ISOFT-Team at NTCIR-12 QALab-2: Using Choice Verification

Soonchoul Kwon, Seonyeong Park, Daehwan Nam,
Kyusong Lee, Hwanjo Yu, Gary Geunbae Lee
Department of Computer Science and Engineering,
Pohang University of Science and Technology
San 31, Hyoja-Dong, Nam-Gu, Pohang-Si, 790-784, South Korea
+82-54-279-5581
{theincluder, sypark322, dhnam, kyusonglee, hwanjoyu, gblee}@postech.ac.kr

ABSTRACT
NTCIR QALab is a competition to computationally solve complex real-world questions. We, ISOFT team, perform the task using a choice verification method. The choice verification method evaluates the truthfulness of each choice by calculating three evidence scores using a knowledgebase, information retrieval, and restriction. We use fundamental natural language processing methods without semantic analysis and minimize the need for manual tagging. We ranked 1st in Phase-1 (71/100) and 6th in Phase-3 (38/100) in QALab-2. The errors are due to the non-existence of named entities and a lack of semantic analysis.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Question answering system, Information retrieval, Knowledgebase

General Terms
Experimentation

Keywords
University Entrance Examination, NTCIR QALab, Question Answering System, Choice Verification

1. INTRODUCTION
A question answering (QA) system is a system that attempts to correctly answer natural-language questions. QA methods developed from information retrieval (IR) methods because QA can provide a specific answer, instead of multiple responses containing possible answers. Classical QA systems focus mainly on simple questions; however, today's QA systems can handle complex questions.

NTCIR QALab is a competition to answer complex real-world questions (Shibuki et al., 2016) using computation. The task question set is taken from the Japanese National Center Test, which is used as a university entrance exam and is designed for high school students. The question set is very difficult: students need a wide range of knowledge, a deep understanding of subjects, and the ability to make complex inferences.

In the first year of QALab, researchers proposed approaches using highly complex systems such as passage ranking (Wang et al., 2014) and semantic parsing (Okita and Liu, 2014) to solve complex problems. These techniques can obtain deep understanding of the question and other valid sentences, but are highly abstract and produce many errors, as compared to fundamental analyses such as named entity linking (NEL) and classical IR.

Our approach exploits the simplest method of solving the training data and applies it to the test data. The central step is choice verification, which verifies the choices using a three-part evidence calculator. The knowledgebase (KB), IR, and restriction are used in this method. Other steps require classical and fundamental analyses such as NEL (Daiber et al., 2013), co-reference resolution (Lee et al., 2012), classical IR queries, and KB methods (such as DBpedia). However our system performance is good: we ranked 1st in Phase-1 (71/100) and 6th in Phase-3 (38/100).

In this paper, we explain our participation in the QALab-2 task. First, we describe the task problem and how we interpret the data (Section 2). Then, we explain the methods in our question solving flow, from question understanding to answer selection, including the important choice verification step (Section 3). Then, we review our results and the errors that occurred in the test data (Section 4). Finally, we conclude our research (Section 5).

2. DATA ANALYSIS
The objective is to provide answers to question sets from Japanese National Center Test (大学入試センター試験) for Phase-1 (Year 1999) and Phase-3 (Year 2011). The subject of the tests is world history; we use the English translations provided by National Institute of Informatics (NII). Each set in the training/test data contains texts that are 5-7 sentences long, and different numbers of questions from 36 to 41. Each question contains various numbers of choices from four to six (typically four). The world history questions require straightforward knowledge of facts, but still require a high level of conceptual knowledge and question understanding (Figure 1). All information including section, underlined references, and answer style is tagged in XML format.

Text

Most of the historical sources in early modern Britain are old handwritten documents. Figure (a) below is an example of a historical manuscript written during the (4) 16th century; as can be seen from this, various records have been left using pen and paper, in handwriting styles that differ from those used today.(…)

Question instruction

Question 4 - From (1)-(4) below, choose the most appropriate sentence concerning society and the economy during the time referred to in the underlined portion (4).

Choices

(1) In the 12th century, the system of domain economies based on compulsory labor (Gutsherrschaft) spread through Western Europe, west of the Elbe River.
(2) In Britain, the first enclosure movement progressed, for the purpose of grazing cattle.
(3) A price revolution took place in Europe due to such developments as an influx of silver from the Americas.
(4) As a result of the Emancipation Reform [of 1861], serfs in Russia were emancipated.

Figure 1. Japanese Center Test example

Each question is a multiple-choice question; the system should choose the most correct (or the most incorrect) choice among the given answers. Most questions have references to the text and ask the respondent to answer the question based on the reference text. We have found that some questions can be solved without using the reference text (see Section 3.1).

Some questions have related pictures or labeled maps. These questions can be answered without any picture information; therefore, the pictures are ignored when answering the question. However the labels on the maps are necessary to solve the question, and thus, we do not answer these questions.

The answer style, i.e., the format of each choice, is tagged in the data. Two thirds of the answers have a style of `sentence`. `(symbol-TF)*2` and `term_person` follow (see Table 1).

3. METHODS

Our approach to solving the National Center Test uses choice verification. We confirm that the choice is true or false by calculating its evidence score. Other steps will be explained in following subsections (Figure 2).

3.1 Question Understanding

The question and choices should be understood to allow correct processing of the question and verification of the choices. We use freely available tools and some amount of manual tagging to interpret the questions.

NEL is the most important question interpretation step in this research. We use DBpedia Spotlight[1] to perform automatic NEL (Daiber et al., 2013). DBpedia Spotlight can located phrases as named entities, and disambiguate and map them to entity uniform resource identifiers (URIs) in DBpedia. In this way, we can utilize properties of the entities and find relationships between them. For some named entities for which DBpedia Spotlight fails, we use manual mapping of questions, texts, and choices to ensure system

Table 1.Common answer styles

Answer style	Example choice
`sentence`	The UK promised India self-government after the war.
`(symbol-TF)*2`	a- Correct, b- Incorrect
`term_person`	Batu Khan

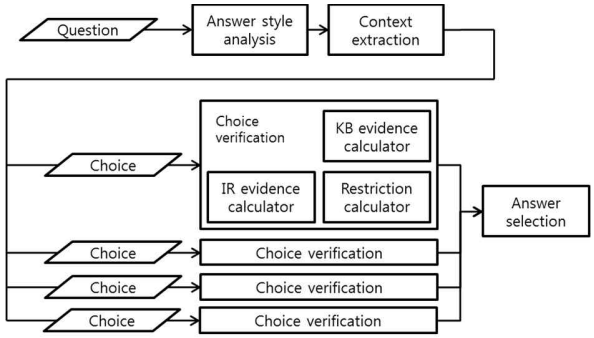

Figure 2. Flowchart to solve a question

accuracy. Manual NEL is the only tagging step in our system; it is critical for system performance. Some mappings are trivial ('the Battle of Plassey' → Battle_of_Plassey), but others are not easy because of synonymous expressions ('Jewish descent' → Who_is_a_Jew?) or different linguistic expressions ('xiangshen' → Landed_gentry_in_China).

Temporal expression is crucial in the restriction calculator (see Section 3.4.3). Some references contain temporal expressions (e.g., 'in the 8th century'), which provide temporal restrictions to verify choices using the restriction calculator. We make rules to extract temporal expressions at the level of centuries, decades, and years. Each time expression is labeled with a start year and an end year (e.g. '8th century' → 701 - 800). If the start or end year is not defined (e.g. 'after 1650'), we label it as negative or positive infinity (e.g. 1650 - ∞).

Most questions ask the respondent to select the most correct choice, but some questions ask to select the most incorrect choice. In the training/test dataset that is translated into English, these questions contain the phrase 'incorrect' or 'least appropriate'. We detect these phrases to determine which choice to select (see Section 3.5).

3.2 Answer Style Analysis

Answer style is a strong clue to determining the strategy used to solve the question. For example, `sentence` choice can be solved by verifying the choice itself. However `(symbol-TF)*2` can be solved by verifying two sentences in the question and selecting the choice that describes the truthfulness of the two sentences. We considered three strategies; 1.) *sentence* to verify `sentence` style questions, 2.) *tf2* to solve `(symbol-TF)*2` style questions, and 3.) *term_other* to solve any term-style choices. We did not answer certain question styles that cannot be solved using our framework or require a deep understanding of maps and symbols. The three strategies differ in detail, but have a very

[1] https://github.com/dbpedia-spotlight

similar approach: they verify the target (choice and/or context) is held in common.

3.3 Context Extraction

We cannot know the truthiness by verifying only the choice itself, because necessary information is included in the instructions or reference text. However, utilizing everything in the instructions and reference text can introduce unnecessary information in (or muddle the differences between) the choices. Context extraction is the step used to extract critical information, avoiding excessive unnecessary information.

In *sentence* and *tf2* strategies, the choice or the sentence in the question to be verified is in the form of a sentence. Most information exists in the target sentence itself; therefore, we use referenced parts as the context only in certain cases: 1.) the reference has a time expression, 2.) the choice has any anaphors, 3.) the reference is a single named entity, or 4.) the reference is shorter than four words (i.e., it is concise). In the t*erm_other* strategy, the choice is a single named entity and important information is in the referenced text and/or the question instruction. We used the reference text and the question instructions as the context.

3.4 Choice Verification

Choice verification is the most important step in this framework. This task requires a wide range of knowledge and multiple strategies to solve. Therefore, we use multiple information sources and strategies. We verify choices to confirm that the evidence indicating truth can be used to solve *sentence* and *term_other* answer style questions. For *tf2* style questions, we verify the two sentences in the question. The three types of evidence from the following evidence calculators are combined to calculate the overall evidence score.

3.4.1 Knowledgebase Evidence Calculator

Knowledgebases such as DBpedia (Daiber et al., 2013, Lehmann et al. 2015) and Wikidata (Vrandečić and Krötzsch, 2014) contain named entities, their properties, and relations among them in the form of triples and triples graphs. Th unit of two entities and their relation is called a triple. The existence of triples of named entities in the choice and the context is strong evidence that the named entities are strongly related.

Using training data analysis, we have found that many *sentence* style choices claim that two or more entities are somehow related. Specifically, two entities from incorrect choices do not have any triples (e.g., 'USA' and 'Tokyo'), rather than explained with an incorrect relation (e.g., 'USA conquered England'). This analysis indicates that finding only a triple between two entities, without any semantic analysis, can lead to correct evidence calculation.

We use DBpedia as the KB in this system. The number of triples among named entity pairs in the choice and the context is used as the evidence score. Duplicated triples between two entities are removed because multiple triples with a large number of common entities will artificially inflate the score. A special relation, `wikiPageWikiLink`, which is labeled when two entities have hyperlinks to each other but the type of relation is undefined, is counted only when there is no triple other than the `wikiPageWikiLink` relation.

Finally, the evidence score is normalized by the number of entities in the choice and the context. Choices with more named entities still show higher evidence values, but this effect is later balanced by the IR evidence calculator.

3.4.2 Information Retrieval Evidence Calculator

IR searches information from unstructured databases such as raw text and indexed documents, rather than from structured databases such as tables and graphs. IRQA has been the closest competitor to KBQA. Compared to KB, IR covers a wider range of information because raw texts are written by different people, without any pre-defined format. On the other hand, the natural language style results of IR are highly variable and sometimes contain false information.

We also focus on IR because the KB has a coverage problem. We use a similar approach in the IR approach, i.e., co-existence of entities is strong evidence. We use a multi-information tagged text database (MITTD) as the text database (Park et al., 2015). Wikipedia text and named entities in MITTD are co-reference resolved and processed by NEL in advance to allow real-time searching. These processes assist in finding more related named entities and related sentences.

We find every named entity in the choice and the context. For the set of entities E, the IR evidence score is

$$\sum_{e \subset E} \log(\max(querycount(title = e, text = E - e), 10))$$

which is the sum of one entity (fixed as the title) and the other entities (searched in the MITTD text). The logarithm and maximum functions are used to normalize and suppress any burst of evidence from a large number of query results. Choices with more named entities usually have lower IR evidence values because there are more search conditions.

3.4.3 Restriction Calculator

Temporal restriction can provide strong evidence in the realm of world history. For example, a choice 'Yan Zhenqing is a calligrapher representative of the Song period' is incorrect because Yan Zhenqing died in the year 785, but the Song period started in 960. According to the temporal information, Yan Zhenqing and the Song period never coexisted and thus, the system can infer that a choice with these entities cannot be correct.

Temporal expression extracted using the rules also works as a temporal restriction (see Chapter 3.1). If the reference contains the expression "in the 8th century," it acts as a restriction; then, every entity in the choices should exist in the 8th century. If any person in the choices is born in the 9th century, the choice is likely incorrect.

Temporal conflict is strong negative evidence. For every entity and temporal expression, we check for temporal conflict. Some temporal properties of named entities are labeled in DBpedia; however, the property names are variable (e.g., `birthdate` or `birthYear` for person, `dateStart` or `foundingYear` for country, and `date` for event). We made simple rules for extracting the start year and end year of each entity. The property name should contain "`date`" or "`year`," where the earliest property of that entity is the start year and the latest entity for the end year.

According to our data analysis, temporal restriction was the most accurate evidence among the three evidence calculators. Thus, we assigned a negative 100 evidence score for temporal conflict. Such a strong score might cause errors, but it did not for the training/test data.

3.5 Answer Selection

The answer selection step is used to select the correct answer choice based on the evidence scores. For *sentence* and *term_other* style questions, we select the choice with the highest/lowest

Table 2. Evaluation result

Phase	Score	Correct	Incorrect	Unanswered
Phase-1	71	28	10	3
Phase-3	38	13	16	7

evidence according to the information from the question interpretation step (see Chapter 3.1). For *tf2* style questions, we select the choice that describes the truthfulness of the two sentences in the question.

4. RESULTS AND ANALYSIS

4.1 Results

We participated in the phase-1 and phase-3 National Center Test tasks. We were first place, out of ten teams, in Phase-1 and sixth place, out of twelve teams, in Phase-3 (Table 2). Our results are significantly better than random guessing (average score of 25), and the result of Phase-1 score is higher than the real student average for 100,000 students (64.13)[2]. The score for Phase-1 is higher than that of Phase-3 because the Phase-3 test dataset (in 2011) required understanding of figures, inferences, and semantic analysis, which are difficult NLP techniques that we do not use.

4.2 Error Analysis

We do not answer ten questions because of their answer styles. For example, we do not answer questions which 1.) need map understanding [`(symbol-term_other)*3`, `image_map`, and `term_other-symbol`], or 2.) need semantic analysis [`o(symbol-symbol-symbol)`, `(symbol-erm_other)*2`, `(term_other-term_other)`, and `(symbol-term_other) (symbol-term_person)`].

We also incorrectly answer 26 other questions. Most misses are due to failure in finding the correct named entities and their evidence. For example, a question fails to find the corresponding named entity in Wikipedia, where the important reference is "xiang ju li xuan." Another question, for which the choices are not composed of named entities, could not be differentiated by the evidence calculators, because the choices are explanations of the reference, written using common, non-named-entity words.

Some errors are due to the lack of semantic analysis in our method. For example, the choice "Kaidu instituted a rebellion against Kublai Khan" is incorrect but was estimated to be correct. We find strong evidence for "Kaidu" and "Kublai Khan" because they have historical conflictions and co-occur multiple times. Their relation was not a "rebellion" but our system failed to catch the difference between "rebellion" and "war."

Among the three evidence calculators, we found that the restriction calculator is the strongest evidence calculator; only one choice in the failed questions had a wrong restriction evidence score. KB and IR evidence calculators had similar performance.

5. CONCLUSION

Our system achieved scores of 71 and 38, and ranked first and sixth in Phase-1 and Phase-3, respectively. Our main approach in solving a question is choice verification, with three evidence calculators: a KB, IR, and restriction. We focus on named entities, instead of deep semantic analysis. Most failures were due to a lack of valid named entities and semantic analysis.

6. ACKNOWLEDGMENTS

This work was supported by the ICT R&D program of MSIP/IITP [R0101-15-0176, Development of Core Technology for Human-like Self-taught Learning based on a Symbolic Approach].

7. REFERENCES

[1] Daiber, J., Jakob, M., Hokamp, C., Mendes, P. N. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In Proceedings of the 9th International Conference on Semantic Systems (I-Semantics). 121-124

[2] Lee, H., Recasens, M., Chang, A., Surdeanu, M., and Jurafsky, D. 2012. Joint entity and event coreference resolution across documents. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 489-500. Association for Computational Linguistics.

[3] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Kleef, P. Auer, S. and Bizer, C. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2), 167-195.

[4] Okita, T. and Liu, Q. 2014. The Question Answering System of DCUMT in NTCIR-11 QALab. NTCIR.

[5] Park, S., Kwon, S., Kim, B., and Lee, G. G. 2015. ISOFT at QALD-5: Hybrid question answering system over linked data and text data. CLEF.

[6] Shibuki, H., Sakamoto, K., Ishioroshi, M., Fujita, A., Kano, Y., Mitamura, T., Mori, T., Kando, N. 2016, Task Overview for NTCIR-12 QA Lab-2. NTCIR.

[7] Vrandečić, D. and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.

[8] Wang, D., Boytsov, L., Araki, J., Patel, A., Gee, J., Liu, Z., Nyberg, E., Mitamura, T. 2014. CMU Multiple-choice Question Answering System at NTCIR-11 QA-Lab. NTCIR.

[2] http://www.dnc.ac.jp/data/suii/h09_h17.html, see 平成 11 年度, 世界史 B