A Model of Vietnamese Person Named Entity Question Answering System

Mai-Vu Tran, Duc-Trong Le
KTLab, University of Engineering and Technology – Vietnam National University, Hanoi
{vutm, trongld}@vnu.edu.vn

Xuan- Tu Tran, Tien-Tung Nguyen
KTLab - University of Engineering and Technology – Vietnam National University, Hanoi
{tutx_52, tungnt_5}@vnu.edu.vn

Abstract

In this paper, we proposed a Vietnamese named entity question answering (QA) model. This model applies an analytical question method using CRF machine learning algorithm combined with two automatic answering strategies: indexed sentences database-based and Google search engine-based. We gathered a Vietnamese question dataset containing about 2000 popular "Who, Whom, Whose" questions to evaluate our question chunking method and QA model. According to experiments, question chunking phase acquired the average F1 score of 92.99%. Equally significant, in our QA evaluation, experimental results illustrated that our approaches were completely reasonable and realistic with 74.63% precision and 87.9% ability to give the answers.

Keywords: Vietnamese question, QA, VPQA, question analysis, answer extraction, question parser

1 Introduction

Numerous researches about Question Answering (QA) systems have been discussed in recent years. Initially, they only answered simple questions; however, currently researches have been focused on methods for more complex questions. Those methods analyze and parse complex questions to various simple questions before using existed techniques to respond. [1]

Automatic question answering – the ability of computers to answer simple or complex questions, posed in ordinary human language – is the most exciting. Building the question answering system is a difficult issue in terms of natural language processing tasks. Presently, automatic question answering systems are revolutionizing the processing of textual information. By coordinating complex natural language processing techniques, sophisticated linguistic representations and advanced machine learning methods, automatic question answering systems can detect exact responses from a wide variety of natural language questions in unstructured texts.

Recent researches demonstrated that the increasing in performance of systems is dependent on the number of probable answers in documents. The exact answer detection is one of the most significant problems in QA systems. For this purpose, our model utilized CRF [5] machine learning algorithm to parse natural questions and some IR strategies to extract answers. The model works on closed domain by extracting human names based on knowledge warehouse and search engines. If answers are not found in database, the question will push into Google search engine. The QA system just supports questions (such as "Who?", "Whom?", "Whose?") in factoid form or one sentence.

The aim of this paper is to design and implement a new classification model, reformulation and answer validation in a QA system. The methodology in our system is to discover correct answer in person domain with NLP techniques, CRF model to parse question, and some strategies to extract answer: knowledge-based, search engine-based and hybrid method. The primary reason of an answer validation component in the system concerns the difficulty of picking up from a document the "exact answer".

Our approach relies on investigating a statistical machine learning method to parse natural question and extract answer candidates by mining the documents or a domain text corpus for their co-occurrence tendency [2]. In the initial phase, questions are parsed by using CRF model. Subsequently, query patterns based on their types

26th Pacific Asia Conference on Language,Information and Computation pages 325–332

are clarified before the search engine detect candidate answer documents and send them to answer processing module to extract correct answers. The system filters candidate answers collection based on their similarities with question and assigns a priority number to the candidate answers. Finally, the system ranks the answers and sends to user for final validation in order to extract the exact answer. Our system modeled in person domain however it could be expanded to open domains in QA systems.

2 Related work

Question answering researches wereclassified by diverse competitive evaluations which are conducted by the question answering track of the Text Retrieval Conference[1], an annual event sponsored by the U.S. National Institute of Standards and Technology (NIST). Starting in 1999, the TREC question answering evaluation initially focused on factoid (or fact-recall) questions, which could be answered by extracting phrase length passages. Some of the TREC systems achieved a remarkable accuracy: the best factoid QA systems can now answer over 70% of arbitrary, open domain factoid questions.

In Webclopedia [6], with each question type, the system provides a set of pattern questions and answers. The system has to determine the type of question based on the similarities between the input question and each of the question patterns. Then the corresponding pattern will be used to find passages containing the answer. Finally, the answer is extracted from the found passages.

The True Knowledge Answer Engine[2] attempts to comprehend a given question by disambiguation from all possible meanings of the words in the question to find the most likely one.It discoverson its database of knowledge of discrete facts. As these facts are stored in a form that a computer can understand, the answering engine attempts to produce an answer according to its comprehended meaning of the input question [8].

Wolfram Alpha[3]is an answering engine developed by Wolfram Research. It is an online service that answers factual questions directly by computing the answer from structured data, rather than providing a list of documents or web pages that might contain the answer as a search engine does, Knowledge Base [9].

In Vietnamese text experiments, Vu M.T, et al [7] proposed a model of question answering system which is based on semantic relation extraction. It is a combination of two methods: snowball of Agichtein, Gravano and the search engine of Ravichandran, Hovy to extract semantic relation patterns from the Vietnamese texts. The experimental system achieves positive results on the domain of tourism and also shows the correctness of the model. However, the statistic relation impacts on the system precision and executed time is depended on network speed.

Nguyen Q.D, et al proposed an ontology-based Vietnamese question answering system that allows users to express their questions in natural language [4]. It includes two components: a natural language question analysis engine and an answer retrieval module. They built a set of relations in the ontology which includes only two person relations. According the system's experimental results are relatively high, the cost for building the database is high, and sometimes the extracted relations cannot cover the data domain.

From these systems, this paper introducesa model of person named entity question answering system in Vietnamese domain with machine learning CRF-based method in question analysis phase; sentences data collection-based and search engine-based strategies in answer extraction phase.

3 System architecture

VPQA model consist of three fundamental modules. The first module (1) focuses on Vietnamese natural language question analysis by CRF. The result set of tagged component in the 3rd step is used in the recommendation sub-module (2). It offers user answers and question patterns by Lucene searching from QA Log Database. Additionally, it is also utilized for the question expansion step and expands queries which are the output for next module.

[1]http://www.trec.nist.gov
[2]http://www.trueknowledge.com
[3]http//www.woframalpha.com/

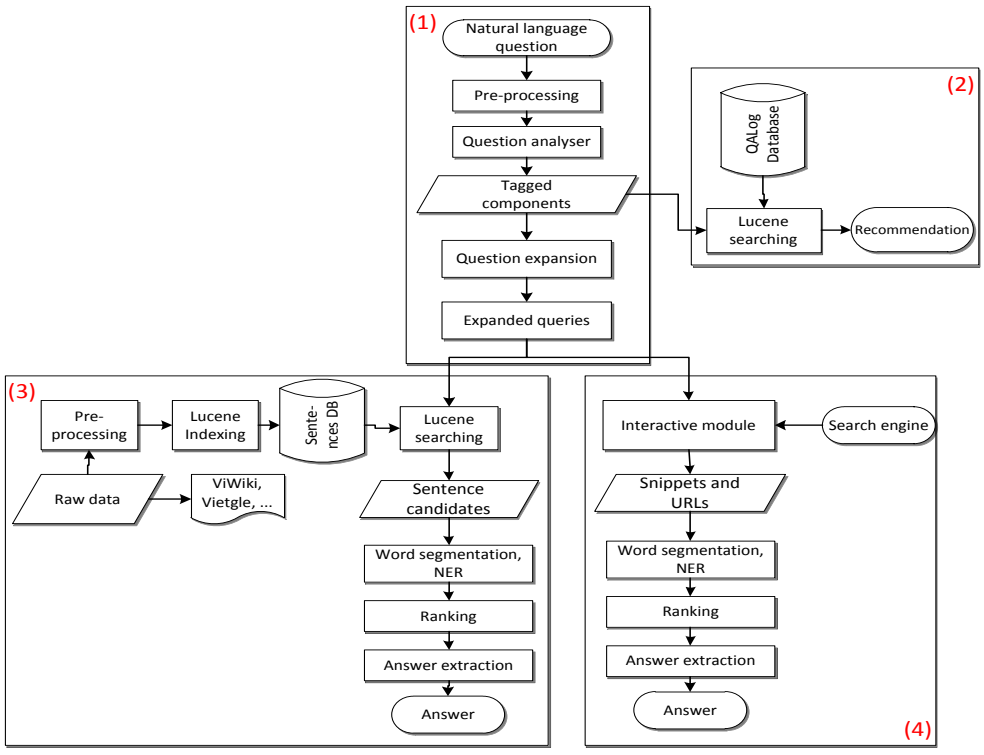

Figure 1: VPQA Model

According to those results, the second module (3) looks its candidates in Lucene[1] indexed sentences' database before determining answer for user by conducting some steps such as: Word Segmentation, NER, Ranking and Answer extraction. Instead of looking in Lucene Database, the last module extracts the set of candidates from snippets returned from Google. The next steps are similar with the 2nd module.

3.1 Question analysis module

3.1.1 "Who, Whom, Whose" question in Vietnamese

Vietnamese linguists have classified Vietnamese sentences by alternative criteria or syntax structure. By Vietnamese "Who, Whom, Whose" questions properties and their mean, they are classified in some forms with four types of component such as: Subject/agent, Verb/action, Object/theme, and Indirect_Ojbect/Co_themyge[6].Commonly, a simple question relate to two forms: two classes of object and three classes of object. Example:

- Relating two classes of objects:
  - ✓ Subject/agent + Verb/action + Object/theme
  - ✓ Object/Theme + Subject/agent + Verb/action
  - ✓ Object/Theme + Verb/action + Subject/agent

Example 1: The question "Who was the Harry Potter book written by?" is same as the Vietnamese question "Cuốn sách Harry Potter được viết bởi ai?"

Above examples have two classes: Tácgiả/Author and Sách/Book

- Relating three classes of objects:
  - ✓ Object/Theme: Indirect_Object/Co_theme+ Verb/action + Subject/agent

Example 2: The Vietnamese question "Ai là tác giả của cuốn Harry Potter xuất bản năm 2004?" is same meaning with "Who is author of the Harry Potter book published in 2004?" include 3 classes: Tác giả/Author, Sách/Book, Năm/Year

[1]http://lucene.apache.org

Label	Meaning	Type of component
WH	Question type	
D_Attr	Feature of job, position	Subject/Agent
D_Time	Feature of time	Idirect_Object/Co_theme
D_Loc	Feature of location	
A_W	Adjective phrase	Verb/Action
V_W	Verb phrase	
N_W	Noun phrase	
Obj	Object	Object/Theme
O	Others	

Table 1: Proposed features and labels

Feature	Meaning	Sign	Example
Lexicon	The existence in Vietnamese dictionary	meaning:0, meaning:±1, meaning:±2	**meaning:-**1:là **meaning:0**:tác+giả
POS tag	Part of speech	pos:N, pos:V, pos:adj, etc.	**meaning:0:**tác+giả **pos:N**
Letter character	Length, capital letter	char:length:n, cap:k:i, cap:k:a	**char:length:11 cap:0:i**
Prefix	The existence of previous word in prefix dictionary	per:prefix	**per:prefix:-2**
Dictionary	Name, location, organization, job dictionary	Per:job, org:i:, etc	**org:0:FPT per:job:-2**

Table 2: Features used in VPQA system

3.1.2 The proposed method

The primary purpose in this module is to determine the feature components of the initial question: Object, Adjective, Verb, Adverb, etc. before making queries for the next modules. This is an automatic chunking problem for natural language question. Its solution is similar with the solution of the POS-tagging problem in information extraction. Using machine learning method CRF (Condition Random Fields) is one of the best solutions in Vietnamese. In many Vietnamese problems, it conduces to satisfactory results, for instances: Word segmentation (93%), POS-tagging (89.69%), Name entity recognition (92.31%), chunking (79.58%), etc.

Through the investigation of data and Vietnamese question features, the model proposed 9 labels and their features respectively. These labels represent four types of component as above in the table 1.

Example 3: Ai là người tìm ra châu Mỹ ? (Who discovered the American?) Ai là (Who)/WH người/O tìm ra (discovered)/V_W châu Mỹ(the America)/Object

In example 3, the set of keywords after implementing the module contain: tìm ra (Discovered)/V_W, châu Mỹ (the American)/Object.

3.1.3 Module processing

The feature selection is the most important step in CRF method. It impacts on the quality of NER and chunking systems. The more careful selection is, the more accurate system is. At a position i of observed data sequence include two parts. The former is data features, the other is respective label. The information of data features helps us determine the information of respective label at an observed data position. It means that labels can be automatically extracted model when has data features. From this point of view, the features used in our system are shown in Table2. From the features in Table 2, the using CRF method for about 2000 tagged questions (Training dataset).

Atthe result, a model which is base for analyzing user question components later is built.

3.2 Answer processing module

Answer extraction module proposes two primary answering strategies: sentences data collection-based and search engine-based. We will address in greater detail each strategy in the following sections.

3.2.1 Sentences data collection-based strategy

First, documents are retrieved and extracted using freely available Wikipedia dumps[1] of Vietnamese editions in XML format in which document contain fields: title, URL, content of article in Wikipedia respectively. Finally, question answering will be conducted follow three steps:

Step 1: Building data collection

The obtained documents are conducted noise reduction and sentence tokenization using JVnTextPro[2] toolkit. After that, we index this new data with some specific fields such as: title, URL, sentences of document using Lucence.

Step 2: Candidate Answer Extraction

Underlying each component of our question answering system is keyword-based document retrieval using Lucene. The system explored two modifications to extract answer: baseline method (Baseline) using word tokenization and CRF method in the question analysis phase (KLB). These strategies are described in greater detail below, and summarized in table4

- Baseline: this is a basic approach to compare with our proposed method which it only uses keywords taken from question to make query for Lucence. To illustrate our method clearer let us observe the example which will use in this paper:
  - ✓ With a question: "Ai là người tìm ra Châu Mỹ?" ("Who discovered theAmerican?")
  - ✓ Keywords: "tìm ra", "Châu Mỹ" ("discovered", "the American")
  - ✓ Query in lucence: +"tìm ra" +"Châu Mỹ" (+"discovered"+"the American")
- KLB: In this section, the system proposed an algorithm to extract answers. Firstly, components of a question have been sent by the question processing phase. These components consist of parts with tag of question, for instance: "Ai là - WH", "người - O", "tìm ra – V_W", "Châu Mỹ - Obj" ("Who - WH", "discovered – V_W", "the American - Obj"). Subsequently, the system chooses potential words to make Lucene query contains labels: "V_W", "A_W", "N_W", "Obj" and other words such as: "D_Time", "D_Loc", "D_Attr" to acquire exact answer by filtering retrieved results from Lucene. Finally, to get more exact answer, the system supplements a query expansion procedure by using a Vietnamese synonym dictionary.

Step 3: Answer selection

Candidate answers collection which has been sent by answer extraction feed in a filteringcomponent. These candidates are ranked by using score formula of Lucene (1). Sentence ranking is based on precision- and recall-like measures. Each question term is assigned by a weight based on its idf. Words that are synonymous according to our lexicons are pooled and their weights summed. The weights of words in the final sentence, and of some other useful terms, are boosted. Synonymous terms from the question are included in the Lucene query as well, each with the pooled weight. We note each document's Lucene DocScore. Finally, answer sentence candidates are recognized person entity answer by using Java open source library VSW[3] and ranked by a formula (2).

In there: rank $_{entity/d}$: rank of answer entity; scored: score of sentence candidate which contain entity; $freq_{entity}$: Frequency of entity in N candidates; N: Number of sentences candidates, δ Threshold

$$score_d = \sum_{tinq} (tf(tind) \times idf(t)^2 \times boost(t.fieldind) \times lengthNorm(t.fieldind) \times coord(q,d) \times queryNorm(q) \quad (1)$$

$$rank_{entity\ /N} = \delta \times score_d \times freq_{entity} + \frac{1-\delta}{N} \quad (2)$$

[1]http://dumps.wikimedia.org/viwiki/20101031/

[2]http://jvntextpro.sourceforge.net/

[2]http://code.google.com/p/vsw/

3.2.2 Search engine-based

In previous section, our system proposed a strategy based on collected data (SEB). The capability of answering in this strategy depends on amount of data warehouse. Therefore, to improve this as well as increase accuracy of answer, we observed other method based on obtained results of search engine. These strategies are described in greater detail follow two step:

Step 1: Snippet Retrieval

Same to previous strategy, after achieve keywords from question processing phase, these keywords will be made Google query by adding wildcard "*" or "**" into keywords. By this way, the system achieve some Google queries form: "*k1 k2…""k1 * k2…", "k1 ** k2…*" (k_i: is ith keyword).

Example: "tìm ra * Châu Mỹ" ("discovered * the American"); "tìm ra" "Châu Mỹ" ("discovered", "the American")

Next, queries will be pushed to Google search engine and obtain candidate snippets by using JSOAP API.

Step 2: Answer extraction

Candidate snippets collection which has been sent by step 1 are recognized person entity answer by using Java open source library VSW and ranked by using frequency of each entity.

4 Experiment and Discussion

In this section, the paper present some achieved results which illustrate that the proposed model as well as our approach is completely reasonable and highly applicable. Our model conducted two main experiments to evaluate system: one to appraise question analysis phase and another one to appraise entire system.

In question analysis phase, initially, we built a question dataset containingabout 2000 popular"Who, Whom, Whose" questions. This dataset was majorly drawn from Yahoo! Answer and some Vietnamese e-newspaper websites with some following requirements: the question must be less ambiguous and meaningful in natural language. After that, we standardized these questions into suitable syntax as well as Vietnamese context and conducted labeling to obtain a standard training dataset. Next, we used 10 fold cross validation in which were dividedthe training data randomly by 9:1 ratio. Then we carried out test and exposed the validated measures: precision, recall and F1 measure as show in table 3.

In Table 3, we presented a chart to compare the measures of 10 folds. The figure shown that the precision of using CRF in question analysis is quite high with F1 measure approximate 93%. This result illustrated that our approach is completely reasonable. However, the chart shown some unexpected results in several sample tests but these will be made well by supplement some specific dictionary as well as strengthen the training data much more.

	Precision	Recall	F1
Fold 1	89.7	90.2	89.95
Fold 2	94.1	95.05	94.57
Fold 3	96.4	96.83	96.61
Fold 4	93.07	94.23	93.64
Fold 5	94.58	96.11	95.33
Fold 6	92.43	93.45	92.93
Fold 7	91.3	92.67	91.98
Fold 8	88.35	89.45	88.89
Fold 9	91.5	92.11	91.80
Fold 10	93.32	95.01	94.15
Average	**92.475**	**93.51**	**92.99**

Table 3: Table of experiment results: 10 foldscross-validation

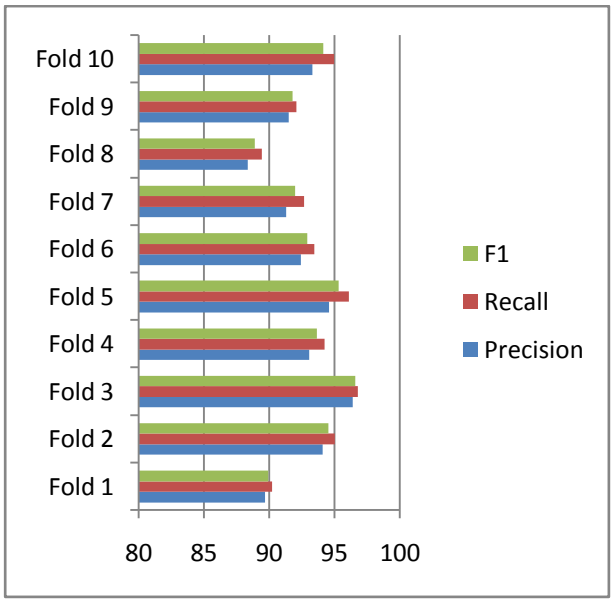

Figure 2: 10-folds cross-validation results chart

	Top 1			Top 3			Top 5		
	ρ	C	T	ρ	C	T	ρ	C	T
Baseline	41.07	54.3	46	42.23	54.7	49	42.29	55.1	52
KLB	79.68	55.6	58	89.39	60.3	59	90.03	60.2	61
SEB	71.44	90	28059	72.18	91.3	29820	73.17	91.7	30123
KLB+SEB	74.63	87.9	11630	79.62	89.3	12657	80.02	91.1	12799

Table 4: The comparisons of KLB, SEB, (KLB+SEB), and Baseline with 3 measures: precision (ρ), capability of answering (C), responded time (*T)*

In the next phase, we evaluated precision and responding time of entire system in which we proposed a method for question analysis as basic system to compare with our system. Here, we used 1000 questions taken from training data. After that we compared obtained result from 3 strategies of answering: knowledge-based (KLB), search engine-based (SEB) and hybrid method of these two strategies (KLB+SEB). Especially, with knowledge-basedstrategy, we carried out one more experiment named Baseline, instead of using CRF we only analyze questions at morphological layer to illustrate the effectiveness of CRF.The result is divided into 3 levels: Top one, three, and five per question, respectively. These obtained results are presented in Table 4.

In this experiment, we used 3 main measures to evaluate. The first one is capability of answering which is defined by $C = \frac{q}{Q}$ (q is amount of questions which system get answers; Q is amount of tested questions). The second one is precision of answers which is defined by $\rho = \frac{q_x}{q}$ (q_x is amount of questions which system get exact answers). And the last one is system performance which is time that system obtains an answer with each question. To evaluate this measure we run system with 1000 loops to answer one question before computing total running time and divided by total of loops. Particularly, it is defined by $\frac{t}{1000}$ (t is total running time 1000 times).

Table 4 presents a chart to compare obtained result per strategy. The chart shows that accuracy of answers and system performance is satisfactory. Top three levels generates the best results, however capability of answering is not really good because of its dependence on covered knowledge warehouse as well as ranking algorithms for returned answer did not achieve highly effectiveresults. Whilst the strategy using search engine has capability of answering as well as its accuracy of answer is acceptable but the running time is too slow. This is not efficient to build a real system, thus we proposed building a two layer system (combine both of above strategy) and achieved result which illustrates that hybrid system is completely reasonable. Additionally, we observed that the result of baseline method and compared it to CRF- based method. Using CRF create results which are much higher than baseline. These shown that the approach based on machine learning algorithms achieved results quite highly as well as illustrated that our proposed system is reasonable and realistic.

5 Conclusion and Future works

In this paper, we proposed and built a model of automatic system to answer questions about name of person in Vietnamese data domain. The achieved results illustrated that our approaches were completely reasonable and realistic. Furthermore, we also built an open framework for building an automatic question answering system. However, the system still remains some limitations due to the lack of amount of training question dataset as well as pessimistic rank algorithms for returned answers.We recommend the knowledge-based method to acquire the most remarkable performance and F1 score. Our future works will focus on building a huge training question dataset, boost a more optimal rank algorithm as well as improve system performance to deploy a real application. Additionally, we'll also extend knowledge warehouse and question domain to build an automatic open domain question answering system.

References

1. Demner-Fushman, Dina, *"Complex Question Answering Based on Semantic Domain Model of Clinical Medicine"*, OCLC's Experimental Thesis Catalog, College Park, Md.: University of Maryland (United States), 2006.
2. Magnini, B., Negri, M., Prevete, R., Tanev, H.: *"Comparing Statisticaland Content-Based Techniques for Answer Validation on the Web"*, Proceedings of the VIII Convegno AI*IA, Siena, Italy, 2002.
3. Boris Katz,*Annotating the World Wide Web using Natural Language*, In Proceedings of the 5th RAIO conference on Computer Assisted information searching on the internet (RIAO'97) 1997.
4. Dai Quoc Nguyen, Dat Quoc Nguyen, Son Bao Pham, *A Vietnamese Question Answering System*, KSE, pp.26-32, 2009 International Conference on Knowledge and Systems Engineering, 2009
5. John D. Lafferty, Andrew McCallum, Fernando C. N. Pereira: *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. ICML 2001: 282-289
6. TuoiT.Phan, ThanhC.Nguyen, ThuyN.T.Huynh. *Question Semantic Analysis in Vietnamese QA System*. The Advances in Intelligent Information and Database Systems book, Serie of Studies in Computational Intelligence, Volume 283, pp.29-40, (2010)
7. Vu Mai Tran, Vinh Duc Nguyen, Oanh Thi Tran, Uyen Thu Thi Pham, Thuy Quang Ha. *An Experimental Study of Vietnamese Question Answering System*. In Proceedings of IALP'2009. pp.152~155
8. Catalin David, Christoph Lange, Florian Rabe: *Interactive Documents as Interfaces to Computer AlgebraSystems: JOBAD and Wolfram|Alpha*; Centre d'Étude et de Recherche en Informatique du CNAM (Cédric) 2010
9. http://corporate.trueknowledge.com/architecture/