XB: A Large-scale Korean Knowledge Base for Question Answering Systems

Jongmin Lee[1], Youngkyoung Ham[1], Tony Lee[1]

[1] Saltlux Inc.
Daewoong Bldg. 689-4, Yeoksam 1 dong,
Gangnam-gu, Seoul, South Korea
{jmlee, ykham, tony}@saltlux.com

Abstract. There are many studies on question answering system which can answer to natural language questions. Diverse techniques are required for building this system, but it cannot be implemented without well-structured knowledge data. For this reason, we construct a large-scale knowledge base in Korean, with the goal of creating a uniquely Korean question answering system.

1 Introduction

Recently, a variety of Question Answering (QA) systems have been developed, such as IBM Watson and Apple Siri. In these systems, a user inputs a query in natural language, and the QA system searches for the corresponding answer, often using inferences from other related search queries, and provides the user with accurate and relevant information. Most QA systems use a knowledge base to store knowledge studied from a multitude of data.

Extremely large knowledge bases, such as YAGO[1] and Wikidata[2], have been constructed using documents written in English, with the contents well known in the world. However, individual countries require individualized QA systems for their own knowledge.

For example, even though the Eulmi Incident is very significant in Korean history, no knowledge of it is found in the English version of Wikipedia. If there is a question about when Eulmi Insident happened, most of existing knowledge resources cannot answer to it. There is no structured knowledge about that question in Korean DBpedia and Korean Wikipedia only has that information in the text. For this reason, it was necessary to construct a large-scale knowledge base in Korean from various knowledge resources, with the goal of creating a uniquely Korean QA system.

The resulting XB was constructed using the dual-spiral method[3], which allows for both automatic conversion and manual construction simultaneously. In addition, the XB implemented knowledge bases like GeoNames[4], Openstreetmap[5], DBpedia[6] and WikiData. Knowledge in the XB is represented as triple(subject/predicate/object). So far, approximately 200 million triples have been constructed. Through the owl axiom inference(rdfs:subClassOf, rdfs:subPropertyOf, owl:Transitive, owl:inverseOf , owl:disjointWith and etc.), the number of triples are increased by 0.4 billion.

2 Development

The XB is a large-scale knowledge base of common sense level for Korean QA systems, utilizing the ontological method to express knowledge. Figure 2 shows a simple process of our question answering scenario. A user inputs a question in natural language form, and it is converted into a SPARQL using various converting techniques. The converted SPARQL finds answers from the knowledge base.

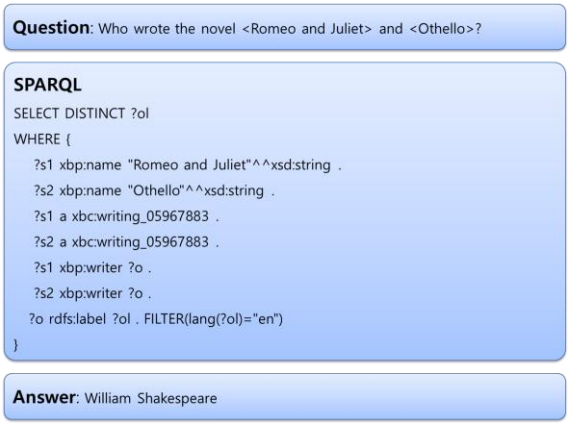

Figure. 1 Part of SPARQL results

The XB is built by the following procedure for the QA scenario. To define classes, we used the hierarchical structure of Korlex[7], WordNet in Korean. Korlex is a lexical database wherein a variety of linguistic relations among synonym, hypernym and hyponym are structured. Classes are chosen by the frequency of searching on each keyword from Korlex and grant relations of higher or not between classes.

Properties refer to YAGO and DBpedia to define key properties based on the frequency of using per property. In addition, a property is added in case it is requested additionally or identified from competency question on the way of constructing the knowledge base.

To build entities, necessary knowledge is extracted from diverse knowledge resources through the rule-based automatic conversion and the curation manually implemented by domain experts, depending on the dual-spiral methodology. Default entities are from Wikipedia pages and are extended, if other resources contain unmapped entities.

The rule-based automatic conversion is a process by which the machine distinguishes between classes and properties through mapping rules between a predefined schema and a knowledge resource to build knowledge.

The curation is a process to additionally verify the automatically converted knowledge or build a new knowledge by human. For example, a main text in a Wiki page written in a natural language is not easily automatically converted. The rule-based automatic conversion and the curation are verified in trade-off for their own results, respectively. Domains that are high-probable to be used in it so that the knowledge related to it can be built primarily, since the core part of knowledge is

constructed based on the Korean Wikipedia. Moreover, the knowledge base has been enlarged with existing knowledge resources such as DBpedia, Wikidata and GeoNames.

Table 1 Knowledge base statistics

Class			Property	
Domain	**URI**	**#Instance**	**URI**	**#Instance**
People	xbc:person_00006026	2,467,831	rdfs:label	19,588,253
Organization	xbc:organization_07523126	972,788	xbp:nation	11,113,066
Event	xbc:event_00025950	407,272	xbp:relatedTerm	7,526,036
Term	xbc:term_05916288	31,339	xbp:description	4875147
Theory	xbc:theory_05637633	1,737	xbp:gender	2,171,672
Literature	xbc:writing_05967883 xbc:book_06013091	579,891	xbp:job	1,974,747
Music	xbc:music_06591368	270,201	xbp:scientificName	1,939,233
Art	xbc:graphic_art_03327573 xbc:work_of_art_04423283	90,930	xbp:bornOn	1,768,723

Table 1 is a part of statistic data about the knowledge base constructed through the above-mentioned processes. Domain refers to the field of knowledge. There are approximately 6,000 classes and approximately 1,000 properties. In addition, there are about 20 million instances that are focused mainly on people, locations, organizations, events, and works.

3 APIs

Generally, a knowledge base based upon ontology uses SPARQL, a standard query language for RDF data. However, it is very difficult for a user who is not familiar with ontology to understand a schema correctly and implement a variety of services utilizing a QA system or a knowledge base through SPARQL. This study provides a variety of APIs other than SPARQL Endpoint to allow a greater number of users to access easily to XB. Table 2 lists the APIs supplied by the XB.

Table 2 List of APIs

API	Description
/api/class	Search class by keywords
/api/classInfo	Get information of a class with its uri
/api/property	Search property by keywords
/api/propertyInfo	Get information of a property with its uri
/api/instance	Search instance by keywords
/api/instanceInfo	Get information of an instance with its uri
/api/instanceTime	Get temporal information of an instance with its uri
/api/instanceSpace	Get spatial information of an instance with its uri
/api/checkType	Check if it is true or false about input instance and class
/api/typeRelation	Inference relationship between two input classes
/api/timeRelation	Inference temporal relationship between two input instances
/api/spaceRelation	Inference spatial relationship between two input instances
/api/shortestPath	Find a shortest path between two input instances

4 Future works

In the near future, additional tools to enhance quality and quantity are expected to be developed.

The knowledge has been completely verified through the curation work, but it is restricted in that a finite number of human ability cannot verify all knowledge in the system. To solve that problem, a crowdsourcing service has been being developed to construct and verify knowledge.

There is also debate as to whether or not to develop massive amounts of knowledge through auto-mapping of a knowledge base featuring a large-scale triploid generated by language processing of knowledge or sentences that are aggregated from different knowledge resources connected with machine learning.

In addition, even if not appearing explicitly in the knowledge base, inferencing rules are defined to analyze relations between pieces of knowledge to generate new knowledge.

The XB has been built mainly with a knowledge resource of Korean language as it is today. However, as most instances are granted with labels and types in English and based on Wikipedia, we believe that it might be relatively easy to extend into Korean if the multi-language link of Wikipedia were used.

The XB will be extended and is expected to be available to public users soon, with a variety of practical applications.

Acknowledge

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP)
(No. R0101-16-0054, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

References

1. Hoffart, J., Suchanek, F. M., Berberich, K., Weikum, G.: YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, Vol 194 (2013) 28-61
2. Vrandečić, D., Markus, K.: Wikidata: a free collaborative knowledgebase. Communications of the ACM (2014) 78-85
3. Kyosung, J., Youngkyoung, H., Kyungil, L.: Dual-Spiral methodology for knowledgebase constructions. International Conference on Big Data and Smart Computing (2016) 477-480
4. Wick, M., Bernard, Vatant.: The geonames geographical database. Available from World Wide Web: http://geonames. Org (2012)
5. Haklay, M., Patrick, W.: Openstreetmap: User-generated street maps. IEEE Pervasive Computing (2008) 12-18
6. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J. et al.: Dbpedia: A nucleus for a web of open data. Springer Berlin Heidelberg (2007) 722-735
7. Yoon, Ae-Sun, et al.: Construction of Korean Wordnet. Journal of KIISE: Software and Applications 36.1 (2009): 92-108.