보기 검증을 통한 일본 센터 시험 문제 해결

권순철○, 남대환, 유환조, 이근배
포항공과대학교, 컴퓨터공학과
theincluder@postech.ac.kr, dhnam@postech.ac.kr, hwanjoyu@postech.ac.kr, gblee@postech.ac.kr

Solving Japanese Center Exam with Choice Verification

Soonchoul Kwon○, Daehwan Nam, Hwanjo Yu, Gary Geunbae Lee
Department of Computer Science and Engineering, POSTECH

요 약

이 논문에서는 한국의 수능 시험에 대응하는 일본 센터 시험의 세계사B 문제를 해결하는 시스템을 만들고 그 성능을 평가했다. 이 시스템은 문제의 각 보기의 신뢰도를 검증하여 어떤 보기가 참인지를 결정한다. 보기 검증을 위해 지식 베이스 기반, 정보 검색 기반, 시간적 제약 기반 검증을 사용하였다. 성능 평과 결과 6개년도 시험 중 5개 시험에서 통계적으로 의미 있는 결과를 얻었다. 이 시스템은 영어를 대상으로 하나, 한국어에도 존재하는 리소스를 사용했기 때문에 한국어에서도 같은 방법론을 적용할 수 있을 것으로 본다. 후속 연구로는 보기의 의미적 분석과 개체명 이외의 정보에 대한 검색이 필요하다.

주제어 : 수능 시험, 신뢰도 검증, DBpedia, NTCIR

1. 서론

질의 응답(question answering; QA) 시스템은 사용자의 자연어 질문에 대해 그 정답을 답해주는 시스템이다. 이러한 시스템의 구축에는 사용자의 질문을 이해하는 자연어 분석 기술, 해당 정보를 데이터베이스에서 검색하는 정보 검색 기술, 정보 검색을 위한 데이터베이스를 구축하는 기술 등이 필요하다.

자연어 질의에는 단순한 사실을 묻는 경우도 있지만, 추상적 사실을 묻거나 추론을 필요로 하는 경우도 있다. 예를 들어, '신라와 경쟁 관계에 있던 국가는?' 이라는 질의는 '경쟁 관계' 라는 추상적인 관계를 이해해야 하며, 이 관계는 '전쟁', '비난' 등의 구체적인 사실에서 추론해야 한다.

복잡한 질의에 대해 정답을 내기 위해서 지금까지는 의미역 표지 부착, semantic parsing 등의 의미 분석 방법을 활용했다. 그러나 이런 방법들은 형태소 분석과 같은 구조 분석 방법보다 정확도가 낮아 오류가 발생하며, 이런 방법을 사용해도 성능이 높지는 않았다.

이 연구에서는 지식 베이스(knowledge base; KB), 정보 검색(information retrieval; IR), 시간적 제약을 이용한 3가지 방법론을 결합해 일본 센터 시험의 복잡한 문제를 해결하였다. 이 방법론은 각 보기의 신뢰도를 검증하는 방법으로, 의미적 분석 없이 개체명 인식(named entity recognition; NER)[1]과 기존에 구축된 지식 베이스, Wikipedia text에 대한 정보 검색을 이용해, 한국어 등 다른 언어에 적용이 간단하다는 장점이 있다.

2. 관련 연구

일본 NTCIR 학회에서는 지난 2014년 QA-Lab task를 통해 일본 센터 시험을 푸는 과제를 제안했다[2]. 해당 과제는 일본어와 영어로 진행됐는데, 일본어에서 최고 성적을 거둔 DCUMT 팀에서는 semantic parsing과 entity type을 이용했고[3], 영어에서 최고 성적을 거둔 CMUQA 팀에서는 semi-phrasal query, passage ranking, source and query expansion 등의 방법을 사용하였다[4]. 그러나 해당 방법들은 데이터와 NER, semantic parsing 등의 자연어 처리 리소스가 충분하지 않은 한국어를 대상으로 적용하기는 힘들다. 한국어 적용을 위해서는 한국어에서 해당 데이터와 리소스가 준비되기 전까지는 한국어에도 존재하는 리소스를 활용한 연구가 필요하다.

3. 데이터 분석

대상 실험 데이터는 한국의 수능 시험에 대응되는 일본 센터 시험(大学入試センター試験)의 세계사B 과목(世界史Ｂ)이다. 센터 시험은 본래 일본어이나, 일본 국립정보학연구소(NII; National Institute of Informatics)에서 NTCIR-11 학회의 QALab task[1]의 진행을 위해 영어로 번역, 마크업, 보기 유형 분류를 한 'Center Shiken Exam Data' 를 사용하였다[2]. 해당 데이터는 XML 파일로 제공되어, 각 대문제와 소문제, 문제 설명, 지문, 참조 대상, 보기, 보기 유형이 마크되어 있다. 데이터는 총 6개년도, 230개의 소문제로 이루어져 있다.

각 연도의 시험은 공통된 주제와 지문을 가진 4개의 대문제와 각 대문제에 딸린 소문제로 이루어져 있다. 각 소문제는 4~6개의 보기를 갖고 있으며, 주어진 지문을 참조하여 풀도록 되어 있다. NII는 각 소문제마다 보기의 유형을 분류하여 정의했는데, *(sentence)* 유형이 총 230문제 중 160 문제로 2/3를 차지하며, 이외에 *(symbol-TF)*2*, *(term_person)* 등의 유형이 그 뒤를 잇는다. 소문제의 과반수를 차지하는 *(sentence)* 보기 유형은 각 보기가 한 문장 형태의 서술로 되어 있으며, 주어진

1) http://research.nii.ac.jp/qalab/

보기 중 역사적으로 맞는 (또는 틀린) 보기를 선택하는 유형이다. 그 다음으로 많은 *(symbol-TF)*2* 유형은 소문제에서 제시된 2개의 지문의 참-거짓 여부를 옳게 나타낸 보기를 선택하는 유형이다.

4. 문제 분류 및 문맥 결정

데이터에서 분석된 보기 유형은 문제 해결 방식에 중요한 단서가 되며, 각 유형마다 다른 방법의 해결 방식이 필요하다. 예를 들어 *(sentence)* 보기 유형은 각 보기를 검증해야 하지만, *(symbol-TF)*2* 유형은 지문을 검증한 뒤 그에 맞는 보기를 선택하는 과정이 필요하다. 우리는 보기 유형을 크게 4종류로 나누어 각각 거의 유사하지만 다른 방법으로 처리했다. 그림, 지도를 활용해야 하거나 처리하기 힘든 유형에는 응답하지 않았다.

각 소문제는 주어진 지문을 참조하여 풀게 되어있으나, 어떤 문제는 필요한 정보가 지문이 아니라 문제 설명에 있으며, 어떤 문제는 오히려 지문을 참조하는 것이 오답을 일으켰다. 따라서 보기와 함께 볼 문맥을 결정하는 것도 문제를 푸는 데 중요한 요소이다. 데이터 분석을 통해 ① 지문이나 문제 설명에 시간 표현이 있을 때, ② it, he, their 등 대명사가 있을 때, ③ 단 하나의 개체명으로 되어있을 때, ④ 3단어 이내로 이루어졌을 때, 해당 참조 부분이나 문제 설명을 보기 검증에 필요한 문맥(context)으로 보고 보기의 일부처럼 처리했다.

5. 보기 검증

센터 시험은 고등학생 수준의 넓은 범위의 지식과 문제 이해 능력이 필요하나, 한 가지 데이터에서 필요한 지식을 찾기도 힘들고, 현재 자연어 처리 기술은 사람 수준의 처리 능력을 갖지 못한다. 따라서 한 가지 데이터와 방법론으로 센터 시험 문제를 해결하는 대신 우리는 크게 두 가지 데이터와 세 가지의 방법론으로 이 문제에 접근했다. 각각의 방법론은 보기가 사실이라고 추측할 수 있는 단서들에 점수를 매기며, 이 점수들의 합으로 보기의 신뢰도를 검증한다.

5.1. KB evidence calculator

KB는 사실을 정형화된 구조로 저장한 것이다. KB의 일종인 DBpedia[2]는 영문 Wikipedia[3]를 기반으로 만들어졌으며, 개체와 개체 사이의 속성을 정의한다[5]. 예를 들어 DBpedia 내에는 <신라, capital, 경주>라는 사실 관계(triple)가 존재한다.

센터 시험의 문제에는 개체와 개체와의 관계를 설명하는 보기가 많이 존재한다. 데이터를 분석한 결과, 틀린 보기들은 triple이 존재하지만 속성만 틀린 경우(예: <신라, destroy, 경주>)보다 아예 존재하지 않는 관계(예: <미국, capital, 경주>)를 나타내고 있는 경우가 많았다. 이는 보기에 등장하는 개체 사이의 관계의 유무만으로 보기가 사실이라는 단서를 평가할 수 있다는 것을 의미한다.

우리는 보기와 문맥 내의 개체 사이에 존재하는 관계의 개수를 단서로 보았다. DBpedia에는 Wikipedia의 페이지에 링크가 존재하나 그 속성은 명확히 정의되지 않은 wikiPageWikiLink라는 속성이 존재하는데, 이 속성은 그 관계의 신뢰도가 불분명하므로 절반의 점수를 주었다. 그 외의 속성의 의미적 차이는 구분하지 않았다.

5.2. IR evidence calculator

KB의 대표적인 단점은 데이터의 범위가 일반 텍스트에 비해 좁다는 것이다. 많은 정보들은 KB에 존재하지 않으며, 따라서 센터 시험 문제를 푸는 데 데이터의 범위가 넓은 일반 텍스트를 검색하는 IR 기술이 필요하다.

IR의 난점은 자연어의 특징인 다양한 표현을 검색하는 것이다. IR evidence calculator는 KB와 마찬가지로 개체와 개체 사이의 관계를 나타내는 단어보다는 개체명의 존재 여부를 주로 이용했으며, 검색 역시 개체명의 검색에 주목했다. 이를 위해 Wikipedia 내의 텍스트에 대해 주위 문맥에서 동일한 개체를 찾아주는 상호참조해결(coreference resolution)[5]과 개체의 다양한 표현을 DBpedia 내의 URI로 mapping해주는 NER[1]을 실행해 검색 재현률을 높이고 빠른 실행속도를 확보했다 상호참조해결으로는 Stanford CoreNLP toolkit[4][6]을, NER로는 DBpedia Spotlight[5][7]과 일부 개체에 대한 수동 태깅을 사용했다[8].

IR에 의해 검색된 문장의 수를 점수로 사용했다. KB triple과 달리 일반 텍스트에는 중복되는 내용과 비슷한 내용이 여러 번 등장하므로, 일부 개체명 사이에서는 너무 많은 문장이 검색된다. 우리는 검색 문장 수에 제한을 두고, 문장 수에 log를 씌워 점수로 활용했다.

5.3 Restriction evidence calculator

센터 시험 데이터의 보기 중에는 시대적 정보로 거짓임을 판별할 수 있는 보기들이 있다. 예를 들어, 데이터 내의 보기 'Qin Hui came into conflict with the party in favor of war, concerning the relationship with the Yuan' 은 오답이다. Qin Hui(진회)는 1155년에 죽었고, Yuan(원나라)는 1271년에 건립되어 두 개체 사이에는 역사적으로 의미 있는 관계가 성립될 수 없기 때문이다. 이런 정보는 지문 내에 시간 조건(예: '15세기에 건국된')이 있는 경우에도 유요하다. 보기의 국가가 935년에 멸망한 '신라'라면 지문의 '15세기에 건국된'이라는 표현 사이의 시대적 제약을 찾을 수 있다.

이러한 시간 정보 역시 DBpedia 내부의 정보를 활용했다. DBpedia의 개체에는 해당 개체에게 중요한 시간 정

2) http://wiki.DBpedia.org/
3) https://en.wikipedia.org/
4) http://nlp.stanford.edu/software/corenlp.shtml
5) https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

표 1 문제 수 및 점수

년도	전체		응답		정답		무작위 기대값		p
	문제 수	총점	문제 수	총점	문제 수	총점	문제 수	총점	
97	40	100	37	92	17	43	10	25	< 0.01
01	41	100	38	94	12	31	10.03	24.56	0.21
03	41	100	35	91	14	34	10.08	24.59	0.07
05	36	100	30	85	17	46	8.75	24.31	< 0.01
07	36	100	32	88	14	37	8.83	24.53	0.02
09	36	100	31	85	14	40	8.92	24.78	0.02

보 triple들이 존재한다. 그러나 개체의 유형마다 시간을 표현하는 속성의 종류가 다르며(예: 인물 - birthYear, 국가 - foundingYear), 같은 유형이라도 개체별로 서로 다른 속성이 혼재되어 있다(예: foundingYear, foundingDate, yearStart). 우리는 이러한 속성을 모두 태깅하는 대신, DBpedia 내에 존재하는 시간표현 중 가장 앞선 시간과 가장 뒤진 시간을 각각 개체의 시작과 끝으로 보았다. DBpedia 내에 시간 표현이 존재하지 않는 개체도 많으나, DBpedia 외의 시간 정보 데이터베이스는 활용하지 않았다.

보기, 지문, 문제 설명 내의 시간 표현 역시 하나의 개체로 보고 개체간의 제약을 검사했다. 시간 표현은 'century', '1050s', 'after-before', 'in the first half of' 등의 표현을 규칙으로 추출하여 활용했다.

이러한 제약, 시대적 모순은 틀린 보기의 신뢰도를 검증하는 데 매우 강력한 단서를 준다. 우리는 이러한 제약이 발견되었을 때 큰 음의 점수를 주어 해당 보기가 거짓으로 판별되도록 했다.

6. 실험 결과 및 분석

실험은 NII에서 번역, 분석한 일본 센터 시험 세계사B 과목 6개년도(97, 01, 03, 05, 07, 09년도)의 230개 소문제를 대상으로 했다. 각 문제의 배점은 센터 시험과 동일하게 했으며, 각 시험은 100점 만점이다.

표 1은 각 연도별 응답한 문제와 정답의 문제 수, 획득 점수를 나타낸 표이다. 각 시험에서 이 시스템은 12~17 문제를 맞혔고, 이는 응답 문제가 아닌 전체 문제에 대해서도 01년의 시험을 제외하면 통계적으로 의미있는 결과이다. 01년 시험은 지도를 활용한 문제가 많았으며, 복잡한 문장 이해와 역사적 지식이 있어야 하는 문제가 많아 다른 시험보다 점수가 낮다.

7. 결론

이 연구에서는 센터 시험과 같이 복잡한 언어 이해와 지식이 필요한 문제에 대해 의미적 분석 없이 문장 내의 개체명 인식과 이를 활용한 IR, KB 검색을 이용한 문제 해결 방법을 제안했다. 이는 역사에 관련된 개체명(인물, 국가, 사건 등)이 Wikipedia에 많이 존재하며, 관련된 문서와 KB가 모두 잘 구축되었으며, 세계사는 언어나 과학 등의 과목보다는 간단한 추론이 필요하기 때문으로 보인다.

또한 이 시스템은 영어를 기본으로 하지만, 활용한 리소스는 Wikipedia text, DBpedia, NER으로 최소한의 리소스만 사용했기 때문에 한국어에도 적용이 가능할 것으로 보인다. Wikipedia text와 NER은 한국어에도 존재하며, DBpedia는 영어를 기본으로 하나, 각 개체마다 한국어 명칭(예: Goguryeo - 고구려)이 병기되어 있다.

이번 연구에서는 의미적 분석을 활용하지 않았고, 개체명만을 검색에 사용하였기 때문에 정답에 한계가 드러났다. 예를 들어, '고구려는 6세기에 백제를 멸망시켰다' 라는 보기는 실제로는 거짓이나, '고구려' 와 '백제' 사이에 KB, IR상 관련이 많고, 연대가 서로 겹치며, '멸망시키다(overthrew)' 의 분석을 하지 않았기 때문에 이 시스템은 해당 보기를 참으로 분석한다. 후속 연구를 통해 이 두 개의 문제점을 개선해 더 높은 점수를 얻을 수 있을 것으로 기대한다.

* 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (R0101-15-0176, Symbolic Approach 기반 인간모사형 자가학습 지능 원천 기술 개발).

참고문헌

[1] David Nadeau, Satoshi Sekine. A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1): 3-26. 2007.

[2] Hideyuki Shibuki, Kotaro Sakamoto, Yoshionobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y. Itakura, Di Wang, Tatsunori Mori, Noriko Kando. Overview of the NTCIR-11 QA-Lab Task. Proceedings of the 11th NTCIR Conference. 518-529. 2014.

[3] Tsuyoshi Okita, Qun Liu. The Question Answering System of DCUMT in NTCIR-111 QA Lab. Proceedings of the 11th NTCIR Conference. 571-578. 2014.

[4] Di Wang, Leonid Boytsov, Jun Araki, Alkesh Patel. CMU Multiple-choice Question Answering System at NTCIR-11 QA-Lab. Proceedings of the 11th NTCIR Conference. 542-549. 2014.

[5] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. Web

Semantics:Science, Services and Agents on the World Wide Web. 7(3): 154-165. 2009.
[6] Wee Meng Soon, Hwee Tou Ng, Daniel Chung Yong Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics. 27(4): 521-544. 2001.
[7] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit.Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 55-60. 2014.
[8] Pablo N. Mendes, Max Jakob, Andres Garcia-Silva, Christian Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. Proceedings of the 7th International Conference on Semantic Systems. 1-8. 2011.
[9] Seonyeong Park, Soonchoul Kwon, Byungsoo Kim and Gary Geunbae Lee. ISOFT at QALD-5: hybrid question answering system over linked data and text data. Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum. 2015.