Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.265 – 270

RESEARCH ARTICLE

BUILDING ONTOLOGY IN A FIELD OF COMPUTER FOR VIETNAMESE LANGUAGE

Tuan. Tran Anh
Faculty of Foreign language – Informatics – Economics
DakLak College of Pedagogy, Vietnam
E-mail: tuanta@dlc.edu.vn

ABSTRACT - The World Wide Web is an enormous repository which concentrates huge volumes of distributed heterogeneous data. Its data is very useful to solve some problems in economics, science, etc. To get data exactly and effectively from this repository is very difficult for human. Computer could do well this task if it understand those data. It is necessary to help computer understand data on the Internet. The paper represents how to build ontology in a field of computer for Vietnamese language. This ontology construction approach is orient object.

Keywords: Ontology for Vietnamese language, OOMP, Ontology

I. INTRODUCTION

Nowadays, Information on the Internet grows rapidly, and it also plays an important role in human life. In addition, the information, which is being shared, developed, and expanded by users, is useful for users and exists mainly in form of text on web pages expressed in natural language. A question is thus how to exploit this information effectively and to help the human being to solve their problems based on the useful information. Therefore, there are many tools which have been built to solve some problems in natural language processing (NLP), Information Extraction, Text Summarization, Information Retrieval, etc.

However, one of the most difficult problems in NLP, Text Summarization, etc. is word – sense disambiguation (WSD). There have been many methods to solve this problem, especially disambiguating a word based on ontology [12]. Words might be disambiguated with respect to computational lexicons and domain ontologies, depending on the meaning they can convey. The word "bank", for instance, has several senses and may refer to "a slope besides a body of water" or to "a bank building". Fortunately, it is also common that the specific sense intended is determined by the textual context of a word. For example, in a geographical ontology, "bank" only means "bank of water [15]. To construct domain ontologies is needed to solve some problems in NLP, Text Summarization, etc.

II. DEFINITIONS

A. Ontology

There have many definitions of ontologies in state – of – the – art. One of the most cited is the one proposed by Gruber [10]: *An ontology is a formal, explicit specification of a shared conceptualization*. Swartout et al. proposed [10]: *An ontology is a hierarchically structured set of terms to describe a domain that can be used as a skeletal foundation for a knowledge base*.

B. *OOMP Ontology*

OOMP is ontology and has hierarchy of semantic concepts based on relationships (Rm, Rp, R_{φ}^{m}, R_{φ}^{p}) [2][3][13]. Figure 1 presents an OOMP ontology example.

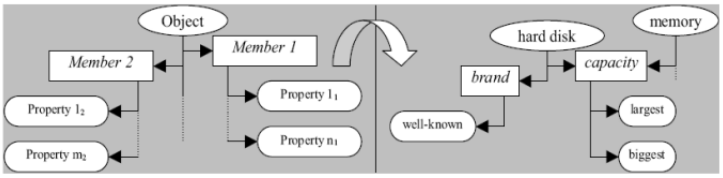

Fig 1. Example of OOMP ontology [2][3][13]

III. A STRATEGY OF DOMAIN ONTOLOGY CONSTRUCTION FOR VIETNAMESE LANGUAGE

This section explains our technique for discovering domain-specific terms and relationships for constructing the ontology. In addition, we present the OOMP ontology organization and how to train it. In this section, we also introduce methods built knowledge in order to train OOMP ontology.

A. *Ontology Structure*

OOMP is domain ontology in the field of computer. This ontology structure was proposed by Thanh.N.C and Tuoi.P.T and was constructed in relational data model [2][3][13]. To support better ontology training in our model, this ontology is improved and presented as follows:

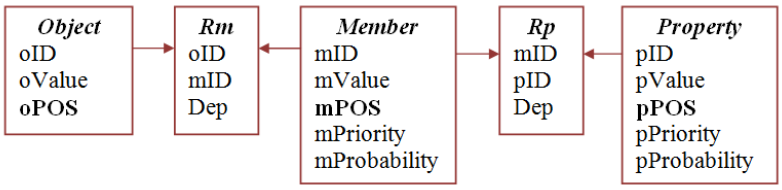

Fig 2: OOMP ontology model

Where each element has some attributes as follows:

- *oID, mID, and pID are the identification of element and primary key.*
- *Value is a content of element.*
- *Probability is used to save the probability $\wp(m_j)$ ($\wp(p_k)$) between Object and Member (between the Member and Property).*
- *Priority presents an element is the more dominant then the other is in a same condition.*
- ***POS** is pos tag of element ex: Np, Nn, An,... This attribute which helps to determine POS tag of element easier is proposed by us.*

B. *Building Patterns of Vietnamese Noun phrase determination*

- Step 1: Based on 27 Vietnamese noun phrase structures which were presented by Chau.N.Q and Tuoi.P.T [1], some of elements in the Vietnamese noun phrase (t1, t2 and s2) are removed because they are not necessary to determine object, member and property. After that, we have patterns as follows:
  - *Pattern 1: T1 T2 s1*
  - *Pattern 2: T1 T2*
  - *Pattern 3: T2 s1*
  - *Pattern 4: T1 s1*
- Step 2: Based on part of speech (POS) of components (T1: Nu|Nn|Ng|Nt; T2: Np|Nc|Na; s1: Aa|An) and the four patterns in step 1, we determine patterns of

Table 1. Pattern of Vietnamese Noun phrase determination

No	Pattern	O	M	P	Example
1	Np Ng Aa	Np	Ng	Aa	<w pos="Ng">bộ xử lý</w> <w pos="Np">Intel Core</w> <w pos="Aa">mới</w>
2	An Np Ng	Np	Ng	An	<w pos="Ng">đĩa cứng</w>

					<w pos="Np">SEAGATE</w> <w pos="An">lớn</w>
…	…	…	…	…	
42	Nc Ng	Ng	Nc		<w pos="Ng">bộ nhớ</w> <w pos="Nc">flash</w>

C. Building Vietnamese WordNet

Relationships (R^m, R^p) in OOMP ontology are very important, because they are used to determine relationships between object, member and property in OOMP ontology training. To determine these relationships, we use relationships of WordNet. For English language, WordNet has been built and developed by Princeton University and latest version is 3.0 [18]. Vietnamese WordNet also has been researched and developed, but it does not support in our OOMP ontology training.

Until now, there are 525 words developed and stored in our Vietnamese WordNet.

D. Ontology Training

This method was adopted and adapted from Thanh.N.C and Tuoi.P.T 's research [2][13]. The three steps are as follows:

- *Step 1: Extract candidate noun phrases (candidate noun phrases are extracted from corpus in Vietnamese language). In this step, the documents will be extracted noun phrases by GATE and Jape built in Chau.NQ and Tuoi.P.T's research [1].*
- *Step 2: Determine the components O, M, and P basing on the pattern (Table 1) from the noun phrase extracted in Step 1.*
- *Step 3: Build relations from O, M, P in Step 2, and then select the best results for ontology.*

IV. APPLICATIONS OF ONTOLOGY

Ontologies can be used to support a great variety of tasks in diverse research areas such as knowledge representation, natural language processing, information retrieval, geographic information systems, etc. [10]

OOMP can be used to add semantics into user's query in query expansion problem for English query [13] and Vietnamese query [14].

V. EXPERIMENT AND EVALUATION

E. Building a corpus for OOMP ontology training

To train OOMP ontology, a corpus in computer field must be built because it currently does not have a standard corpus for Vietnamese language. The corpus is extracted from websites: Network administration 1.571 files – 111 MB [17], PC World Magazine 1.070 files – 80,8 MB [16]. The steps building the corpus are presented as follows:

- *Step 1: Vietnamese documents are normalized to Unicode and saved in text files (txt).*
- *Step 2: Extract Vietnamese noun phrases (NPs) from the normalized documents in step 1 and Select suitable NPs which do not have one of some characters ~ ! @ # $ % ^ & * ( ) + = - [ ] { } \ / : ; ? > < . ,. A result is presented in the following table:*

Table 2. Number of unsuitable and suitable NPs

Length of NP (1)	Number of extracted NPs (2)	Number of un-suitable NPs (3)	Number of suitable NPs (4)	Ratio (%) between (4) and (2)
1	199.259	8.685	190.574	95,7
2	223.744	9.141	214.603	95,9
3	40.223	3.680	36.543	90,6
4	2.261	142	2.119	93,7
5	31	0	31	100
Total	**465.518**	**21.648**	**443.870**	

- *Step 3: Remove some components which are not necessary are in suitable NPs. The results of this step are stored in file text and used to train OOMP ontology. The results are showed as belows:*

Table 3. Results of Number of suitable NPs

Length of NP	Number of suitable NPs	Number of suitable NPs after removed some components
1	190.574	190.574
2	214.603	246.311
3	36.543	6.985
4	2.119	0
5	31	0
Total	**443.870**	**443.870**

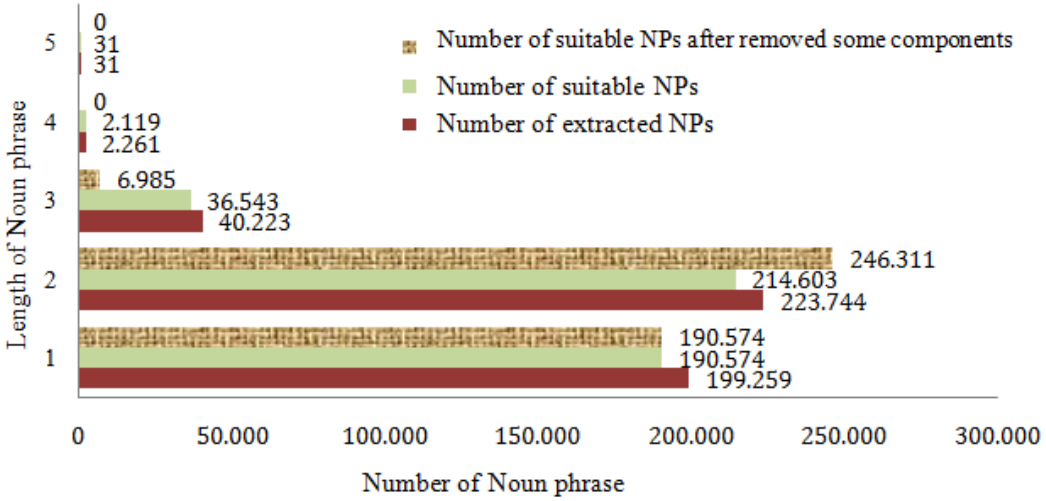

Fig 3: The results of extracted NPs

F. Experiment of ontology training

The Vietnamese noun phrases saving in the corpus are input data for the training of ontology. Criteria for selection:

- *Object, Member: positive frequency; field of computer; correct POS.*
- *Property: positive frequency and correct POS.*
- *Rm, Rp: Dep is positive; oID, mID, pID of Rm, Rp must exist in table corresponding.*

Training results are presented in the following table and figure:

Table 4. Extracted components of ontology from the corpus

No	Type	Number of candidates	Number of chosen candidates	Ratio (%)
1	Object	7.669	1.058	*13,8*
2	*Member*	6.228	2.084	*33.5*
3	*Property*	82	44	*53.6*
4	*Relations Rm*	84.525	4.325	*5.1*
5	*Relations Rp*	3292	1.381	*42*

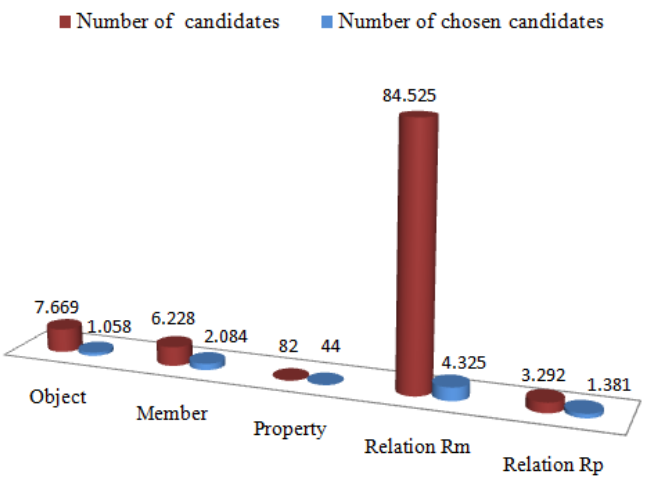

Fig 4: The result of ontology training

VI. CONCLUSION

In the paper, we propose how to build domain ontology for Vietnamese. To train ontology, patterns of Vietnamese noun phrase determination and Vietnamese WordNet in computer field were built. The results of Vietnamese noun phrase determination and the Vietnamese WordNet are 42 patterns and 525 words, respectively. Although ratio of components in OOMP is not high in experimental result, it supported to solve semantics of users' query. Ratio is not high in the experimental result because the experimental corpus which is built by authors is static and number of documents is not still large.

REFERENCES

Vietnamese:

[1] Chau, N.Q and Tuoi, P.T. 2008. Vietnamese Key Phrase Recoginition. Journal on ICT of Ministry Information and Communications, Pp 64 – 73, No 19.

[2] Thanh, N.C and Tuoi, P.T. 2008. Information Retrieval: An Ontology Solution for Query Complete. Journal on ICT of Ministry Information and Communications, Pp 84 – 92, No 19.

[3] Thanh, N.C. 2010. Query Expansion Model Construction in Text Information Retrieval. Doctoral Dissertation, HCMC University of Technology.

English:

[4] Andreou, A. 2005. Ontologies and Query Expansion, Master Thesis, School of Informatics University of Edinburgh.

[5] Fensel, D. 2000. The Semantic Web and its Languages, IEEE Intelligent Systems, Vol 15, No. 6, November/December, pp. 67-73.

[6] Fu, G., Jones, C.B., Abdelmoty, A.I. 2005. Ontotology-based Spatial Query Expansion in Information Retrieval. International Conference on Ontologies, Databases and Applications of Semantics, Agia Napa, Cyprus.

[7] Chen, G. 2010. Ontology – based Query Expansion in biomedicine domain. Master Thesis, Roskilde University.

[8] Imran, H., Sharan, A. 2009. Thesaurus and Query Expansion, International Journal of Computer Science and Information Technology (IJCSIT), Vol 1, No 2, Pp.89 – 97.

[9] Nagvili, R. and Verladi, P. 2003. An Analysis of Ontology based Query Expansion Strategies. 14th European Conference on Machine Learning (ECML 2003), September 22-26.

[10] Nieto M. A. M. 2003. An Overview of Ontologies, Center of Research in Information and Automation Technologies Technical Report, Puebla, Mexico.

[11] Revuri, S., Upadhyay, R.S., and Kumar, P.S. 2006. Using Domain Ontologies for Efficient Information Retrieval. Proceedings of the 13th International Conference.

[12] Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. ACM Comput. Surv. 41, 2, Article 10 (February 2009), 69 pages.

[13] Thanh, N.C, Tuoi, P.T. 2007. An Ontology - based approach of Query Expansion. Proceedings of the 9th iiWAS2007, Jakarta, Indonesia.
[14] Tuan,T.A, Tuoi,P.T., and Thanh,N.C. (2012). Query Expansion based on Ontology for Vietnamese Query, In Proceedings of the 14thInternational Conference on Information Integration and Web-based Applications & Services (iiWAS2012), ACM December 3-5, 2012, Bali, Indonesia, p 332-335.
[15] Xia Wang, Vassilios Peristeras (2008). oWSD: A Tool for Word Sense Disambiguation, National University of Ireland, Galway, Digital Enterpise Research Institute, Galway, Ireland.

Web page:

[16] Vietnamese Pcworld, http://www.pcworld.com.vn/articles/san-pham/.
[17] Network Administrator, http://www.quantrimang.com.vn/phancung/index.aspx
[18] WordNet 3.0, http://wnsqlbuilder.sourceforge.net.
[19] WordNet, http://wordnet.princeton.edu/wordnet/.

