Vietnamese Text Retrieval: Test Collection and First Experimentations

Ho Bao Quoc
Vietnam National University
Ho Chi Minh City School of Natural Sciences
227 Nguyen Van Cu – Q5 – Ho Chi Minh City – Vietnam
hbquoc@fit.hcmuns.edu.vn

Abstract

In this paper we present the Vietnamese specialities in word boundary, morphology, part of speech that must be addressed in information retrieval relative tasks. Our experiments have shown how different types of Vietnamese index terms: "tiếng", words, compound words, combination of word and compound word contribute to Vietnamese text processing and retrieval. We also introduce our Vietnamese test collection on which experimentations have been done and report the method used to construct this test collection.

1. Vietnamese specialities

Vietnamese is a monosyllabic language which uses a Latin alphabet with accents on the vowels to create new tonalities such "ă", "â", "ê", "ô", "ư". Vietnamese have six different tons which modify the meaning of the words, for example: ma (phantom), má (cheek), mà (but), mả (tomb), mã (code), mạ (rice seedling). Therefore, we can not use ASCII to encode Vietnamese characters. Instead, there are many character-sets have been using in Vietnamese electronic text such as: ABC, TCVN, VNI, UTF-8…and UFT-8 is the most common nowadays. Consequently, we may need a normalization of encoding prior to the phase of indexing.

Vietnamese has a special linguistic unit called "tiếng" (equivalent to hanzi of Chinese) which is similar to traditional morphemes in respect of content and similar to traditional syllables in respect of form [7]. A Vietnamese word consists of one or more "tiếng" separated by space, for example: "sách" (book), "dữ liệu" (data), "xã hội chủ nghĩa" (socialist) etc. Therefore, the whitespaces can not be used to identify the word boundary. This is a challenge for both Vietnamese Natural Language Processing (NLP) in general and Vietnamese text retrieval in particular. We will discus in details how different kinds of Vietnamese index terms contribute to the precision and recall of IR system in the experimentation section.

Vietnamese word is morphologic invariant: The word form is unchanged to its different grammatical roles in the sentence like that in Euro-Indian languages. Therefore, the lemmatization in index phase is not necessary for Vietnamese words. However, there are some exceptions in the processing of which morphologic normalization is needed. These exceptions are raised by two cases: the first is, the usage of vowels *i* and *y* is interchangeable in some circumstances such as "bác sĩ" and "bác sỹ", both of them correctly mean "doctor". The second is, the position of the tons may be variant, for example, "hòa bình" and "hoà bình" are acceptable. Though prfix and suffix can be seen in Vietnamese texts, they are used infrequently, for instance, the prefix "sự" transform a verb the verb "lựa chọn"

(choose) to a noun "sự lựa chọn" (choice), yet "lựa chọn" itself is also a noun with the meaning of "choice", on the other hand, the suffix "hóa" transform a noun "hiện đại" (modern) to a verb "hiện đại hóa" (modernization)

Unlike in morphologic variant language, the part of speech (grammatical category) of Vietnamese word can't be recognized from word form. It dependent, however, on the context of word:

"Thành công (success) của dự án đã tạo tiêng vang lớn"
"The success of the project makes a big echo"

"Anh ta đã ***thành công*** (succeed) trong nghiên cứu khoa học"
"He have succeed in scientist research"

"Buổi biểu diễn đã ***thành công*** (successful) "
"The show was successful"

The word ***Thành công*** in the first sentence is a noun, whereas in the second, it is a verb and in the third one, it is an adjective.

With the mentioned specialities above, we suppose that to get a high precision in Vietnamese text retrieval systems, NLP techniques should be applied to extract index terms that well represent the content of the documents. At least, Vietnamese Word Segmentation should be incorporated to identity Vietnamese words correctly. This hypothesis has been tested and results have been shown under experiments section.

2. Test collection

We have been constructing a Vietnamese test collection for our experimentations to identify the better index term for Vietnamese text retrieval. We used the pooling method to construct such collection.

As well known, a test collection for IR system test consist three parts: document collection, topic set and relevance assessments for each topic. The choice of search topics is important since better topics yield better reliability of the test collection. The search topics are chosen base on characteristic of language, size (in number of words) and the search domain. The relevance assessment constructing is the most tedious and time consuming phase. Of cause, we can't judge the relevance of all documents in the collection. Therefore we have been used the polling method [5] to build the relevance assessment file. We construct our test collection as following:

2.1 Document collection

Our text collection contains two parts: the first part is set of Vietnamese well known news papers (tuổi trẻ, thanh niên …) given by "Centre of Information and Prohibition of Ho Chi Minh City" (VN1). The original encoding of this collection is in TCVN character-set, we have transformed this part to UTF-8 character-set. This collection consist 11.398 documents of about 30Mb. The documents are tagged in SGML-like format.
The second part is the set of Vietnamese text (VN2) extracted from Vietnamese - English text collection. It contains 25.215 documents of approximately 69MB. This bilingual collection we had mined from the web site VOA [8], it contained about 1000 document pairs English – Vietnamese.

Collection	Num of docs	Size
VN1	11.398	30Mb
VN2	25.215	69Mb

2.2 Search topics

We have been constructing 14 search topics based on the themes of the documents in our document collection. These 14 topics would

like to cover the different types of topics: short topics, long topics, topics containing simples words, topics containing compound words…The set of topic is organized in TREC topics format. Each topic contains a narrative part giving how to judge whether a document is relevance to the topic. This information makes a guideline for the human assessor.

<TOP>
<NUM> 10</NUM>
<TITLE>
Thương mại Việt Mỹ
</TITLE>
<DESCRIPTION> Các chính sách và hoạt động liên quan đến thương mại giữa Việt nam và Mỹ
</DESCRIPTION>
<NARRATIVE>
Các chính sách mới trong quan hệ thương mại hai nước, các cuộc tiếp xúc của các tổ chức thương mại của hai bên, các báo cáo về kết quả của sự hợp tác thương mại giữa hai nước. Các bài báo nói về các vấn đề trên được cho là liên quan.
</NARRATIVE>
</TOP>

Fig 1. An example of search topics:

<TOP>
<NUM> 10</NUM>
<TITLE>
Vietnam America Trading
</TITLE>
<DESCRIPTION>
The policies and activities relates to trading of Vietnam and America
<NARRATIVE>
The new policies in trading of two countries, the events are organized of trading organizations of two contries, the reports of trading cooperation Vietnam – America, the documents relate the subjects above are judged relevance.
</NARRATIVE>
</TOP>

Fig 2. Translation of topic in Fig 1

2.3 relevance assessment

We have used pooling method to constructing the relevance assessment. We use SMART, Lemur, and Terrier to make the pool. For each system and for each search topics, we use 50 top relevance documents. These 50 documents are judged by human assessors.

We are continuing to add more topics and judges the relevance documents for new topics. We are intention to having 25 topics with relevance assessments in the next month.

3. Experimentations

3.1 Indexing units for Vietnamese IR

As mentioned above, word is the basic unit of indexing in traditional IR. Vietnamese sentences is composed of continuous "tiếng" separated each others by white space, each "tiếng" being a string of Latin characters with some special accents. A single "tiếng" may have no meaning by itself: most of Vietnamese word is composed with two "tiếng"[4]. For example, in *ngôn ngữ* the latter is meaningful (*linguistics*) but the former is not, and both "tiếng" together have also a meaning (*language*). Another specific characteristic in Vietnamese document is that a "tiếng" considered separately may have a different meaning than combining

with two or three contiguous "tiếng" together. For example, *trang trí* means "*décor"* (if used as a noun) or "*to decorate"* (if used as a verb), but "*trang"* and "*trí"* independently mean respectively "*page"* (noun) / "*to shift"* (verb) and "*mind"* (noun). So, to determine correct words for indexing consists of detecting not simply meaningful words but also words suitable meaning. In the following, "term" will designate meaningful word.

There are two methods of indexing [3]:

a) The first one relies on linguistic knowledge and consists of ***dictionary-based word segmentation***. Sentence will be segmented into terms which are identified from dictionary entries. When there are word segmentation ambiguities, the longest-matching strategy is used to select the best term. For example:
"công nghệ thông tin"("information technology") can be segmented in three ways with 7 possible terms – {"*công"*, "*nghệ"*, *"thông"*, *"tin"}*, *{"công nghệ"*, *"thông tin"}*, and *{"công nghệ thông tin"}*- all of these are meaningful but the latter is chosen since it is longest meaningful word.

   Two main problems are raised from this technique are:
   - The **loss in recall**, this problem is identical to the one in Chinese IR [3]: when the longest matching is used, only the longest term is identified as an index. However, a long term may contain shorter terms, as indicated in the above example, the term "*công nghệ thông tin"* contains 6 others terms, and documents indexed by "*công nghệ thông tin"* can also be referred under two others terms such as "*công nghệ" (technology)* and "*thông tin" (information)* . Since these two last terms are included in *công nghệ thông tin – information technology*, they are not considered as independent indexes for IR.
   - The **Unknown word** problem, especially proper nouns, new political words, abbreviations, etc… These words are less likely to appear in the dictionary.

b) The second method is ***n-grams*** which is a non based-linguistic technique. Usually, uni-grams or bi-grams are often chosen for its reasonable memory cost and performance. And uni-grams or bi-grams also fit well to Vietnamese meaningful words. Longer words are compounded from n-grams of length of one or two. This method is very powerful for resolving the above two problems above.

   - Regarding the loss in recall, in order to detect shorter terms in a long term, full segmentation of the long term into bi-grams is done. Bi-grams which have a meaning in Vietnamese language can be determined by scanning from left to right, and never by selecting two "tiếng" appearing in the middle of the long term. Therefore, for the term "*công nghệ thông tin"(Information technology)*, two selected bi-grams are "*công nghệ"(technology)* and "*thông tin"(information)*, yet never "*nghệ thong"* since it is nonsense. Thus in Vietnamese text, we do not have the cross-word segmentation phenomenon as in Chinese documents [3].

- Concerning proper noun, such as, *Hoàng Liên Sơn* (name of a mountain in North Vietnam), segmentation based on bi-grams will split this term into "*Hoàng Liên"* and *"Liên Sơn"*. If both bi-grams occur in the same document, there is a higher probability that the document concerns *Hoàng Liên Sơn* than those with three uni-grams. This technique can also be used to detect new political terms or abbreviations.

Finally, the step of ***removing stop words*** in Vietnamese documents needs specific process, besides common technique as used in European language for removing prepositions, pronouns. We used a given stop list to remove stop words as often seen, and employ heuristic rule to detect stopwords which are not in stop list. For example, a possible rule used is: if a bigram is in form XX (two word are the same) is it is a stopword [4] : *lâng lâng* , *chiều chiều* .

3.2 Experiments

The SMART system [1] is used for the experimentation.
The indexing results for a document are vector of weights:

$$D_i \rightarrow (d_{i1}, d_{i2}, ..., d_{im})$$

where d_{ik} ($1 \le k \le m$) is weight of the term t_k in the document D_i, and m is the size of the vector space. The weight d_{ik} of a term in a document is calculated by *ltc* weight scheme of SMART according to formula

$$d_{ik} = \frac{[\log(f_{ik})+0.1]*\log(N/n_k)}{\sqrt{\sum_j [\log(f_{jk})+0.1)*\log(N/n_k)]2}}$$

where f_{ik} is the occurrence frequency of the term tk in the document D_i, N is the total number of documents in the collection; n_k is the number of documents that contain the term t_k
A query is indexed in a similar way, and a vector is also obtained for a query

$$Q_j \rightarrow (q_{j1}, q_{j2}, ..., q_{jm})$$

Similarity between Di and Qj is calculated as the inner product of their vectors, that is:

$$Sim\ (D_i, Q_j) = \sum_k (d_{ik} * q_{jk})$$

Four kinds of test have been carefully examined so that a comparison among these results can be made in order to choose the best way for indexing. In all four method below, we removed stopwords :

1. using single word as indexes
2. using bigram
3. mixing single word and dictionary-based segmentation
4. using dictionary-based segmentation to find out units indexes

3.2.1 Single "tiếng" (uni-gram):

In the first examination, we indexed a test collection using single "tiếng" (uni-gram) as index terms. The result of using single word is imprecision but it may provide a basic on which one can measure improvements by other representation methods. The average precision 11-pt for this case is **0.3636**

3.2.2 Using bigram

In the second, we used bigrams as indexes. In this method, the average precision is augmented to **0.3778**, but lost of precision for high recall

3.2.3 Mix uni-gram and dictionary-based segmentation

In the third, we mixed 1-gram with the application of dictionary-based segmentation. In fact, we constructed compound words in scanning from a

lexicon. Moreover, we also kept 1-gram of these segments. The average precision for 11-ptr is **0.4989.**

3.2.4 Dictionary-based segmentation

In the last one, we used a small machine readable Vietnamese dictionary about 30 000 units. We have done a pre-processing test collection by scanning from left to right and looking up in the dictionary in order to find a good segmentation. When it had been found, we connected its words by "under score" characters[1]. After this pre-processing, we used the processed collection to run SMART. The average precision for 11-pt is improved to **0.5625**

The detail results of four methods of representation are following:

	1-gram	bigram	1-gram & lexicon	lexicon
Number of queries	14	14	14	14
Retrieved documents	280	280	280	280
Relevant documents	64	64	64	64
Rel_ret	52	31	59	58
Trunc_ret	228	231	174	167
Recall – Precision Averages:				
at 0.00	0.5792	0.6157	0.7321	0.7411
at 0.10	0.5792	0.6115	0.7321	0.7411
at 0.20	0.5173	0.5937	0.6964	0.7411
at 0.30	0.4387	0.4865	0.6119	0.6804
at 0.40	0.4000	0.4685	0.5109	0.5615
at 0.50	0.3944	0.4649	0.4920	0.5615
at 0.60	0.3145	0.2813	0.4024	0.4995
at 0.70	0.2878	0.1710	0.3849	0.4825
at 0.80	0.2364	0.1662	0.3473	0.4507
at 0.90	0.1272	0.1662	0.3089	0.3845
at 1.00	0.1244	0.1305	0.2692	0.3443
Average precision for all point				
11-pt Avg	**0.3636**	**0.3778**	**0.4989**	**0.5625**

[1] "Under score" characters are used in order that SMART will treat as a normal word.

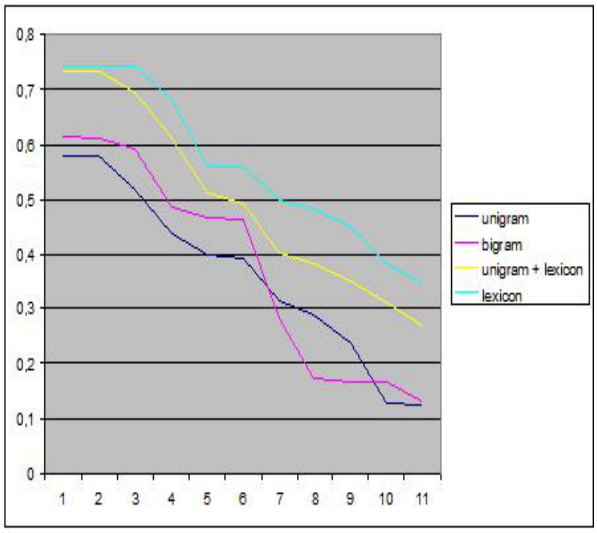

Fig 3. Recall – precision graphs

4. Concluding remarks and future works

This paper is an overview of specific problems of indexing for Vietnamese IR. Accepted some problems which are proper to Vietnamese documents (bi-grams selection, stop words), most of methods used are those already experimented in Chinese IR. Evaluation the performance of three methods mentioned above has proven to be effective of using dictionary-based segmentation method for Vietnamese IR.

We are trying application of statistic methods to find out compound words that have been not exit in our dictionary and using linguistic knowledge to deal with unit indexes more complex such as noun phrase or verb phrase.

This research is carried out jointly with a French team from the laboratory CLIPS of IMAG and the University of Joseph Fourier (Grenoble, France).

We are continuing to construct our Vietnamese test collection by adding more topics and modifying the relevance assessments.

References

[1] Gerard Salton, Michael J. McGill. *Introduction to modern Information Retrieval System*. McGraw-Hill, 1980.

[2] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, United Kingdom, 1979.

[3] Jian-Yun Nie, Jiangfeng Gao, Jian Zhang, Ming Zhou. *On use of Words and n-grams for Chinese Information Retrieval*. Proceeding of the 5th International Workshop Information Retrieval with asia languages. 1997.

[4] Nguyễn Kim Thản. *Nghiên cứu ngữ pháp tiếng Việt*. Nhà xuất bản Giáo Dục. 1997.

[5] Gilbert G and Sparck Jones. *Statistical bases of relevance assement for the 'Ideal' information retrieval test collection*. BL R&D Report 5481, Cambridge, England, 1979

[6] Doulag W. Oard. *A survey of multilingual text retrieval*. UMIACS-TR-96-19. 1996

[7] Dinh Dien, Hoang Kiem. *Vietnamese Word Segmentation*. NLPRS2001 - Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium - November 27-30, 2001 –Tokyo, Japan

[8] Van B. Dang, Bao-Quoc Ho. *Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining*. RIVF 2007 – Internaltional Conference on Research, Innovation and Vision for the Future – March 05-09, 2007 – Hanoi, Vietnam.