TTA Standard

정보통신단체표준(국문표준)

TTAK.KO-10.1098 제정일: 2018 년 12 월 19 일

오픈 도메인 자연어 질의 응답을 위한 질문 분석 메타데이터

Metadata for Question Analysis for Open-domain Question Answering

TTA 한국정보통신기술협회
Telecommunications Technology Association

표준초안 검토 위원회 메타데이터 프로젝트그룹(PG606)

표준안 심의 위원회 소프트웨어/콘텐츠 기술위원회(TC6)

	성명	소 속	직위	위원회 및 직위	표준번호
표준(과제) 제안	최미란	한국전자통신연구원	책임연구원	PG606 위원	TTAK.KO-10.1098
표준 초안 작성자	허정	한국전자통신연구원	책임연구원	-	TTAK.KO-10.1098
	김현기	한국전자통신연구원	책임연구원	-	TTAK.KO-10.1098
	최미란	한국전자통신연구원	책임연구원	PG606 위원	TTAK.KO-10.1098
사무국 담당	김재웅	TTA	단장	-	TTAK.KO-10.1098
	박정혜	TTA	수석연구원	-	TTAK.KO-10.1098

본 표준 발간 이전에 접수된 지식재산권 확약서 정보는 본 표준의 '부록(지식재산권 확약서 정보)'에 명시하고 있으며, 이후 접수된 지식재산권 확약서는 TTA 웹사이트에서 확인할 수 있습니다.
본 표준과 관련하여 접수된 확약서 외의 지식재산권이 존재할 수 있습니다.

발행인 : 한국정보통신기술협회 회장
발행처 : 한국정보통신기술협회
13591, 경기도 성남시 분당구 분당로 47
Tel : 031-724-0114, Fax : 031-724-0109
발행일 : 2018.12.

서 문

1 표준의 목적

본 표준의 목적은 오픈 도메인 질의 응답을 위한 질문 분석에서 요구되는 주요 항목인 메타데이터를 정의하는 것이다.

질문 분석은 기계가 사용자의 자연어 질문을 이해하고 정답을 찾기 위한 단서들을 분석하여 질문의 의미를 정형화하는 기술이며 오픈 도메인 질의 응답 시스템을 위한 가장 기반이 되는 기술이다.

최근 인공 지능의 기술적 성과에 대한 기대와 관심이 전세계적으로 커지고 있다. 인공 지능의 다양한 분야 중, 기계가 사람과 소통하기 위한 기초 기술인 언어 지능에 대한 관심도 크게 높아지고 있다. 언어 지능의 핵심은 사람의 자연 언어를 기계가 이해하고 사람과 원활하게 의사소통 할 수 있도록 자연 언어를 분석하는 것이다. 특히, 사람의 질문에 기계가 정보를 제공하기 위해서는 자연 언어로 된 질문을 정확하게 이해하는 것이 중요하다. 본 표준에서는 사람의 자연어 질문을 기계가 이해할 수 있도록 정형화하는 것을 목적으로 하며 질의 응답뿐만 아니라 대화 처리와 같은 다양한 언어 지능 분야에서 유용성이 클 것으로 판단된다.

2 주요 내용 요약

본 표준에서는 오픈 도메인 질의 응답을 위한 질문 분석을 통해 정형화되는 다양한 정보들과 메타데이터 구조를 표준화하고, 개별 정보에 대한 인식 지침을 정립하고자 한다.

질문의 의미를 정형화하기 위한 주요 정보는 아래와 같다.

a) 입력된 자연어 질문과 언어 분석된 질문 결과를 저장하는 질문 분석 기본 정보
b) 복합 질문(complex question)을 최소 단위의 질문으로 분할하고, 분할 질문 간의 관계성을 정형화하는 질문 분할 정보
c) 질문 별로 다양한 정답 전략 수립을 목적으로 다각적인 관점에 따라 질문을 분류하는 질문 분류 정보
d) 질문 분석을 위한 중요한 단서인 의문사 정보에 기반한 질문 유형 분류 정보
e) 자연어 질문에서 정답을 지칭하는 부분과 연관된 질문 초점 정보

f) 질문에서 찾고자 하는 정답의 의미적인 유형을 제약하는 정답 유형 정보
g) 질문의 대상이 되는 주요한 개체를 인식하는 질문 토픽 정보
h) 정답을 제약하는 질문 내 단서들은 인식하는 정답 제약 정보

본 표준에서는 위의 언급된 주요 정보를 기반으로 질문의 의미를 정형화하는 JSON 형식의 메타데이터와 메타데이터의 값을 자연어 질문으로부터 인식하기 위한 기준을 정립한다.

3 인용 표준과의 비교

해당 사항 없음.

Preface

1 Purpose of the standard

The standard is to define the metadata which is required as major items in the question analysis module for open-domain question answering system.
The question analysis is a technology that formalizes the meaning of the questions for machines to understand the user's natural language questions and to provide answers by analyzing cues in the questions. It is one of the most essential technologies for the open-domain question answering system.
The achievements and expectation by AI technologies are spreading all over the world. Thus, of the various fields of artificial intelligence, interests in language intelligence, which is the basic technology for machines to communicate with people, are increasing. At the core of language intelligence exists the analysis of the natural language so that the machine understands the natural language of a human and communicates smoothly with him. In particular, in order for a machine to provide information to a human question, it is important to correctly understand the question in its natural language. The standard is to formalize the natural language questions so that the machine can understand it. It is considered to be useful in the various language intelligence fields such as dialogue processing as well as question answering.

2 Summary of the standard

The standard defines the different information elements and metadata structures which are formalized by the question analysis for open-domain question answering systems and provides the information recognition guidelines for each element.

The major information elements to formalize the meaning of the questions are as follows:

a) The basic question analysis information which stores the input question in natural language and the results of the analyzed question
b) Question decomposition information that decomposes the complex question into questions of minimum units and formalizes the relation between the

decomposed questions.

c) Question classification information based on multiple perspectives to establish the various answering strategies for each question.

d) Information on the classification of question types based on WH-question words, which is an important clue for question analysis.

e) Question focus information related to the part of the natural language question that indicates the correct answer

f) Answers type information that limits the semantic type of the correct answers searched.

g) Question topic information that identifies the entity for the object of the question.

h) Answer constraining information that recognizes the clues that constrain correct answers.

The standard defines the metadata in JSON format that formalizes the question meaning based on the major information elements defined above for the question analysis and establishes the criteria for recognizing the values of the metadata from the natural language questions.

3 Relationship to Reference Standards

- None.

목 차

오픈 도메인 자연어 질의 응답을 위한 질문 분석 메타데이터 (Metadata for Question Analysis for Open-domain Question Answering)

1 적용 범위

본 표준에서는 오픈 도메인 질의 응답을 위한 질문 분석을 통해 정형화되는 다양한 정보들과 메타데이터 구조를 표준화하고, 개별 정보에 대한 인식 가이드라인을 정립하고자 한다.

자연어 질의 응답은 기존 포털에서 서비스하고 있는 정보 검색보다 진일보한 기술로서, 모든 검색 서비스 업체에서 기술 투자를 하고 있는 분야이다. 또한, 로봇 및 모바일 에이전트 관련 산업도 새롭게 조명을 받고 있는 상황이다. 이와 같은 인공 지능의 최종 응용 시스템들의 대부분은 인간과 기계의 의사소통을 기본으로 하며, 이를 위해서는 자연어 처리 기술이 선행되어야 한다. 또한, 사용자의 요구(질문)에 대해 기계가 응답을 할 수 있어야 한다. 이를 위해서는 질의 응답 기술이 기반 기술로 적용되어야 하는 것이다. 향후, 다가올 IT 환경에서 기계와 인간의 원활한 의사소통을 위해서는 자연어 처리와 질의 응답, 대화 처리 기술 등이 기반 기술로 반드시 활용되어야 하기 때문에 IT 산업의 전반적인 분야에 미치는 영향은 점차적으로 높아질 것으로 판단된다.

본 표준에서는 사람과 기계의 소통을 위해서 사람의 질문을 기계가 이해할 수 있는 단위로 정형화하는 질문 분석의 개별 의미 정보들에 대한 메타데이터 정보와 해당 정보에 대한 인식 기준을 제시한다. 이를 통해, 사람의 자연어를 인터페이스로 사용하는 모든 분야에서 언어를 이해하여 기계가 사람과 소통할 수 있도록 함으로써 보다 친숙하게 기계를 활용할 수 있도록 환경을 조성할 수 있을 것으로 기대한다. 본 표준에서는 사람의 자연어 중, 오픈 도메인 질의 응답에서 요구되는 질문만을 대상으로 하기 때문에 정보 검색적 측면의 의미 정형화로 그 범위가 제한된다.

2 인용 표준

해당 사항 없음.

3 용어 정의

3.1 개체(entity)
인물, 기관, 지역, 작품명 등과 같이 주요한 고유 명사들을 의미

3.2 분할 질문
분할된 최소 단위의 질문

3.3 분할 질문 관계
분할 질문들 간의 의미적 관계

3.4 어휘 정답 유형
질문 내에서 정답을 제약하는 개념어

3.5 의미 정답 유형
이미 정의된 의미적 카테고리로 정답의 유형을 분류

3.6 정답 유형
사용자가 요구하는 정답의 유형

3.7 정답 제약
정답 후보를 제약하는 질문 내의 다양한 단서들

3.8 질문 분석
사용자의 자연어 질문을 기계가 이해할 수 있도록 분석하여 정형화하는 기술

3.9 질문 분할
복문으로 구성되거나, 단문이지만 정답을 지지하는 둘 이상의 단위 질문으로 분할될 수 있는 질문으로 나누는 기술

3.10 질문 유형
질문에서 요구하는 정답을 가장 잘 표현하는 의문사에 기반한 질문 분류

3.11 질문 초점
사용자의 질문에서 정답을 지칭하는 부분

3.12 질문 토픽
질문 내에서 가장 대표적인 개체

3.13 SPO 관계

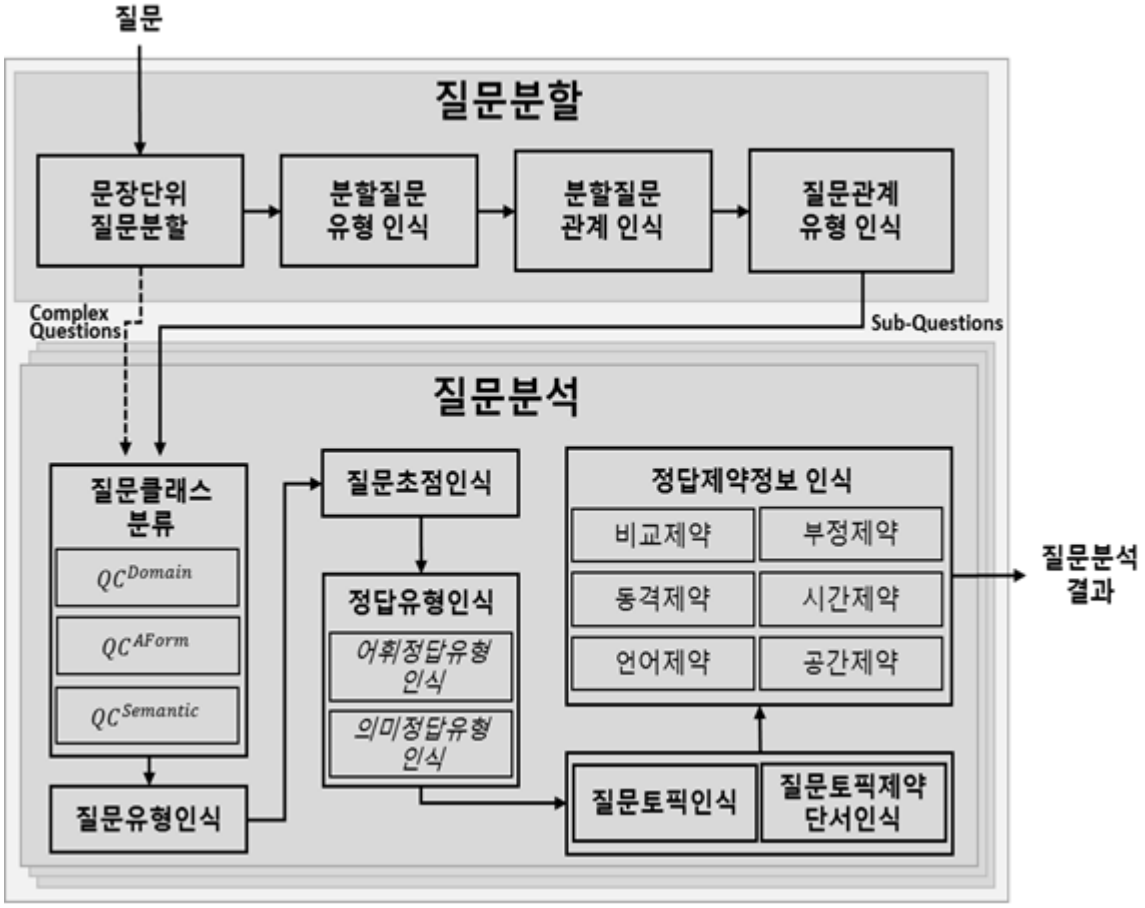
주요한 개체(subject), 개체 속성(property)과 속성값(object)의 관계

4 약어

SPO Subject Property Object

5 질문 분석 메타데이터

질의 응답을 위한 질문 분석의 구성도는 (그림 5-1)과 같다. 자연어 질문은 언어 분석기를 통해 형태소 분석, 개체 분석, 구문 구조 분석 등을 수행한다. 언어 분석된 결과에 기반하여 (그림 5-1)과 같은 흐름으로 자연어 질문을 분석하게 된다. 본 표준에서는 (그림 5-1)의 질문 분석 구성에 따라 메타데이터를 구분한다.



(그림 5-1) 질문 분석기 구성도

본 표준에서 제시하는 질문 분석 메타데이터의 구성은 <표 5-1>과 같다. 메타데이터는 JSON 형태로 계층적인 구조로 되어 있다. 계층적 구조는 들여쓰기와 심볼(↳)로

표현하였다.

<표 5-1> 질문 분석 메타데이터의 구성

구분	메타데이터 이름	설명	비고
질문 분석 기본 정보 (orgQInfo)	strOrgQuestion	질문 스트링	
	orgUnit	원질문 정보 객체	
	↳strQuestion	원질문 스트링	
	↳strTaggedQ	질문 태깅 결과 스트링(lexico-semantic pattern) 매칭을 위한 언어 분석 결과 출력	
	↳ndoc	질문의 언어 분석 결과 객체	
질문 분할	qDecompType	질문 분할 유형	
	↳vSubQRelation	분할 문장 간의 관계 객체 배열	
	↳ iID1	첫 번째 문장 ID	
	↳ iID2	두 번째 문장 ID	
	↳ relation	두 문장의 관계	
	↳vSubQInfo	분할 질문 정보 객체 배열	
	↳iID	분할 질문 ID	분할 질문 번호
	↳ strSubQ	분할 질문 스트링	분할 질문
	↳ subQType	분할 질문 유형	
	↳ qAnalUnit	분할 질문에 대한 분석 결과 객체(원질문의 분석 결과 구조가 반복)	
질문 분류	QClassification	질문 분류 정보 객체	
	↳vQDomain	질문 도메인 객체 배열	
	↳ qType4Chg	질문 유형 코드	질문 유형
	↳ dWeightCQT	인식 신뢰도	
	↳ansQType	정답 형태에 따른 질문 유형 객체	vQDomain의 하부 구조와 동일함
	↳vSemQType	의미적 질문 유형 객체	vQDomain의 하부 구조와 동일함
의문사 기반 질문 유형	vQTs	의문사 기반 질문 유형	
	↳qt	질문 유형 코드(숫자)	
	↳strQTClue	의문사 스트링	
질문 초점	vQFs	질문 초점 객체 배열	

구분	메타데이터 이름	설명	비고
	↳strQF	질문 초점 스트링	질문 초점
	↳dWeightQF	질문 초점 신뢰도	
어휘 정답 유형 (LAT)	vLATs	어휘 정답 유형 객체 배열	
	↳strLAT	어휘 정답 유형	LAT
	↳vCompoundLATs	어휘 정답 유형의 복합 명사 형태 스트링 목록	
	↳strID	어휘의 코드(의미 또는 개념망 ID)	LAT 의미 코드
	↳dConfidenceLAT	인식의 신뢰도	
의미 정답 유형 (SAT)	vSATs	의미 정답 유형 객체 배열	- vSAT는 세분류 SAT
	↳atSAT	의미 정답 유형 코드	
	↳dConfidenceSAT	인식 신뢰도	SAT 신뢰도값
지식 베이스 타이틀	vTitles	주요한 개체 정보 배열	
	↳strEntity	개체 스트링	온라인 백과사전의 Title
	↳atEntityType	개체의 NE 코드	
	↳vEntityInfo	개체의 지식 베이스 정보	
	↳ strNormEntity	지식 베이스 타이틀로 정규화된 개체 스트링	
	↳ strID	지식 베이스 타이틀 ID	
	↳ strExplain	타이틀의 정의문(homonymy에 있는)	
	↳ dWeightEn	모호성 해소 가중치	
질문 토픽	vQTopic	타이틀들 중 가장 중요한 타이틀	vTitles 구조와 동일
정답 제약	answerConstraint	정답 제약 정보 객체	
	↳vAcTime	시간 제약 객체 배열	
	↳ TExpression	시간 표현 스트링	
	↳ valueAt	정규화된 시점의 시간 표현	
	↳ valueBegin	정규화된 기간의 시간 표현 (시작 시간)	
	↳ valueEnd	정규화된 기간의 시간 표현 (종료 시간)	
	↳vAcLoc	공간/장소 제약 객체 배열	
	↳ LExpression	공간/장소 표현 스트링	
	↳ kbURI	정규화된 공간/장소 표현	

5.1 질문 분석 기본 정보

기본 정보에서는 사용자의 자연어 질문과 해당 질문에 대한 언어 분석된 결과가 저장된다. 개별 Key(메타데이터 이름)별 설명은 아래와 같다.

a) strOrgQuestion: 질의 응답 시스템에 입력되는 스트링 정보가 저장
   질문의 유형은 주관식과 객관식으로 구분될 수 있다. 객관식의 경우, 질문과 함께 보기 정보도 함께 질문 분석의 입력으로 주어진다.
   예) 김구의 호는 무엇인가? [1. 백범, 2. 충무, 3. 다산]
b) orgUnit: 순수 질문 정보와 질문에 대한 언어 분석된 정보를 저장
   1) strQuestion: 보기와 같은 정보를 제외한 질문 스트링만 저장
      예) 김구의 호는 무엇인가?
   2) strTaggedQ: 언어 분석 결과가 부착된 질문을 저장
      예) <PS_NAME:김구/NNP>+의/JKG 호/NNG+는/JX 무엇/NP+이/VCP+ㄴ가/EF+?/SF
   3) ndoc: 질문의 언어 분석된 결과 상세 정보를 저장하는 객체

5.2 질문 분할 정보

복합 문장으로 구성된 질문은 문장 간의 관계성을 분석해야 정확한 정답을 추출할 수 있다. 따라서, 복합 문장을 단일 문장으로 분할하고, 단일 문장들 간의 관계성을 분석하는 것은 중요하다. 또한, 분할된 단일 문장들은 별도의 질문으로 정답을 찾는 전략 수립에 중요한 단서를 제공한다. 질문 분할 정보의 상세 설명은 아래와 같다.

a) vSubQInfo: 분할된 개별 질문들에 대한 정보를 저장하는 객체 배열
   1) iID: 분할 질문의 질문 번호
   2) strSubQ: 분할된 질문의 스트링
   3) subQType: 분할된 개별 질문들의 유형을 저장. 분할된 단위 질문이 다음 중 어떤 유형에 포함되는지 분류함(분할 질문 유형)
      (a) Question: 질문 문장인 경우(질문 초점이 포함된 문장)
          * 질문이 분할되지 않는 경우, 전체를 Question 으로 처리
      (b) Fact: 질문이 아니고, 단순한 사실을 기술한 문장
      (c) Inner-Question: Question 인데, 전체적으로 내포형 질문으로 선행 질문에 해당하는 질문인 경우
      (d) Outer-Question: Question 인데, 전체적으로 내포형 질문의 후행 질문에 해당하는 경우

4) qAnalUnit: 분할된 개별 질문에 대한 분석 결과를 저장하는 객체(원질문의 분석 결과 구조가 반복됨)

b) vSubQRelation: 분할된 문장 쌍들 간의 관계 정보를 저장하는 배열
   1) iID1: 분할 문장 쌍의 첫 번째 질문 번호
   2) iID2: 분할 문장 쌍의 두 번째 질문 번호
   3) relation: 분할 문장 쌍의 관계 정보. 분할된 단위 질문의 분할 질문 유형이 결정되면, 분할된 단위 질문들 간의 관계를 인식함(분할 질문 관계)
      (a) And 관계: 일반적으로 Fact, Question 들 간에는 And 관계를 가짐
         - 예) 이곳은 용암 동굴로 동굴 안에 다양한 종유석과 석순이 흩어져 있으며 총 길이는 약 7.4 km 에 달한다. 김녕굴과 함께 천연기념물 제 98 호로 지정된 제주도의 동굴은 무엇일까?
         - 분할 문장 1: 이곳은 용암 동굴로 동굴 안에 다양한 종유석과 석순이 흩어져있으며 총 길이는 약 7.4 km 에 달한다.
         - 분할 문장 2: 김녕굴과 함께 천연기념물 제 98 호로 지정된 제주도의 동굴은 무엇일까?
         - (분할 질문 1) AND (분할 질문 2)
      (b) Depend-on 관계: Inner-Question 과 Outer-Question 은 서로 Depend-on 관계를 가짐
         - 예) 이 사람은 한글을 창제한 왕이다. 이 왕은 조선의 몇 대 왕인가?
         - 분할 문장 1: 이 사람은 한글을 창제한 왕이다. (inner-question)
         - 분할 문장 2: 이 왕은 조선의 몇 대 왕인가? (outer-question)
         - (분할 문장 2) DEPEND-ON (분할 문장 1)

c) qDecompType: 분할된 질문들 간의 관계에 따라 복합 문장의 질문에 대한 질문 분할 유형을 지정. 질문 분할 유형은 크게 병렬형과 내포형으로 구분
   1) 병렬형: 분할된 문장 내에 Question 이 두개 이상이고, 동일한 정답을 요구하는 경우
      - 예) 이것은 보통 텔레비전 리모컨에서 이용하는 전자기파의 한 종류다. 눈으로는 볼 수 없고 일반적으로 공기 분자에 산란되기 어려워 대기를 잘 통과한다. 1800 년경 영국의 천문학자 허셜이 발견한 이것은 무엇일까?
      - 분할 문장 1: 이것은 보통 텔레비전 리모컨에서 이용하는 전자기파의 한 종류다. (Question)
      - 분할 문장 2: 눈으로는 볼 수 없고 일반적으로 공기 분자에 산란되기 어려워 대기를 잘 통과한다. (Fact)
      - 분할 문장 3: 1800 년경 영국의 천문학자 허셜이 발견한 이것은 무엇일까? (Question)

2) 내포형: 분할된 문장 내에 Question 이 두개 이상이고, 다른 정답을 요구하고, 선행된 Question(inner-question)의 정답을 찾아야, 후행 Question(outer-question)의 정답을 찾을 수 있는 경우

5.3 질문 분류 정보

질문에 대한 정답을 정확히 찾기 위해서는 질문을 다양한 관점에서 분류하여야 하고, 분류된 정보에 기반하여 각기 특화된 전략으로 정답을 추출하는 것이 효과적이다. 이를 위해서는 질문을 특정 기준에 맞춰서 분류할 필요가 있다. 본 표준에서는 질문 분류를 세가지 관점에 따라 대분류하고 있고, 개별 관점 별로 통계에 기반하여 세부 분류하고 있다. 질문 분류의 상세 정보는 아래와 같다.

a) QClassification: 질문 분류 정보를 저장하는 객체
b) vQDomain: 질문을 도메인에 따라 분류하고 정보를 저장하는 객체 배열
   1) qType4Chg: 도메인에 따른 질문 분류 코드 정보(모든 질문 분류에 포함)
      - 도메인에 기반한 질문 분류 기준은 아래와 같음

<표 5-2> 질문 도메인 분류 카테고리 정의

도메인	설명	도메인	설명
건축	건축물과 관련된 질문	예술	예술(미술, 음악 등)과 관련된 질문(작품을 제외한 악기, 사조, 장르, 방식 등)
스포츠게임	스포츠나 게임과 관련된 질문	인물	인물에 대한 질문
과학	물리, 화학, 지구과학, 생물, 기술 등과 관련 질문	작품	예술 작품과 관련된 질문
법	법과 관련된 질문	종교	종교에 대한 질문
사회문화	사회문화, 전통과 관련된 질문	지리	지리와 관련된 질문
언어	언어적 지식에 대한 질문	상식	시사와 관련된 상식 문제
역사	역사적 사실 및 사건과 관련된 질문	Unknown	도메인을 결정하기 어려운 유형의 질문(예: 연상형 질문과 같은 경우)

   2) dWeightCQT: 질문 분류의 신뢰도 정보(모든 질문 분류에 포함)
c) ansQType: 질문에서 요구하는 정답의 형태에 따른 질문 분류 정보 객체
   1) qType4Chg: 정답 형태에 따른 질문 분류 코드 정보

<표 5-3> 정답 형태에 따른 질문 분류 카테고리

분류	설명
가부형	Yes/No의 대답을 요구하는 경우
단답형	단답 형식의 명사(구) 또는 어휘로 정답을 제시해야 하는 경우
서술형	주관식 문장이나 개조식으로 정답을 제시하는 경우
나열형	정답이 1개 이상인 형태의 질문
순서형	정답이 순서대로 제시되어야 하는 질문

d) vSemQType: 질문을 구성하는 개체와 정답의 의미적 관계에 따른 질문 분류 정보를 저장하는 객체 배열

1) 의미적 질문 분류는 개체(subject), 속성(property), 속성값(property value, object)의 SPO 관계에 따른 질문의 의미적 분류 카테고리임.

2) qType4Chg: 의미적 관계에 따른 질문 분류 코드 정보

(a) 코드는 대분류 2 자리, 중분류 2 자리, 소분류 2 자리로 구성된 6 자리 숫자로 표현됨.

(b) 질문 분류는 사용자의 설정에 따라 대분류/중분류/소분류를 선택하고, 해당하는 코드에 값이 채워진 6 자리의 코드 정보가 기재됨.

<표 5-4> 의미적 질문 분류 카테고리

대분류	중분류	소분류	설명
정의형	용어 요청형		정의문이 제시되고 용어를 찾는 질문
	의미 요청형		용어가 제시되고 정의문을 찾아 정답을 제시하는 질문
사실 관계형	속성값 요청형		개체와 속성을 제시하고 속성값을 찾는 문제
추론형	연산 추론형	속성 비교형	속성값에 대한 비교 연산이 수행되어야 하는 질문
		시간 비교형	속성값 중, 시간 정보에 대한 연산이 수행되어야 하는 질문.
		계산형	속성값에 대한 사칙연산이 수행되어야 정답을 제시할 수 있는 질문
정보 요청형			특정한 정보(non-factoid)에 대한 검색을 요청하는 질문
OOD형 (Out-Of-Domain)			멀티미디어 정보와 함께 제시되는 질문. 본 표준은 텍스트 질의 응답에 대한 것이므로 테스트만을 대상으로 함.

5.4 의문사 기반 질문 유형 분류 정보

질문에 포함되어 있는 의문사는 질문의 유형 및 정답 유형을 인식하기 위한 중요한 단서이다. 영어의 5W1H 와 같이 한국어에서 중요한 의문사는 10 가지로 구분된다. 의문사에 기반한 질문의 유형은 아래의 정보와 같이 정리된다.

a) vQTs: 의문사 기반의 질문 유형 정보를 저장하는 객체 변수

1) qt: 의문사 기반의 질문 유형

<표 5-5> 질문 유형 분류 카테고리와 예문

의문사	패턴	예문
누구		미국 대통령은 누구인가요? 은행 금리 조정은 누가하나요? 이순신은 누구인가요?
몇	몇+(의존)명사	서울에는 구가 몇 개 있나요? 튤립은 몇 월에 피나요? 구청은 몇 시부터 오픈하나요?
무슨	무슨+명사	두메산골에서 두메는 무슨 뜻인가요? 풍속과 풍랑은 무슨 관계가 있나요?
어느	어느+명사	불국사는 어느 시에 있나요? 은행 금리는 어느 기관에서 관리하나요?
무엇	명사+무엇 무엇+명사	위지보드가 무엇입니까? 카푸치노는 무엇에서 유래되었나요?
어디		넥타이는 어디에서 기원을 두나요? 가장 인구가 많은 나라는 어디인가요? 와이브로를 개발한 연구기관은 어디인가?
언제		독일은 언제 통일 되었나요? 우체국은 언제 오픈하나요?
얼마		지리산은 얼마나 높은가요? 만리장성을 완공하는 데 얼마나 걸렸나요?
왜		엘리뇨는 왜 생기는 거지요?
어떠하 어찌하		간장게장은 어떻게 만드나요? 팜므파탈은 어떤 의미인가요? 한국은 어찌하여 분단되었나요?

2) strQTClue: 의문사 기반 질문 유형을 결정하는 질문 내 의문사 스트링

5.5 질문 초점 정보

질문 초점은 질문 내에서 정답으로 대체될 수 있는 어휘나 어구를 의미한다. 이는 정답 후보를 질문 초점과 대체한 후, 근거 검색을 수행하기 위함이다. 또한 질문 초점으로 인식된 어구의 중심어는 대부분 어휘 정답 유형으로 볼 수 있으므로, 어휘 정답 유형을 인식하기 위한 단서로서도 활용 가능하다. 질문 초점 인식을 위한 기준은 아래와 같다.

a) 정답을 지칭하는 지시 대명사와 함께 쓰인 (복합)명사 → 지시 대명사+(복합)명사가 질문 초점임.
   - 예 1) 채무자 자신의 재산을 함부로 처분할 가능성이 큰 경우 채무자의 재산을 임시로 확보하는 조치가 필요하다. 이 조치는 무엇일까? → 질문 초점: 이 조치
   - 예 2) 이 그림을 그린 화가는 문명 세계를 벗어나고자 남태평양의 타히티 섬으로 떠나 작품 활동을 했다. 대표 작품으로는 '타히티의 여인들'이 있는데 후기 인상파 화가인 이 사람은 누구일까? → 질문 초점: 이 사람

   주의) '이 그림'은 장학퀴즈에서 제시된 그림을 지칭하는 것으로 정답을 지칭하는 것이 아니므로 '지시 대명사+(복합)명사'이지만 질문 초점으로 인식하지 않음.
b) 정답을 지칭하는 지시 대명사 '이것'
   - 예 1) 이것은 14 행의 짧은 시로 이루어진 서양 시가로 대표적인 작가는 페트라르카, 셰익스피어 등이 있다. → 질문 초점: 이것
   - 예 2) 이것은 사람을 포함한 거의 모든 생물에서 중요한 에너지원으로 쓰인다. 탄수화물이 많은 음식물을 먹은 후에는 혈액 중 이것의 농도가 증가해 한동안 평소보다 높게 유지되는데 이것은 무엇일까? → 질문 초점: 이것, 무엇
c) 질문 내에 포함되어 있는 의문사: 누구, 어떤, 무엇, 몇, 무슨, 어느, 어디, 언제, 얼마나, 얼마
   1) '몇', '어떤', '무슨', '어느'는 뒤의 명사와 함께 질문 초점이 됨.
      - 예) 현재 1 초의 정의는 어떤 원자의 복사선이 진동하는 데 걸리는 시간을 의미할까? → 질문 초점: 어떤 원자
   2) 1)을 제외한 의문사는 의문사만 질문 초점이 됨.
      - 예) 형사 재판에서 원고는 누구일까? → 질문 초점: 누구

위에서 언급한 기준에 따라 인식된 질문 초점 정보는 아래와 같이 JSON 포맷에 저장된다.

a) vQFs: 질문 초점 정보가 저장되는 객체 배열
   1) strQF: 질문 내에 질문 초점에 해당하는 부분의 스트링
   2) dWeightQF: 질문 초점 정보에 대한 신뢰도값

5.6 정답 유형 정보

정답 유형은 질문에서 요구하는 정답의 의미적 개념을 제약하는 정보이다. 정답 유형은 어휘 정답 유형과 의미 정답 유형으로 구분된다.

5.6.1 어휘 정답 유형

어휘 정답 유형은 질문 내에서 정답의 유형을 제약하는 (복합)명사이다. 어휘 정답 유형과 정답 후보들은 어휘 지식의 관계 정보(상하 관계, 동의 관계 등)에 기반하여 정답 후보들을 제약할 수 있다. 어휘 정답 유형의 예는 아래와 같다.

- 예 1) 열을 가장 잘 전달하는 물질은 무엇일까? → 어휘 정답 유형: 물질
- 예 2) 찰리 채플린이 공장에서 나사 조이는 일을 반복하는 공장 노동자로 등장해 현대 산업 사회의 인간 소외를 코믹하게 그려낸 이 영화의 제목은 무엇일까? → 어휘 정답 유형: 영화

어휘 정답 유형에 대한 정보는 아래와 같이 JSON 포맷에 저장된다.

a) vLATs: 어휘 정답 유형 정보가 저장되는 객체 배열
   1) strLAT: 어휘 정답 유형의 스트링 정보
   2) vCompoundLATs: 어휘 정답 유형이 복합 명사일 경우, 가능한 복합 명사의 조합 스트링 목록
      - 예) 대한민국의 대표하는 음악 감독은? → 어휘 정답 유형: 감독, vCompoundLATs: 음악 감독, 음악, 감독
   3) strID: 어휘 정답 유형의 어휘 지식 내 코드 정보(어휘의 의미 정보)
      - 어휘 정답 유형은 어휘 지식에 포함되어 있는 어휘만 대상임(어휘 지식에 기반하여 정답 후보를 제약함으로)
   4) dConfidenceLAT: 어휘 정답 유형의 신뢰도 정보

5.6.2 의미 정답 유형

의미 정답 유형은 질문에서 요구하는 정답의 의미적인 유형으로 “개체명 태그 세트 및 태깅 말뭉치(TTAK.KO-10.0852)” 표준[1]을 기준으로 한다. 즉, 질문에서 요구하는 정답의 개체명이 무엇인가를 인식하여야 한다.
의미 정답 유형 정보는 아래와 같은 JSON 포맷으로 저장된다.

a) vSATs : 의미 정답 유형 정보가 저장되는 객체 배열
   1) atSAT: 개체명 태그 세트 및 태깅 말뭉치(TTAK.KO-10.0852) 표준의 개체명 태그 코드 정보

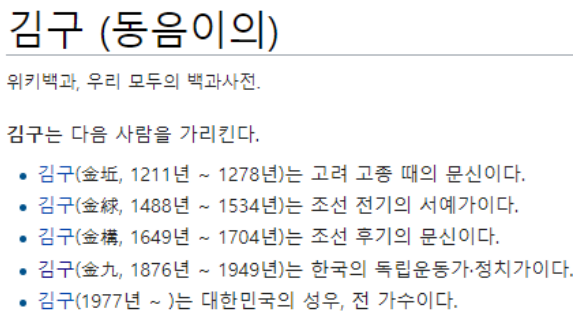
2) dConfidenceSAT: 의미 정답 유형의 신뢰도 정보

5.7 지식 베이스 타이틀과 질문 토픽 정보

질문 내에서는 정답과 연관된 중요한 개체들(지식 베이스의 타이틀들)이 포함되어 있다. 이는 정답을 찾기 위한 대상을 축소시켜 주고, 찾아야 할 문서를 제약하는 중요한 정보를 제공한다. 따라서, 질문에 존재하는 개체들이 여러 개일 경우, 어떤 개체를 핵심 개체로 인식해야 할지는 중요한 문제이다. 이 핵심 개체를 질문 토픽으로 정의한다. 토픽의 대상이 되는 개체들은 사람, 기관, 장소, 나라, 작품 등이 있다. 인식된 토픽이 모호성이 발생할 때는 토픽의 모호성을 해소하는 토픽 제약 정보를 인식하여 토픽의 모호성을 해소한다.

예) 백범 김구는 언제 사망하였나? → 토픽: 김구, 토픽 제약 정보: 백범

위의 예의 경우, 토픽으로 인식된 김구는 동명이인으로 온라인 백과사전에서는 (**오류! 참조 원본을 찾을 수 없습니다.**)와 같이 5 명이 존재한다. 따라서, 어떤 김구인지를 모호성을 해소하기 위해서 '백범'이라는 토픽 제약 정보를 인식하고, 이 정보를 기반으로 (**오류! 참조 원본을 찾을 수 없습니다.**)과 같이 온라인 백과사전의 관련 정보를 기반으로 독립운동가 김구라는 것을 인식하여야 한다.


김구 (동음이의)

위키백과, 우리 모두의 백과사전.

김구는 다음 사람을 가리킨다.

- 김구(金坵, 1211년 ~ 1278년)는 고려 고종 때의 문신이다.
- 김구(金絿, 1488년 ~ 1534년)는 조선 전기의 서예가이다.
- 김구(金構, 1649년 ~ 1704년)는 조선 후기의 문신이다.
- 김구(金九, 1876년 ~ 1949년)는 한국의 독립운동가·정치가이다.
- 김구(1977년 ~)는 대한민국의 성우, 전 가수이다.

(그림 5-2) 온라인 백과사전의 김구(동음이의) 페이지

김구

위키백과, 우리 모두의 백과사전.

이 문서에는 여러 문제가 있습니다. 문서를 편집하여 수정하시거나 토론 문서에서 의견을 나눠주세요. [숨기기]

- 이 글의 중립성에 대한 이의가 제기되었습니다. (2016년 1월)
- 이 문서는 전체적인 내용을 이해하기 어려울 정도로 너무 길게 쓰여 있습니다. 문서의 본 주제에 맞도록 해당 부분들을 요약 정리하거나, 필요하다면 문서를 분할해주세요. 문서에 대한 의견은 토론란에서 나누어 주세요. (2016년 1월 24일에 문서의 요약이 요청되었습니다.)

같은 이름을 가진 다른 사람에 대해서는 김구 (동음이의) 문서를 참조하십시오.

백범은 여기로 연결됩니다. 다른 뜻에 대해서는 백범 (동음이의) 문서를 참조하십시오.

김구(金九, 1876년 8월 29일(음력 7월 11일) ~ 1949년 6월 26일)는 일제강점기 독립운동가이자 대한민국의 종교인, 교육자, 통일운동가, 정치가이다. 한인애국단을 이끌었고 대한민국 임시 정부 주석을 역임하였으며 1962년 '건국훈장 대한민국장'이 추서되었다.

몰락 양반가의 후손[1] 으로 태어나 과거에 응시하였으나 실패, 이후 동학농민운동에 참가하였고, 한때 불교 승려로 활동했으며 이후 개신교(감리교)에 귀의하였다.

자(字)는 연하(蓮下), 처음 이름은 창암(昌巖)이고, 호(號)는 백범(白凡), 연상(蓮上)이다. 호는 미천한 백성을 상징하는 백정의 '백(白)'과 보통 사람이라는 범부의 '범(凡)' 자를 따서 지었다.[2][3] 19세 때 이름을 창수(昌洙)로 바꾸었다가, 37세(1912년)에 거북 '구'(龜)였던 이름을 아홉 '구'(九)로 바꾸었다. 그 밖에 환속 이후의 이름인 김두래(金斗來), 피난 시기에 사용한 가명인 장진(張震), 장진구(張震球)도 있었다. 젊어서 동학교도 였고, 불교에 귀의해서 법명 원종(圓宗)을 얻은 승려였으며[4], 신민회에서 활동하면서 기독교 신자가 되었다.[5] 양산학교, 보강학교 등에서 교육자로 교편을 잡기도 했고, 해서교육총회 학무총감으로도 활동했다. 교육·계몽 운동 중 일본 제국 경찰에 연행되어 수감되기도 하였다. 김방경의 25대손으로 본관은 구 안동이며, 황해도 해주 출신이다.

1919년 이후 상하이에서 대한민국 임시 정부에 참여하여, 의정원 의원, 경무국장, 내무총장, 국무총리 대리, 내무총장 겸 노동국 총판 등을 지냈다. 외교 중심의 독립운동이 성과를 얻지 못하자 1921년 임시 정부 내 노선 갈등 이후 일부 독립운동가들이 임시 정부를 이탈하고, 만주 사변 이후에 일본의 중국 침략이 본격화되면서 중국 관내 여러 지역으로 임시 정부를 옮겨다녔으며, 1924년에는 만주 대한통의부 박희광(朴喜光) 등을 통한 친일파 암살 및 주요공관 파괴, 군자금 모집 등을 비밀리에 지휘하였고, 이후 한인애국단을 조직하여 이봉창의 동경 의거, 윤봉길의 흥커우 의거 등을 지휘하였다.

1926년 12월부터 1927년까지 1930년부터 1933년까지 임시정부 국무령을, 이후 국무위원, 내무장, 재무장 등을 거쳐 1940년 3월부터 1947년 3월 3일까지 임시정부 국무위원회 주석을 지냈다. 1945년 광복 이후에는 임시정부 법통 운동과, 이승만, 김성수 등과 함께 신탁 통치 반대 운동과 미소 공동위원회 반대 운동을 주도하였으며 1948년 1월부터 남북 협상에 참여했다.

김구
金九

(그림 5-3) 백범 김구 선생의 페이지

지식 베이스 타이틀과 질문 토픽 정보는 아래와 같이 JSON 포맷에 저장된다.

a) vTitles: 질문에서 인식된 타이틀들의 정보를 저장하는 객체 배열
   1) strEntity: 인식된 개체의 스트링
   2) strEntityType: 인식된 개체의 개체명 태그 정보
   3) vEntityInfo: 인식된 개체에 대한 지식 베이스 정보 객체 배열
      (a) strNormEntity: 지식 베이스 타이틀로 정규화된 개체의 스트링
         - 온라인 백과사전의 redirection 정보와 같이 이형태(synonym) 타이틀들은 하나의 대표 타이틀로 연결됨. 대표 타이틀이 정규화된 개체의 타이틀임.
      (b) strID: 지식 베이스 타이틀의 ID 번호
      (c) strExplain: 지식 베이스 타이틀의 정의문 스트링
      (d) dWeightEn: 타이틀의 모호성 해소에 대한 스코어
b) vQTopic: 인식된 타이틀들(vTitles) 중에서 가장 중요한 타이틀 정보
   1) 내부 정보는 vTitles 와 동일한 구조임

5.8 정답 제약 정보

정답 제약은 정답을 직간접적으로 제약할 수 있는 질문 내의 단서를 의미한다. 정답 제약의 가장 일반적이고 중요한 단서는 시간과 공간 제약 정보이다. 시간과 공간 정보 제약에 대한 기준은 아래와 같다.

a) 시간 제약
   1) 질문에서 정답을 제약하는 시간 정보
   2) 질문 초점과 어휘 정답 유형과 연관된 시간 정보만 대상으로 함.
      - 예) 이것은 1904 년에 송수만, 심상진 등이 "국가의 존망이 달린 것이므로 조그마한 땅도 양여할 수 없다"는 목표와 의지를 내세워 창설한 항일 단체다. 일본의 황무지 개간권 요구에 반대해 저지시킨 이 단체는 무엇일까? → 시간 제약 정보: 1904 년
b) 공간 제약
   1) 질문에서 정답을 제약하는 공간 정보
   2) 질문 초점과 어휘 정답 유형과 연관된 공간 정보만 대상으로 함.
      - 예) 이곳은 국제 습지 조약 보존 습지로 경상남도 창녕군에 있는 대표적인 자연 습지다. 이곳은 어디일까? → 공간 제약 정보: 경상남도 창녕군

인식된 시간 정보의 정규화는 연월일의 8 자리 숫자와 시분초의 6 자리 숫자가 연결된 14 자리 숫자로 표현된다. 또한, 시간 정보는 특정 시점을 표현하는 경우와 기간을 표현하는 경우로 구분된다. 공간 정보는 공간 정보 지식 베이스와 연관된 URI 로 정규화된다.

시간과 공간 제약 정보는 아래의 JSON 포맷으로 저장된다.

a) answerConstraint: 정답 제약 정보(시간과 공간 제약 정보)를 저장하는 객체
   1) vAcTime: 시간 제약 정보를 저장하는 객체 배열
      (a) TExpression: 질문 내 시간 제약 정보에 해당하는 스트링
      (b) valueAt: 정규화된 14 자의 특정 시점의 시간 정보
      (c) valueBegin: 특정 기간에서 정규화된 14 자리의 시작 시간 정보
      (d) valueEnd: 특정 기간에서 정규화된 14 자리의 종료 시간 정보
   2) vAcLoc: 공간 제약 정보를 저장하는 객체 배열
      (a) LExpression: 질문 내 공간 제약 정보에 해당하는 스트링
      (b) kbURI: 지식 베이스에 기반한 공간 정보의 정규화 URI 정보

부 록 Ⅰ-1

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

지식재산권 확약서 정보

해당 사항 없음.

※ 상기 기재된 지식재산권 확약서 이외에도 본 표준이 발간된 후 접수된 확약서가 있을 수 있으니, TTA 웹사이트에서 확인하시기 바랍니다.

부 록 Ⅰ-2

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

시험인증 관련 사항

해당 사항 없음.

부 록 Ⅰ-3

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

본 표준의 연계(family) 표준

해당 사항 없음.

부 록 Ⅰ-4

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

참고 문헌

[1] TTAK.KO-10.0852, 개체명 태그세트 및 태깅 말뭉치, 2015

부 록 Ⅰ-5

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

영문표준 해설서

해당 사항 없음.

부 록 Ⅰ-6

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

표준의 이력

판수	채택일	표준번호	내용	담당 위원회
제1판	2018.12.19	제정 TTAK.KO-10.1098	-	메타데이터 PG (PG606)