INTISARI

Why-question adalah pertanyaan *non-factoid* yang memerlukan jawaban berupa penjelasan. Jawaban-jawaban itu biasanya tersebar dalam beberapa dokumen. Sehingga, metode yang sesuai untuk menjawab *why-question* adalah metode QA yang berbasis IR. Tetapi, ada dua masalah utama dalam metode QA yang berbasis IR, yaitu masalah *word mismatch* dan masalah jawaban yang berulang/berlebihan (*redundant*) and terfragmentasi *(fragmented*).

Untuk menyelesaikan masalah *word mismatch*, diusulkan metode QA yang berbasis semantik yang dikombinasikan dengan metode deteksi kausalitas. Selanjutnya, untuk menyelesaikan masalah jawaban yang berulang/berlebihan and terfragmentasi, diusulkan metode KI yang berbasis text yang dikombinasikan dengan metode IR yang berbasis ontologi dalam rangka mengintegrasikan jawaban-jawaban yang diperoleh.

Terdapat tiga fase utama dalam metode yang diusulkan. Pertama, *why-question* dianalisa untuk mengubahnya ke dalam bentuk representasi *triples* yang sesuai dengan skema ontologi, dan mengekspansi pertanyaan menggunakan *SPARQL query processing*. Fase ini menghasilkan anotasi semantik dari pertanyaan (i.e., OSA, ASA, dan CA). Kedua, dengan menggunakan anotasi semantik dari pertanyaan, dicari dokumen-dokumen yang mengandung jawaban. Ketiga, dokumen-dokumen tersebut disegmentasi menjadi kalimat-kalimat. Kalimat-kalimat tersebut diskoring dan dipilih untuk memperoleh kalimat-kalimat jawaban. Dan selanjutnya, kalimat-kalimat jawaban tersebut diintegrasikan menjadi satu jawaban yang terintegrasi menggunakan metode integrasi pengetahuan berbasis teks.

Evaluasi dilakukan pada setiap fase dari metode yang diusulkan. Pada fase *question analysis*, metode representasi semantik dari pertanyaan menghasilkan nilai evaluasi yang bagus yaitu 0.98 (*Precision*), 0.98 (*Recall*), 0.15 (*Undergeneration*), dan 0.1 (*Overgeneration*). Pada fase *document retrieval*, terlihat adanya perbaikan yang signifikan hasil dari metode yang diusulkan terhadap metode berbasis text, pada nilai-nilai *MRR* (81 kali), *P@1* (9,4 kali), *P@5* (7 kali), dan *P@10* (6,4 kali), dan terhadap metode berbasis ontology, pada nialai-nilai *MRR* (80%), *P@1* (153%), *P@5* (45%), dan *P@10* (33%). Pada fase *sentence extraction*, terlihat adanya perbaikan terhadap metode *Monge-Elkan* dengan fungsi internal 0/1 pada nilai-nilai *MRR* (16%), *P@1* (15%), *P@5* (14%), and *Recall* (19%). Terakhir, pada fase *answer integration* juga menunjukan hasil yang cukup bagus, pada nilai-nilai *Precision* (0.83), *Recall* (0.78), *Undergeneration* (0.14), dan *Overgeneration* (0.09).

Keywords: *why-question*, *lexico-syntactic patterns*, *proximity-based causality detection*, *ontology-based information retrieval*, *text-based KI*

ABSTRACT

Why-question is a complex (i.e., non-factoid) question that needs a textual explanation answer. The answers usually scatter over a document collection. Thus, a method that is suitable to answer the why-question is the IR-based QA method. However, there are two main problems in IR-based QA method, including the word mismatch problem, and redundant and fragmented answers.

Therefore, to solve the word mismatch problem, the semantic-based QA using ontology-based IR method combined with causality detection is proposed. Furthermore, to solve the redundant and fragmented answers (i.e., not integrated answers), a text-based KI method is incorporated into the ontology-based IR for integrating the answers.

There are three main phases in the proposed method. Firstly, a why-question is analyzed to convert it into triples-based representation compliant with the domain ontology, and then expand the question by executing the *SPARQL* of the why-question over the KB. The phase returns semantic annotations of the question (OSA, ASA, and CA). Secondly, a list of documents is retrieved based on the OSA, ASA, CA, and the semantic index of the document collection constructed based on the domain ontology-lexicon. Thirdly, the documents are segmented into sentences, the sentences are scored, and selected to obtain a list of sentence answers, and then the sentences are integrated using the text-based KI method into an integrated natural language answer.

The evaluations are conducted in each phase of the method and show some results. Firstly, the triples construction method shows the good performance of the semantic representation (i.e., the set of semantic triples and the set of semantic annotations) of why questions. The semantic representation can be effectively identified, where the value of *Precision*, Recall, *Undergeneration,* and *Overgeneration* is 0.98, 0.98, 0.15, and 0.1, respectively. Secondly, the proposed document retrieval method shows the significant improvement over the baseline method (i.e., the phrase-based method) in term of *MRR* (81 times), *P@1* (9.4 times), *P*@5 (7 times), and *P*@10 (6.4 times), and over the ontology-based approach in term of *MRR* (80%), *P@1* (153%), *P*@5 (45%), and *P*@10 (33%). Also, the proposed sentence extraction method shows the improvement over the Monge-Elkan with the 0/1 internal function method in term of *MRR* (16%), *P*@1 (15%), *P*@5 (14%), *and Recall* (19%). Finally, the proposed answer integration method shows the good results in term of *Precision* (0.83)*, Recall* (0.78)*, Undergeneration* (0.14)*,* and *Overgeneration* (0.09).

Keywords: why-question, lexico-syntactic patterns, proximity-based causality detection, ontology-based information retrieval, text-based KI