Multi-source multi-strategical machine reading for KB development

Key-Sun Choi
KAIST
Machine Reading Lab.@SWRC
Nature of Multi-source

Text source: 343 out of 384 (89.32%) 

= Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “complete sentence(s)”. 

Bullet statement: 24 out of 384 (6.25%) 

= Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “a word” or “a series of words”. 

Hidden or predictable: 17 out of 384 (4.42%) 

= Questions whose answers “cannot” be found in Korean DBpedia and Wikipedia.
Type 1: Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “complete sentence(s)”. 

Subtype 1-1 : 95 out of 384 (24.73%)

= Questions whose answers “can” be found in Korean DBpedia and Wikipedia “directly” from “one single sentence”. 

e.g.,

Ko.NLQ#81:
대한민국의 제16대 대통령은 누구인가?
Who was the 16th President of the Republic of Korea?

Answer in Ko.DBpedia & Wikipedia:
노무현 (盧武鉉, 1946년 9월 1일 ~ 2009년 5월 23일)은 대한민국의 제16대 대통령이다.
Roh Moo-hyun (盧武鉉, 1 September 1946 - 23 May 2009) was the 16th President of the Republic of Korea.
Subtype 1-2 : 223 out of 384 (58.07%)

= Questions whose answers “can” be found in Korean DBpedia and Wikipedia “directly” from “two or more sentences”.

e.g.,

Ko.NLQ#230:
Who was the person that dropped out of Reed College, founded Apple Inc., and fought against pancreatic cancer?
리드 대학교를 중퇴하고, 애플을 창업하였으며, 췌장암으로 투병한 사람은 누구인가?

Answer in Ko.DBpedia & Wikipedia:

Steven Paul "Steve" Jobs (February 24, 1955 – October 5, 2011) was an American pioneer of the personal computer revolution of the 1970s. Jobs briefly attended Reed College in 1972 before dropping out, … He co-founded Apple in 1976 … In mid-2004, he announced to his employees that he had a cancerous tumor in his pancreas.
Questions whose answers “can” be found in Korean DBpedia and Wikipedia “indirectly” where “some processes of reasoning are required around several sentences”.

e.g.,

**Ko.NLQ#12**:
Who was the last princess of Joseon?
조선의 마지막 황녀는 누구인가?

**Answer in Ko.DBpedia & Wikipedia**:
Princess Deokhye (德惠翁主, 25 May 1912 - 21 April 1989) was the daughter of Emperor Gojong and the noblewoman Yang. Gojong was the 26th and the last king of Joseon. Gojong used to have 9 sons and 4 daughters, but only 3 sons and 1 daughter survived, so that Princess Deokhye was actually the only daughter of Gojong.

덕혜옹주(德惠翁主, 1912년 5월 25일 ~ 1989년 4월 21일)는 조선의 제26대 왕이자 대한제국의 초대 황제였던 고종과 귀인 양씨 사이에서 태어난 딸이다. 고종은 조선의 제26대이자 마지막 임금이다. 고종은 총 9남 4녀의 자녀가 있었지만 3남 1녀만이 성년이 될 때까지 생존하여 덕혜옹주가 사실상 유일한 딸이 되었다.
Type 2: Answers in a fragment or \{word\}^+

- Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “a word” or “a series of words”
- 24 out of 384 (6.25%)
- Questions whose answers “can” be found in Korean DBpedia and Wikipedia in the form of “a word” or “a series of words”.

• 韩国战略与金融部下所属的组织是什么？
• 대한민국 기획재정부의 산하 기관들은 무엇인가?
• What are the affiliated organizations of the Ministry of Strategy and Finance of South Korea?
  • Ko.NLQ#168:
• When was the year that the Korea Underwater Association changed its name to “Korea Scuba Diving Association”, joined the Confédération Mondiale des Activités Subaquatiques (CMAS), joined the Korean Olympic Committee, and changed its name to “Korea Underwater Association” respectively?
History

- 1968년 3월 6일 : 한국스쿠버다이빙클럽 설립
- 1969년 7월 20일 : 제1회 전국수중경기대회 개최 및 매년 민수경, 수중방학경기, 수중체험, 수중수영 경기 개최
- 1973년 3월 2일 : 한국스쿠버다이빙협회 개칭  2 March 1973; changed its name to "Korea Scuba Diving Association"
- 1981년 11월 20일 : 세계수중연맹(CMAS) 가입을 위해 한국수중협회로 개칭
- 1982년 6월 5일 : 세계수중연맹(CMAS) 가입  5 June 1982; joined the CMAS
- 1988년 12월 27일 : 아시아수중연맹(AUF) 가입 및 대한체육회 준회원 가입 및 대한수중협회 개정
- 1992년 2월 24일 : 대한체육회 정식 가입  24 February 1992; joined the Korean Olympic Committee
- 1997년 4월 7일 : 사단법인 대한수중협회 설립 개회
- 2010년 5월 3일 : 사단법인 대한수중협회로 명칭 변경  3 May 2010; changed its name to "Korea Underwater Association"
• 小说“哈利波特”的副标题是什么？

• 소설‘해리 포터’의 부제목들은 무엇인가?
• What are the subtitles of the novel ‘Harry Potter’?

• Ko.NLQ#218
해리 포터

미국판 소설에서 처음 사용된 (해리 포터) 로고

저자 : 조앤 K. 롭링

국가 : 영국

출판사 : Bloomsbury Publishing

출판일 : 1997-6-26

소설 사례

[표]

| 출간 일자 | 제목 | 제목(한국어로)
|----------|-----------------|----------------|
| 1997년 6월 26일 | 해리 포터의 불변의 돌 | 해리 포터와 이사람들이 시작해
| 1998년 7월 3일 | 해리 포터와 비밀의 방 | 해리 포터와 비밀의 방
| 1998년 7월 3일 | 해리 포터와 이사람들이 시작해 | 해리 포터와 이사람들이 시작해
| 2000년 7월 8일 | 해리 포터와 황금의 관 | 해리 포터와 황금의 관
| 2003년 7월 3일 | 해리 포터와 불의 주인 | 해리 포터와 불의 주인
| 2005년 7월 8일 | 해리 포터와 원석의 신 | 해리 포터와 원석의 신
| 2007년 7월 12일 | 해리 포터와 죽음의 임금 | 해리 포터와 죽음의 임금

소설 사례

(끝)
Wikipedia: The Rich Source of Knowledge

• Wikipedia describes such information with various ways:

  **Articles:**

  Alan Turing
  From Wikipedia, the free encyclopedia
  
  "Turing" redirects here. For other uses, see Turing (disambiguation).
  Alan Mathison Turing, OBE, FRS ("Turing"); 23 June 1912 – 7 June 1954) was a British pioneering computer scientist, mathematician, logician, cryptanalyst, philosopher, mathematical biologist, and marathon and ultra distance runner. He was highly influential in the development of computer science, providing a formalisation of the concepts of algorithm and computation with the Turing machine, which can be considered a model of a general purpose computer. Turing is widely considered to be the father of theoretical computer science and artificial intelligence.[5] During the Second World War, Turing worked for the Government Code and Cypher School (GCHQ) at Bletchley Park, Britain's codebreaking centre. For a time he led Hut 8, the section responsible for German naval cryptanalysis. He devised a number of techniques for breaking German ciphers, including improvements to the pre-war Polish bombe method, an electromechanical machine that could find settings for the Enigma machine. Turing played a pivotal role in cracking intercepted coded messages that enabled the Allies to defeat the Nazis in many crucial engagements, including the Battle of the Atlantic; it has been estimated that this work shortened the war in Europe by as many as two to four years.[6]

  → Related researches: Learning by Reading

  **Infoboxes:**

  ![Alan Turing Infobox]

  → Related researches: DBpedia, Yago

  **Categories:**

  Our target

  → Related researches: Yago, Catriple

  Categories: Alan Turing | 1912 births | 1954 deaths
  20th-century mathematicians | 20th-century philosophers
  Academics of the University of Manchester Institute of Science and Technology
  Alumni of King's College, Cambridge
  Artificial intelligence researchers | Atheist philosophers
  Bayesian statisticians | British cryptographers
  British long-distance runners | British male athletes
  British people of World War II | Computability theorists
  Computer designers | Deaths in Cheshire | English atheists

  → Related researches: Yago, Catriple
Subtype 3-1  :  14 out of 384 (3.64%)

= Questions whose answers “cannot” be found in Korean DBpedia and Wikipedia “partially”, because “some portions of information are lacked in the KBs”. e.g.,

Ko.NLQ#114 :
남이장군, 겨울연가, 연가지가와 관련 있는 섬은 무엇인가?
What is the island related to General Nami, Winter Sonata and Yeongajiga?

Answer in Ko.DBpedia & Wikipedia :
Namisum is a tiny half-moon shaped island located in Chuncheon, South Korea, ⋯. Its name originated from General Nami, ⋯. The island, especially the Metasequoia path, was one of the main filming locations of the Korean Broadcasting System 2002 television drama series Winter Sonata, starring Bae Yong-joon and Choi Ji-woo.

Korean DBpedia and Wikipedia are not inclusive of any information about “Yeongajiga”.

Multi-strategy for KB (Active WiseKB)

Knowledge Base

Knowledge Population Engine
- L2K
- B2K
- C2K
- K2K

KB Completion Control Engine
- Entity Summarization, Active learning
- Distant supervision-based IE learning: Iterative evolving process
- Human-Intervention

Incomplete frame-surface-knowledge graph engine:
- FRDF(Frame-based)
- SRDF(graph-based)

Corpus (Web)

SPARQL
Distant supervision-based IE learning

Web (Wikipedia)

KB

Local Data Construction

Plain text (fully unstructured)
백남준은 한국 태생의 미국 미술가, 작곡가, 전위예술가이다.

Itemized text (Wikipedia bullet-list)
백남준

학력
- 서독 뮌헨 대학교 대학원 음악사학과 졸업 (음악학 석사)
Relation Extraction Learning Method

L2K

(Plain) Local Data ➔ 구성요소 자동 추출 ➔ Rel-Syntactic Patterns (parse) ➔ High-precision Human Intervention ➔ Text extraction patterns ➔ Decision Tree Model ➔ Labels

Rel-Lexical Patterns (words)

Features ➔ 자질 추출

Can be tailored to specific domains

Fast, no extra NLP resource/tool

B2K

(Item) Local Data ➔ 구성요소 자동 추출 ➔ Label-Syntactic Patterns (wiki/html structure) ➔ Classifier type (m-SVM) ➔ Labels

Label-Lexical Patterns (words)

Features ➔ 자질 추출

Features

Decision Tree Model

Labels

Evidence Integration

Can be tailored to specific domains

Fast, no extra NLP resource/tool
Frame KBQA

Who is the first person reached the South Pole?

- **Pseudo-QUERY**
  - frame: Arriving
    - fe:goal: South Pole
    - fe:theme
  - frame: People
    - fe:experience: First_experience
    - fe:name: ?x

- **Frame semantic-based Pseudo SPARQL query (QAF)**
  - linguistic-level semantics
    - ≈ bottom-up knowledge grounding
  - event-central representation
    - cause & effect, emotion, opinion motion, problem & solution, ...
      - more than KB T-Box
  - Well-known semantic parsing approach to understand natural language