Building Knowledge Base for Vietnamese Information Retrieval

Thanh C. Nguyen
CSE Faculty, HCMC UT
268 Ly Thuong Kiet, 10th district, HCMC, Vietnam
thanh@cse.hcmut.edu.vn

Hai M. Le, Tuoi T. Phan
CSE Faculty, HCMC UT
268 Ly Thuong Kiet, 10th district, HCMC, Vietnam
{lmhai, tuoi}@cse.hcmut.edu.vn

ABSTRACT

At present, Vietnamese knowledge base (vnKB) is one of the most important focuses of Vietnamese researchers because of its applications in wide areas such as Information Retrieval (IR), Machine Translation (MT) etc. There have been several separate projects developing vnKB in various domains. The training in vnBK is the most difficulty because of quantity and quality of training data, and lacking of available Vietnamese corpus with acceptable quality. This paper introduces an approach, which first extracts semantic information from Vietnamese Wikipedia (vnWK), then trains the proposed vnKB by applying support vector machine (SVM) technique. The experimentation of the proposed approach shows that it is a potential solution because of its good results and proves that it can provide more valuable benefits when applying to our Vietnamese Semantic Information Retrieval system.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*. H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *user issues.*

General Terms

Algorithms, Experimentation.

Keywords

Term extraction, topic classification, knowledge base, SVM, vnKB

1. INTRODUCTION

A knowledge base can serve as the core of any information system such as information retrieval, semantic web, question answering or text summarization ... It provides many semantic information which affects outcome quality of those systems.

Recently Vietnamese researchers have been focusing on developing vnKBs to use them in various applications on natural language processing area. However there are several difficulties of vnKB development task according to those researches. First of all, there is not existent a common vnKB which any system and solution can refer to. Also, the complexity of Vietnamese grammar impacts on extracting and retrieving correct data from Vietnamese documents. Other reason is a lack of available Vietnamese corpus with acceptable quality.

To develop a good vnKB, those difficulties must be resolved with consideration for effects on its development task. The first difficulty can be solved by defining core structure of vnKB then train it continuously. The second can be done by developing good quality Vietnamese word segmentation and using dictionaries of experts thus can affect data analysis and retrieval. Without a large-scale quality Vietnamese corpus, the last is hard to close.

In our research, solutions for the first and the second difficulties are proposed in section 4 and thereafter developed step-by-step. For the third, there are two options- one using manual built Vietnamese documents and the other using data popular or well-known websites. Since there is no many Vietnamese softcopy documents for corpus collecting, using vnWK is the best solution to solve the third difficulty in our research. By this way, an approach of extracting data from vnWK to train our vnKB is introduced in the paper as below.

The second section of the paper gives an overview on the results of Wikipedia exploitation from other researches. The third section provides a summary of vnWK including its database organization. Next section summarizes the SVM. Our vnKB development and training solution in section 5 is the key of the paper. The experiment in section 6 provides results of implementing and practicing the vnKB model. Last section is our conclusion which mentions the good, the bad of our approach and necessary improvements.

2. RELATED WORKS

There are several researches and toolkits from other groups in the world regard Wikipedia data extraction.

First of all, research of J. Isbell group [1] (from Digital Media Systems Laboratory of Hewlett-Packard Development Company) focused on structure data and template of English Wikipedia page such as WikiText markup and Infobox template including many sub-fields then convert raw data of that WikiText into Abstract Syntax Trees and finally use regular expressions to extract information as RDF [2] from English Wikipedia pages.

iiWAS2009, December 14–16, 2009, Kuala Lumpur, Malaysia.
Copyright 2009 ACM 978-1-60558-660-1/09/0012...$10.00.

The research of S.Auer et al [3] investigated the Wikipedia page template structure then "separating valuable information from less important one", and "extracting this information from templates in wiki texts" and "converting it into RDF under usage of unified data types", after that "querying and browsing this information even though its schema is very large and partly rudimentary structured" (from [3]). From roughly 10GB raw data, their extracted result was 106,049 categories, 111,548 classes, 647,348 instances, 8,091 properties and 8,415,531 triples.

Other research from S. Chernov [4] analyzed regular semantic relationship among terms of pages in English Wikipedia then extracted a core set of pages which have a common topic also their belonged categories. Their experiment test data based on measures of Number of links between categories and Connectivity Ratio also proposed new measure named the Semantic Connection Strength measure.

Lastly, some commercial products provide features of Wikipedia data extraction in [5] and [6].

Regarding SVM technique, there are many recent researches which applying SVM to solve issues of natural language processing (NLP) area such as text classification (categorization), sentence extraction or phrase recognition.

The approach of N.L.Minh [7] group from JAIST used SVM Ensemble to extract sentences in order to support their text summarization research. They used 7 features regarding sentence information such as location, length, relevant to title … to build training model and to test data. Their experiment performed on 500 Vietnamese documents by SVM-Light then achieved 0.53 F-measure in compression with individual SVM 0.51 F-measure.

As similar as above research, the result of C. Kruengkrai et al [8] focused on some new features such as TF-IDF, paragraph and section structure. Then they performed experiments in order on Ziff-Davis and the cmp-lg datasets with four kernel functions (Linear, Polynomial, Gaussian and Sigmoid). Their achieved maximum result were 0.515 for precision (Sigmoid), 0.962 for recall (Sigmoid) and 0.659 for F-measure (Sigmoid) on Ziff-Davis dataset; and 0.408 for precision (Gaussian), 0.937 (Linear and Sigmoid) and 0.487 (Gaussian) on cmp-lg datasets. They are better result for other researcher's reference to.

Final project report of Alex Cheng [9] provided overall picture of applying SVM to chunk base noun phrase in NLP in which IOB Tag and Open/Close Bracket were used to represent two major classes of chunk representation. There were seven experiments performing on training data (consisting of sections 15-18 of the WSJ part of the Penn TreeBank) and test data (section 20) based on SVM-light with kernel functions (2nd degree polynomial vs. 3rd degree polynomial) and cost parameter C constant (1). The max precision, recall and F of those experiments were in order 94.65%, 94.13%, 94.14%. They showed good result of approach also SVM ability.

Other report of W.St.Charles group [10] provided an evaluation of roles and benefit of applying current techniques on noun phrase extraction issue. It also shown comparison among those techniques which were used by many research groups in the world. Their experiment was trained and tested on a common dataset for the CONLL (Conference on Natural Language Learning) 2000 shared task. It shown that performance measure 94.39 with SVM parsing method on Perl, C++, Python of group Taku Kudoh, Yuji Matsumoto in published "Chunking with Support Vector Machines" was best in the list.

The research of Silvia Baptista [11] which focused on applying SVM based on GATE NLP toolkit [13] to do noun phrase chunking for medical data in Unified Medical Language System. Their experiment's comparison shown that, SVM results of the GATE program were not better than results of CaRE system.

The last is research of N.Q.Chau [12] who is a member of our research group. He focused on Vietnamese key noun phrase extraction based on applying SVM technique. The final result of his experiment was not so high but it was very valuable due to high complexity and irregular of Vietnamese.

3. VIETNAMESE WIKIPEDIA

In general Wikipedia, each page defines a meaningful concept. By that, one term might link to many pages in same topic or other topics. Therefore, Wikipedia can be considered as a semantic network of concepts. Semantic information may be extracted from this network.

More late than English Wikipedia, vnWK was created in 2003. It had about 70,000 pages in February 2009 from just 40,000 pages in June, 2008. By this amount, vnWK is useful to support applications and researches on natural language processing area. vnWK database is organized in primary entities such as *page*, *pagelinks*, *categorylinks*, *redirects*, *externallinks* as Fig.1.

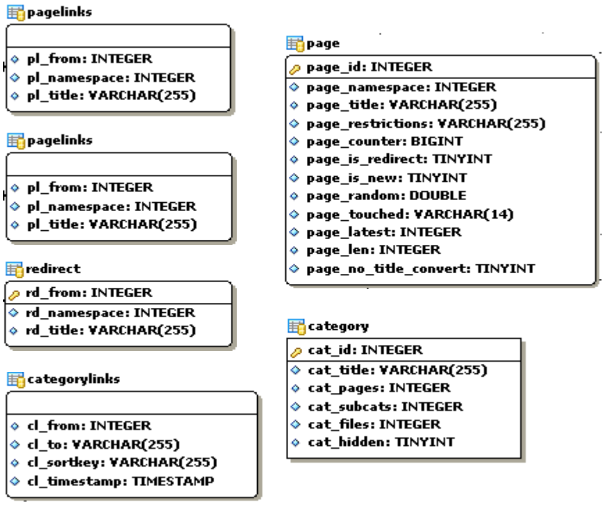

Figure 1. Primary data entities of vnWK

4. SUPPORT VECTOR MACHINE

SVM techniques are the most complex machine learning ones but they are also the most accurate and computationally inexpensive. SVM can be binary classifiers or regression (Support Vector Regression). It can be applied in many sub-areas of NLP such as classifying documents by separately topics or extracting phrase from sentences, etc.

Its core solution is that configuration information of the problem is determined first, and then it is mapped into a n-dimensional space, where n is the number of features that the configurations should be compared on. Usually, two colors (such as red and blue) will be assigned to represent its comparison result in "yes" and "no" answer (or positive and negative "half-space" which is similar to "hyper-plane" term). When classification task completes on dataset, the system attempts to separate the red items from the blue ones, then to create two n-1 dimensional hyper-planes. This means that, hyper-planes of 2-dimensional space are separated by a line; and hyper-planes of 3-dimensional space are separated by a plane. Support vectors in 2-dimensional space are illustrates in Fig.2 as below.

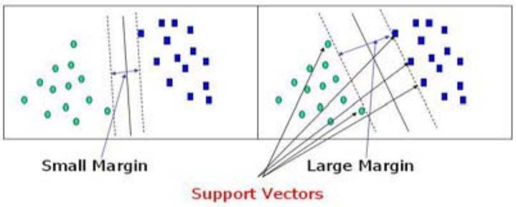

Figure 2. Support Vectors in 2-dimensional space

5. VIETNAMESE KNOWLEDGE BASE (vnKB)

5.1 Structure

Basically our vnKB is designed based on OWL-RDF format in Protégé toolkit. Its organization is introduced as below.

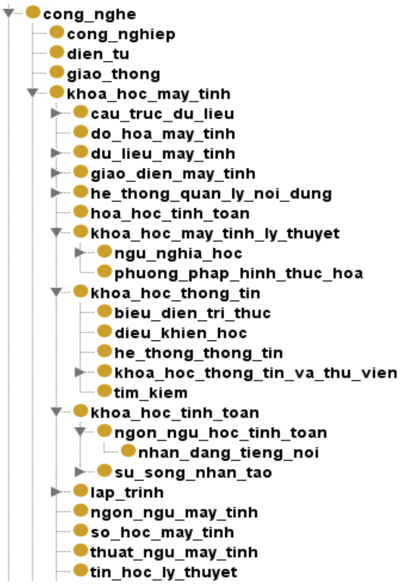

Figure 3. Proposed vnKB organization

In the organization, each leaf node is a sub-class which links to group of relevant terms that are selected in extracted terms from vnWK. These selected terms create a sub-semantic network including meaning full sense in sub-domain of that node. Other high level node relates to parent class of its children also link to more-general domain.

From the proposal model of vnKB, the approach of exploiting data from vnWK for its training is introduced in the model of "vnKB" system in Fig.4 below.

In this model, the first module "Term & Link Extraction" is responsible for processing each page of vnWK to look at and extract its candidates of term (in page data entity) and relevant links (in pagelink and redirect data entity). After that these candidates are re-evaluated based on Vietnamese dictionary which created by linguistic experts. This step will help to keep candidates which having really full meaning also appropriate links, and remove the others. The measures of *precision* and *recall* will be used to analyze result of selected data also. Steps in first module execute automatically based on library from GATE [13] also toolkit of Vietnamese word segmentation and POS tagger [14], and SVM classifier technique for phrase extraction [14].

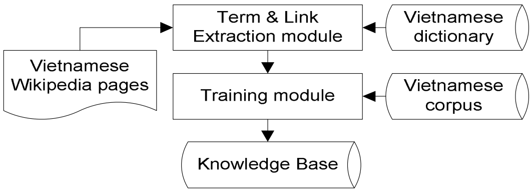

Figure 4. Proposed model of vnKB system

Last module in the system processes selected candidates (terms and their links) to train the vnKB to appropriate sub-class and sub-domain. The main purpose of this is to remove unsatisfied data such as duplicated, unclear or ambiguous data. An important support of Vietnamese corpus in execution steps of the module is to provide probability of occurrence each output candidate also probability of depending each others, thereby the system can recognize dominant candidates which having higher probability values to keep for next processing step by SVM. Here SVM technique is applied to recognize correspondent topic that a term or link should be belonged to. After above steps, the senses of each concept (term, selected candidate) are also stored in vnKB along with it. That will help subsequent query task.

5.2 Algorithm of term & link extraction

At present, there have been several methodologies to extract data from Vietnamese documents but most of them are rule-based also low flexible for irregular case. Thus, applying SVM to extract Vietnamese data from vnWK is a suggestion from our research to utilize SVM's advantages in machine-learning field.

The conditions of data extraction are (a) it is *phrase*, (b) its *length* (number of word) is *less than 3*, (c) it is *neither a proper noun nor stop-word*. By this, data should be noun or noun phrase format which can be easily extracted from terms and links of vnWK.

To do this term extraction, the system's model first must be trained by learning classes from manual extracted key phrases including features of training documents (in kind of scientist paper, vnWK page, e-book …). The features here are *phrase information*, phrase's *location* ("yes" for terms in title of document, and "no" vice versus), *length* ("yes" for length 1, 2, 3 and "no" for vice versus), *part-of-speech* tag ("yes" for noun phrase and "no" for the others), its *finding* in dictionary ("yes", "no"), *frequency* ("yes" for positive number and "no" for zero) by applying SVM classifier technique. These features will build a 6-dimensional vector space which includes many phrase configuration tuples. After that, base on training model, new phrases and their configuration tuples will be signed "yes" or "no" accordingly retrieved annotated result from SVM's running. Then, "yes" phrases including relevant features will be selected as enough as result of first module. Thereby, links of documents which those "yes" phrases (terms) belong to will be recognized and selected for result of link extraction.

5.3 Algorithm of term's topic classification

The SVM classifier technique is still applied in this algorithm but not for data extraction purpose. The main target here is how to recognize correct topic(s) to which a term should belong. To do that, as same as first algorithm steps, new training model is built by learning trained data including relevant features. From this, a n-dimensional vector space is built also which n is number of features in the model. New used features are *topics* in leaf nodes of the vnKB tree in Fig.3 that feature's value is "yes" for correct topic detection and "no" for none. Depending on sub-area that our vnKB supports, the topic list can be changed then the features of this algorithm will be changed accordingly to.

From this training model, the system will analyze practice data to determine their topics then store some topic-recognizable data to relevant dataset as algorithm result.

Our research in initial time just uses SVM linear, polynomial kernel functions for above algorithms. The other kinds of kernel functions such as Gaussian or Sigmoid will be considered more in near future.

6. EXPERIMENT

Our vnKB model is developed in MS SQL Server 2005 database management system which has many advantages. Test and training data are collected from the lastest version of vnWK (dumped on Jun 27 2009). It includes huge data as follows:

- page.sql.gz (8.5 MB) including 320,160 pages
- pagelinks.sql.gz (68.3 MB)
- categorylinks.sql.gz (4.4 MB)
- interwiki.sql.gz (7 KB)
- category.sql.gz (532 KB)
- redirect.sql.gz (430 KB)

For SVM library, the DotNet version (SVM.NET 1.6 from http://www.matthewajohnson.org/software/svm.html) is used to build the practice models of two algorithms. Our initial experiment performs on 2 runs of our dataset that containing 500 pages and their links only in which 80% for training and 20% for test. Also, our evaluation is based 4 signals:

- FP (false positive): it is a negative and its classified label is a positive,

- FN (false negative): it is a positive and its classified label is a negative,

- TP (true positive) and TN (true negative): it is same as its classified label,

and 3 measures Recall (R), Precision (P) and F for the test set that *R = TP/(TP+FN)*, *P = TP/(TP+FP)* and *F=2*R*P/(R+P)*

The first experiment performs on above data for term & link extraction algorithm. Table 1 describes its result in summary. The result is also 1,250 extracted phrases (terms) and about 3,000 links. They are not a large dataset but can be increased more in near future regarding this algorithm improvement.

Table 1. Summary of term & link extraction

		P (%)	R (%)	F (%)
Linear kernel (LK)	(LK) Run #1	45.35	41.31	43.23
	(LK) Run #2	44.37	46.12	45.23
	(LK) Average	*44.86*	*43.71*	*44.23*
Polynomial kernel (PK)	(PK) Run #1	46.39	50.30	48.26
	(PK) Run #2	45.10	46.15	45.62
	(PK) Average	*45.74*	*48.23*	*46.94*

and illustrated chart as below.

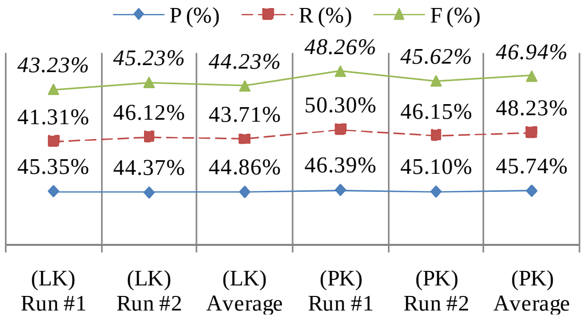

Figure 5. Comparison of first experiment results

Next, the experiment of second algorithm performs on result of first step (1,250 phrases) to put them to correspondent location in list of 16 topics.

Table 2. Summary of term's topic classification

		P (%)	R (%)	F (%)
Linear kernel (LK)	(LK) Run #1	28.01	26.45	27.21
	(LK) Run #2	40.44	42.11	41.26
	(LK) Average	*34.22*	*34.28*	*34.23*
Polynomial kernel (PK)	(PK) Run #1	37.87	40.93	39.34
	(PK) Run #2	35.52	30.13	32.61
	(PK) Average	*36.69*	*35.53*	*35.97*

and illustrated chart as below.

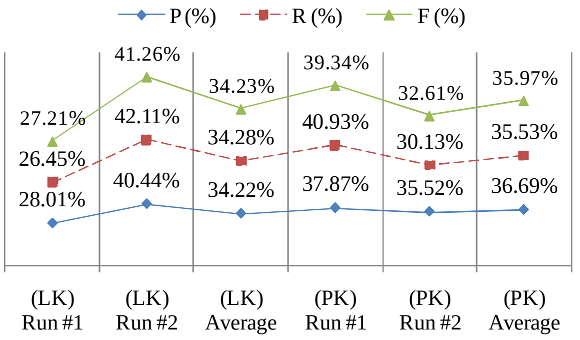

Figure 6. Comparison of second experiment results

The initial results of these algorithms are not actually high due to high complexity of Vietnamese. From these, the main task of furthering our research is to improve steps of algorithms to increase quality of training and testing data. The meaning of these results shows that the system vnKB can be implemented and our solution is possible to build up a vnKB.

7. CONCLUSION

The paper introduces the approach of exploiting data from Vietnamese Wikipedia (vnWK) to develop and train Vietnamese Knowledge Base (vnKB). In the paper, the structure of vnKB, which mapped to appropriate topics of vnWK, is proposed. Besides, the model of the system vnKB is introduced to show the way text data in vnWK will be extracted to train our vnKB. Applying SVM in two steps of vnKB training solution is the primary focus of the research team in expectation of good result for the whole progress. The initial experiment performed on Vietnamese documents (Vietnamese Wikipedia) is actually not good but it shows that the proposed approach is possible for implementation and it can be upgraded in further research.

8. ACKNOWLEDGMENTS

Our thanks to all members in the BK-NLP group of HCM City University of Technology Vietnam for their enthusiastic collaboration, especially Professor Dr. Tuoi Phan. This work was supported by the key project KC.01/06-10/02 “Research and development multimedia information retrieval systems supporting Vietnamese” of the group.

9. REFERENCES

[1] J. Isbell and M.H. Butler, “Extracting and Re-using Structured Data from Wikis”, in Digital Media Systems Laboratory of Hewlett-Packard Development Company, Bristol, HPL-2007-182, 14th November, 2007

[2] http://www.w3.org/TR/rdf-primer

[3] S. Auer and J. Lehmann, “What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content”, The Semantic Web: Research and Applications, pages 503-517, 2007.

[4] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhuo, “Extracting semantic relationships between wikipedia categories”, In 1st International Workshop: SemWiki2006 – From Wiki to Semantics (SemWiki 2006), co-located with the ESWC2006 in Budva, Montenegro, June 12, 2006.

[5] http://www.knowlesys.com/products/wikipedia_data_extract

[6] http://www.evanjones.ca/software/wikipedia2text

[7] M. L. Nguyen, A. Shimazu, X.H.Phan, T.B.Ho and 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, Nov. 14-17, 2119, Kobe, Japan, Symposium 5, Session 2, http://hdl.handle.net/10119/3909

[8] C. Kruengkrai, C. Jaruskulchai , “Using OneClass SVMs for Relevant Sentence Extraction”, Proceedings of the 3rd International Symposium on Communications and Information Technologies (ISCIT-2003), September 3-5, 2003. Songkhla, Thailand, http://www.tcllab.org/canasai/pubs/iscit-03-one-class.html

[9] A.Cheng, “Base Noun Phrase Chunking with Support Vector Machines”, Final Project Report, Cornell University, Ithaca, NY

[10] W.St.Charles, “Noun Phrase Extraction - An Evaluation and Description of Current Techniques”, Department of Computer Science The University of Tennessee at Chattanooga

[11] S.Baptista, “Integrating Medical Text Extraction Tools”, technical report, Department of Electrical Engineering and Computer Science Massachusetts Institute of Technology January 31, 2008

[12] C.Q.Nguyen, T.T.Phan, “A Pattern-based Approach to Vietnamese Key Phrase Extraction”, In Proceedings of The 5th International IEEE Conference on Computer Sciences-RIVF’07 (2007).

[13] GATE, http://gate.ac.uk

[14] C.Q.Nguyen, T.T.Phan, “A Hybrid Approach to Vietnamese Part-Of-Speech Tagging”, In Proceeding of The 9th International Oriental COCOSDA 2006 Conference - O-COCOSDA’06 (12/2006), Malaysia, pp.157-160.