Ontology-Based Sentence Extraction for Answering Why-Question

A. A. I. N. Eka Karyawati
Department of Computer Science
Faculty of Mathematics and Natural Sciences, Udayana University
Bali, Indonesia
eka.karyawati@cs.unud.ac.id

Azhari
Department of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences, Gadjah Mada University
Yogyakarta, Indonesia

Edi Winarko
Department of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences, Gadjah Mada University
Yogyakarta, Indonesia

Agus Harjoko
Department of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences, Gadjah Mada University
Yogyakarta, Indonesia

***Abstract*—Most studies on why-question answering system usually used the keyword-based approaches. They rarely involved domain ontology in capturing the semantic of the document contents, especially in detecting the presence of the causal relations. Consequently, the word mismatch problem usually occurs and the system often retrieves not relevant answers. For solving this problem, we propose an answer extraction method by involving the semantic similarity measure, with selective causality detection. The selective causality detection is applied because not all sentences belonging to an answer contain causality. Moreover, the motivation of the use of semantic similarity measure in scoring function is to get more moderate results about the presence of the semantic annotations in a sentence, instead of 0/1. The semantic similarity measure employed is based on the shortest path and the maximum depth of the ontology graph. The evaluation is conducted by comparing the proposed method against the comparable ontology-based methods, i.e., the sentence extraction with Monge-Elkan with 0/1 internal similarity function. The proposed method shows the improvements in term of MRR (16%, 0.79-0.68), P@1 (15%, 0.76-0.66), P@5 (14%, 0.8-0.7), and Recall (19%, 0.86-0.72).**

Keywords— why-question answering; ontology-based sentence extraction; semantic similarity measure; Monge-Elkan similarity; sentence scoring

I. INTRODUCTION

Why-question is a complex (i.e., non-factoid) question. Different from a factoid question that has a short answer (i.e., some phrases), the why-question needs a textual explanation answer (i.e., sequence of some sentences).

Most studies on why-question answering system usually used the keyword-based approaches. They rarely involved domain ontology in capturing the semantic of the document contents, especially in detecting the presence of the causal relations. Consequently, the word mismatch problem usually occurs and the system often retrieves not relevant answers.

For solving this problem, we propose an answer extraction method by involving the semantic similarity measure, with selective causality detection. The selective causality detection is applied because not all sentences belonging to an answer contain causality.

Thus, the main contribution of this research is a scoring method for extracting the sentences that contain answers to a why question. The scoring method employs the semantic similarity measure between the semantic annotations of sentences and of a why-question, with selective causality detection. Semantic annotations of a sentence or a question are semantic entities that annotate the sentence or the question, respectively. The selective causality detection is applied because not all sentences belonging to the extracted answer candidates (i.e., paragraphs) contain causality.

Moreover, the motivation of using the semantic similarity in the scoring function is to get more moderate results about the presence of the semantic annotations in a sentence, instead of 0/1. A semantic entity is said to be present in a sentence, not only if there are exactly similar terms occur but also if there are other terms that semantically similar occur in the sentence. In this research, the exactly similar terms are defined as the terms that label the similar semantic entity. The semantically similar terms are defined as the terms that have the semantic similarity value between them less than a specific threshold.

In our research, the sentence extraction is the second step of answer extraction phase of the three phase why-QA method (i.e., question analysis [2], document retrieval, and answer extraction). The first step of the answer extraction is a paragraph extraction that has a scoring formula similar to the document ranking formula.

The proposed method is applied in a specific domain (i.e., Text Retrieval domain). The Text Retrieval domain ontology used is this research has been built by [2]. The domain lexicon of the Text Retrieval ontology is used as a basis for identifying

978-1-5386-0549-3/17/$31.00 ©2017 IEEE

the semantic annotations of sentences. Moreover, the ontology schema is employed to measure the semantic similarity between two semantic entities.

The evaluation is conducted by comparing the proposed method against the comparable ontology-based methods, i.e., the sentence extraction with Monge-Elkan with 0/1 internal similarity function, and against the baseline method, i.e., the sentence extraction with semantic-annotation-detection-based ranking.

The rest of this paper is structured as follows. The related works in answer extraction for why-question answering is presented in Section 2. Section 3 explains the proposed ontology-based sentence extraction method. The results and discussion are explained in Section 4. Finally, the conclusion of this paper is presented in Section 5.

II. Related Works

Most of the proposed why-question answering methods based on document collection are keyword-based methods, see Table 1. Mori et al. (2008) identified the answer candidates (ACs) as the longest series of sentences corresponding to a seed that satisfies the conditions, such as the seed is in the series sentences and every sentence has a score greater than a threshold. Nakakura & Fukumoto (2008) identified the ACs by extracting all paragraphs in the top-30 documents that match with the extraction patterns.

In the scoring task, since the appropriateness of long ACs can be estimated by a combination of, at least, the two measures including the relevance of the ACs to the topic of the question and the appropriateness of the writing style of the ACs [5], some researchers [3, 6, 7] have tried to define scoring function by combining the two measures. Soricut & Brill, (2006) used the N-gram co-occurrences statistics to define the relevance score and used the statistical question/answer translation to define the appropriateness of writing style score.

TABLE 1. Summary of why-question answering approaches

Reference	Answers Extraction	Scoring
(Soricut & Brill, 2006)	Segmenting documents into sentences.	Combination of relevance and appropriateness score
(Mori et al., 2007)	Combination of lexical chains and lexico-syntactic patterns.	Combination of relevance and appropriateness score
(Murata et al., 2007)	Segmenting all one-, two and three continuous paragraphs into sentences	TF, distance between the question terms and length of AC of terms
(Shima & Mitamura, 2007)	Segmenting documents into sentences	machine learning
(Higashinaka & Isozaki, 2008)	Segmenting documents into sentences	SVM ranker
(Mori et al., 2008)	The longest series of sentences corresponding to a seed	Maximal sentence score of the series
(Nakakura & Fukumoto, 2008)	Matching extraction patterns	Cosine similarity measure
(Verberne et al., 2010)	Disjoint and sliding passages	Several machine learning techniques
(Oh et al., 2012)	Segmenting documents into sentences	Murata methods and SVM
(Oh et al., 2013)	Similar to Oh et al., 2012	Murata methods and re-ranking using SVM

On the other hand, Mori et al., (2007) used the document score from the search engine to define the relevance score and used the lexico-syntactic patterns and extracted clue predicates to define the appropriateness of writing style score. However, Mori et al., (2008) used the document score measured by using Web search engine results to define the relevance score and using correlation value to define appropriateness of writing style score. Nakakura & Fukumoto, (2008) used cosine similarity measure between the retrieved ACs and the question to rank the ACs.

Other researchers [8-12] used the relevance of a document with respect to the question measure as a basis for ranking the ACs by utilizing the machine learning techniques.

Similar to the most previous researches (see Table 1), we extract answers by segmenting paragraphs within the retrieved documents into sentences. In contrast to some previous researches, the sentence scoring only employs a relevance score, without involving the appropriateness measure for each sentence. The reason is that not all of the sentences relating to an answer contain causality. Consequently, it is not effective to apply the appropriateness of writing style (i.e., causality style) for each sentence.

Different from the previous researches that did not involve a semantic similarity, the relevance score of a sentence of our proposed method is based on the semantic similarity measure. The semantic similarity is the similarity between the semantic annotations of the question and the semantic annotations of the sentence, with selective causality detection.

In the answer selection task, most of researchers only selected one AC, the best scoring of the ACs, as a final answer [3, 6-8, 10, 11]. Other researchers presented the N-top ACs as the final answers [4, 9, 13]. Different from the studies, our proposed method selects the sequences of sentences from some passages (i.e., ACs). The proposed sentence selection method uses two specific threshold values to determine whether a paragraph contains answers and whether a sentence in the selected paragraph is extracted as an answer or not.

III. The Proposed Method

A. Task Definition

Given a question $\mathbf{q}$ and its corresponding paragraphs $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$. The answer extraction is defined as a task to select relevant answers from the paragraphs. Specifically, some sentences that are relevant to the question are selected from each paragraph. The sentence is assumed to be relevant if its scoring value is greater than a specific threshold. The scoring sentence considers the semantic similarity measure between semantic annotations of a question and semantic annotations of the sentence (SenSA).

There are three sets of semantic annotations of a question, which are a set of original semantic annotations (OSA), a set of additional semantic annotations (ASA), and a set of causality annotations (CA) [2]. The OSA is identified from the original question (i.e., inputted question), and the ASA is identified through the query expansion process. Elements of CA annotate causality expressions contained in the question.

The answer is concatenation of sequences of the selected sentences regarding the sentences order got from each paragraph. For example, suppose the selected sentences from $\mathbf{p}_1 = \{\mathbf{s}_{11}, \mathbf{s}_{12}, \ldots, \mathbf{s}_{1m_1}\}$, from $\mathbf{p}_2 = \{\mathbf{s}_{21}, \mathbf{s}_{22}, \ldots, \mathbf{s}_{2m_2}\}$, …, from $\mathbf{p}_t = \{\}$ (i.e., there is no relevant sentence from paragraph $\mathbf{p}_t$), …, and from $\mathbf{p}_n = \{\mathbf{s}_{n1}, \mathbf{s}_{n2}, \ldots, \mathbf{s}_{nm_n}\}$. Thus, the answer is $A = \{\mathbf{s}_{11}, \mathbf{s}_{12}, \ldots, \mathbf{s}_{1m_1}, \mathbf{s}_{21}, \mathbf{s}_{22}, \ldots, \mathbf{s}_{2m_2}, \ldots, \mathbf{s}_{n1}, \mathbf{s}_{n2}, \ldots, \mathbf{s}_{nm_n}\}$.

B. Method

The proposed sentence scoring method based on semantic similarity measure defines a scoring formula as a linear combination of three semantic similarities. The first similarity is a semantic similarity between a set of question focuses (i.e., identified from OSA) and SenSA. The second similarity is a semantic similarity between ASA and SenSA. The third similarity is a similarity between CA and SenSA.

Suppose Sim is the presence of causality in a sentence. Sim($\mathbf{s}_{SA}$, CA) is equal to 1, if the sentence contains a causality, and equal to 0 if the sentence does not contain a causality. SemSim is the semantic similarity between two set of concepts. The semantic-similarity-based scoring function is given by,

$$\text{score}(\mathbf{s},\mathbf{q}) = \text{sim}(\mathbf{s}_{SA},\mathbf{q}_{SA}) = \lambda_1 \text{SemSim}(\mathbf{s}_{SA},\mathbf{q}_f) + \lambda_2 \text{SemSim}(\mathbf{s}_{SA},\text{ASA}) + \lambda_3 \text{Sim}(\mathbf{s}_{SA},\text{CA}) \quad (1)$$

where $\lambda_i \in [0, 1]$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. It is set that $\lambda_1 = 0.35$, $\lambda_2 = 0.50$, and $\lambda_3 = 0.15$ because they work well in our experiments. The terms $\mathbf{s}$, $\mathbf{q}$, $\mathbf{s}_{SA}$, $\mathbf{q}_{SA}$, and $\mathbf{q}_f$ stand for a sentence, a question, a set of semantic annotations of the sentence, a set of semantic annotations of the question, and question focuses, respectively. SemSim is given by,

$$\text{SemSim}(\mathbf{x},\mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} \max_{1\le j\le N} \left(\text{NSemSim}(x_i, y_j)\right) \quad (2)$$

$$\text{NSemSim}(x_1,x_2) = \frac{\text{SemSim}(x_1,x_2)}{\log(2\max\_depth)} \quad (3)$$

$$\text{SemSim}(x_1,x_2) = -\log\left(\frac{|\text{min\_path}(x_1,x_2)|+1}{2\,max\_depth}\right) \quad (4)$$

where semantic similarity between two concepts (i.e., SemSim(x_1,x_2)) uses the semantic similarity measure based on the shortest path and the maximum depth [14]. The shortest path (*min_path*) can be measured by calculating the minimum number of edges separating both concepts [15]. NSemSim(x_i,y_j) is a normalized semantic similarity between x_i and y_j.

In this research, the shortest path between concept (i.e., referring to as class in the domain ontology) *A* and concept *B* is defined as the minimum number of *rdfs:subClassOf* relation that link concept *A* and concept *B*. The *SPARQL* query processing is used to identify the shortest path between concepts (e.g., concept *A* and *B*). In this research, maximum depth of the domain ontology is 7. Thus, the *SPARQL* query is constructed to identify whether the shortest path is 0, or 1, or …, or 7, or there is no *rdfs:subClassOf* relation that link concept *A* and concept *B*.

Equation (2) is a hybrid similarity that combines token-based similarity and internal similarity function (e.g., the semantic similarity measure) for finding the best match of each token (i.e., concept). This approach is similar to Monge-Elkan similarity [16], where the internal function is the normalized semantic similarity function.

The proposed sentence extraction method uses two threshold values. The first threshold value (0.49) is used to determine which paragraphs will be selected and the second threshold value (0.25) is used to determine which sentences belonging to the paragraph will be selected.

A paragraph is selected if there is a sentence belonging to the selected paragraph that contains causality (i.e., causality score = 1) and has similarity score greater than the first threshold value. Fig 1. presents the pseudo-code to extract some sentences from an extracted paragraph. A sentence is extracted if the similarity score of the sentence is greater than the second threshold value (see step 7 in Fig. 1).

```
Input: ParagraphID, Paragraph (i.e., an extracted paragraph)
       Qfocus (i.e., Question Focus), ASA, and CA
       λ1, λ2, λ3

Output: SelectedSentences
        //SelectedSentences has four attributes:
        //<SentenceID, Sentence, SimilarityScore,
        //CausalityScore>

1 : //Split the paragraph into sentences, and save
    //in array Sentences
    Sentences=splitSentence(Paragraph)
2 : for (i=1; i<=Sentences.length) do
3 :    //Set SentenceID
       SentenceID = setSentenceID(i,ParagraphID)
4 :    //Identify the semantic entity (SE) of each sentence
       SEOfSentence = IdentifySE(Sentences[i])
5 :    //Compute the semantic similarity between QFocus and
       //SEOfSentence SimilarityScore1), and between ASA and
       //SEOfSentence (SimilarityScore2)by using Equation (3)
       SimilarityScore1=estimateSemanticSimilarity(QFocus,
                                                   SEOfSentence)
       SimilarityScore2=estimateSemanticSimilarity(ASA,
                                                   SEOfSentence)
6 :    //Determine the causality score of the Sentences[i]
       CausalityScore = 1
       if CA ∈ SEOfSentence, and 0 otherwise
7 :    //Compute the similarity score of the Sentences[i]
       SimilarityScore = λ1*SimilarityScore1 +
       λ2*SimilarityScore2 + λ3*CausalityScore
8 :    //Select the sentences that has similarity score
       //greater than Threshold2
       if (SimilarityScore>Threshold2) then
9 :        add <SentenceID, Sentences[i], SimilarityScore,
           CausalityScore> to SelectedSentences
10:    end if
11: end for
```

Fig. 1. Pseudo-code of the select sentence function.

C. Experiments

The evaluation is conducted by comparing the proposed method against a comparable ontology-based methods that is the sentence extraction with Monge-Elkan with 0/1 internal similarity function (i.e., the alternative method), and against a baseline method that is the sentence extraction with semantic-annotation-detection-based ranking.

The Monge-Elkan with 0/1 internal similarity function method: The scoring formula of this method is similar to the semantic-similarity-based ranking (see (1)), but the internal similarity function (the normalized semantic similarity, i.e., NSemSim function in (2)) is replaced by Sim function (equals to 1 if two strings are exactly similar and 0 otherwise).

The semantic-annotation-detection-based ranking: The scoring formula of this method is defined as a linear combination of three similarities: first, a similarity between QFocus and SenSA (i.e., equal to 1 if all elements of QFocus are present in SenSA, and 0 otherwise); second, a similarity between ASA and SenSA (i.e., equal to 1 if at least one of elements of ASA is present in SenSA, and 0 otherwise); and third, a similarity between CA and SenSA (i.e., equal to 1 if at least one of elements of CA is present in SenSA, and 0 otherwise). Thus, the semantic-annotation-detection-based scoring formula is given by,

$$\begin{aligned}&\text{score}(\mathbf{s},\mathbf{q}) = \text{Sim}(\mathbf{s}_{SA},\mathbf{q}_{SA})\\&= \lambda_1 \text{Sim1}(\mathbf{s}_{SA},\mathbf{q}_f) + \lambda_2 \text{Sim2}(\mathbf{s}_{SA},\text{ASA}) + \lambda_3 \text{Sim2}(\mathbf{s}_{SA},\text{CA})\end{aligned} \quad (5)$$

where $\lambda \in [0,1]$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. In this research, the λ_1, λ_2, and λ_3 are set as follows: $\lambda_1 = 0.35$, $\lambda_2 = 0.50$, and $\lambda_3 = 0.15$. It is because they perform well in our experiment.

Similar to the semantic-similarity-based sentence selection, the Monge-Elkan with 0/1 internal similarity function and the semantic-annotation-detection-based sentence extraction method also use two threshold values. The first threshold value is used to determine which paragraphs will be selected, and the second threshold value is used to determine which sentences belonging to the paragraph will be extracted. The first threshold value is also set to be 0.49 and the second threshold value to be 0.25.

The proposed sentence extraction method is evaluated by using the dataset of the pairs of a question and the corresponding a list of relevant answers. The list of relevant answers is a list of sentence ID. The sentence ID is constructed by concatenating the document ID, the paragraph ID, and the sentence ID, sequentially. For instance, a sentence ID DOC0039PAR0779SEN0001 is the ID of the first sentence of the paragraph that has ID DOC0039PAR0779. On the other words, the document ID is DOC0039, the paragraph ID is DOC0039PAR0779, and the sentence ID is DOC0039PAR0779SEN0001.

The evaluation is performed by conducting some experiments to measure the effectiveness and efficiency of the methods. The effectiveness of the methods is estimated by calculating the five standard evaluation measures, *MRR* (Mean Reciprocal Rank), *P@1*, *P@5*, *Precision*, and *Recall* of each method [17-19]. Moreover, the efficiency of the methods is estimated by calculating the runtime of the system when the method is executed.

In this research, because the sentences that contain answer are more than one, and the system should retrieve all the relevant answers, the system should position all of the relevant answers at "the first rank". Consequently, in this research, "the first rank" is defined as not only the actual first position but also other positions depend on the number of the relevant answers. If the number of the relevant answers is *M*, the answer that is in the 1st, the 2nd, …, or the *M*th position has the ranking at "the first rank". If the answer is in the $(M+1)^{th}$, the $(M+2)^{th}$, …, or the $(M+i)^{th}$ position, the answer is ranked at *"*the second rank*"*, *"*the third rank*"*, …, or *"*the $(i+1)^{th}$ rank*"*, respectively. Hence, the formula of *MRR* and *P@k* [17-19] are modified as,

$$MRR = \frac{\sum_{i=1}^{M} \frac{\sum_{j=1}^{N} MRR_{i,j}}{N}}{M} \quad (6)$$

$$P@k = \frac{\sum_{i=1}^{M} \frac{\sum_{j=1}^{N} P@k_{i,j}}{N}}{M} \quad (7)$$

where *M* is the number of questions, *N* is the number of relevant answers (i.e., the value of *N* is variety depend on the questions). $MRR_{i,j} = \frac{1}{R_{i,j}}$, where $R_{i,j}$ stands for the rank of the j^{th} relevant answer for the i^{th} question. $P@k_{i,j}$ stands for the precision at *k* for the i^{th} question of the j^{th} relevant answer, which is 1 if the j^{th} relevant answer is found in the top-*k* answers (i.e., the answers to the i^{th} question), and 0 otherwise.

The evaluation performances are the average values of each measure. The formula of the average measure, $\overline{f}$, is given by,

$$\overline{f} = \frac{\sum_{i=1}^{M} f_i}{M} \quad (8)$$

where *M* stands for the number of iterations (i.e., 20), f_i is the function of *MRR*, *P@k* (k = 1, 5, 10), and *RunTime* of the i^{th} iteration.

The experiments are conducted by generating randomly 10, 20, 30, and 40 questions from the why-question collection in 10 iterations, where the total number of questions available is 5921 why-questions. The question collection is constructed through three steps, firstly, why-questions (i.e., the general domain questions) is collected from the Web; secondly, the why-questions are analyzed to identify general patterns of the why-questions; and thirdly, the why-questions in a specific domain (i.e., Text Retrieval) are generated using the general patterns [2]. The general domain why-questions are collected from the Verberne's data collection [20]. This collection of why-questions is based on the Webclopedia question collection [21]. Other Verberne's data collection used is user-generated why-questions that obtained from the Microsoft RFP data set [22].

In the evaluation process, a question is executed over a list of experimented relevant paragraphs. The lists of experimented relevant paragraphs are obtained from the paragraph extraction process of the why-question answering, which is our previous research.

IV. Results and Discussions

Table 2 shows the evaluation results of the proposed sentence extraction against the two other methods. Values in bold correspond to the best results for the corresponding metrics. As can be seen, for all variety number of data used, the average values of *MRR* of the proposed method (i.e., around 0.79) are higher than the average values of *MRR* obtained from the alternative and the baseline method (i.e., around 0.68). In average, the relevant sentences extracted by using the proposed method, are at position 1/0.79 = 1.26 or in average at the 1st or the 2nd rank position (i.e., ideal result should be at the first rank).

As shown in Table 2, the *P@1* value resulted by the proposed method is around 0.76. It means 76% of the sentences extracted are at the first rank position. On the other words, if there are *N* relevant sentences, 0.76**N* sentences that have a position in the rank 1 until the rank *N* of the ranked results are relevant. Compared to the alternative method, the proposed method can improve the *P@1* about 15% (0.76-0.66). The baseline method shows the *P@1* (0.59) smaller than the alternative method.

Furthermore, the proposed method shows that 80% of the total sentences extracted by the proposed method are the top-5 sentences (*P@5*=0.80), better than the alternative method, which only around 70% of the sentences extracted, are the top-5 sentences. The baseline method returns P@5 (0.76) better than the alternative method.

TABLE 2. Comparison results of the proposed sentence extraction methods against the other methods

	Metrics	The Proposed Method	The Alternative Method	The Baseline Method
Data=10	MRR	**0.770353**	0.663619	0.685405
	P@1	**0.736341**	0.625354	0.617107
	P@5	**0.799198**	0.709012	0.761058
	Precision	**0.707390**	0.777973	0.580176
	Recall	**0.856699**	0.709012	0.819953
	RunTime (s)	207.022280	8.227720	8.715520
Data=20	MRR	**0.801243**	0.716835	0.653711
	P@1	**0.775452**	0.689924	0.575951
	P@5	**0.813567**	0.748332	0.754535
	Precision	**0.715970**	0.789894	0.614831
	Recall	**0.865512**	0.753332	0.815653
	RunTime (s)	204.647150	8.333785	8.698520
Data=30	MRR	**0.788261**	0.675622	0.648712
	P@1	**0.767409**	0.656700	0.580917
	P@5	**0.794559**	0.694586	0.729630
	Precision	**0.699685**	0.829231	0.586305
	Recall	**0.851411**	0.724586	0.810445
	RunTime (s)	180.590090	8.034453	8.431870
Data=40	MRR	**0.817463**	0.680311	0.696240
	P@1	**0.796530**	0.657758	0.635496
	P@5	**0.827133**	0.704942	0.770414
	Precision	**0.729733**	0.805082	0.624732
	Recall	**0.872966**	0.719942	0.827983
	RunTime (s)	191.402290	8.265127	8.582880

The proposed method shows 19% improvements in term of Recall (0.86-0.72) over the alternative method. The proposed method also shows the improvements both in term of Precision (15%, 0.70-0.61) and in term of Recall (6%, 0.86-0.81) over the baseline method. However, the alternative method is better than the proposed method in term of Precision (14%, 0.8-0.7).

Based on the experiment results, it can be said that the use of the semantic similarity for estimating the relevance of a sentence with respect to a question can improve the performance of the sentence extraction system. The reason for that is because there are usually more sentence terms which are semantically similar to the question terms than the sentence terms which are only labeledly-similar (i.e., they label the same semantic entity) to the question terms. The two terms are semantically similar if they do not only label the same semantic entities but also label different semantic entities that semantically similar (i.e., the semantic similarity value between the semantic entities is less than a specific threshold).

However, the exact (0/1) internal similarity function (i.e., the alternative method) returns the best performance in term of *Precision* (0.80). The reason for that is that if the alternative method returns a sentence (with small *Recall*, about 0.72), the sentence contains entities that are labeledly-similar to most entities in OSA and ASA. Thus, the possibility of retrieving relevant sentences is high (*Precision* is about 0.80).

The overall of the experiment results of the proposed sentence extraction method are good (mostly>0.75) but Precision (*Precision* = 0.70). It is because the questions used in the experiments are in well-ordered forms, the question patterns, and the concepts and relations contained in the questions are recognized by the system.

However, the *Precision* of the proposed method is only 0.70; it means that the system extracts fairly high amount (30%) of non-relevant sentences. It shows that the sentence ranking approach still cannot effectively select only the relevant sentences. Thus, even though most of the relevant sentences extracted (can be seen from the Recall value that around 0.86), but the significant number of the non-relevant sentences also extracted. The scoring method should be modified to get more effective approach.

Besides estimating the effectiveness of the proposed methods by comparing the methods based on the five performance metrics (i.e., *MRR*, *P@1*, *P@2*, *Precision*, and *Recall*) as explained above, the efficiency of the methods is also estimated by comparing the average values of runtimes among the four methods. As can be seen in Table 2, the efficiency of the proposed method is the lowest returning runtime about 200 seconds, compared to the alternative and the baseline method returning *runtime* about 8.2 and 8.5 seconds, respectively. It means that the proposed method consumes time around 24 times longer than the others.

The reason why the runtime of the proposed method is extremely high is that the method involves the semantic similarity, where the implementation of the semantic similarity computation is time-consuming. As have been

explained, the semantic similarity is estimated by using repeated *SPARQL* query processing over the knowledge base of the domain ontology underlying the system.

V. Conclusion and Future Works

In the sentence extraction, the use of the semantic similarity can improve the performance of the system, but the *Precision*. The best results in term of *MRR* (0.79), *P@1* (0.76), *P@5* (0.80), and *Recall* (0.86), are returned by the proposed sentence extraction method which employs a scoring method that involves a semantic similarity measure with selective causality detection. However, the highest result of Precision (0.80) is returned by Monge-Elkan with the 0/1 internal function method. The reason is that if the method (i.e., the alternative method) returns a sentence, the sentence will contain entities that are labeledly-similar to most entities in OSA and ASA. Consequently, the possibility of retrieving relevant sentences (i.e., w.r.t. precision) is high.

Besides the strengths of the proposed sentence extraction method, there are some noticeable weaknesses of the proposed method. The performance overhead of the proposed method is 200s, as the method involves semantic similarity computation. The semantic similarity is estimated by using repeated *SPARQL* query processing over KB, which is time-consuming.

Out of the drawbacks, this research has proved that the semantic similarity measure can be employed to improve the sentence extraction performance of a why-QA system.

To reduce the performance overhead, the method for estimating the semantic similarity measure can be modified by using other techniques, instead of the shortest path algorithm based on *SPARQL* query processing (i.e., by using Dijkstra algorithms or investigating other algorithms).

References

[1] D. Molla. and J. L. Vicedo, "Special section on restricted-domain question answering", Journal of Computational Linguistics, Vol. 33, No. 1, pp. 41-61, 2007.

[2] A.A.I.N.E. Karyawati, E. Winarko, Azhari, and A. Harjoko, "Ontology-based why-question analysis using lexico-syntactic patterns", International Journal of Electrical and Computer Engineering (IJECE), Vol. 5, No. 2, pp. 318-332, 2015.

[3] T. Mori, M. Sato, and M. Ishioroshi, "Answering any class of Japanese non-factoid question by using the Web and example Q&A pairs from a social Q&A website", In Proceedings of 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 59-65, Sydney, Australia, 2008.

[4] S. Nakakura and J. Fukumoto, "Question answering system beyond factoid type questions", In the 23rd International Technical Conference Circuits/Systems (ITC-CSCC 2008), pp. 617-620, Yamaguchi, Japan, 2008.

[5] K.-S. Han, Y.-I. Song, and H.-C. Rim, "Probabilistic model for definitional question answering", In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06), pp. 212–219, 2006.

[6] R. Soricut and E. Brill, "Automatic question answering using the Web: beyond the factoid", Journal of Information Retrieval - Special Issue on Web Information Retrieval, Vol. 9, No. 2, pp. 191–206, 2006.

[7] T. Mori, M. Sato, M. Ishioroshi, Y. Nishikawa, S. Nakano and K. Kimura, "A monolithic approach and a type-by-type approach for non-factoid question-answering", In Proceeding of NTCIR-6 Workshop Meeting, pp. 469-476, Tokyo, 2007.

[8] H. Shima and T. Mitamura, JAVELIN III: answering non-factoid questions in Japanese", In Proceeding of NTCIR-6 Workshop Meeting, pp. 464-468, Tokyo, 2007.

[9] R. Higashinaka and H. Isozaki, "Corpus-based question answering for why-questions", In Proceedings of International Joint Conference on Natural Language Processing (IJCNLP), pp. 418-425, Hyderabad, 2008.

[10] S. Verberne, L. Boves, N. Oostdijk, and P. Coppen, "What is not in the bag of words for why-QA?", Computational Linguistics, Vol. 32, No. 2, 229–245, 2010.

[11] J. –H. Oh, K. Torisawa, C. Hashimoto, T. Kawada, S. De Saeger, J. Kazama, and Y. Wan, "Why question answering using sentiment analysis and word classes", In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural language Processing and Computational Natural language Learning, pp. 368-378, Jeju Island, Korea, 2012.

[12] J.–H. Oh, K. Torisawa, C. Hashimoto, M. Sano, S.D. Saeger, and K. Ohtake, "Why-question answering using intra- and inter-sentential causal relations", In Proceeding of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 1733-1743, Bulgaria, 2013.

[13] M. Murata, S. Tsukawaki, T. Kanamaru, Q. Ma, and H. Isahara, "A system for answering non-factoid Japanese questions by using passage retrieval weighted based on type of answer", In Proceeding of NTCIR-6 Workshop Meeting, pp. 477-482, Tokyo, 2007.

[14] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification", In WordNet: an electronic lexical database, MIT Press, pp. 265–283, 1998.

[15] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and application of a metric on semantic nets", IEEE Transactions on Systems, Man and Cybernetics, Vol. 19, No. 1, pp. 17–30, 1989.

[16] Monge and C. Elkan, "The field-matching problem: algorithm and applications", In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 267-270, Portland, Oregon, 1996.

[17] C.D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval, Cambridge University Press, New York, 2008.

[18] R. Baeza-Yates and B. Ribeiro-Neto, 1999, Modern information retrieval, ACM Press, New York.

[19] J.A. Thom and F. Scholer, 2007, "A comparison of evaluation measures given how users perform on search tasks", In Proceedings of the 12th Australasian Document Computing Symposium, pp. 100-103, Melbourne, Australia.

[20] S. Verberne, L. Boves, N. Oostdijk, and P. Coppen, "Evaluating discourse-based answer extraction for why-question answering", In Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'07), pp. 735-736, Amsterdam, 2007.

[21] E. Hovy, U. Hermjakob, and D. Ravichandran, 2002, "A question/answer typology with surface text patterns", In Proceedings of the Human Language Technology confer-ence (HLT), San Diego, CA.

[22] S. Verberne, "Learning to rank for why-question answering", Journal of Information Retrieval , Vol. 14, pp. 107–132, 2011.