A Question Answering Model Based Evaluation for OVL (Ontology for Vietnamese Language)

Dang Tuan Nguyen and Tri Phi-Minh Nguyen

***Abstract*—In this paper, we focus on using a Vietnamese Question Answering system to evaluate the OVL (Ontology for Vietnamese Language). The OVL has been evaluated using its effects as real component in question answering systems. This evaluation method helps us estimate the competence of OVL through following answered questions: which reasoning mechanism can be built up on OVL's structure and knowledge domain, which types of question can be answered from the reasoning mechanism, how to evaluate the experimental results.**

***Index Terms*—Ontology, Ontology Evaluation, Question Answering.**

I. Introduction

In the aim of contributing to Vietnamese language processing, we are now developing OVL (Ontology for Vietnamese Language) in a research project of the Natural Language and Knowledge Engineering Group at the University of Information Technology (VNU-HCM). The OVL's structure and its related details have been mentioned in [7]. OVL is now published as the first and open ontology [13] with the domain in various fields such as news, society, economy, culture, sport and life…

In recent paper [8], we have taken advantage of Pellet [11] reasoning services [14] to check its correctness. In the paper [8], a simple reasoning method is also developed for some WH-question such as *what, where, when, who* to OVL ontology. At that time, Pellet reasoned is also used to execute SPARQL query and ensure OVL was consistency. After applying reasoning strategies on OVL, our systems are now supporting more complex types of question such as *why, how.*

In this paper, we focus on using a Vietnamese Question Answering (QA) system to evaluate the OVL. In such systems, reasoning mechanisms based on the typical structure of OVL are well implemented to find answer for common type of questions.

II. Building a Vietnamese Question Answering System to Valuate OVL

After checking the correctness of the OVL using Pellet reasoning services, a QA system is implemented in order to estimate the OVL's answer competence.

Some related works are introduced in [1], [2], [3], [4], [5], [6]. In these publications, a Vietnamese ontology in e-books library domain has been exploited to develop a searching tool based on QA model. This searching tool is built by using some natural language processing techniques. In this paper, we do not use the syntactic and semantic processing techniques like the techniques mentioned above.

The QA system to evaluate OVL executed as follows:

1) Determining keywords from user's query. This step is supported by pre-defined templates and a Vietnamese WordNet (WNVL: WordNet for Vietnamese Language) [9].
2) Generating SPARQL query from above keywords.

In general, there are two groups of questions are performed by our system:

Group 1: This group consists of question types such as *who, what, where, when.* In this group, the users can simply retrieve related information from OVL.

Group 2: This contains more complex types of question which require reasoning mechanisms from ontology such as *why, how* questions.

The method to answer questions in Group 1 (*who, what, where, when*) is basically based on determining keywords from Vietnamese questions. Finally, these keywords are passed to SPARQL query to retrieve triples from ontology. In OVL, a triple includes three items: *subject, predicate, object.* Therefore, to answer a question in the first group, it is necessary to find the most suitable item when receiving two remaining items in advance.

These specific steps to answer *what*, *where*, *when*, *who* questions are described in the following steps [8]:

Step 1: This is aimed at determining type of received question from user. Clarifying in question types basically depends on the appearance of some query words which are distinct for each type of question.

Step 2: Getting rid of meaningless words, extracting keyword which plays a role as a *subject* in a *triple* from query question.

Step 3: This is to extract keywords that are noticed as *predicate* item (representing for relationship between *subject* and *object* in *triple*). *Subjects* from previous step will be removed, the remaining items are considered as *predicate.* It is necessary to develop a procedure to list *predicate*s which are in OVL matching with *predicate* from inputting question (in OVL, *predicate* is also called as *object property* or *datatype property*). This step follows:

Step 3.1: From OVL data, collecting *predicates* which match with the *predicate* from query. These *predicates* will be stored in set *matchedPredicates*={}.

Step 3.2: Sorting elements in *matchedPredicates* with descending order in string's length.

Manuscript received July 24, 2010; revised February 7, 2011.

Dang Tuan Nguyen, Tri Phi-Minh Nguyen, Faculty of Computer Science University of Information Technology Vietnam National University – Ho Chi Minh City(email: dangnt@uit.edu.vn, trinpm@gmail.com).

Step 4: SPARQL's query will be initialized with *keyword* from Step 2. The result of this query is a set of triples T (including *subject, predicate, object*) in which *subject* contains *keyword* from Step 2.

Step 5: This step is to traverse *matchedPredicates* set. A *subject* from Step 2 and each of appropriate elements in *matchedPredicates* form a pair. This pair is compared with each pair (*subject, predicate*) in T to retrieve corresponding *objects.* If the result is absolutely compared (two strings are totally similar), it is stored in *highPriorityAns*={}. On the other hand, if the result is relatively compared (two strings are nearly similar), it is stored in *lowPriorityAns*={}. *Objects* from *highPriorityAns* set are usually the exact answers to inputting question, and the answers from *lowPriorityAns set* are usually marked as responds for reference.

In Group 2, spine technique to find answers is based on reasoning on the default components of Ontology Web Language (OWL) [12]. The result of this reasoning will be the *r* relationship between two individuals in OVL. Obviously, this *r* relationship has not yet defined in the OVL.

In general, there are three main reasoning mechanisms to inspect the *r* relationship to give answers for second group question (*why, how*). Whether, it is based on the characteristics of property in OWL (functional property or FP; inverse functional property or IFP; transitive property, TP; and symmetric property, SP) as shown in Figure 1; or relied on the domain and range of property; or finally, based on the definition of a class in ontology.

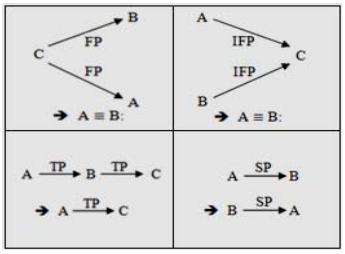

Figure 1: Characteristics of property and their functionality in OWL language [12].

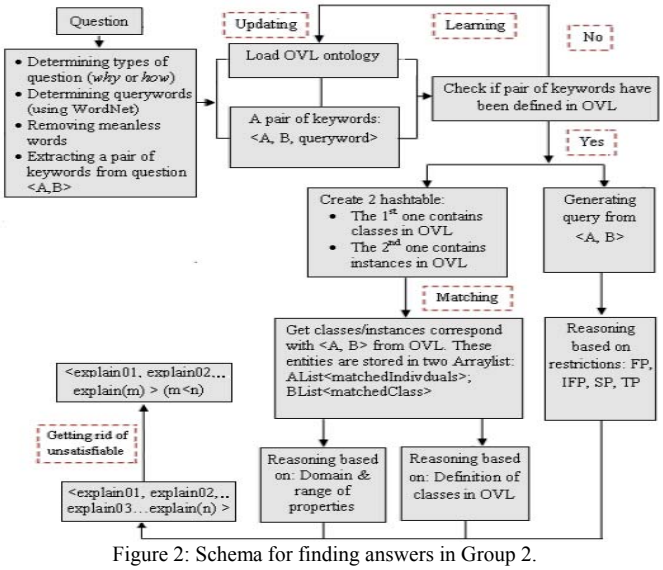

Figure 2: Schema for finding answers in Group 2.

Schema, as illustrated in Figure 2, has been used to find answers with a *why* or *how* question on OVL. There are four main steps in such a scheme:

Step 1: To determine what type of question from user's question *q*. The system will remove meaningless words and find query word from question *q*. These procedures will be supported from WNVL's data synsets (WNVL [9]). At the end of this step, the inputting question from user is separated into three main items: *kword_a, kword_b, queryword_c*. From defined question templates and *queryword_c*, the system will notice whether the inputting question can be answered or not.

Step 2: To check whether a pair of <*kword_a, kword_b*> exists in ontology OVL or not. If not, the learning mechanism allows user to add data such as defining entities, classes, related relationships from question *q* into the ontology. This method will enhance the answer competence of system in time. The necessary condition to continue next steps is that the pair <*kwords_a, kwords_b*> must be defined in OVL. If this condition is satisfied, the auto-generated SPARQL will be constructed including the above <*kwords_a, kwords_b*>. The result from this query is stored in data table *t*.

Step 3: In this step, reasoning strategies are applied. This technique is based on default components of OWL language standard to find answer from question query *q*. The QA systems will traverse all of strategies until they find appropriate answers. Mechanisms of the reasoning include:

1) Reasoning mechanism relied on characteristics of property: based on the above data table, the system applies traits of property between individuals to seek for suitable answer. The next two strategies aim at inspecting "is-a" relationship between an individual and particular class in OVL. They will operate as following: verifying classes, individuals in OVL which are corresponding to *kword_a* and *kword_b (matching technique)*.
2) Reasoning mechanism based on the domain and range of properties: searching properties in OVL which relate to class name extracted from *q,* and putting them into the set *matchedProp*={}*;* and inspecting the set *matchedProp* in order to generate SPARQL query which includes individual name from query question *q* and element in set *matchedProp.* The answer will be given by the system if the question returns the number of positive row.
3) Reasoning mechanism referred to the definition of a class: a SPARQL query will be formed by combining properties from the definition of a class with individuals extracted from query question *q* respectively. If the result from query matches with definition of class, individuals and classes which are extracted from question *q* will be noticed as existing "is-a" relationship. In this case, the explanation of the result is recorded from the above SPARQL query.

Step 4: This is to remove unsatisfiable answers to sort the most suitable ones. This step checks the ratio of the similarity between pairs of keywords from question and answers. For each of reasoning strategy, the system will have its own different mechanism to check each of the ratios of the similarity. For example, in case of reasoning based on functional property: The *object* from the first triple and the *object* from the second one have to contain *kword_a* or *kword_b*.

III. Developments

TABLE 1. Experiment Results for Questions in First Group.

Form	Question	HPA	LPA	Sum	Time (ms)
What	Tỉnh Nghệ An có website chính thức là gì? *(What is official website of Nghe An province?)*	1	43	44	1666
	Nghệ An có biển số xe các loại là gì? *(What are car's registration numbers in Nghe An province?)*	4	21	25	3884
	Thành phố Buôn Ma Thuột có quy mô dân số nội đô là bao nhiêu? *(What is population scale in Buon Ma Thuoc city?)*	1	5	6	1219
	Đà Nẵng có mã điện thoại là gì? *(What is telephone code in Da Nang city?)*	1	33	34	3030
	Thủ đô Hà Nội có mã bưu chính là gì? *(What is Ha Noi's postal code?)*	4	104	108	1654
	Huế có tỉ lệ lao động phi nông nghiệp là bao nhiêu? *(What is labour ration in Hue province?)*	1	35	36	1329
	Thành phố Đồng Hới Thuộc Việt Nam là thành phố nào? *(What is Dong Hoi in Vietnam province?)*	0	2	2	9675
	Thành phố Huế có mật độ dân số là bao nhiêu? *(What is population density in Hue city?)*	4	16	20	2802
Where	Mã điện thoại 710 ở đâu? *(Where is telephone code 710?)*	0	1	1	2446
	Quận Tân Bình ở đâu? *(Where is Tan Binh district?)*	0	1	1	2329
	Mã bưu chính 59 ở đâu? *(Where is postal code 59?)*	0	1	1	1876
	Biển số 99 ở đâu? *(Where is car registration number 99?)*	0	1	1	1550
When	Việt Nam gia nhập WTO năm nào? *(When did Vietnam join WTO?)*	1	62	63	1852
	Hà Nội giải phóng năm nào? *(When was Hanoi emancipated?)*	1	59	60	6240
	Chủ tịch Hồ Chí Minh sinh năm nào? *(When was Ho Chi Minh predident born?)*	0	1	1	1869
	Ngân hàng Đông Á có ngày thành lập là? (When was Dong A bank established?)	1	5	6	1639
Who	Cao Đức Phát là người nào? *(Who is Cao Duc Phat?)*	0	2	2	9058
	Ca khúc "Một cõi đi về" được sáng tác bởi ai? *(Who is "Một cõi đi về" song composed by?)*	1	0	1	9972

In our QA systems, a list of about 200 questions which are formed from 28 pre-defined templates has been tested. The content of these questions basically relates to the defined knowledge in OVL and inputs randomly to systems. Processing procedures such as detecting types of question, extracting words and key words, etc. will be logged along with the final result.

Some tested questions in first group's question have listed in Table 1, its columns are described as following:

1) Form: Types of questions,
2) Question: Vietnamese questions,
3) High Priority Answer (HPA): Sum of the most suitable answers for input question,
4) Low Priority Answer (LPA): Sum of related answers for input question, which is used as references to question,
5) Sum: total retrieved answers from a question,
6) Time: time to get all answers from inputting question, in milliseconds.

The procedure to detect whether an answer will be classified into HPA set or LPA set is done automatically by systems through checking the similarity between the question and answers. The most appropriate answers corresponding to question are often stored in High Priority Answer set. In case that the system does not find out any answer in this set, the reason is that the *predicate* item from question does not match to any *predicate* from OVL's triples, so the user must track to find out the answer from LPA set. For example, with the question "*Quận Tân Bình ở đâu?",* there is no *predicate* such as "*ở*" or "*ở đâu*" in OVL ontology, but *predicates* such as "*thuộc*" and "*trực thuộc*" accompany with keyword "*Quận Tân Bình*". Therefore sum of answers in HPA is 0 while LPA's is 2 (Table 2). In this situation, although answers in LPA have lower priority but they are still acceptable for this inputting question.

Table 2 includes templates for questions in second group along with their corresponding reasoning mechanism.

Items such as <A>, <B> in templates are entities which are input from user. According to our observations from experimental results, the anwers from systems will be the most appropriate if the similarity between item <A>, item <B> and individuals in OVL are highest.

With questions in the second group's templates, our systems will traverse all implemented reasoning strategies until it can return suitable answers. In some cases, if the systems cannot solve the answer because of missing entities's definition, its learning mechanism will advocate to add data and its description to OVL.

TABLE 2. Second Group's Question Templates for Vietnamese QA Systems

No.	Template	Reasoning mechanism	Number of tested questions
001	<Tại sao> ngân hàng A và ngân hàng B có phương thức hoạt động tương tự nhau? (*Why do bank A and bank B have the system similar to each other?)*	Characteristic of property (Functional Property)	9

002	<Tại sao> nhân vật/người A và nhân vật/người B là cùng một người? *(Why do person A and person B indicate the same person?)*	Characteristic of property (Functional Property)	9
003	<Tại sao> địa danh A và địa danh B là cùng một nơi ở Việt Nam? *(Why do place A and place B indicate the same place in Vietnam?)*	Characteristic of property (Inverse Functional Property)	7
004	<Tại sao> công ty A và công ty B có chung ngành hàng kinh doanh? *(Why do company A and company B have the same enterprise?)*	Characteristic of property (Inverse Functional Property)	13
005	<Tại sao> doanh nghiệp A và doanh nghiệp B có nguồn lực tài chính tương đương nhau? *Why do company A and company B have the financial competence equivalently?*	Characteristic of property (Inverse Functional Property)	9
006	<Tại sao> cơ quan A và cơ quan B là hai cơ quan ngang cấp nhau ? *(Why is agency A equivalent to agency B?)*	Characteristic of property (Inverse Functional Property)	8
007	<Tại sao> môn thể thao A và môn thể thao B có phương pháp tính điểm tương tự nhau? *(Why do sport subject A and sport subject B have the same score methodology?)*	Characteristic of property (Inverse Functional Property)	8
008	<Tại sao> đơn vị/ cơ quan A thuộc quản lý của đơn vị/ cơ quan B? *(Why does agency A belong to agency B?)*	Characteristic of property (Transitive Property)	7
009	<Tại sao> thiết bị A là thành viên/bộ phận của lớp B? *(Why does instance A belong to class B?)*	Domain and Range of Property	7
010	<Tại sao> khách sạn A được xếp vào khách sạn n sao tại Việt Nam? *(Why is hotel A classified into n star hotel in Vietnam?)*	Class Definition	10
011	Chất lượng dịch vụ của công ty A <như thế nào>? *(How is the service quality of company A?)*	Class Definition	10
012	A và B có quan hệ như thế nào với nhau tại việt nam? *(How is the relationship between object A and object B in Vietnam?)*	Characteristic of property (Transitive Property)	7
013	Quan hệ giữa cá thể A và lớp B là <như thế nào>? *(How is the membership between instance A and class B?)*	Class Definition (Class Definition)	7

IV. Conclusions

The OVL has been evaluated using its effects as real component in question answering systems. This evaluation method helps us estimate the competence of OVL through following answered questions: which reasoning mechanism can be built up on OVL's structure and knowledge domain, which types of question can be answered from these reasoning mechanism, how to evaluate the experimental results.

TABLE 3. Some Experimented Questions from Second Group's Questions

No.	Template	Reasoning mechanism	Evaluated answers (by user)
001	Tại sao ngân hàng ANZ Việt Nam và ngân hàng HSBC có phương thức hoạt động tương tự nhau? *(Why do ANZ bank and HSBC bank have the system similar to each other?)*	FP	✓
002	Tại sao Nguyễn Ái Quốc và Hồ Chí Minh là cùng một người? *(Why do Nguyen Ai Quoc and Ho Chi Minh indicate the same person?)*	FP	✓
003	Tại sao Bến Nghé và Gia Định là cùng một địa danh ở việt nam? *(Why do Ben Nghe and Gia Dinh indicate the same place in Vietnam?)*	IFP	✓
004	Tại sao công ty AIGI và công ty Duy Tân có chung ngành hàng kinh doanh? *(Why do company A and company B have the same enterprise?)*	FP/IFP/TP/SP	
005	Tại sao doanh nghiệp Việt Nhật và tập đoàn FPTW có nguồn lực tài chính tương đương nhau ? *(Why do company A and company B have the financial competence equivalently?)*	FP/IFP/TP/SP	
006	Tại sao Khách sạn Caravelle được xếp vào khách sạn 5 sao tại Việt Nam?) *(Why is Caravelle hotel classified into five star hotel in Vietnam?)*	Domain and Range of Property	✓
007	Tại sao bộ môn Bóng Bàn và bộ môn Cầu Lông có phương pháp tính điểm tương tự nhau? *(Why do sport subject A and sport subject B have the same score methodology?)*	IFP	✓
008	Chất lượng dịch vụ của công ty Vietravel như thế nào? *(How is the service quality of Vietravel?)*	Class definition	✓
009	Tại sao công ty PVEP Algeria trực thuộc Tập Đoàn Dầu Khí Quốc Gia? *(Why does PVEP Algeria agency belong to Petro Vietnam Corp.?)*	TP	✓
010	Quan hệ giữa cá thể Vỏ 2759 và lớp Phụ Tùng Xe Máy là như thế nào? *(How is the relationship between tire 2759 and class motobike equipment?)*	Domain and Range of Property	

Two different techniques to answer two groups of questions are described in details. They both take advantage of Vietnamese Wordnet (WNVL [9]) and defined templates to detecting types of question, extracting key words and useful query words from user. The first group of questions mainly queries information from OVL, while with the second group, some different reasoning strategies are applied to find

the answers. The OVL's tester will input question based on templates and keep track of return answers. Besides, the procedure to check the correctness of OVL through Pellet reasoned is mandatory before using QA systems and exploiting OVL's application in the future.

According to our experiments: questions in Group 1 (*what, who, where, when*) will be solved if the user provided two out of three items of corresponding triples; question in Group 2 (*why, how*) required related entities have to defined in details along with restrictions, relationships.

The next important task is to apply Vietnamese grammar analyzing methods and semantic processing techniques - along with increase data for OVL up to many times, to enhance more complex Vietnamese QA systems which can support complicated types of queries such as sub-query, negative query, etc. At that time, our QA system will evaluate and exploit the OVL ontology easily and more efficiently.

REFERENCES

[1] Dang Tuan Nguyen, Ha Quy-Tinh Luong, "A framework for building Vietnamese language query processing components in e-library searching systems", International Conference on Information Technology (ICIT 2009), Venice, Italy, October 28-30, 2009. Proceedings of World Academy of Science, Engineering and Technology, pp. 1354-1358, Vol. 58, October 2009. ISBN: 2070-3724.

[2] Dang Tuan Nguyen, Ha Quy-Tinh Luong, Tuyen Thi-Thanh Do, "Building a Vietnamese language query processing framework for e-library searching systems", International Journal of Computer Science and Information Security (IJCSIS), pp. 092-096, Vol. 6, No. 1, October 2009. ISSN: 1947-5500.

[3] Dang Tuan Nguyen, Ha Quy-Tinh Luong, "Document searching system based on natural language query processing for Vietnam Open Courseware library", International Journal of Computer Science Issues (IJCSI), pp. 7-13, Vol. 6, No. 2, November 2009. ISSN (online): 1694-0784, ISSN (print): 1694-0814.

[4] Dang Tuan Nguyen, Tuan Ngoc Pham, Quoc Tan Phan, "A Semantic Model for Building the Vietnamese Language Query Processing Framework in e-Library Searching Application", Proceedings of the 2nd International Conference on Machine Learning and Computing (ICMLC 2010), pp. 179-183, February 9-11, 2010, Bangalore, India. ISBN: 978-0-7695-3977-5. Editions IEEE.

[5] Dang Tuan Nguyen, Huy Vu Nguyen, Quoc Tan Phan, "Using the Vietnamese Language Query Processing Framework to Build a Courseware Searching System", The 2010 International Conference on Telecom Technology and Applications (ICTTA 2010), Proceedings of the 2010 Second International Conference on Computer Engineering and Applications, Vol. 2, pp. 117-121, March 19 - 21, 2010, Bali Island, Indonesia. ISBN: 978-0-7695-3982-9. Editions IEEE.

[6] Dang Tuan Nguyen, Ha Quy-Tinh Luong, Quoc Tan Phan, "An e-Library Searching System Based on the Vietnamese Language Query Processing Framework", The 2010 International Conference on Information and Emerging Technologies (ICIET 2010), Proceedings of the 2010 Second International Conference on Computer Engineering and Applications, Vol. 2, pp.517-520, Bali Island, Indonesia, March 19 - 21, 2010. ISBN: 978-0-7695-3982-9. Editions IEEE.

[7] Dang Tuan Nguyen, Tri Phi-Minh Nguyen, "Building and Evaluation of a Universal Ontology for Vietnamese Language", Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering 2010 (ICACTE 2010), Vol. 6, pp. 227-230, August 20-22, 2010, Chengdu, China. Editions IEEE.

[8] Dang Tuan Nguyen, Tri Phi-Minh Nguyen, "Pellet Reasoner Based Evaluation of OVL (Ontology for Vietnamese Language)", Proceedings of the 2010 2nd International Conference on Software Technology and Engineering (ICSTE 2010), Vol. 2, pp. 180-184, October 3-5, 2010, San Juan, Puerto Rico, USA. Editions IEEE.

[9] Dang Tuan Nguyen, Tu Ngoc Nguyen, "Building a WordNet for Vietnamese Language", Proceedings of the 2010 International Conference on Intelligence and Information Technology (ICIIT 2010), Vol. 1, pp. 605-609, October 28-30, 2010, Lahore, Pakistan. Editions IEEE.

[10] Nguyen Phi Minh Tri, Nguyen Ngoc Tu, "Xây dựng công cụ truy vấn thông tin hướng tri thức-ngữ nghĩa trên ontology tiếng Việt", B.Sc. Thesis in Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2010.

[11] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, Yarden Katz, "Pellet: A practical OWL-DL reasoner", Web Semantics: Science, Services and Agents on the World Wide Web, v.5 n.2, p.51-53, June, 2007.

[12] http://www.w3.org/TR/owl-features/

[13] http://ovl-open.sourceforge.net

[14] http://clarkparsia.com/pellet