



ESTIMATION QUESTION TYPE ANALYZER FOR MULTI CLOSE DOMAIN INDONESIAN QAS

Iping Supriana, Ayu Purwarianti, Wiwin Suwarningsih*

School of Electronic Engineering and Informatics, Bandung Institute of Technology, Indonesia

Research Center for Informatics*, Indonesian Institute of Science, Indonesia

DOI: 10.5281/zenodo.583975

Keywords: multi close domain, question classification, estimation question type, question answering system.

Abstract

We propose an automated estimation scheme to analyze question classification in Indonesian multi closed domain question answering systems. The goal is to provide a good questioning classification system even if using only available language sources. Our strategy here is to build a pattern and rule to extract some important words and utilize the results as a feature for classification estimation of automated learning-based questions. Scenarios designed in automated learning estimates: (i) Analyzing questions, to represent the key information needed to answer user questions using target focus and target identification; (ii) Classify the type of question, construct a taxonomy of questions that have been coded into the system to determine the expected answer type, through some question processing patterns and rules. The proposed method is evaluated using datasets collected from various Indonesian websites. Test results show that the classification process using the proposed method is very effective.

Introduction

Question classification is an important phase in most question answering systems. There are three approaches for question classification [1]: rule based, language modeling and machine learning based. The questions classification of rule-based uses a set of standard heuristic rules based on taxonomy. Rule-based approaches classify questions using rules created manually by experts. Rule-based classification uses rules to detect keyword questions and utilize Word Net to map target categories [2]. Rule-based approaches do not support other domains or different languages because it is difficult to create a new set of rules framework. Rule-based approach performs well in specified data sets and not on the degradation of new data set performance [3]. Rule-based approaches are accurate in predicting certain categories of questions. However, this is not measurable for a large number of questions and syntactic structures. Rule-based questioning classification studies focus more on using the syntax and relationship rules between words [5][6], rules of relation and determines layer taxonomy by determining coarse class [7] [8].

Machine learning focuses on developing computer programs that can teach themselves to grow and change. Machine learning provides potential solutions across these domains and more [4]. There have been many research classification questions with approaches to machine learning such as Purwarianti et al [9] proposing a shallow parser to take some important words and utilize the results as a feature for classification of SVM-based questions. Skowron [10] uses the SVM algorithm with a composite feature of word categories and focus questions using a syntactic semantic structure. Zhang [11] and Mishra [13] compared various methods of machine learning to classify questions such as Nearest Neighbors, Naive Bayes, Decision Tree, Wicked Networks from Winnows (SnoW) and Support Vector Machine (SVM) using bag-of-word and n-gram features.

The questions classification is analyzed by various approaches from the point of view of previous researchers to improve class classification for the better. Although the learning engine approach is better but the rule-based approach has its own challenge: how to improve the system feature automatically to produce some complex information so that the classification becomes true. Based on that in this paper will propose the scheme of



INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

classification of questions using a combination approach of template patterns with the rules of relations between words. The generation of inter-word relationship rules uses the learning concept of knowledge based on the variation of the close domain. The proposed contributions are new scheme for Indonesian multi close domains that can be used to such as generate question template patterns and Expected Answer Type (EAT) variations.

The rest of the paper is organized as follows: Section. 2 describes several related work on question classification rule-based; Section 3 proposed method an estimation automatic learning for question analysis and question classification; Section. 4 discusses the learning issues involved in QC and presents our learning approach.

Related Work

The research on the question classification of intensive has been done by many researchers but the question classification is still a trend of continuous research review. This is based on several considerations such as how large the system is able to define the expected type of response [4][6], how large the system is capable of sorting words that include stop-world automatically in the classification of questions [12] [14] and how the system is able to handle variations of question forms to produce accurate high [14]. Especially rule-based classification has its own challenge: how to improve system features automatically to generate some complex information so that classification becomes true.

There are many question classification researches with rule-based questioning typically use word extraction functions, syntax patterns and build relationships between words. Such as Te, *et al* [13] focuses on word extraction and establishing the rules of relations between words using Tibetan. The proposed method is to collect information by comparing keywords. This strategy is done to reduce the search space and improve the efficiency of the word search process. The results of this approach are able to initialize the knowledge base with the relation rules for the Tibetan language question structure.

Sarrouti *et al* [1] proposed an effective and efficient method for Biomedical QTs Classification. They have classified the Biomedical Questions into three broad categories and defined the Syntactic Patterns for particular category of Biomedical Questions. Therefore, using these Biomedical Question Patterns, they have proposed an algorithm for classifying the question into particular category. The proposed method was evaluated on the Benchmark datasets of Biomedical Questions. The experimental results show that the proposed method can be used to effectively classify Biomedical Questions with higher accuracy.

Riloff [6] develop a rule based system Quarc that can read a short story and find the sentence in the story that best answer a given question. Quarc uses heuristic rules that look for lexical an semantic rules in the question and the story. Each rule awards a certain number of points to a sentences. After all of the rules have been applied, the sentence that obtains the highest score is returned as the answer.

Haris & Omar [5] describes a rule-based approach to analyze and classify written examination questions through natural language processing for computer programming subjects. In general, Bloom's Taxonomy or the Taxonomy of Educational Objectives (TEO) acts as a main guideline in assessing a student's cognitive level. However, academicians need to design the appropriate questions and categorize it to the cognitive level of TEO manually.

Biswas et al [6] proposed a compact and effective method for question classification. Here rather than using a two layered taxonomy of 6 course grain and 50 fine grained categories developed by Li and Roth [7]. They have classified the questions into three broad categories studied the syntactic structure of the question and suggest the syntactic patterns and expected answer type for particular category of questions. Using these question Patterns they have also suggested an algorithm for classifying the question into particular category. They have also studied the syntactic structure of the question and suggest the syntactic patterns and expected answer type for particular category of questions



Proposed Work

Here, we would like to show our framework of Question processing with method an estimation automatic learning for question analysis, question classification and automatic extraction learning. Framework of proposed method that we have built can be seen in Figure 1.

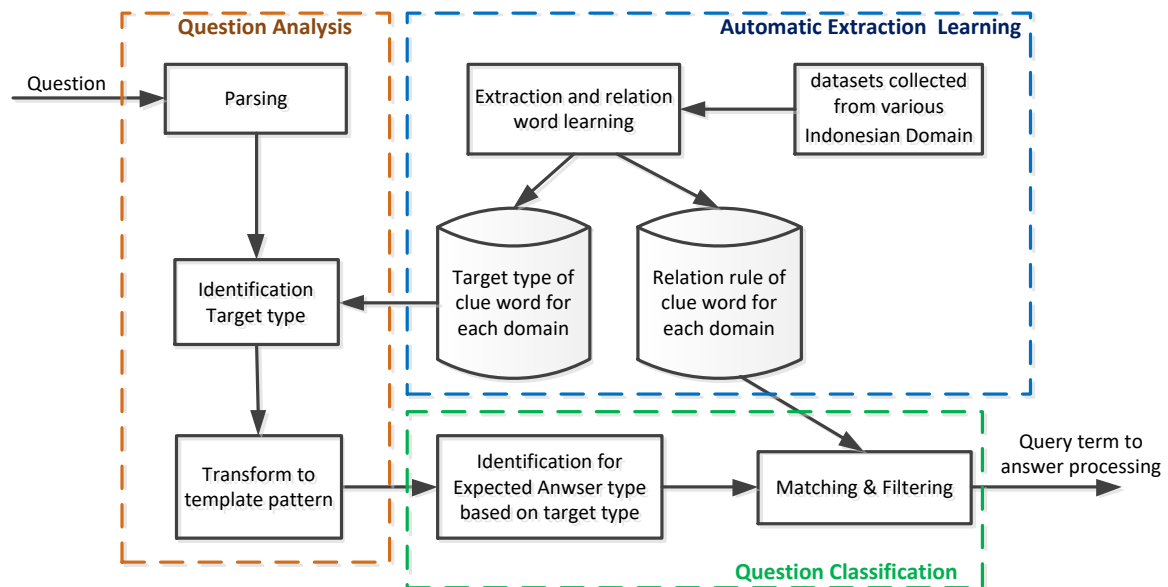


Fig. 1 Proposed Work

Question Analysis

In Indonesian question always contains some special question pronouns such as *siapa* (who), *apa* (what), *kapan* (when), *dimana* (where), *bagaimana* (how). Question pronouns instead of unknown which inquirer most wants to know. Correspondingly, the position of question pronoun is where question target type. In addition, in some special questions with multiple question pronouns, there is more than one question target type and focus to describe domain types determined of kind of domain. In this paper will discuss with multiple close domain such as medical, sport, law and religion.

Question Classification

The purpose of question classification is to set the questionnaire label based on the type of expected answer (EAT). This is done to facilitate searching and generating using target types of sentence queries based on defined domains. Because the type of answer expected for each domain is different. The process of generating the expected types of answers takes the approach of extraction of the rules of relations between words in a sentence. Expected Answer Types such as: Yes/No, Factoid and Definition. When EAT is already defined the feeding stage to filter is done by using the concept of pattern matching. The final result of this filtering process will be used for the next step in the QAS of searching for answers

Automatic Extraction Learning

This stage is a process to collect a number of clue words based on domain groups. This stage is done automatically by using the learning approach based on the rules of relations between words. The word used as clue word will be the benchmark to see the relation of every word in the sentence. The results of this stage are a number of words that belong to the target type and set of relation rules used for the process of question analysis and question classification.



INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

Result and Discusses The Learning Issues

In this section we will discuss the results of the proposed method using Indonesian sentence examples. The process of parsing for the question sentence through several stages of the tokenize process then the sentence question, the labeling process by using POS tagger Indonesian, stop word elimination, phrase identification done to find the existence of phrase from the sentence question. Identify the question word and the clue word. After the clue word is found then identification of the target type by searching in the database to determine the domain type of the question sentence. This is done to determine the group of sentences and words that are co-occurrence.

Identification Target Type

The Indonesian question type has a clear sign. For example, the pronoun of questions like *siapa* (who), *apa* (what), *kapan* (when), *dimana* (where), *bagaimana* (how) is a specific question. What's more, pronouns of this question not only identify the type of question but also mark the target type of the domain. In the yes-no question, the question word always uses the "kah" particles followed by the word clue word in "benar" or "salah". Correspondingly, "benar" or "salah" is a sign of positive-negative questions and there are other words that answer the question. In other words clue is always a hallmark of alternative questions. In addition, a set of predefined question words was also used in our study. In the Indonesian sentence, the word question is divided into three types: pronoun question, adverb questions and question particles. (the pattern of indonesian question see table 1) To overcome this issue, we list all regular expression patterns of each question that are used in our experiments as follow.

Transformation To Template Pattern

This transformation stage serves to extract the question sentence into the semantic-based template pattern. The questionnaire pattern in this dissertation is made manually based on observations from some medical consultation webs. Patterns built using semantic-based templates. This semantic-based template combines and leverages the SRL with the versatility of the templates. The illustration of the transformation process from the question into the template pattern is described as follows. Examples of sentences used are sentence_tanya_1. Based on the result of the EAT classification, the EU's final sentence_1 is 'CAUSE' with the question 'WHAT' and the instruction word is 'cause'.

Example : *Apa penyebab badan lelah?* (1)
(English : What causes body tired?)

If it refers to table IV.3, we get the pattern "APA + cause + <PROBLEM> | <POPULATION>?". This pattern can be broken down into two different patterns: "APA + *penyebab* + <PROBLEM>" or "APA + cause + <POPULATION>?". The "|" sign means or the identified pattern can be used one.

In order to use this semantic-based template pattern, the sentence_1 must identify the existing word or phrase into the PROBLEM or POPULATION element. Referring to Table IV.2 the phrase 'tired body' belongs to the PROBLEM element. So the resulting pattern shape for sentence_1 is "APA + *penyebab* + <PROBLEM>?".

The resulting pattern for the sentence_1 is used to search the answer in the case database. The search for answers using template patterns is done by matching questionnaire patterns to get the pair of answer sentence patterns. Once the template pattern answers obtained the system will search for words or phrases that correspond to PICO elements recorded in the database.

Table 1. Sample of Pattern of question sentence for question word 'what' and 'how'

| Question Word | Pattern of question sentence |
|---------------|---|
| | Apakah + karena + verb(event) + sehingga + noun(person) + verb(event)? |
| | Apa + ada + faktor + noun(property) + yang + bisa + noun(comparison) ? |
| | Apa + yang + verb(direction) + bila + obyek-noun (Person)+ noun(comparison) ? |
| | Apakah + tanda-tanda+ subyek-noun(person) + yang + memiliki + |



INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

| | |
|--|---|
| Apa (What) | noun(comparison) ? |
| | Apakah + faktor + noun(comparison) ? |
| | Apakah + penyebab + verb(event) ? |
| | Apakah + efek + verb(event) + obyek-noun + noun(comparison) + dan noun(comparison) ? |
| | Apakah + noun(property) + parah sekali? |
| | Apakah + subyek-noun(person) + harus + verb(event) + obyek-noun (Person) ? |
| | Langkah + apa + noun(person) + verb(event) ? |
| | Apakah + ket.waktu + noun(person) + verb(event) ? |
| Bagaimana (how) | Bagaimana + supaya + bisa + noun(comparison) + lagi? |
| | Bagaimana + noun(comparison) + untuk + obyek-noun (Person) ? |
| | Bagaimana + supaya + noun(person) + verb(event) + lagi ? |
| | Bagaimana + dengan + noun(comparison) + noun (property) + obyek-noun (Person) ? |
| | Bagaimana + noun (property) + bekerja ? |
| | Bagaimana + noun (comparison) + dikendalikan ? |
| | Bagaimana + upaya + noun(person) + mengendalikan + noun (problem) ? |
| | Bagaimana + cara + verb(event) + noun(comparison) + untuk +obyek-noun (Person)? |
| Bagaimana + noun(time) + masih + noun(comparison) ? | |

Identification To Expected Answer Type

In this section we analyze the structure of sentence patterns based on grammatical structure with examples from natural language. The grammatical rules used to distinguish strings from symbols that are sentences received from existing examples. For example a simple Indonesian sentence with a subject structure followed by a predicate. The subject takes the form of a single noun, while the predicate is a verb or adjective or number followed by an object.

Our main goal is to classify the multi close domain Questions into three broad categories: Yes/No, Factoid and Definition Questions. To achieve this goal, we propose several Syntactic Patterns of each domain. Table 1 show that *apa* (what) and *bagaimana* (how) types of questions could belongs to Factoid and definition Questions, while *Why Type* of questions belongs to only Summary Questions, etc. To detect Yes/No Questions we used the regular expression (see pattern (1)), where the questions should start with three types of words. We found that this method is significantly efficient (see example sentences for pattern at tabel 2)

Table 2. Answer type based on category

| Question | Categori | Answer |
|---------------------------|------------|---|
| <i>Bagaimana</i> (how) | Factoid | Domain entity names, Number |
| | Definition | Phrase, paragraph |
| <i>Apa</i> (What) | Factoid | Domain entity names, Number, short expression |
| | Definition | Phrase, paragraph |
| Yes/No | Yes/No | <i>Ya</i> (yes) or <i>tidak</i> (No) |

The purpose of question classification is to set the questionnaire label based on the expected answer type. The process of defining the type of EAT done in this paper is to inventory the type of question sentences collected from the online literature sources. EAT for the medical domain are proposed on 7 types, namely: SYMPTOMS (defining symptoms of disease), DIAGNOSES (interpreting diagnoses for patients), TREATMENT (preventive measures to reduce the impact of disease), BENEFITS (benefits of taking a drug or performing certain treatments), DIRECTION (instructions on use and consumption of drugs), CAUSES (causes of health problems), PREVENT (treatment to prevent health problems). For law domain with 6 EAT such as



INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

DEFINITION, ACTION, TIME, LOCATION, CRIMINAL and CIVIL. For sport domain with 5 EAT ACTION, TIME, LOCATION, TYPE_OF_SPORT, RULES_OF_PLAY

Filtering

The filtering module serves to filter out some irrelevant answers that are produced based on simple thresholds. While the valuation module is an important phase because the answers arrive at by the system should be fairly reliable to be presented to the user. One is to apply a thorough evaluation process to the answers given after the filtration phase. The aim is to achieve in collecting supporting evidence for each candidate's answer and apply various assessment techniques to evaluate supporting evidence.

Automatic extraction learning

At this stage, an automated extraction rule has been established to form target and EAT types. The way that is done by the transformation of sentences from declarative sentences into interrogative sentences. The method used for making the transformation rule in this dissertation is to utilize the outcome sentence from the dependence between Named Entity. It is assumed that sentences that have a dependent relationship share a semantic role (semantic role) as long as it is raised in the same or related sentence. This is driven by many studies in lexical semantics in which the hypothesis of word behavior, especially with regard to the expression and interpretation of rules to a large extent determined by meaning.

Each slot in the semantic-based template pattern, examined its semantic role and adjusted in the transformation rules. The role of the text corresponding to the transformation rules is extracted and changed according to the slot. The extraction results are inserted into the question sentence.

Table .3 Rules of Transformation of the Declarative sentence

| Rule | Bentuk kalimat tanya |
|----------------------------------|--|
| If System find role = <COMPARE> | Apa + manfaat + <COMPARE> + <?> |
| If System find role = <PROBLEM> | Bagaimana + <INTERVENSI> + <PROBLEM> + <?> |
| If System find role = <LOCATION> | Diman + <EVENT>+ terjadi+<?> |
| If System find role = <DRUG> | Berapa + <DRUG> + dikonsumsi + <?> |
| If System find role = <OUTCOME> | Apa + indikasi + <OUTCOME> + <?> |
| If System find role = <CONTROL> | Mengapa + <CONTROL> + diperlukan + <?> |
| If System find role = <PASIENT> | Siapa + <PATIENT> (?) |

Each slot in the semantic-based template pattern, examined its semantic role and adjusted in the transformation rules. The role of the text corresponding to the transformation rules is extracted and changed according to the slot. The extraction results are inserted into the question sentence.

Conclusion

In this paper we have presented a novel method for classifying the multi close domain Questions into two broad categories that summarize all possible cases of the Expected Answer Types: Yes/No, Factoid and Summary Questions. We presented the rules that are represented by a set of patterns for each domain. In analysis using our proposed work prove that the multi domain classification problem can be solved quite accurately using our proposed method.

References




- [1]. Mourad Sarrouiti, Abdelmonaime Lachkar and Said El Alaoui Ouatik, Biomedical Question Types Classification using Syntactic and Rule based Approach, In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR, pages 265-272. ISBN: 978-989-758-158-8.
- [2]. Hovy E, et al. A question/answer typology with surface text patterns. Proc 2nd Int Conf Human Language Technology Research. HLT'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 2002. p. 247-51.



INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

- [3]. S. Jayalakshmi and Ananthi Sheshasaayee , Question Classification: A Review of State-of-the-Art Algorithms and Approaches, Indian Journal of Science and Technology, Vol 8(29), November 2015
- [4]. Ali Mohamed Nabil Allam and Mohamed Hassan Haggag, The Question Answering Systems: A Survey. International Journal of Research and Reviews in Information Sciences (IJRRIS) Vol. 2, No. 3, September 2012
- [5]. Haris, S.S., Omar, N., 2012. A rule-based approach in Bloom's Taxonomy question classification through natural language processing, in: *Computing and Convergence Technology (ICCCT), 2012 7th International Conference on*. pp. 410–414
- [6]. Biswas, P., Sharan, A., Kumar, R., Sciences, S., 2014. Question Classification Using Syntactic And Rule Based Approach. *Int. Conf. Adv. Comput. Commun. Informatics* 1033–1038.
- [7]. Li X, Roth D. Learning question classifiers. COLING 2002
- [8]. Harish Tayyar Madabushi & Mark Lee, High Accuracy Rule-based Question Classification using Question Syntax and Semantics, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1220–1230, 2016.
- [9]. Ayu Purwarianti, Tsuchiya Masatoshi, Seiichi Nakagawa, Estimation of Question Types for Indonesian Question Sentence, www.anlp.jp/proceedings/annual_meeting/2006/B2-8.pdf
- [10]. Skowron, Marcin and Kenji Araki, "Effectiveness of Combined Features for Machine Learning Based Question Classification", Journal of Natural Language Processing 2005, pp. 63-83, 2005.
- [11]. Zhang, Dell and Wee Sun Lee, "Question Classification using Support Vector Machine", ACM SIGIR 2003.
- [12]. Gupta, P., Gupta, V., 2012. A Survey of Text Question Answering Techniques. *Int. J. Comput. Appl.* 53, 1–8.
- [13]. Rou Te, Research on question classification method of Tibetan online automatic question-answering system, 2011 Fourth International Conference on Intelligent Networks and Intelligent Systems.
- [14]. Ramprasath, M. & Hariharan, S., 2012. A Survey on Question Answering System. *International Journal of Research and Reviews in Information Sciences (IJRRIS), United Kingdom*, p.171-179.
- [15]. Ellen Riloff & Michael Thelen, A rule based question answering system for reading comprehension test, <https://www.cs.utah.edu/~riloff/pdfs/quarc.pdf> .
- [16]. Megha Mishra ,Vishnu Kumar Mishra and Dr. H.R. Sharma, Question Classification using Semantic, Syntactic and Lexical features , International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.3, July 2013



| | |
|---|--|
|  | <p>Iping Supriana Suwardi received the docteur-ingenieur degree in Informatics from the Institute National Polytechniques, Grenoble, French in 1985. He is currently a Professor of the School of Electrical and Informatic Engineering, Bandung Institute of Technology, Indonesia.</p> <p>His research interest include: information automation, dynamic computer graphic, image analysis, recognition system, and image interpretation. He has authored or coauthored over 50 published articles.</p> <p>He is the inventor and the implementer of the Gigital Mark Reader (DMR). DMR is software employed for evaluating examination results automatically using computer scanning. DMR is widely used in Indonesia education instituion.</p> |
|  | <p>Ayu Purwarianti. She was graduated from her bachelor and master degree at Informatics Program, Bandung Institute of Technology. She got her doctoral degree from Toyohashi University of Technology, Japan. Since 2008, she has become a lecturer at School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia. Her research interent is on computational linguistics, mainly on Indonesian natural language processing and Indonesian text mining. She is now active as the education officer at IEEE Indonesia.</p> |
|  | <p>Wiwin Suwarningsih. She was obtain his Magister degree from Informatics department, Bandung Institute of Technology. Currently she still studying at School of Electrical Engineering and Informatics, Bandung Institute of Technology as doctoral student since 2013. Currently she is working as a researcher at Indonesia Institute of Science. Her research interest are on text minning and question answering system.</p> |