A Few Survey of Developments and Challenges Arising on General and Indonesian Question Answering System

Hernawan Sulistyanto, Azhari S.N
Dept. of Computer Science
Gadjah Mada University
Yogyakarta, Indonesia
email: hnsltyanto@gmail.com

***Abstract*—Natural language based interaction between human and computer is an emerging research area with a wide range of possibilities for various types of application. For instance, Question Answer System (QAS) is being intensively researched and then broadly applied in the various ways. The QAS interfaces human and computer in retrieving the specific information from the collection of documents to natural language question queried by human. Generally speaking, the QAS has differed with the usual search engine on the given answer. On the one hand the search engines give a list of relevant documents to user query. On the other hand a short and exact answer is returned by QAS. This paper, as a small review of the QAS, would like to address the development trends, and then coupled with challenges arising on Indonesia QAS. The aim of this paper is to help in identifying the approaches and methods already used, and also to sign the challenges would be faced.**

Keywords-quenstion answering system (QAS), search engine, retrieval information

I. INTRODUCTION

Advancement in computer technology has facilitated to retrieve the information easily by using single mouse click. Retrieving the information is performed by a search engine which employs information retrieval (IR) system to return a list of relevant documents their user need. In conventional manner, retrieving the information from huge documents is based on a specific keyword. Unfortunately, this searching method could not satisfy the user's needs to get the exact information. In addition, the user should be responsible to read all of retrieved documents to discover the required answer.

Question Answering System (QAS) can be used to tackle this issue. QAS is a system that returns short text as an exact answer by retrieving the answer from its corpus to a question written in natural language [4]. In essence the QAS is presented with natural language questions and the expected output is the exact answer identified either in a text or in small text fragments containing the answer. In IR, engine's query language represents the input query through some keywords and the output is conveyed as a list of documents that are presumably relevant to the user query. On the contrary, QAS allows the user to ask his/her question directly to the system with natural language. An answer then will be provided by the QAS in the form of exact and short answer extracted from a source document. Therefore, QA technology is different with conventional IR technology.

Question Answering System (QAS) has been an exciting research topic and has consumed much attention in recent year. A number of researches dealing with QAS development have been introduced in various presentations by many researchers. Recently, semantic-approach based QAS has been intensively proposed by [12][17][19][20][22] to facilitate the use of special query languages or application which allow user to query the semantic data repository using natural language. Developing the semantic approach cannot be separated with ontology in which the ontology plays a key role in providing vocabularies for corresponding word that can be used by application to understand words meaning as in [19][25] and close gap between semantic web and data bases in tractable query answering [1]. Moreover, according to [13] ontology can be also used to enhance the system intelligence degree which can automatically generate the related knowledge areas of more comprehensive questions and corresponding answers. The user query that cannot be retrieved from ontology will be recovered by QAS from web documents through cross-document technique [22] and cross-lingual [5]. Meanwhile, in [17] the natural language user query is converted to Resource Description Framework (RDF) on finding the answer. The answer returned to user by QAS is usually in the form short and exact answer. By [6] the given answer is also completed with its explanation to construct the modified QAS that is called the real word question type system.

Proposing and implementing some different approaches have enhanced the QAS development level in particularly its performance. The most important thing of QAS discussion is at paying attention on some challenges arising, such as use other than English corpus, combine the existing techniques, propose new approaches possibilities, etc. Our main contribution in this paper is at addressing the trends of QAS research and exhibiting the awareness from further challenges in this area, in which this report is based on the recent researches published in between 2008 and 2013.

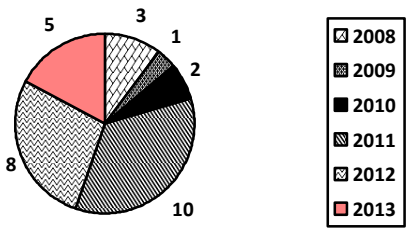

Figure 1. The composition of references that are reviewed in this report

Overall, the rest of paper is organized into these sections. Current QAS architecture would be described in section 2. In section 3, it is going to present direction of QAS developments. Challenges arising on general followed by Indonesian QAS are introduced in section 4. Finally, section 5 releases some valuable conclusion in the research area of QAS.

II. General Architecture of QA System

Basic task of QAS is to retrieve an exact and correct answer from a collection of documents to the question given in natural language. The prototypical QAS consists of four main modules, those are question analysis, document retrieval, document analysis, and answer selection.

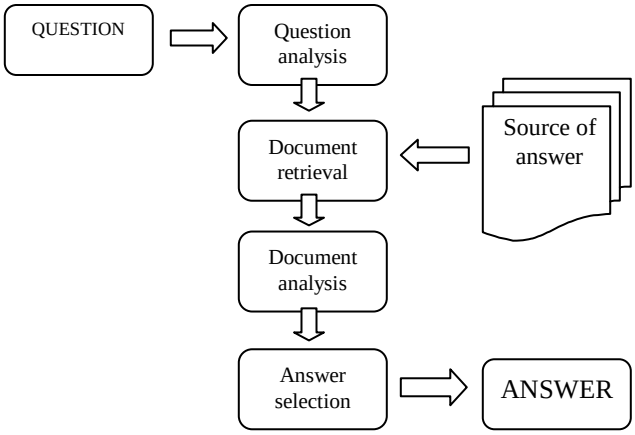

Figure 2. The general approach to QAS

A. Question Analysis

Analyzing the natural language question as input to QAS is to be first step toward at finding the answer. The question analysis aims to understand the purpose of the question which can be identified by analyzing the question in many ways, such as morpho-syntactic analysis. This analysis is to indicate whether a word is a verb, singular noun, plural noun, etc. By indicating the word it would be useful information to judge the kind of information from the question is asking for. Mainly there are three types of question analyzer related to QAS using the document based approach [6]: lexical based, syntactical based, and semantic based. The question keyword and expected answer type (EAT) is given by question analyzer as a result in the lexical based QAS [27][28]. In the syntactical based QAS, the question analyzer derives the question syntactical tree. Both question keyword and question syntactical tree are used to retrieve the relevant document or passage.

Question classification has main task at predicting the entity type of the answer from a natural language question by mapping a question to a predefined category that specifies the entity type of the expected answer [10]. Identifying the semantic type of the question is also an important step for understanding what the question asks for. By classification, question is put into several semantic categories and looked at the key question word, such as when seeks a date/time, where a location, who a person, etc. The accuracy of question classification is very important for the overall performance of the QAS if it goes wrong then it will affect the working of next process. Once the question is classified into proper type of question it derives expected answer type, extracts most relevant keyword, and reformulates a question into its semantically equivalent multiple questions [24]. The hierarchies of question type that based on the types of answer sought are constructed by question classification system and then the input question is put into appropriate category in the hierarchy. Categorizing the question can be implemented in the various ways and approaches. Pattern matching is one of the simplest way yet quite effective. For this reason, [2] use it to enhance the effectiveness of the Question classification system by ranking the similar document instead of traditional distance metrics. Meanwhile, the heuristic rule-based algorithm requires write some heuristic rules manually for question classification. In other side, a machine learning approach can automatically build a high performance question classification program and learned classification program is more flexible than a manual one since it can be easily adapted to a new domain. It is proven that most of the successful question classification research recently is done with Support Vector Machine (SVM) [8][11][21]. However, SVM as a typically question classification technique using machine learning approach has a main problem, in which this problem is the high dimensionality of the feature space. Hence, [10] applies latent semantic analysis technique by comparing SVM with Back-propagation Neural Network (BPNN) based classifier and success to reduce the large feature space of question derived by typically machine learning techniques. Similar problem is solved by [23] by using online QAS that employ matching rule. In their research, [11][23] success not only enhance the efficiency of question classification time but also improve the classification accuracy by removing the redundant features. Final finding from [11] is BPNN perform better compared to SVM.

B. Document Retrieval

Finding the answer is started at retrieving the document/information which is to identify documents that are likely to contain the answer. The mission of document retrieval is to extract relevant documents from the corpus of interest. In an effort to pin-point relevant information more accurately, the documents are split into several passage, and passages are then treated as documents. Selecting and filtering the passages that are considered relevant to the input question are then done in order to narrow the search space for the answer. For one thing, the sorting method can also be used for ordering the passages as it will give the most appropriate document where the exact answer for the question is assumed to be available. Another thing, indexing the database can be also basis for information retrieval in which all of the search requests are answered by indexing process based on the keyword index in knowledge, document ID records the questions and answers to the corresponding ID numbers, synonyms and inverted table content [26]. Conversely, a database independent technique is introduced by [15] implementing the distributional semantic model (DSM) approach which combines several techniques in the question cube framework to retrieve passage containing the exact answer for natural language question.

C. Document Analysis

The document analysis analyzes the documents selected by document retrieval to identify phrases that are of the appropriate type. Ordinarily, by using a named entity identifier, the multiword strings as names of include person, dates, location, etc can be recognized and classified.

D. Answer Selection

Matching between representation of the question and the candidate answer-bearing texts will produce a set of candidate answer. Typically, the matching process may require first to match up the text unit's semantic from a candidate answer text with the semantic type belonging to the expected answer. As it is the final component in QA architecture, the answer selection module is responsible at identifying, extracting, and validating answers from the collection of ordered passage passed to it. Generally, validating the answer successfully is performed by an answer validation system based on machine learning approach which is responsible in deciding the answer correctness, in term whether the answer of a question answering system is correct or not, as done by [7] and [18]. According to [18] the answer validation system must return a judgment for the selected answer, validated or rejected, from the received set of triplets consisting of question, answer, and supporting text. Furthermore, [7] compare the validation performance between validating the selected answer on text-based and web-based.

III. TRENDS ON INDONESIAN QAS DEVELOPMENT

QAS is an information retrieval system based on natural language question input to get the short and exact answer. A number of QAS model have created by many researchers in which those QAS works on either the multilingual in [5] or the specific language based development, such as Tibetan based [23], Turkish based [2], Chinese based [13][25], and French based [7]. Moreover, Indonesian based is utilized by [3][6][9][12][14][28][29].

Indonesian language (Bahasa Indonesia) spoken by over 167 million people is the official language in Indonesia. In fact according to our knowledge, the Indonesian language (herein after simply 'Indonesian') is also being learned by people in the some country around the world. Based on this given fact, it is very important to make more intensive development in the Indonesian QAS researches. Hence, a research trend in Indonesian QAS needs to be addressed. Typically QAS may be performed in many ways and approaches. The Indonesian QAS has been done with semantic analysis-based by [9][12] to obtain semantic representation of Indonesian sentences that are used by a question answering module which stores declarative sentences as fact in a knowledge–based. Extending the semantic analysis is performed by [12] with adding a number of axioms designed to encode useful knowledge for answering question. By adding axioms, the QAS presented in [12] is more robust than in [9] where the system has had capability to answer questions which previously could not have been answered.

The biggest challenge in QAS is at categorizing the question into a particular type. General approach to question categorization is by matching the pattern of each question type based on the position of question words and various question keywords. Both the question words and question keywords may represent the question type. The questions can be grouped into factoid (person, organization, location, date-time, quantity) and non-factoid (definition, reason, method) question type [28]. Accordingly, [28] construct QAS to handle both of the factoid and non-factoid question by using monolingual approach and then extended by [29] with using phrase-based approach in machine learning based factoid. However, a small modification on QAS architecture has been done in [6] where cased-based QAS there is only question analysis and case retrieval component. As the answer is already available in the case base, [6] replaces the passage retriever component with the case retriever and the answer finder disappeared. Besides, in an Indonesian sentence usually the verbs as predicate do not appear in the original word form yet usually is attached by the suffix. Affixing the suffix (such as"–kan") in the Indonesian sentence will affect the sentence syntactically. Variation of the sentence syntactic is handled by [14] with the word class induction employing hierarchical agglomerative clustering. Initialized in [28] phrase based answer finding is proposed by [29] to solve the problem arising in their previous

research in which the major problem is mainly in answer scoring technique deficiency.

IV. CHALLENGES ON INDONESIAN QAS

A. QAS in General Scope

The QAS development is identified by four components, namely topics talked, approach used, features employed, and database utilized. Generally, there is no focusing on specific topic. Nevertheless, developing the intelligent QAS and using the cross-documents over multi-lingual are topic trends in the last five years. In addition, developing the QASs have implemented the approaches toward semantic-ontology by using various features and web based documents. A new challenge arising in question analyzing is range query. For the most part, range queries are utilized to get the answers of the questions which are about between two values. These range queries differ than what IR does in which the general IR approach is exact match. It will be particular challenge for further research of QAS. Furthermore, another one is about question understanding. Assume there is a question "What is the name of person who has more than 1 billion rupiahs yearly revenue?" In this case, the question understanding need considered for next question answering research.

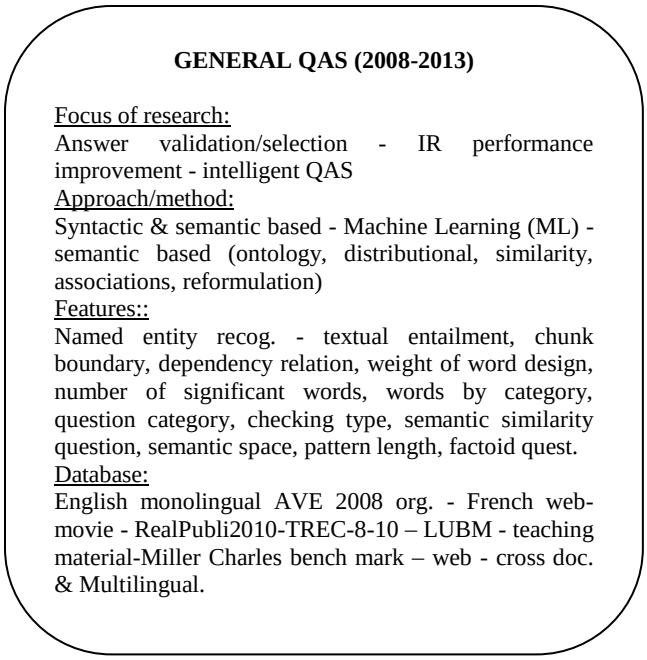

Figure 3. Illustration of general QAS development

B. QAS in Indonesian Scope

Comparatively, Indonesia QASs have been developed toward achieving the question categorization performance in determining the question type. Indonesian language is still an under-resourced language when it comes to natural language processing. Singularly, Indonesian is spoken by more than 200 castes (such as Java, Sunda, Batak, Bali, etc) having different dialect spread in more than ten hundreds islands thus it will affect the Indonesia question grammar structure. In addition, there are many Indonesian word consisting of two words is just one meaning, such as "Siapakah pemenang nomor lari 100 meter putra?". Here, "nomor lari" in Indonesian is deemed one word but in English may can be deemed two words consequently will be placed in different phrase.

Besides, Affixing of Indonesian with prefix "me-, ber-, di-, ter-", infix"-el-, -em- , -er-" and suffix '"-kan, -i" will be also particular challenge in question parsing. Moreover, presence the compound word and others making the syntactically and semantically sentences more complex certainly will make the bigger challenges on Indonesian Question answering system, especially in question parsing.

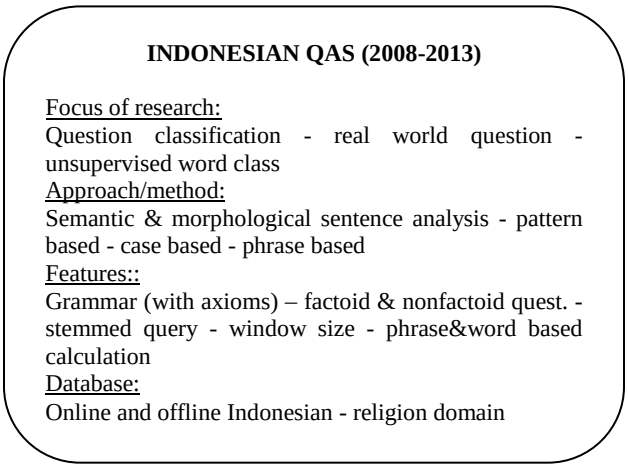

Figure 4. Illustration of Indonesia QAS development

V. CONCLUSION

A survey on QAS is presented in this paper. Both general QAS and Indonesian QAS basically have similar development trends. A number of approaches introduced by the reviewed papers successfully are implemented on QAS. However, Indonesian QAS may need more attention in specially question analyzer. In this stage, it is not only just need parsing the question sentence but also understanding the semantically and syntactically sentence regarded to many different speaking dialect in Indonesian.

ACKNOWLEDGMENT

The author is a lecturer in Muhammadiyah University of Surakarta who is pursuing the doctoral degree at School of Computer Science, Gadjah Mada University, Yogyakarta.

REFERENCES

[1] Cali, A., Gottlob, G., and Lukasiewicz, T., "A general datalog-based framework for tractable query answering over ontologies", Journal of Web semantics: services and agents on the world wide web, Elsevier, Vol. 14, 2012, pp. 57-83, doi: 10.10.16/j.websem.2012.03.001.

[2] Celebi, E., Gunel, B., and Sen, B., "Automatic question answering for Turkish with pattern matching", IEEE Trans., 2011, pp. 389-393

[3] Ferdian, F., and Purwarianti, A., "Implementation of semantic analysis in Indonesian text-understanding evaluation system", In Proceedings of IEEE International Conference on Computational Intelligent and Cybernetics, Bali, 2010.

[4] Ferrández, O., Izquierdo, R., Ferrández, S., and Vicedo, J.L., "Addressing ontology-based question answering with collections of user queries", Journal of Information Processing and Management, vol. 45, 2009, pp. 175–188, doi: 10.10.16/j.websem.2011.01,002.

[5] Ferrandez, O., Spurk, C., Kouylekov, M., Dornescu, I., Ferrandez, S., Negri, M., Izquierdo, R., Tomas, D., Orasan, C., Neumann, G., Magnini, B., and Vicedo, J.L., "The QALL-ME framework:a specifiable-domain multilingual question answering architecture", Journal of web semantics: science, services and agents on the world wide web, Elsevier. Vol. 9, 2011, pp. 137-145.

[6] Fikri, A., and Purwarianti, A., "Case based Indonesian closed domain question answering system with real word questions"' In Proceedings of IEEE 7th the International Conference on Telecommunication Systems, Services, and Applications (TSSA 12), 2012, pp. 181-186.

[7] Grappy, A., Grau, B., Falco, M.H., Ligozat, A.L., Robba, I., and Vilnat, A., "Selecting answers to question from web documents by a robust validation process", In Proceedings of IEEE International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Press, 20011, pp. 55-62, doi: 10.1109/WI-IAT.2011.210.

[8] Huang, Z., Thint, M., and Qin, Z., "Question classification using head word s and their hypernyms", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 08), IEEE Press, 2008, pp. 927-936.

[9] Larasati, S.D., and Manurung, R., "Towards a semantic analysis of bahasa Indonesia for question answering", In Proceeding of the 10th Conference of the Pacific Association for Computational Linguistics, IEEE Press, 2008, pp. 273-280.

[10] Loni, B., Khoshnevis, S.H., and Wiggers, P., "Latent semantic analysis for question classification with neural networks", In Proceedings of IEEE International Conference on ASRU, 2011, pp. 437-442.

[11] Loni, B., Tulder, G.V., Wigger, P., Tax, D.M.J., and Loog, M., "Question classification by weigthed combination of lexical , syntactical and semantic feature", 2011.

[12] Mahendra, R., Larasati, S. D., and Manurung, R., "Extending an Indonesian semantic analysis-based question answering system with linguistic and world knowledge axioms", In Proceedings of IEEE The 22nd Pacific Asia Conference on Language, Information, and Computation, Cebu, Philippines, 2008, pp. 262-271.

[13] Mei, L.H., "Intelligent question answer system of research based ontology on excellent course", In Proceedings of IEEE 4th International Congference on Computational and Information Science, IEEE Press, 2012, pp. 784-787, doi:10.1109/ICCIS.2012.177

[14] Mistica, M., Lau, J.H., and Baldwin, T., "Unsupervised word class induction for under-resourced language: a case study on Indonesian", In Proceedings of The 6th International Joint Conference on Natural Language Processing (IJCNLP 13), Nagoya, Japan, 2013.

[15] Molino, P., Basile, P., Caputo, A., Lops, P., and Semeraro, G., "Exploiting distributional semantic model in question answering", In Proceedings of IEEE 6th International Conference on Semantic Computing, IEEE Press, 2012, pp. 146-153.

[16] Muthukrishnan Ramprasath and Hariharan, S. 2012. Using ontology for measuring semantic similarity for question answering system. In Proceedings of IEEE International Conference on Advanced Communication Control and Computing Technology (ICACCCT 12), pp. 218-223.

[17] Najmi, E., Hashmi, K., Khazalah, F., and Malik, Z., "Intelligent semantic question answering system", In Proceeding of IEEE International Conference on Cybernetics, 2013, pp. 255-260.

[18] Pakray, P., Pal, P., Bandhyopadhay, S., and Gelbukh, A., "Automatic answer validation system on English Language", In Proceeding of IEEE 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE 10), IEEE Press, 2010, pp. 329-333.

[19] Ramprasath, M., and Hariharan, S., "Using ontology for measuring semantic similarity for question answering system", In Proceedings of IEEE International Conference on Advanved Communication Control and Computing (ICACCCT 12), IEEE Press, 2012, pp. 218-223.

[20] Ramprasath, M., and Hariharan, S., "Improving QA performance through semantic reformulation", In Proceedings of IEEE International Conference on Engineering, Nirma University, Dec. 2013, pp. 1-4.

[21] Silva, J., Coheur, L., Mendes, A., and Wichert, A., "From symbolic to sub-symbolic information in question classification", Artificial Intelligence Review, vol. 35, no.2, 2011, pp. 137-154.

[22] Tartir, S., Arpinar, I.B., and McKnight, B., "Semantic QA: Exploiting semantic associations for cross-document question answering", In Proceedings of IEEE 4th International Symposium on Innovation in Information and Communication Technology, 2011, pp. 1-6.

[23] Te, R., "Research on question classification method of Tibetan online automatic question answering system", In Proceedings of IEEE 4th International Conference on Intelligent Networks and Intelligent System, IEEE Press, pp. 211-213, doi: 10.1109/ICINIS.2011.42.

[24] Walke, P.P., and Karale, S., "Implementation approach for various categories of question answering system", In Proceeding of IEEE Conference on Information and Communication Technology (ICT 2013), pp. 402-407.

[25] We, Z., Xuan, Z, Wei, Z., and Junjie, C., "Design and implementation of influenza question answering system on multi-strategies", In Proceedings of IEEE International Conferences, IEEE Press, 2012, pp. 720-722.

[26] Yunjuan, L., Lijuan, M., Lijun, Z., and Qinlin, M., "Research and application of information retrieval techniques in intelligent question answering system", In Proceedings of IEEE International Conference, 2011, pp. 188-190.

[27] Zhang, G., Zhang, W., Bai, Y., Kang, S., and Wang, P., "An open domain question answering system for NTCIR-8 C–C task", In Proceedings of IEEE NTCIR-8 Workshop Meeting, Tokyo, Japan, 2010.

[28] Zulen, A. A., and Purwarianti, A., "Study and implementation of monolingual approach on Indonesian question answering for factoid and non-factoid question", In Proceedings of IEEE 25th Pacific Asia Conference on Language, Information , and Computational (PACLIC 2011), 2011, Singapore.

[29] Zulen, A.A, and Purwarianti, A., "Using phrase-based approach in machine learning based factoid Indonesian question answering", In Proceedings of CISAK, 2013.