다중소스 기반 질의 응답 시스템

박선영○, 권순철, 최준휘, 유환조, 이근배
포항공과대학교, 컴퓨터공학과
{sypark322, theincluder, chasunee, hwanjoyu, gblee}@postech.ac.kr

Multi-source based Question Answering System

Seonyeong Park○, Soonchoul Kwon, Junhwi Choi, Hwanjo Yu, Gary Geunbae Lee
Department of Computer Science and Engineering, Pohang University of Science and Technology

요 약

본 논문에서는 지식베이스와 다중 소스 레이블 문서를 동시에 활용한 다중소스 기반 오픈 도메인 질의 응답 시스템에 대해 소개한다. 제안하는 질의 응답 시스템은 자연어처리를 기반으로 한 질의 분석 모듈, SPARQL (Simple protocol and RDF Query Language) query 생성 및 검색 부분, 다중 소스 레이블 문서 검색 부분으로 이루어져 있다. 정확도가 높은 지식베이스 기반의 질의 응답 시스템으로 정답을 우선 탐색한다. 지식베이스 기반 질의 응답 시스템으로 정답을 찾는 데 실패하거나, SPARQL query 생성에 실패하면, 다중 소스가 레이블된 문서 검색을 통해 정답을 찾는다. 제안하는 질의 응답 시스템은 지식베이스만 사용한 질의 응답 시스템보다 높은 성능을 보인다.

주제어: 질의 응답 시스템, SPARQL, 다중소스 레이블 문서 검색

1. 서론

구글, 네이버와 같은 검색 엔진이 사용자의 질의와 관련된 문서들을 찾아주는 것과 다르게, 질의 응답 시스템은 자연어 질문에 짧은 답을 출력하는 시스템이다. 질의 응답 시스템의 이러한 특징은 빅 데이터 시대에 필요한 정보를 찾고자 하는 사용자의 욕구를 충족시킬 수 있다. 질의 응답 시스템은 정보 검색 기반 질의 응답 시스템 (Information retrieval based Question Answering, IRQA)과 지식베이스 (Knowledgebase, KB)를 기반으로 한 질의 응답 시스템 (Knowledgebase based Question Answering, KBQA)으로 나뉜다.

IRQA [1, 2]은 정답 유형을 분석하고 질의를 어휘적, 구문적, 의미적으로 분석하는 모듈, 분석된 결과를 활용하여 문서 검색을 위한 쿼리를 생성하는 모듈, 문서 집합에서 쿼리와 관련된 단락과 문장을 찾는 모듈, 정답 후보를 추출하고 순위화하여 최종적으로 정답을 출력하는 모듈로 이루어진다. 최근에는 DBpedia와 Freebase와 같은 거대하고 [3, 4], 구조화된 지식베이스가 등장했다. 이러한 지식베이스는 개체 단위의 지식이 연결되어 있는 linked form을 따르고 있으며, <entity, relation, entity> 로 이루어진 Triple 단위로 구성되어 있다. 일반적으로 가공되지 않은 문서집합보다 데이터베이스의 크기는 작지만 정제과정을 거쳤고, 구조화되어 있고 semantic inference가 가능하다는 등의 장점이 있다. Template-based SPARQL Learner (TBSL)[5] 과 ontology-based Question answering System Pythia [6] 등 대부분의 KBQA는 질의 분석 단계, 자연어 질의에 있는 개체(entity)와 서술어(predicate)을 지식베이스의 entity와 relation에 매핑하는 단계, 매핑결과를 기반으로 자연어 질의를 SPARQL query로 변환하는 단계, SPARQL query 검색 후 정답을 출력하는 단계로 이루어져 있다. 일반적인 KBQA는 semantic parsing등을 통해 질의를 Logical Form 으로 구조화하여 SPARQL query로 변환하는 것이 중요하다. 또한 문서 검색을 기반으로 한 IRQA에서는 문서 집합에서, query와 관련된 단락, 쿼리와 관련된 문장을 찾는 과정, 정답 후보들에서 정답을 랭킹하는 과정이 중요하고, 이 때 정답이 가진 문맥을 알 수 있다는 장점이 있다. 이 정답이 어떤 문서에서 검색되었고, 어떤 단락에서 나타났는지 등의 정보는 정답을 찾는데 주요 정보로 활용된다. 반면에, KBQA에서는 이런 정보가 활용되지 않고, SPARQL query의 결과가 정답이 된다. 기존의 연구는 문서 검색을 기반으로 한 IRQA 또는 지식베이스를 기반으로 한 KBQA와 같은 단일 소스를 기반으로한 질의 응답 시스템이 연구되어 왔다. [1,2,5,6]

본 논문에서는 지식베이스와 문서 검색의 장단점을 인지하고, 두 데이터를 모두 활용한 다중소스 기반의 질의 응답 시스템을 제안한다.

2. 시스템 구조

본 논문에서 제안하는 다중소스 기반 질의 응답 시스템 (그림 1) 은 질문에서 주요 정보를 추출하여, SPARQL query를 생성하여, 지식베이스에서 먼저 정답을 찾는다. SPARQL query 생성에 실패하거나, SPARQL query 검색에 실패하면 다중 소스 레이블 문서 검색을 통해 정답을 찾는다. SPARQL query 생성부는 질의 분석부분, slot과 SPARQL query template 추출 부분, resource (named entity) 와 class (type of named entity) URI(uniform resource indicator)를 식별하는 부분, 자연어 질의에서 서술어와 지식베이스의 property URI와 매핑하는 부분 (자연어 패턴 검색, property URI 후보 추출, property URI 후보들과 자연어 서술어의 의미적 유사도를 측정하

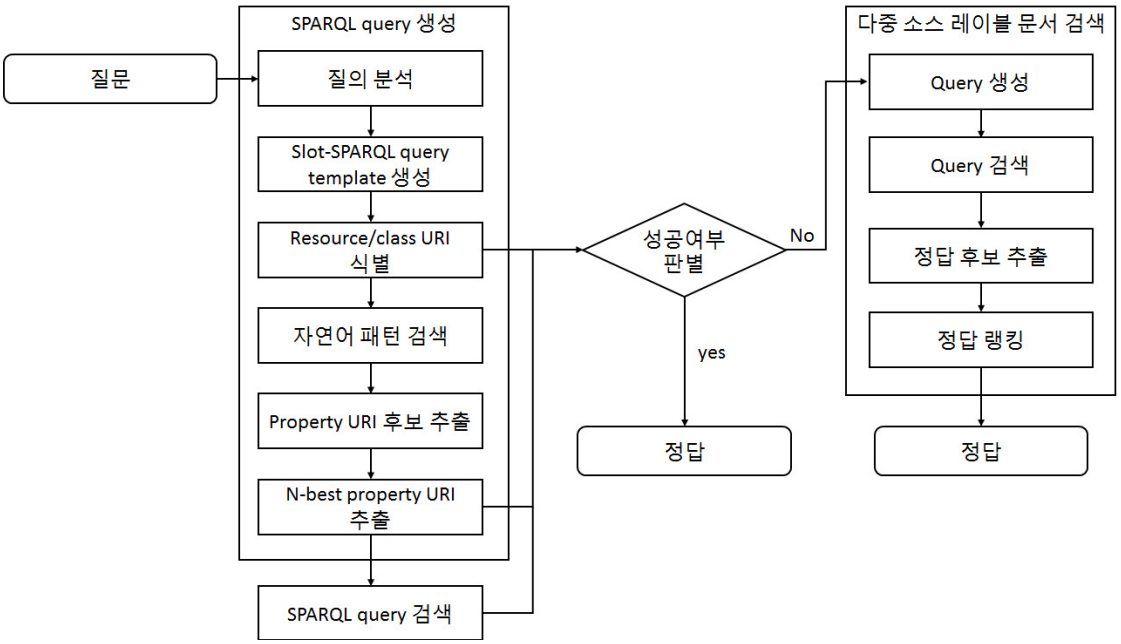

그림 1. 전체 시스템 구조 및 흐름도

여 유사도가 높은 순으로 property URI를 추출하는 부분)을 통해 SPARQL query를 생성하여, 지식베이스의 일종인 DBpedia[1]에서 정답을 찾는다. 이 때, SPARQL query 생성에 실패하는 경우 (resource URI 식별 실패 또는 property URI 매핑하는 부분에서 유사도 임계값을 넘지 않는 경우) 또한 SPARQL query 생성에는 성공했으나, DBpedia에서 결과를 찾지 못한 경우에는 다중 소스 레이블 문서 검색을 통해 정답을 찾는다. 시스템 구조는 크게 2.1 SPARQL query 생성부 2.2 다중 소스 레이블 문서 검색부로 나뉜다.

2.1. SPARQL query 생성부

- 질의 분석 모듈: “who", "what", "where" 과 같은 의문사를 검출하고, POS, Dependency Parse Tree, Semantic Role label, Chunk, named entity를 추출한다. 또한 각 의문사와 언어처리 결과를 활용하여 휴리스틱으로 class를 추출한다. class는 lexical answer type을 의미하며 대부분 명사이다. class는 질의에서 나타나는 단어로 이루어져 있으며, 추출되지 않을 수 있다.

- Slot-SPARQL query template 모듈: 질의 분석 결과를 이용해 휴리스틱 룰을 활용하여 Slot과 SPARQL template을 추출한다. Slot과 Sparql query template은 TBSL[5]에서 정의된 개념이다. Slot은 *"In which country does the Nile start?"* 라는 질의가 있을 때, *<?x, resource, Nile>, <?c, class, country> <?p, property, start country>* 와 같이 <최종적으로 치환될 DBpedia의 URI, URI의 type, 자연어 질의에서 나타난 단어>로 이루어져 있다. 다음 모듈에서 named entity는 resource 와 매핑되고, 본동사와 본동사의 argument는 property와 매핑된다. SPARQL query template은 query의 형식 구조로 질의에서 lexical word를 검출하여 휴리스틱으로 추출한다. How many~, How, why등을 제외한 의문사로 시작하는 질의는 *SELECT DISTINCT ?x WHERE{?x ?p ?y.}*, 사실여부를 묻는 질의는 *ASK WHERE{?x ?p ?y.}*의 형식으로 SPARQL query를 생성해야 한다.

- resource/class URI 식별 모듈: Slot extraction에서 검출된 결과를 이용하여 부분 매칭을 통해 resource/class URI를 식별한다. class는 질의 분석 모듈에서 검출 될 수도 있고 검출 되지 않을 수 있다. 하지만, resource URI 검출에 실패하면 SPARQL query 생성이 불가능함으로 다중소스 레이블 문서 검색을 통해 정답을 찾는다.

- 자연어 패턴 검색 모듈: <서술어 - DBpedia property URI> 의 구조를 따르는 PATTY[2] 패턴 저장소에서 Slot에서 추출된 서술어를 검색한다. 검색시 DBpedia property URI에 매핑된 결과가 없으면 의미적 유사도를 통해 DBpedia property URI 매핑을 시도한다.

- Property URI 후보 추출 모듈: 이전 모듈에서 추출한 resource URI와 class URI를 활용하여, 해당 resource 또는 class와 연결된 property URI를 추출하여 매핑할 property URI 대상 후보로 한다. DBpedia 전체 Property URI들과 질문에서 추출한 서술어와 모두 의미적 유사도를 비교하기에는 그 범위가 크기 때문이다. 아래의 SPARQL query를 사용하여 property URI 후보들을 추출한다. *SELECT DISTINCT ? property WHERE {{<IDENTIFIED_RESOURCE_URI> ?p []} UNION {[] ?p <IDENTIFIED_RESOURCE_URI>.}}*

- n-best property URI 추출 모듈: 질의에서 서술부를

1) http://dbpedia.org/

2) http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/patty/

추출하고, 이 서술부와 property URI후보들 각각의 의미적 유사도를 측정하여 가장 유사도가 높은 2개의 property URI를 이용하여 query를 만든다. n-best property의 수는 학습 데이터를 통해 조절하였다. 의미적 유사도를 측정하기 위해서, Explicit Semantic Analysis (ESA) [7]에서 사용한 방법을 활용하였다[3]. 타겟 스트링을 위키피디아 개념들에 대한 weighted vector로 변환한다. 이 vector간의 거리를 측정하여 질문의 서술어와 property URI의 유사도를 측정한다. 앞서 언급한 유사도가 임계값을 넘지 못하면 SPARQL query 생성에 실패하여 다중 소스 레이블 문서 검색을 통해 정답을 찾는다. 2014년도에 개최된 Question Answering over linked data (QALD) 경연대회[4] 에서 제공하는 학습 데이터를 통해 임계값을 조절하였다.

2.2. 다중 소스 레이블 문서 검색부

- 다중 소스 레이블 데이터베이스: 본 논문에서 제안하는 시스템에서 사용되는 다중 소스 레이블 데이터베이스는 영어 위키피디아 원 문서에서 발생하는 정보 손실을 막기 위해 NLP tool 및 named entity linking tool [8] 을 활용하여 다중 소스가 레이블된 데이터베이스를 활용하였다. 예를 들면, 위키피디아 원문에 "Kim moved to Gunpo when she was six years old and later attended Suri High School." 라는 문장에서 다양한 정보를 태깅하여 아래와 같이 의미 손실은 막는 형태의 문장들로 구성된 데이터베이스를 구축하는 것이다. She가 의미하는 것이 무엇인지 Co-reference resolution을 활용하여 레이블링하고, Spotlight tool [8]을 활용하여, Named entity가 DBpedia의 어떤 URI form인지 레이블링하고 해당 DBpedia URI의 type들을 레이블링하여 정보 손실을 막는다.태깅 결과는 아래와 같다. “Kim[co:Kim Yuna DE:Kim_Yu-Na Type:Agent -> Person -> Athlete -> FigureSkater] moved to Gunpo [Type:Place -> PopuatedPlace -> Settlement->City] when she [Co:Kim Yuna D.E:Kim-Yu-Na Type:Agent -> Person -> Athlete -> FigureSkater] was six years old and later attended Suri High school."

- Query 생성 및 검색 모듈: 앞의 질의 분석 모듈을 활용하여 named entity와 명사 동사 등의 주요 키워드에 weight를 주고 Wordnet에서 유사어를 추출하여 Query를 extension하여 문서 검색을 한다.

- 정답 후보 추출 모듈: named entity가 표시된 다중 소스 레이블 데이터 베이스의 속성을 활용하여, Query와 관련된 문서의 문장들에서 named entity를 추출하여 정답 후보로 간주한다.

- 정답 랭킹 모듈: 오픈소스 Ephyra [9] 를 활용하여, 기계학습 기반 방법과 룰 기반 방법을 하이브리드한 정답 유형 검출기를 이용하고, 정답 유형을 정답 랭킹에 활용한다. 또한 정답 후보의 다중 소스 레이블 데이터베이스에서 검색 가능한 문맥 정보를 활용하여, 정답을 찾는다. 정답 후보가 나온 문장, 정답 후보가 나온 문서의 Title 정보와 질의문과의 Cosine similarity, Wordnet Similarity를 활용한다.

3. 실험 및 결과

시스템 성능 평가 방법으로 QALD 경연대회에서 사용한 성능 측정 방법 중 질의에 대한 응답율이 100%일 때의 recall, precision, F-1 score를 측정하였다. 최근 질의 응답 시스템 연구 [10, 11] 에는 precision, recall, f-1 score 등을 성능 측정에 활용한다. (표 1) 질의 셋은 공인된 QALD 경연대회 평가셋을 이용하였다. 평가셋은 총 50 질의이다. 검색 대상 데이터베이스로는 지식베이스는 DBpedia 3.9 와 텍스트 데이터베이스로는 영어 위키피디아를 사용하였다. Method에 P는 property URI 매핑에 PATTY 를 이용한 패턴 매칭만 사용한 결과이다. P+ESA는 property URI 매핑에 PATTY와 ESA를 활용한 의미적 유사도를 이용한 결과이다. P+ESA+IR은 PATTY와 ESA를 활용했음에도 SPARQL query를 생성하지 못하거나 답을 찾지 못했을 때 추가적으로 정보 검색을 활용하여 답을 찾는 방법을 취했을 때 결과이다. PATTY만을 이용하여 property URI 매핑을 시도했을 경우, property URI 매핑에 실패하여 SPARQL query를 생성할 수 없었음으로 recall, precision, F-1 score는 0이다.

표 1. QALD-4에 대한 평가 결과

Method	recall	precision	F-1 score
P	0.0	0.0	0.0
P+ESA	0.26	0.21	0.23
P+ESA+IR	0.27	0.23	0.25

50개의 질의 중에서 P+ESA 방법으로 28개의 SPARQL query 생성에 성공하였다. SPARQL query 생성 실패 원인은 아래와 같다.

- resource URI 식별 실패

ex)How many James Bond movies are there?

"James Bond"가 Entity로 식별되었으나, 정답을 찾기 위한 resource인 “JamesBondFilms"과 매핑되기 위해서는 “James Bond Movie"로 식별되어야 한다.

- property URI 매핑 실패

ex) Does the Isar flow into a lake?

"flow into"가 “riverMouth"로 매핑된다. 정답을 찾기 위해서는 "inflow"로 매핑되어야 한다.

4. 결론

언어처리 툴을 이용하여, SPARQL query를 추출하기 위한 Slot과 Template을 추출하였다. 또한 패턴 매칭과

3) http://www.cs.technion.ac.il/~gabr/resources/code/esa/esa.html

4) http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=home&q=home

ESA를 이용한 의미적 유사도 비교를 사용하여 질의에서 검출된 서술어와 DBpedia의 property URI 매핑하였다. 지식 베이스 기반 검색을 통해 답을 찾지 못하는 경우 뿐만 아니라. resource 식별에 실패할 경우, property URI 매핑에 실패할 경우 지식베이스 기반 검색을 시도하지 않고 문서 검색을 통해 정답을 찾는 방법을 활용하였다. 결과적으로 지식베이스만을 활용한 시스템보다 SPARQL query 생성 실패시, 문서검색을 시도한 방법의 성능이 높았다. 본 시스템은 DBpedia를 지식베이스로 사용하였지만, 도메인에 의존적이지 않은 방법이기 때문에 다른 지식베이스에서도 동일하게 적용될 수 있다. 현 시스템은 베이스라인 시스템임으로 각 모듈의 성능이 높고, IRQA의 성능이 높으면, 기존의 KBQA에 비하여 더 큰 성능 향상을 기대할 수 있다.

사사

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [R0101-15-0176, Symbolic Approach 기반 인간모사형 자가학습 지능 원천 기술 개발]

참고문헌

[1] Harabagiu, S. M., Maiorano, S. J., & Pasca, M. A.,Open-domain textual question answering techniques. Natural Language Engineering, 9(3), 231-267, 2003

[2] Lee, G. G., Seo, J., Lee, S., Jung, H., Cho, B. H., Lee, C., ... & Kim, K.,SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP. In TREC, 2001

[3] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. DBpedia a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal,2003

[4] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J, Freebase: a col-laboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250). ACM, 2008

[5] Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A. C., Gerber, D., & Cimiano, P., Template-based question answering over RDF data. In Proceedings of the 21st international conference on World Wide Web (pp. 639-648). ACM, 2012

[6] Unger, C., & Cimiano, P. Pythia: Compositional meaning construction for ontology-based question answering on the Semantic Web. In Natural Language Processing and In-formation Systems (pp. 153-160). Springer Berlin Heidelberg, 2011

[7] Gabrilovich, E., & Markovitch, S., Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In IJCAI (Vol. 7, pp. 1606-1611), 2007

[8] Mendes, M. Pablo., Jakob, M., García-Silva, A., and Bizer,C., DBpedia Spotlight: Shedding Light on the Web of Documents. Proceedings of the 7th International Conference on Semantic Systems. 1-8, 2011

[9] Schlaefer, N., Ko, J., Betteridge, J., Sautter,G., Pathak, M., and Nyberg, E., Semantic Extensions of the Ephyra QA System for TREC 2007. Proceedings of the Sixteenth Text RE-trieval Conference, 2007

[10] Fader, A., Zettlemoyer, L., & Etzioni, O., Open question answering over curated and extracted knowledge bases. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1156-1165). ACM, 2014

[11] Unger, C., Forascu, C., Lopez, V., Ngomo, A. C. N., Cabrio, E., Cimiano, P., & Walter, S., Question answering over linked data (QALD-4). In Working Notes for CLEF 2014 Conference., 2014