A Basic framework to Build a Test Collection for the Vietnamese Text Catergorization

Viet Hoang-Anh, Thu Dinh-Thi-Phuong, Thang Huynh-Quyet
Hanoi University of Technology, Vietnam
vietha-fit@mail.hut.edu.vn, thanghq-fit@mail.hut.edu.vn

Abstract

The aim of this paper is to present a basic framework to build a test collection for a Vietnamese text categorization. The presented content includes our evaluations of some popular text categorization test collections, our researches on the requirements, the proposed model and the techniques to build the BKTexts - test collection for a Vietnamese text categorization. The XML specification of both text and metadata of Vietnamese documents in the BKTexts also is presented. Our BKTexts test collection is built with the XML specification and currently has more than 17100 Vietnamese text documents collected from e-newspapers.

1 Introduction

Natural Language Processing (NLP) for such popular languages as English, French, etc. has been well studied with many achievements. In contrast, NLP for unpopular languages, such as Vietnamese, has only been researched recently. It means that expecting international scientists to care about our problems is not feasible in the near future. In this paper, we present our research results on that field, especially on Vietnamese test collections for Vietnamese text categorization. This paper will be organized as follows. Section 2 proposes our research on the requirements, models and techniques to build a Vietnamese test collection for researches and experiments on Vietnamese text categorization. Section 3 presents our results with BKTexts test collection. Lastly, the focus of our ongoing research will be presented in section.

2 Model of building a test collection for the Vietnamese text categorization

Until now, there has not been a Vietnamese standard test collection for Vietnamese text categorization. Vietnamese documents used in previous studies of Vietnamese researchers are gathered by themselves and were not thoroughly checked. Moreover, all over the world, there have been a lot of test collections in many different languages, especially in English such as the Reuters-21578, the RCV1 and the 20NewsGroup[1]. Therefore, we intend to build a Vietnamese standard test collection for the Vietnamese text categorization. We defined a framework for building Vietnamese test collections as follows. Basic requirements for a Vietnamese test collection text categorization

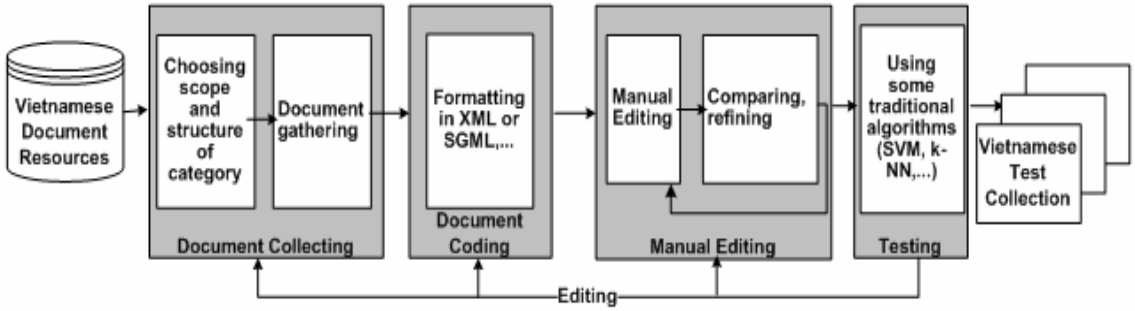

Figure1. The system architecture to build a Vietnamese test collection for text categorization

Our model to build a Vietnamese test collection for text categorization is accomplished in four stages: collecting, auto coding, manual editing, and testing (Figure 1).

[1] Available from http://kdd.ics.uci.edu/

From available resources, we gather Vietnamese documents for the test collection in accordance with the scope and the structure of categories. Researchers usually use documents collected from e-newspapers because these documents are pre-processed and less ambiguous. Then an auto system tags documents in the XML (or SGML) formatting specification.

After being coded, documents are manual edited by editors. The editors would assign the categories they felt applicable. They also edit specification tags of formatted documents in order to completely and more precisely describe attributes of documents. Lastly, to assess the accuracy of the test collection, we use some famous categorization algorithms such as SVM, k-NN, etc. Performing the test and correction several times, we will gradually obtain a finer and more precise test collection. The process ends when errors are below a permitted threshold.

3 The BKTexts test collection for Vietnamese text categorization

With the model mentioned above, we are constructing the BKTexts test collection for the first version. We collected about 17100 documents for the BKTexts from two e-newspapers *http://www.vnexpress.net* and *http://www.vnn.vn*. Categories are organized in a hierarchical structure of 10 main classifications and 37 sub-classes. Documents are marked up with XML tags and given unique ID numbers. The XML specification of a document in the BKTexts test collection is described in Figure 2. Building a successful Vietnamese test collection for text categorization has a significant meaning. It will be a useful material for any study on text categorization and Vietnamese processing in the future because it reduces a lot of manual work and time, as well as increases the accuracy of experimental results.

4 Conclusion and future work

We have presented our research results on defining requirements, the model and techniques to build a Vietnamese test collection for researches and experiments on Vietnamese text categorization. Currently, we continue building the BKTexts on a larger scale for publishing widely in the near future. This test collection enables researchers to test ideas and to objectively compare results with published studies.

```
<?xml version="1.0" encoding="Unicode" ?>
<BKTEXTS ID="" SPLIT="">// Identifying the document ID
and whether it belongs to the training or test set.
<METADATA>
    <TOPICS></TOPICS> // Categories of the document
    <DATE></DATE> // the date of the document
    <VNFONT></VNFONT> // the Vietnamese font of the
document
    <SIZE></SIZE> // the size of the document
    <UNKNOWN_TEXT> </UNKNOWN_TEXT> // the noisy
characters in the document
<SOURCE> // the source of the document
          <DATELINE></DATELINE>
          <ORGS></ORGS>
          <COUNTRIES></COUNTRIES>
 </SOURCE>
 <AUTHOR> // authors of the document
          <FULLNAME></FULLNAME>
          <ORGS></ORGS>
          <COUNTRIES></COUNTRIES>
 </AUTHOR>
<CODER> // the editor coding the document
          <FULLNAME></FULLNAME>
          <ORGS></ORGS>
          <COUNTRIES></COUNTRIES>
          <NOTES></NOTES> // some notes of the editor
 </CODER>
</METADATA>
<TEXT>
    <TITLE></TITLE> // Title of the document
    <SUMMARY></SUMMARY> // Summary of the document
    <HEADLINE></HEADLINE> //
    <BODY></BODY> // the main text of the document
</TEXT>
<COPYRIGHT></COPYRIGHT>
</BKTEXTS>
```

Fig.2. The XML specification of the BKTexts

References

David D. Lewis, Reuters-21578 Text Categorization Test Collection, www.daviddlewis.com, 1997.

David D. Lewis, Yiming Yang, Tony G.Rose, Fan Li, "RCV1: A new Benchmark Collection for Text Categorization Research", in: *Journal of Machine Learning Research 5*, pp.361-397, 2004.

Huynh Quyet Thang, Dinh Thi Phuong Thu. Vietnamese text categorization based on unsupervised learning method and applying extended evaluating formulas for calculating the document similarity. *Proceedings of The Second Vietnam National Symposium on ICT.RDA, Hanoi 24-25/9/2004, pp. 251-261 (in Vietnamese)*

Dinh Thi Phuong Thu, Hoang Vinh Son, Huynh Quyet Thang. Proposed modifications of the CYK algorithm for the Vietnamese parsing. *Journal of Computer Science and Cybernetics, Volume 21, No. 4, 2005, pp. 323-336 (in Vietnamese)*