Information Processing and Management 000 (2016) 1–18

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Review on the advancements of disambiguation in semantic question answering system

Sofian Hazrina[a,*], Nurfadhlina Mohd Sharef[a], Hamidah Ibrahim[a], Masrah Azrifah Azmi Murad[a], Shahrul Azman Mohd Noah[b]

[a] *Department of Computer Science, Faculty of Computer Science and Information Technology, University Putra Malaysia, Selangor, Malaysia*
[b] *Knowledge Technology Research Group, Center for Artificial Intelligent Technology, Faculty of Information Science & Technology, National University of Malaysia, Selangor, Malaysia*

ARTICLE INFO

Article history:
Received 7 September 2015
Revised 12 June 2016
Accepted 19 June 2016
Available online xxx

Keywords:
Natural language question
Semantic question answering
Ambiguity types
SQA disambiguation solution

ABSTRACT

Ambiguity is a potential problem in any semantic question answering (SQA) system due to the nature of idiosyncrasy in composing natural language (NL) question and semantic resources. Thus, disambiguation of SQA systems is a field of ongoing research. Ambiguity occurs in SQA because a word or a sentence can have more than one meaning or multiple words in the same language can share the same meaning. Therefore, an SQA system needs disambiguation solutions to select the correct meaning when the linguistic triples matched with multiple KB concepts, and enumerate similar words especially when linguistic triples do not match with any KB concept. The latest development in this field is a solution for SQA systems that is able to process a complex NL question while accessing open-domain data from linked open data (LOD). The contributions in this paper include (1) formulating an SQA conceptual framework based on an in-depth study of existing SQA processes; (2) identifying the ambiguity types, specifically in English based on an interdisciplinary literature review; (3) highlighting the ambiguity types that had been resolved by the previous SQA studies; and (4) analysing the results of the existing SQA disambiguation solutions, the complexity of NL question processing, and the complexity of data retrieval from KB(s) or LOD. The results of this review demonstrated that out of thirteen types of ambiguity identified in the literature, only six types had been successfully resolved by the previous studies. Efforts to improve the disambiguation are in progress for the remaining unresolved ambiguity types to improve the accuracy of the formulated answers by the SQA system. The remaining ambiguity types are potentially resolved in the identified SQA process based on ambiguity scenarios elaborated in this paper. The results of this review also demonstrated that most existing research on SQA systems have treated the processing of the NL question complexity separate from the processing of the KB structure complexity.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Semantic question answering is the means to retrieve information (Shekarpour et al., 2014) or formulate answers based on an NL question (Hakimov et al., 2013). The NL question that is posed by users is mapped to the Simple Protocol and RDF

[*] Corresponding author.
E-mail addresses: hazrina@gmail.com (S. Hazrina), nurfadhlina@upm.edu.my (N.M. Sharef), hamidah.ibrahim@upm.edu.my (H. Ibrahim), masrah@upm.edu.my (M.A.A. Murad), shahrul@ukm.edu.my (S.A.M. Noah).

http://dx.doi.org/10.1016/j.ipm.2016.06.006
0306-4573/© 2016 Elsevier Ltd. All rights reserved.

Query Language (SPARQL) query structure, and the answer is retrieved from linked resources (Lopez et al., 2010a, b). The answers are presented to the user in NL, or any other required form.

According to Kaufmann and Bernstein (2007), users prefer an SQA system that accepts an NL question to one that uses keywords, phrases and a graphical interface. Furthermore, keyword-based query systems do not accept well-formed sentences; greatly reducing the potential to automatically resolve linguistic ambiguity (Gracia & Mena, 2009). Recent research (Park, Shim & Lee, 2014) also has shown that keyword-based queries lack a clear specification of the relations among words; which leads to user inconvenience because they must perform additional tasks to convey the correct context for their question. This statement is supported by Yahya et al. (2013a), who found that SQA systems promotes user convenience by discovering relevant information in knowledge bases (KBs) or LOD. Based on a previously published in-depth literature review, the ease with which an SQA system can be applied depends on its ability to: (1) handle complex NL queries, (2) automatically disambiguate ambiguity, (3) automatically aggregate data from heterogeneous domains and (4) automatically construct SPARQL queries across heterogeneous resources (Lopez et al., 2012; Shekarpour et al., 2014).

First, an SQA system that accepts NL questions faces the complexity of questions posed by users. There are several types of sentences, i.e., declarative, imperative, interrogative and exclamatory (Hooper, 2007). However, SQA focuses only on imperative sentences and interrogative sentences because they are used by a user to ask questions. Both types of sentence can vary in length and complexity. A number of SQA systems have been developed for processing NL questions before translating them to SPARQL to query the KB(s). However, those systems are generally limited to simple questions (Ferre, 2013).

Second, there are evidences that an SQA system performs three types of disambiguation: linguistic disambiguation, conceptual disambiguation and heterogeneous resource disambiguation. According to Sharef and Mohd (2012), an NL question may have multiple meanings. Therefore, an SQA system must be equipped with a linguistic disambiguation solution to identify the context of the question. For example, consider the NL question 'What is the most delicious cherry?' . The linguistic disambiguation solution must identify the context of the question, which is the name of a fruit, instead of a keyboard model or a movie name (Luo et al., 2014). On the other hand, conceptual disambiguation is needed because the terminology used in an NL question may be different from the terminology used in the KB. For example, one of the questions in a QALD-2 question is “How many inhabitants does Maribor have?” The conceptual ambiguity occurs because the concept used in the KB is “totalPopulation”, instead of “inhabitant” . A conceptual disambiguation solution is required to disambiguate the KB concept while searching for matches from homogeneous or heterogeneous resources (Xu & Pottinger, 2014). Heterogeneous resource disambiguation is a process to select the best KB(s) in the event that multiple KB-concept matches are encountered from multiple resources.

Third, some NL questions that require data to be aggregated from heterogeneous domain and heterogeneous resources (Lopez et al., 2012; Sharef, Noah, & Murad, 2013). For example, an aggregative query such as ’identify thief’ requires data from three domains: criminal record, medical record, and credit card record. A better quality result is produced when the facts from the heterogeneous domains and the heterogeneous resources are considered (Lopez et al., 2010 a,b).

Fourth, a query may require knowledge retrieval from heterogeneous KBs with different schemas. Additionally, each resource may only contain a partial answer. In this type of situation, users rely on the SQA system to construct a SPARQL query across different resources (Lopez et al., 2012).

As summarized above, there are four SQA-system capabilities responsible for users’ easy reliance on the SQA system to query the KBs. However, as mentioned by Kaufmann and Bernstein (2010), some of these capabilities are the major problems in the SQA system, such as, the complexity of the disambiguation solution, the current SQA systems that mostly are built for homogeneous domains and lastly, the complexity of the formal query construction. This is the motivation for this study to understand the existing research in SQA systems. In Section 2, the existing SQA processes are described in detail and the SQA conceptual framework is proposed based on the literature. In Section 3, the complexity of the NL question posed by user is described. In Section 4, the existing works on ambiguity in English and the identification of ambiguity types from multidisciplinary literature is summarized. Once the ambiguity types are identified, in Section 5, the ambiguity scenarios and the ambiguity types that have been resolved in the SQA system are identified. In Section 6, a review of the advanced disambiguation solutions for SQA systems is presented, and in Section 7, the processing of NL question complexity and the processing of KB structure complexity are discussed. Section 8 presents suggestions for future research on the SQA system.

2. Semantic question answering system

An SQA system enables a non-expert user to pose a question in a form of natural language sentence to retrieve the answer from one or more KB(s) or LOD. The SQA system provides ease of use to the user because it can conceal the complex processes used to generate an accurate and comprehensive answer. In recent years, many SQA systems have been developed with different sets of processes based on the primary focus of the study. This section aims to understand the existing SQA processes that had been applied to query KB(s) or LOD.

In this study, an in-depth study on 13 existing SQA systems is performed to understand the scope of the SQA processes and to highlight the existing disambiguation solutions for SQA systems. The selection criteria for this review are the SQA systems that elaborate their entire SQA processes and provide disambiguation solutions. Fig. 1 illustrates the SQA conceptual framework which is formulated based on the existing processes of the SQA systems that were developed by experts in the field.

Download English Version:

https://daneshyari.com/en/article/4966501

Download Persian Version:

https://daneshyari.com/article/4966501

Daneshyari.com