you finish this work$_{1.2}$]. [He will go to Saigon tomorrow.$_{1.3}$]

The score of a span is evaluated based on the discourse tree of the input text. A summary of the text is generated by taking the spans with highest scores. For example, the summary of the text in Example 1 with the compress ratio of 30% is "*Bạn nên đến gặp Thành hôm nay*". Therefore, a crucial problem in our approach to text summarization is to implement a discourse parser for Vietnamese text. This problem will be discussed in detailed in Section III.

978-0-7695-4288-1/10 $26.00 © 2010 IEEE
DOI 10.1109/IALP.2010.8

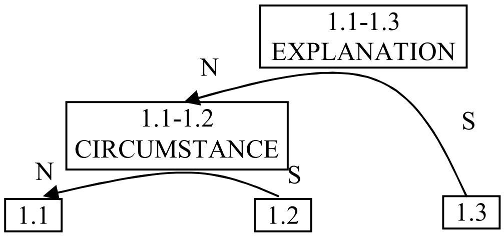

Figure 1. The Discourse Tree of Example 1

III. Analysis the Discourse Structure of Text

Research on generating discourse structures (e.g., [3, 6, 7, 9]) use cue phrases, syntactic structures, cohesive devices, noun-phrase cues and verb-phrase cues as a signal to segment text and to recognize discourse relations. As far as we know, no discourse parser for Vietnamese has been published. Therefore, this research develops a discourse parser for Vietnamese text, inherited the work of [6]. Since Vietnamese and English have different characteristics, we will analyze the differences between the two languages that affect the discourse analysis and propose our solution to this problem.

To construct the discourse structure of a text, the following tasks should be performed: (i) segmenting text into edus; (ii) recognizing discourse relations between spans; and (iii) selecting and combining discourse relations created in step (ii) to form a discourse structure that covers the entire text. The first task of our discourse parser – discourse segmentation – is presented next.

A. Difficulty in Discourse Segmentation for Vietnamese

The purpose of discourse segmentation is to split a text into edus. The discourse segmentation process includes of two steps :

- Divide text into paragraphs and sentences. This step can be done quite simply based on punctuation marks.
- Divide sentences into edus.

Most of research on RST for English bases on cue phrases such as *because, but, although,* etc. to segment text [3, 7]. For example, the sentence "We cannot be sure the product is safe *although* we have tested it." can be splitted into two edus "We cannot be sure the product is safe" and "*although* we have tested it.", based on the cue phrase *although*. However, the segmentation process for Vietnamese cannot rely on cue phrases as English. Because of the characteristic of Vietnamese, it requires more complicated treatment, as described next.

1) Discourse segmentation for Vietnamese

The simplest method to detect edus in English text is to base on cue phrases. However, the use of cue phrases in Vietnamese is not simple as in English. Since Vietnamese is a monosyllabic language, a cue phrase may be recognized incorrectly as a part of another word. For example, let us consider the example shown below:

(2) a. Tôi đến muộn/$_{\text{I am late}}$ vì/$_{\text{since}}$ xe tôi bị hỏng/$_{\text{my car was broken}}$.

b. Tôi hỏi/$_{\text{I asked}}$ Ba/$_{\text{Ba}}$ vì sao/$_{\text{why}}$ cậu không đến/$_{\text{you didn't come}}$.

b'. Tôi hỏi/$_{\text{I asked}}$ Ba vì/$_{\text{Ba vi}}$ sao/$_{\text{why}}$ cậu không đến/$_{\text{you didn't come}}$.

c. Ba/$_{\text{three}}$ vì sao/$_{\text{stars}}$ sáng/$_{\text{bright}}$ trên bầu trời Hà Nội/$_{\text{in the sky of Hanoi}}$.

In Example 2a, the word "vì/$_{\text{since}}$" is a cue phrase. In Example 2b, "vì" is not a word, but "vì sao/$_{\text{why}}$" is. The word "vì sao" is a cue phrase in Example 2b. There are two ways of tokenizing the phrase "Ba vì sao", as shown in Examples 2b and 2b'. The first case - the correct one - is "Ba/$_{\text{Ba}}$" (the name of a person) and "vì sao/$_{\text{why}}$" (Example 2b), in which "vì sao/$_{\text{why}}$" is a cue phrase. The second case is "Ba vì/$_{\text{Ba vi}}$" (the name of a place) and "sao/$_{\text{why}}$" (Example 2b'), in which "sao/$_{\text{why}}$" is a cue phrase.

Because of the ambiguity phenomenon, a word may be recognized incorrectly as a cue phrase. In Example 2c, the phrase "Ba vì sao" is tokenized in the same way as in Example 2b. However, it is understood differently. "Vì sao/$_{\text{why}}$" in Example 2b is a cue phrase; whereas "vì sao/$_{\text{stars}}$" in Example 2c is not a cue phrase. "Vì sao" is a conjunction in Example 2b, whereas it is a noun in Example 2c. This ambiguity situation can be solved by using word category.

Even though a cue phrase has been detected, syntactic information is still needed to split sentence into edus. Let us consider the following example:

(3) a. Vì/$_{\text{because}}$ trời/$_{\text{it}}$ mưa/$_{\text{rains}}$ nên/$_{\text{therefore}}$ đường/$_{\text{the road}}$ trơn/$_{\text{slippery}}$.

b. Bác Hồ/$_{\text{Uncle Ho}}$ làm/$_{\text{do}}$ mọi việc/$_{\text{every thing}}$ đều vì/$_{\text{for}}$ nước/$_{\text{the country}}$ vì/$_{\text{for}}$ dân/$_{\text{people}}$.

Both words "vì" in Examples 3a and 3b are conjunction words. However, the word "vì" in Example 3a is a cue phrase in the structure [vì <clause> nên <clause>]. It can be used to split the sentence into edus. The word "vì" in Example 3b contributes to the object phrase of the verb "làm". It cannot be used to split a sentence into edus. Instead, syntactic information is needed to do this task.

2) Discourse Segmentation for Vietnamese

From the characteristics of Vietnamese mentioned above, it is clear that discourse segmentation using only cue phrases is not accurate. Therefore, we propose to use other signals, including: (i) punctuation marks, quotation marks ; and (ii) syntactic structure.

Examples of the rules that combine the above signals to segment text into edus are:

❶ Combine punctuation marks and syntactic structure:

Pattern 1: <clause>{,|<clause>}+.

In the above rules:

{}+ means the string inside {} can appear more than once.

| means the sentence should be splitted at this point.

<clause> means there is a clause at this place.

An example of Rule 1 is:

(4) [Trời mưa,][sân trơn,][bóng ướt.]

[It rained,] [ground was slippery,] [ball was wet.]

❷ Syntactic structure:

Pattern 2: <subject <noun phrase> | <subordinate clause> | > <predicate>.

In the above rule, the <subordinate clause> is recognized as an embedded unit and is considered as an edu. It is splitted from the sentence. Since the <noun phrase> and the <predicate> do not have a complete meaning, the combination of them is an edu. A SAME-UNIT relation is used to combine these two spans. Example of this pattern is:

(5) [Ngôi nhà [tôi mới xây] rất đẹp.]

[The house [that I have just built] is very beautiful.]

❸ Cue phrases:

Pattern 3: Vì/$_{\text{Since}}$ <span> | nên/$_{\text{therefore}}$ <span>.

(6) [**Vì** trời mưa][**nên** đường trơn.]

***[Since** it rained] [**therefore** the road is slippery.]*

❹ Combine cue phrases, sentential marks and syntactic structure:

Pattern 4: Khi/$_{\text{when}}$ <verb phrase>,| <clause>.

(7) [**Khi** được dự báo trước đợt rét lạnh,][bà con nông dân sẽ chuẩn bị đối phó hiệu quả, không bị mất mùa.]

*[**When** being informed about a cold spell,][farmers will prepare to face it effectively, without losing crop.]*

To segment a text into edus, the system first syntactically parses the input text. Then it recognizes cue phrases and sentential marks in the text. Finally, it segments text into edus using segmentation rules.

B. *Identify discourse relations*

In this research, the set of discourse relations is taken from [6]. As reported in [6], cue phrases can be used to signal discourse relations in English text. Cue phrases can also signal discourse relations in Vietnamese. The process of identifying discourse relation is carried out at three levels: between clauses, between sentences, and between paragraphs.

1) Identify discourse relations between clauses

At this level, cue phrases, syntactic information and segmentation patterns are used to identify discourse relations between edus. A rule for recognizing discourse relations includes of the following fields:

Marker: cue phrases or sentential marks (in some cases, it can be any phrase).

Pos1: position of the cue phrase in the first clause. Possible values are B(begin), M(middle) or E(end).

Pos2: position of the cue phrase in the second clause. Possible values are B(begin), M(middle) or E(end).

Pattern: the pattern of the sentence, which corresponds to one of the segmentation patterns.

Type: the type of discourse relation, which can be S-N, N-S or N-N.

Rel: the name of the discourse relation

Score: score of the rule, corresponding to the certainty of the discourse relation. Score is a real value between 0 and 1.

For example, Example 6 "**Vì** trời mưa **nên** đường trơn." satisfies the segmentation pattern 3. Therefore it is splitted into two edus, "**Vì** trời mưa" and "**nên** đường trơn.". The cue phrases "vì/$_{\text{since}}$" and "nên/$_{\text{therefore}}$" stand at the begin of the edus (both Pos1 and Pos2 are B). The discourse relation between the two edus is CAUSE_EFFECT. The first edu is a satellite. The second one is a nuclei. The certainty of this relation is 100%.

In case a sentence satisfies the segmentation pattern 2 (e.g., Example 5), the <subordinate clause> and the related <noun phrase> has an ELABORATE relation, in which the <subordinate clause> is the satellite and the <noun phrase> is the nuclei. The <noun phrase> is not a clause without the <predicate>. The combination of them is an edu. There is a SAME-UNIT relation between these edus. SAME-UNIT is not a discourse relation, but a relation to indicate that two segments must be connected to have a meaning.

In case there is no cue phrase in the sentence; syntactic information and segmentation patterns do not help much, other signals have to be used. We propose to use word pairs that are semantically related. For example, since the words "dài/$_{\text{long}}$" and "ngắn/$_{\text{short}}$" in Example 8 are antonym, a CONTRAST relation holds between the two clauses of this example.

(8) [Tháng năm ngày **dài**,][tháng mười ngày **ngắn**.]

*[The day of May is **long**,][the day of October is **short**.]*

2) Recognizing discourse relations at the sentence level and paragraph level

At the sentence and paragraph levels, cue phrases can also be used to signal discourse relations. For example, the cue phrase "Tuy nhiên/$_{\text{however}}$" in Example 9 signals a CONTRAST relation between the two sentences.

(9) Tìm kiếm thông tin trên mạng không khó. **Tuy nhiên** tìm được đúng thông tin mong muốn không phải lúc nào cũng đơn giản.

*[Finding information in Internet is not difficult.][**However**, finding exactly the desired information is not always simple.]*

However, not all sentences and paragraphs have cue phrases. Syntactic information does not help in recognize discourse relation. Therefore, beside cue phrases, other signals have to be used to identify relations. We propose to identify the discourse relation between two texts based on their contexts, using methods described below.

❶ Using word pairs that are semantically related. For example, the two words "đẩy/$_{\text{push}}$" – "ngã/$_{\text{fall}}$" in Example 10 have a CAUSE-EFFECT relation. Therefore, a CAUSE-EFFECT relation holds between the two sentences. This method is also used at the clause level.

(10) Hùng bị **ngã**. Dũng vừa **đẩy** nó.

*[Hung **felt**.][Dung **pushed** him.]*

This method requires a thesaurus that consists of information about semantic constraints among words (like WordNet). Since there is no available WordNet for Vietnamese, a Vietnamese semantic tree [13] is used for this purpose.

❷ Using cosine equation to evaluate the similarity between two spans: In this method, the first span is considered as the fundamental vector, the second span is vectorized based on the first vector. The similarity between the two vectors is computed by the following formula:

$$\cos(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum (x_i)^2} * \sqrt{\sum (y_i)^2}}$$

in which x_i and y_i are vectors corresponding to the two spans.

When the similarity between the two spans is larger than a threshold, the relation between the two spans will be considered as ELABORATION. The relation will be JOINT otherwise. The accuracy of this method is proportional to the length of the span.

C. *Constructing Discourse Trees*

The process of constructing discourse tree is carried out at three levels: sentence-level, paragraph-level and text-level.

Constructing the discourse tree at sentence-level is already performed at the segmentation process. At a higher level, the discourse tree is constructed from the set of discourse relations in the text, using a bottom-up strategy similar to the CYK algorithm [5] in syntactic parsing.

IV. Text Summarization Using Discourse Structure

The summary of a text is generated from a set of salient units of the text. The score of a span is calculated by the formula proposed by Marcu [9]. The text summarizing algorithm is shown below:

```
Input :  A text T and a number p ( $1 \le p \le 100$)
Output: The most important p% of the edus of T
Algorithm  :
    1. Construct the discourse tree of T.
    2. Compute the score of edus and sort them by decreasing order.
    3. Cluster edus by scores
    4. Select the first n edus from the sorted list to generate a summary. n is computed such as the length of the summary is nearest to p% of the input text. Edus that belong to the same cluster have to be selected or deselected altogether.
```

Figure 2. The text summarizing algorithm

V. Experimental Results

To carry out experiments with the summarizing system, we collected 140 documents related to computers from the website PCWorld (http://www.pcworld.com.vn/). Each document has approximately 25 sentences. The average words of a sentence is 30. These documents are manually summarized with the compress ratios of 20% and 30%.

To evaluate the system performance, the summaries created by human sentences were compared with the one generated by the system. Precision, recall and F-score measures are calculated as

$$P = \frac{\text{number of correct phrases derived by system}}{\text{total number of phrases derived by system}} \tag{12}$$

$$R = \frac{\text{number of correct phrases derived by system}}{\text{total number of phrases extracted by human}} \tag{13}$$

$$F - score = \frac{2 * P * R}{P + R} \tag{14}$$

The F-scores of the system are 47.4% and 47.7% when the compress ratios are 20% and 30%, respectively. As far as we know, most of text summarization system extracts sentences from text [e.g., 2,10]. The English text summarization system [2] receives the highest precision of 45% at the compression of 30%, whereas the Vietnamese one [10] reported the F-score of 53%. Since the most important clauses are selected by our system, we cannot compare these systems directly. However, these numbers show that our approach is promising in solving the text summarization task.

VI. Conclusions

This paper presents our approach to text summarization, concentrating on analyzing discourse structures for Vietnamese text. By investigating the characteristics of Vietnamese, we have proposed methods to segment text into edus and to recognize discourse relations between spans. The discourse tree is constructed by an algorithm like the CYK algorithm in syntactic parsing. The score of spans in the tree is calculated by the formula proposed by Marcu [9]. The summary of the input text is created from the set of spans that have highest scores. Our experiments with 140 documents from the website PCWorld achieve the F-scores of 47.4% and 47.7% when the compress ratios are 20% and 30%, respectively. These numbers show that our approach is promising in solving the text summarization task.

To increase the system performance, our future works include: (i) creating a more complete set of discourse segmenting rules; (ii) investigating other methods to recognize discourse relations between spans; and (iii) integrating other text summarizing techniques (e.g., position based method) to this approach, in order to increase the performance of the system.

Acknowledgment

This paper is completed under the sponsor of Vietnam's National Foundation for Science and Technology Development (NAFOSTED).

References

[1] P. Do and K. Hoang, "Extracting Main Ideas in Vietnamese Documents. Supporting Content Summarization", In Journal of Posts, Telecommunications and Information Technology, 2007.

[2] M.A. Fattah and F. Ren, "Automatic Text Summarization", World Academy of Science, Engineering and Technology 37, 2008.

[3] K. Forbes, E. Miltsakaki, R. Prasad, A. Sarkar, A. Joshi and B. Webber, "D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar", Journal of 12(3), 2007, pp..261-279.

[4] E. Hovy and C. Lin, "Automatic Text Summarization in SUMMARIST", In I. Mani and M. T. Maybury, editors, Advanced in automatic text summarization. The MIT Press, 1999, pages 81-94.

[5] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing", Computational Linguistics and Speech Recognition, Prentice Hall, 2008.

[6] H.T. Le, G. Abeysinghe, and C. Huyck, "Generating Discourse Structures for Written Texts", In Proceedings of the International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland.

[7] Linguistic Data Consortium, "The Rhetorical Structure Theory" Discourse Treebank Publication, catalog number LDC2002T07 and ISBN 21-58563-223-6.

[8] I. Mani and M. T. Maybury, "Advanced in automatic text summarization", The MIT Press, 1999.

[9] D. Marcu, "The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts", PhD Thesis, Department of Computer Science, University of Toronto, 1997.

[10] M.L. Nguyen; A. Shimazu; X.H. Phan; T.B. Ho; S. Horiguchi, "Sentence Extraction with Support Vector Machine Ensemble", In Proceedings of the First World Congress of the International Federation for Systems Research : The New Roles of Systems Sciences For a Knowledge-based Society, 2005.

[11] T. Nomoto and Y. Matsumoto, "A New Approach to Unsupervised Text Summarization", SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

[12] G. Salton, A. Singhal, M. Mitra and C. Buckley, "Automatic Text Structuring and Summarization", In I. Mani and M. T. Maybury, editors, Advanced in automatic text summarization. The MIT Press, 1999, pages 341-355.

[13] Vietlex Semantic Tree, DOI=http://www.vietlex.com/resources/semanticTree.html, 2009.