KorQuA: Answer Extraction by Flexible Matching, Filtering and Interpretation

Kyung-Soon Lee[1], Jae-Ho Kim[2], Key-Sun Choi[2]

[1] NII (National Institute of Informatics)
2-1-2 Hitotsubashi Chiyoda-ku Tokyo 101-8430 Japan
kslee@nii.ac.jp

[2] Division of Computer Science, KORTERM,
KAIST (Korea Advanced Institute of Science and Technology),
373-1 Kusung Yusong Daejeon 305-701 Korea
{jjaeh,kschoi}@world.kaist.ac.kr

Abstract. This paper describes a **Kor**ean **Qu**estion **A**nswering (KorQuA) system. For flexible text retrieval, we represent terms in a question as five data types, and retrieve passages by matching terms according to these data types. In answer extraction, we filter non-fact answers with negative or uncertain contexts, and interpret a relative expression as an absolute date answer for date type question. Through component analysis, we show that terms need to be matched flexibly according to their characteristics, answer filter make a system more reliable, and answer interpretation make a system more intelligent. We also describe construction of a Korean question answering test collection.

1 Introduction

Information retrieval systems have concentrated on retrieving documents for user's query. In many cases, a user has a question to require a specific answer such as "Who invented the paper clip?" In this case, ranked answers with links to supporting documents are more useful than ranked documents. Question Answering (QA) is a task that combines techniques from information retrieval, information extraction, and natural language processing. Many researches have been introduced to extract specific answers from a document collection at the Text Retrieval Conference (TREC) QA track which was started in 1999 (TREC-8) [10]. General approach to question answering problem consists of question classification, passage retrieval, and answer extraction: The system attempts to classify a question according to the type of its expected answer. Next, the system retrieves passages of documents based on conventional information retrieval for a question. The system performs a shallow parse for the retrieved documents to detect the same entity type as the expected answer type. If an entity of the expected answer type is found near to terms in the question, the system returns it as an answer. Otherwise, the system fall back to best-matching-passage techniques. Recent many works are about classifying questions as to the expected answer type and using a wider variety of methods for finding the expected answer types in retrieved passages [12, 1, 4].

In our observation, QA systems need to consider these things to be more confident and intelligent: Terms in a text can be classified as the characteristics of matching. For example, range expression doesn't need exact term matching and proper noun such as book title needs exact term matching. To support an answer for a question, the answer should not have negative or uncertain contexts. And date expression has a tendency to be represented as a relative expression in a news article such as today, last month, and 2 years ago. Human can easily interpret it as the absolute date expression.

In this paper, we present a Korean question answering (KorQuA) system which deals those features. The major points of KorQuA are as follows:

(1) For flexible text matching, terms in a question are represented as five data types: date, range, core, keyword, and expected answer type. The similarities of passages are calculated by matching terms according to these data types.
(2) We filter non-fact answers modified by negative or uncertain contexts.
(3) We interpret answers of relative date expressions as absolute date expressions for date type question.

We also built a **Kor**ean **Q**uestion **A**nswering **Te**st **C**ollection (KorQATeC-I). The major features of KorQATeC-I are as follows:

(1) We regarded unambiguously interpretable answers from document contexts as answers.
(2) We selected candidate documents to be judged by using several document retrieval systems for a question itself and a question including answers already known.
(3) In judging answers, we excluded answers that showed the inconsistency or uncertainty for fact in total stream of a document.

2 KorQuA: Answer Extraction by Flexible Matching, Filtering and Interpretation

Our KorQuA system takes a natural language question as an input and produces a list of answers ranked. Our system consists of three components: question interpretation, flexible passage retrieval, and answer extraction by filtering and interpretation.

2.1 Question Interpretation

This step is to interpret a natural language question to data types and to determine the expected answer type for a question. For flexible text matching, terms in a question are represented as five data types: date, range, core, keyword, and question type. To determine an expected answer type, we analyze terms in a question with clue words and Korean thesaurus. The expected answer type is divided into three categories:

predefined answer category, semantic category for undefined answers, and ANY category for unknown question type.

2.1.1 Question Representation

The five data types are as shown in Table 1. Each term in a question is represented as one of the data types.

Table 1. Data types defined for question representation

Question Type and Representation Format	Example	
	Term	Representation
DATE <DATE, number, Qdate>	1991 년 (1991 year)	<DATE, 1991, Qyear>
RANGE <RANGE, Qmin, Qmax>	50 이상 (above 50) 50 이하 (below 50) 70 ~ 80	<RANGE, 50, *> <RANGE, *, 50> <RANGE, 70, 80>
CORE <CORE, term or phrase>	“로미오와 줄리엣” (“Romeo and Juliet”)	<CORE, 로미오와 줄리엣>
KEYWORD <KEYWORD, term; segmented terms>	법정제한액 (“Legal, Restriction. Amount of money”)	<KEYWORD, 법정제한액: 법정,제한,액>
QTYPE <QTYPE, Qtype, Atype>		<QTYPE, who, person>

In the *DATE type*, the value of Qdate can be one of the Qyear, Qmonth, and Qday, which represent the year, the month, and the day for a term in a question, respectively. The *RANGE type* represents range of number with the lower and the upper limit. The ‘*’ symbol means infinity. The *CORE type* recognizes important words which need exact matching described with <>, ‘ ’, and "" symbols in a question. The *KEYWORD type* involves non-stop terms with part-of-speech tag such as noun, verb, adjective, and adverb except for terms taken as DATE, RANGE, and CORE type. The *QTYPE* is a question type which represents what category of the entity a user is asking for.

2.1.2 Question Classification

We determine a question type, QTYPE, based on clue words and semantic category for terms in a question. The expected answer types are divided into three categories: predefined answer type, semantic category type for undefined answer type, and ANY type for unknown question type.

Predefined answer type according to clue words: we parse a question and determine the answer type for a question type based on clue words such as 누가 (who), 어디 (where) and 언제 (when), which show the focus of a question. Predefined answer types are as follows: PERSON, ORGANIZATION, LOCATION (country, city, natural place), TIME (year, month, date, season), NUMBER, NUMBER+UNIT, WAY, and REASON. The answer type such as TIME is divided into several specific answer types such as YEAR, MONTH, DATE, and SEASON.

Answer type with a semantic category itself: The semantic category itself can be represented as a question type and an answer type. In Korean, the ellipsis of a verb and an interrogative in an interrogative sentence is a general phenomenon. In this case, the semantic category of the last noun X is used as an answer type. For which, what, and ellipsis type of question, the answer type is determined depending on the semantic category of the noun X. If the semantic category of the noun X does not belong to the predefined answer type, the answer type is determined as the semantic category of the noun X. For example, in an elliptical question sentence "햄릿(Hamlet)의 저자(author)는? (The author of Hamlet?)", the answer type is determined as PERSON because the semantic category of the last noun '저자 (author)' belongs to PERSON category. In another example, "… 질병(disease)은?", the semantic category of '질병 (disease)' is not defined as the predefined answer type. In a Korean thesaurus, the semantic category for '질병 (disease)' is defined as <2419>. Therefore, the semantic category <2419> is determined as an answer type for the question.

ANY type: When the semantic category information for a noun does not exist in the thesaurus, the answer type is ANY type which has answers with arbitrary words.

2.2 Flexible Passage Retrieval

We use a Korean version of SMART system [9, 5] to retrieve documents which are likely to contain the answers to the question. The top 50 documents are passed as an input to passage retrieval.

Passages are defined as overlapping sets consisting of a sentence and its two immediate neighbors [1]. The similarity of passage *i* is calculated by term matching and question inclusion ratio as follows:

$$Sim_{psg_i} = (\frac{1}{4}S_{i-1} + \frac{1}{2}S_i + \frac{1}{4}S_{i+1}) \cdot (\frac{N_p}{N_q}) \qquad \textbf{(1)}$$

where S_i is the score for sentence *i* which is the sum of inverse document frequency of matched terms for terms in a question. N_q is the total number of terms in a question, and N_p is the number of matched term in a passage. The *question inclusion ratio* is calculated by (N_p/N_q), which prefers a passage including various terms for a question.

For flexible passage matching, five data types are considered to match terms in each passage. For ***DATE type*** represented as <DATE, NUM, Qdate> in a question, the absolute date is calculated from a relative date representation in a passage based on the base date (or written date) of each document. And the date matching is conducted by checking whether Qdate is equal to the absolute date of a term in a passage. The base date of a document is defined as '<DATE> Dyear Dmonth Dday </DATE>' for each document. The temporal expressions such as '올해 (this year)' refers to the year of the base date of a document. In this case, the absolute year for that representation is equal to the base year of a document. We can get the absolute date by mapping relative representation to the base date with number of addition or subtraction. For ***RANGE type*** represented as <RANGE, Qmin, Qmax>, the range matching is whether a number, Dnum, in a passage belongs to the range of (Qmin < Dnum) and (Dnum <

Qmax) or not. For example, a term ‘80 이상 (above 80)’ is represented as range type <RANGE, 80, *>, and is matched successfully in a term ‘87’ in a passage. For ***CORE type***, exact matching is meaningful when the value of CORE type is a phrase. A term of CORE type has more important weight than terms of other data types. In Korean, a compound noun is freely represented with or without space. For a compound noun represented as ***KEYWORD type***, it can be segmented into several single nouns. If a compound noun is not matched with a term in a passage, we try partial matching with a single noun. For example, “법정제한액” can be divided into three single nouns such as ‘법정 (legal)’, ‘제한 (restriction)’, and ‘액 (amount of money)’. In this case, a passage has lower value according to the degree of matching than that of full compound noun matching.

2.3 Answer Extraction

Answer extraction involves three components: entity extraction based on several language resources, filtering non-fact answers modified by negative or uncertain contexts, and answer derivation for date type question.

2.3.1 Entity Extraction

To extract candidate answers corresponding to the expected answer type, we identify named entities in passages. The named entities are as follows: Person, Organization, Location (country, city, place), Date (season, year, age), Quantities, Durations, and Linear measures, and semantic categories. Language resources used for entity extraction are a proper noun dictionary, lexical patterns, a Korean thesaurus, the error logs of proper noun which general nouns are misclassified to, and verb’s case frame. The number of entries is 44143 for the proper noun dictionary, 1926 for lexical patterns, 1361 for error log of proper noun, 9256 for verb case frames, and 23275 for a Korean thesaurus which consists of 12 levels and 2502 concept nodes.

2.3.2 Filtering Answers Modified by Negative or Uncertain Context

We exclude candidate answers modified by negative or uncertain contexts for extracted named entities from answers. The negative and uncertain contexts contain negative verbs, uncertain verbs, and uncertain background domains of a document.

- Negative context: 아니다 (no), 않다 (not) ,못하다 (no, not), 부인하다 (deny) , 부정하다 (negate)
- Uncertain context: 가정하다 (assume), 추측하다 (suggest), 추정하다 (estimate), 짐작하다 (guess), 상상하다 (imagine), and future tense
- Uncertain background domain: 영화 (movie), 꿈 (dream), 소설 (fiction), 희곡 (play), 드라마 (drama), 시나리오 (scenario), 동화 (children’s story)

The answers should be fact with supporting information for a question. Though a candidate answer has the supporting information for a question, it is not acceptable to be an answer if it is described in an uncertain domain, and it is modified by a negative verb, an uncertain verb, an uncertain noun, or a verb of uncertain tense or aspect such as a future tense or a progressive form.

2.3.3 Answer Interpretation

We interpret a new answer representation from a derivable representation which is not in a document for the date type answer. To derive an answer for the date type, we calculate an absolute date from a relative date representation in a text based on the base date of each document. Answer interpretation can be calculated by following expression:

$$\text{Derived answer date} = \text{Ddate} \pm \text{Tdate} \quad \textbf{(2)}$$

where *Ddate* is the written date of a document and *Tdate* is a relative date expression in a text.

For example, when the base date of a document is 1995 and there is a relative expression , '10 년 전 (10 years ago)', the value of *Dyear* is '1995' and the value of *Tdate* is -10. Therefore, we calculate the absolute date, 1985, by calculating '1995 – 10'. The derived answer is presented as a final answer for date type question to the user.

2.3.4 Answer Ranking

Candidate answers in a passage are ranked according to the similarities of answers. The similarity of answer, Sim_{ans}, is calculated as:

$$Sim_{ans} = Sim_{psg} \cdot F_{atype} \quad \textbf{(3)}$$

where Sim_{psg} is the similarity of a passage and F_{atype} can have 0 or 1 as a value according to whether the category of answer matches the expected answer type. Answers in the same passage are preferred as to the position of occurrences of a given answer.

3 Construction of a Korean Question Answering Test Collection

We constructed a **Kor**ean **Q**uestion **A**nswering **Te**st **C**ollection (KorQATeC) [6, 7], KorQATeC-I and KorQATeC-II, in 2000 and 2001. Although KorQATeC-II can evaluate QA system in more various aspects, in this paper we describe KorQATeC-I on which our system is evaluated. In building the KorQuATeC, the guidance of TREC QA is referenced [10, 2]. Some distinct characteristics from TREC QA test collection are as follows: (1) We included a derivable answer from document contents as an answer. (2) The answer has the following format: [Question-ID Document-ID:Boolean-expression-for-relevance {<A> answer string </A> }* {<DA> derived answer string </DA>}*]. (3) We selected documents to be judged by using various document retrieval techniques. Two kinds of query are used: One is original questions and the other is expanded questions including already known answers. (4) In answer judgment, we excluded answers that represent the inconsistency or uncertainty for fact in total stream of a document.

To briefly introduce the KorQATeC-II, it has these characteristics: (1) Including entity type, list type, summarization type, and context-based answers. (2) Including questions which are not guaranteed to have an answer. (3) Including twelve types questions classified by Graesser (e.g concept completion, quantification, verification, comparison, definition, etc.) [2]. (4) Including derivable answers such as date, number, linear measure, and rank type answers.

3.1 Documents

The document set consists of 207067 documents (360MB) that are taken from articles of three newspaper companies from January of 1992 to May of 1995. The documents are tagged as SGML format. In <DATE> part of the document, the date expressed when each document was written. As background country of a document is Korea, the term '우리나라 (our country)' represented in a document means 'Korea'.

3.2 Questions

The questions are taken from Quiz database (17461 questions) on internet and the document collection to be retrieved. The number of question is 100. The 10 questions are variant questions for 90 questions. The base date of questions is August 2000 when we made the questions. Although the tense of a question is the past, it can be expressed with the present tense in a document. Table 2 shows the distribution of question types for the KorQATeC-I.

Table 2. The distribution of question types in a Korean QA test collection (KorQATeC-I)

Question Type	The Number of Question	Answer Type
WHO	17	PERSON
WHERE	20	LOCATION (country, city, location, place, and etc.)
WHEN	14	TIME (date, year, age, season)
HOW	20	QUANTITIES (number, height, money, %, and etc.)
WHY	4	REASON
WHAT/ WHICH	25	UNDEFINED ENTITIES (animal, plant, mineral, disease, event, religion, and etc.)

3.3 Selecting Candidate Documents to be judged

We used a pooling method to assemble the candidate documents to be judged by human assessors. We retrieved documents using the two types of questions. One is original questions and the other is expanded questions including already known answers. The 16 runs were created by several information retrieval schemes such as morpheme indexing, bi-gram indexing, Boolean retrieval, vector space retrieval, relevance feedback, and various weighting schemes. And then we took the top 70 documents retrieved in each run for a given question. Totally 22898 documents are judged in answer judgments by human assessors.

3.4 Answer Judgments

The 10 human assessors accomplished evaluation and two persons per question took an evaluation independently each other. Human assessors read each whole document, extracted answer strings, and included derivable answer strings. The result format of answer judgment is as follows:

[<*Qid*> <*Did*>*: <Boolean-value--for-relevance> <answer-string>*]*

The *Boolean-value-for-relevance* is taken 1 or –1 as to whether answers are included in the document *Did* for the question *Qid*. If it is 1, the answer strings for the question are followed with expression {<A> answer-string </A>}+. When the answer was not expressed directly but was derived from a document, the answer string was expressed as {<DA> derived answer string </DA>}* and {<CA> clue string for derived answer string </CA>}*.

4 Experiments

We evaluated the performance of KorQuA system on the Korean QA test collection, KorQATeC-I. Our system produces five ranked answers for each question. The output was scored automatically by relevance information.

The MRAR (Mean Reciprocal Answer Rank) and accuracy rate are used as evaluation scheme. The MRAR is used to compute the overall performance of system.

$$MRAR = \frac{1}{n}(\sum_{i=1}^{n}\frac{1}{r_i}) \quad \textbf{(4)}$$

where n is the number of questions and r_i is the rank assigned by a system at which a correct answer is found for question i, or 0 if no correct answer was found.

The accuracy rate means percentage of questions whose the correct answers were in the top k answers returned by the system.

4.1 Results

Table 3 shows the results of KorQuA system on MRAR and accuracy rate. The performance was evaluated for five answer units: answer, passage 50 bytes, passage 250 bytes, whole passage, and whole document. The output of whole passage and whole document were evaluated for comparisons of performance in passage retrieval and document retrieval system.

Table 3. Performance of KorQuA system on MRAR and accuracy rate

Answer Unit	Mean Length	MRAR (rank 5)	Accuracy Rate
Answer	18 Byte	0.270	36.7%
Passage 50	48 Byte	0.476	57.8%
Passage 250	185 Byte	0.526	63.3%
Passage	429 Byte	0.626	75.6%
Document	1807 Byte	0.593	75.6%

With the answer unit, our system showed correct answers on 36.7% of the questions, with a MRAR score of 0.270 in the top five answers. The score of MRAR was 0.476 for 50 bytes and 0.526 for 250 bytes. In our question answering system based on document retrieval, the average accuracy rate of correct document for all questions was limited at 74.6% for the top 50 documents.

4.2 Component Analysis

Passage matching: To investigate passage matching for date type, we analyzed Question 13: “1991 년 노벨평화상을 탄 사람은 누구인가? (Who won the Nobel Peace Prize in 1991?)” The year ‘1991’ is represented as ‘1991’ directly in 9 documents, and is expressed as temporal expression such as ‘지난해’ (last year) or ‘작년’ (a year ago) in 9 documents among 17 relevant documents including answers. For range type, in Question 83: “인도 인구의 80%이상은 어떤 종교를 믿는가? (Which religion is believed by above 80% of India's population?)”, all of the relevant documents with answers are represented as ‘82’ or ‘83’ instead of ‘80’. Therefore, we can see that consideration of date and range type is necessary for passage matching from the above observations.

Filtering non-fact answers: A document, HRM920105-40, has supporting information for a question 12 and an answer string. But, the domain of this document explains the story of movie ‘JFK’. Therefore, the answer extracted from this document is unreliable.

Q12: 미국 대통령 케네디의 암살범은?

<title> 올리버스톤 감독 새영화 `JFK' </title>

<text> … 오스왈드는 존 피츠제럴드 케네디의 암살범………</text>

Q12: Who assassinated President of U.S. Kennedy?

<title> A new movie ‘JFK’ directed by Oliver Stone </title>

<text> … Os-wald is an assassinator of John F. Kennedy ………..</text>

Answer interpretation: For Question 39 “에이즈가 세계 최초로 발견된 해는? (In which year was AIDS discovered in the first of the world?)”, the relevant document with an answer, CSM931202-81, describes the answer as derivable form ‘12 년 지나 (12 years passed)’ instead of a direct answer ‘1981’. We calculated ‘12 years ago’ to direct answer form based on the written date of the document. Without answer derivation processes, in the above example of passage matching, if the variation of Question 13 is “What year was Aung San Suu Kyi awarded the Nobel Peace Prize?”, we could not get answers from 9 documents. Therefore, answer interpretation is necessary for date representation.

5 Conclusions

We have described the Korean question answering system. For flexible text matching, a question is represented as five data types: date, range, core, keyword, and question type. Answer extraction stage involved entity extraction, filtering non-fact answers modified with negative and uncertain contexts, and answer interpretation for the date type question. The performance of our system on MRAR is 0.476 and 0.526 for 50 bytes and 250 bytes answer units for top five answers, respectively. Through component analysis, we can conclude that terms need to be matched flexibly according to their characteristics, answer filter makes a question answering system more reliable, and answer interpretation makes a system more intelligent.

We built a Korean QA test collection which included derived answers as answers and excluded answers that represent the inconsistency or uncertainty for fact in total stream of a document in judging answers.

Acknowledgments

This research was conducted while I was a PhD student at KAIST (Korea Advanced Institute of Science and Technology) and supported by the Korean Ministry of Science and Technology.

References

1. Abney, S., Collins, M., Singhal, A.: Answer Extraction. In Proceedings of the 8th Conference on Applied Natural Language Processing (2000)
2. Burger, J., Cardie, C., Chaudhri, V. et al.: Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). NIST (2001)
3. Harabagiu, S., Moldovan, D., Pasca, M., et al.: FALCON: Boosting Knowledge for Answer Engines. In Proceedings of the 9th Text Retrieval Conference (TREC-9) (2000)
4. Ittycheriah, A., Franz, M., Zhu, W-J., Ratnaparkhi, A.: IBM's Statistical Question Answering System. In Proceedings of the 9th Text REtrieval Conference (TREC-9) (2000)
5. Lee, J.H., Ahn, J.S.: Using n-Grams for Korean Text Retrieval. In Proceedings of 19'th ACM SIGIR International Conference on Research and Development in Information Retrieval. (1996) 216-224
6. Lee, K-S., Kim, J-H., Choi, K-S.: Construction of Test Collection for Evaluation of Korean Question Answering System. In Proceedings of the 12th Conference on Hangul and Korean Language Information Processing (2000) 190-197 (written in Korean)
7. Kim, J-H., Lee, K-S., Oh, J-H., Chang, D-S., Choi, K-S.: KorQATeC-II: Construction of Test Collection for Evaluation of Question Answering System. In Proceedings of the 13th Conference on Hangul and Korean Language Information Processing (2001) (written in Korean)
8. Radev, D.R., Prager, J., Samn, V.: Ranking Suspected Answers to Natural Language Questions Using Predictive Annotation. In Proceedings of the 8th Conference on Applied Natural Language Processing (2000)
9. Salton, G.: The Smart Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Inc., Englewood Cliffs, NJ (1971)
10. Voorhees, E.M., Tice, D.M.: The TREC-8 Question Answering Track Evaluation. In Proceedings of the 8th Text Retrieval Conference (TREC-8) (1999)
11. Voorhees, E.M., Tice, D.M.: Building a Question Answering Test Collection. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2000) 200-207
12. Voorhees, E.M.: Overview of TREC-9 Question-Answering Track. In Proceedings of the 9th Text REtrieval Conference (TREC 9) (2000)