자연어처리와 정보검색을 이용한 질의응답 시스템

김민정[0] 신우석 김다영 김동건 박세영 윤희근
경북대학교 컴퓨터학부
min7859@gmail.com mell03@naver.com uoi504@naver.com gun528@naver.com seyoung@knu.ac.kr hkyoon@sejong.knu.ac.kr

Question and Answer System using Natural Language Processing and Information Retrieval

Minjeong Kim[0], Wooseok Shin, Dayoung Kim, Donggun Kim, Seyoung Park, Heegeun Yoon,
School of Computer Science and Engineering, Kyungpook National University

요 약

질의응답 시스템이란 사용자로부터 자연어 질문을 입력받아 해답을 제공해주는 시스템이다. 이러한 질의응답 시스템은 국가와 기업의 미래 경쟁력을 좌우할 것으로 보인다. 본 논문에서는 자연어로 주어진 질의어를 분석하고 정답후보들의 지식을 추출하여 각 후보들의 유사도를 계산하여 정답을 구하는 방법을 제안한다. 그리고 실험을 통하여 본 논문에서 제안하는 방법을 평가한다.

1. 서 론

정보검색 기능은 사용자들에게 많은 편의를 제공하였고, 이를 바탕으로 더욱 더 진보된 정보화 사회를 만들어 냈다. 한발 더 나아가 현대 정보화 사회는 소셜 미디어나 모바일 기기 등을 통해 수많은 데이터가 끊임없이 생성되고 있다. 따라서 이런 수많은 데이터에서 의미 있는 값을 분석하고 처리하는 인공지능 기술 개발 확보가 국가와 기업의 미래 경쟁력을 좌우한다.

본 논문에서 제안할 것은 이러한 기술 중 하나인 질의응답 시스템이다. 질의응답 시스템은 사용자로부터 질문을 입력받아 해답을 제공해주는 시스템이다. 본 논문에서 개발한 질의응답 시스템은 먼저 사용자로부터 자연어 객관식 질의를 입력받아 질의어를 분석한다. 그 다음 객관식 질의와 주어진 정답후보의 순위를 계산하여 높은 순위의 정답후보를 정답으로 결정한다.

그러나 질의응답 시스템을 개발하는 것에는 문제점이 있다. 컴퓨터는 지능과 지식을 가지고 있지 않기 때문에 자연어 문장이 가지는 의미를 정확히 파악하지 못한다는 점이다. 따라서 자연어 문장이 가지는 의미를 파악하기 위해서, 필요한 지식은 한국어 위키피디아 백과사전[1]을 사용하고 정답 결정 알고리즘을 구축한다.

본 논문에서는 2장에서 질의응답 시스템에 대해 설명하고, 3장에서는 실험을 통해 이 시스템의 성능을 분석하고, 4장에서 결론 및 향후 연구에 대해서 논의한다.

2. 질의응답 시스템

본 논문에서 제안할 것은 객관식 문제를 푸는 질의응답 시스템이다. 본 논문의 질의응답 시스템은 그림 1과 같이 구성된다. 우선 필요한 지식데이터를 구축한다. 그리고 질의어가 들어오면 해당 질의어를 분석하고 질의에 대한 정답 후보가 주어져 있을 때, 구축한 지식데이터를 이용하여 정답후보에 대한 지식을 추출하고 이를 통해 후보 순위를 결정하여 정답을 결정한다.

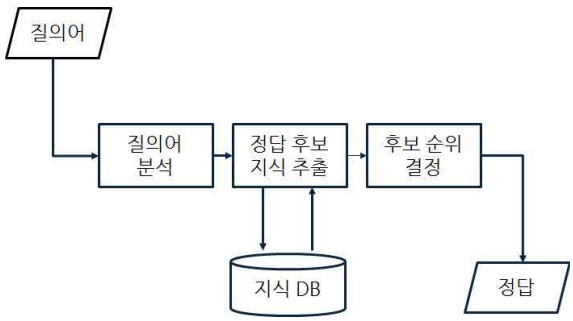

그림 1 <시스템 설계도>

2.1 지식 데이터 구축

컴퓨터는 정답을 판별할 지식을 가지고 있지 않기 때문에 지식데이터를 구축하여야 한다. 지식데이터 구축에는 위키미디어에서 제공하는 한국어 위키피디아 백과사전의 데이터베이스 파일[2]을 사용한다. 이 데이터베이스 파일은 문서의 제목과 그 문서의 쌍으로 구성되어있다. 마찬가지로 지식데

1) http://ko.wikipedia.org

2) http://dumps.wikimedia.org/backup-index.html

이터도 이와 같이 구성한다.

2.2 질의어 입력

질의응답 시스템은 사용자가 자연어로 질의한 문제에 대해 응답하는 시스템이다. 본 논문에서 개발하는 질의응답 시스템은 객관식 문제를 푸는 질의응답 시스템이기 때문에 입력은 자연어로 된 문제와 키워드 형식의 정답후보들로 주어진다.

2.3 질의어 분석

정답을 구하기 위해서는 질의어와 정답후보들의 지식을 비교해야한다. 질의어를 정답후보들의 지식과 비교하기 위해 형태소 단위로 자른다. 왜냐하면 문장 구조와 조사가 달라 음절 단위로 문자열을 비교하는 것은 의미가 없기 때문이다.

예를 들어, 질의어로는 “임진왜란 시기에 거북선을 만든 조선의 장군은?”가 주어지고, 지식으로는 “조선의 장군이며 임진왜란 시기에 거북선을 만들었다.”라고 주어졌을 경우 음절 단위로 비교를 하면 같은 의미라고 판별하지 못 한다.

형태소 단위로 자르기 위해 HMM을 사용하는 코모란(komoran) 형태소 분석기를 사용한다 [1].

2.4 정답 후보 지식 추출

객관식 질의로 문제와 정답후보들이 주어졌을 때, 이 정답후보들은 키워드의 형태로 나타난다.

정답을 구하기 위해 정답후보들의 지식과 질의어의 지식을 비교해야 한다. 그러나 현 상태에서 정답후보들의 지식은 그들의 키워드밖에 없으므로 질의어의 지식과 비교하기에 정보가 부족하다. 따라서 2.1에서 구축한 지식데이터에서 주어진 키워드로 정답후보에 대한 지식을 추출한다. 이 지식데이터는 위키피디아 문서의 제목과 위키피디아 문서의 쌍으로 이루어져있으므로 정답후보를 문서의 제목들과 비교하여 해당하는 문서를 가져온다.

예를 들어, “임진왜란 시기에 거북선을 만든 조선의 장군은?” 과 같은 질의어와 “이순신”, “권율”라고 키워드가 주어졌을 때, “이순신”은 제목이 “이순신”인 위키피디아 문서[3]로 사상(寫像)되고 “권율”은 제목이 “권율”인 위키피디아 문서[4]로 사상된다.

이 추출한 지식도 2.3장과 같은 이유로 형태소 분석을 한다.

2.5 정답후보들의 순위 결정

2.4장의 결과로 정답후보들의 지식은 문서이다. 정답후보들의 순위를 결정하기 위해 질의어와 정답후보들의 문서를 비교하여 각 유사도를 측정해야 한다. 유사도는 두 문서가 얼마나 비슷한지를 나타낸다. 질의어와 유사도가 가장 높은 문서를 정답으로 결정한다.

이 유사도를 계산하기 위해서 일반적으로 정보검색에서 널리 사용되는 벡터 모델을 사용한다. 벡터 모델은 문서를 단어 색인 등의 식별자로 구성된 벡터로 표현하여 각 단어에 대해 점수를 매겨 유사도를 측정할 수 있다 [2] [3]. 이 측정된 값으로 정답후보들의 순위를 결정한다.

아래의 식은 벡터 모델을 표현한 식이다.

$$d_j = (w_{1,j}, w_{2,j}, \cdots, w_{t,j})$$

d = document, w = word

그러나 단순히 벡터모델만 사용한다면 단어의 빈도수(term frequency)와 문서 빈도수(document frequency)를 고려하지 않아 모든 문서에 자주 등장하는 흔한 단어와 그 문서에만 나타나는 유일한 단어를 같은 가중치로 두는 문제점이 있다.

따라서 이 문제점을 해결하기 위해 TF-IDF를 사용한다. TF-IDF는 한 문서에는 자주 나오지만 그 외의 다른 문서에서는 거의 나오지 않는 단어에 높은 가중치를 둔다.

TF-IDF에서 TF(단어 빈도, term frequency)란, 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값이다. TF값이 높을수록 해당 단어의 가중치가 높다.

$$tf(t,d) = \frac{freq(t,d)}{\max(freq(t,d))}$$

t = term, d = document

그리고 DF(문서 빈도, document frequency)란, 특정 단어를 포함하는 문서의 수를 의미한다. DF값이 높을수록 다른 문서에도 해당 단어가 많이 나온다는 뜻으로 단어의 가중치가 낮다.

$$df(t,D) = \frac{freq(t,D)}{\max(freq(t,D))}$$

D = set of documents

TF와 DF를 이용하여 TF-IDF를 계산한다.

$$tf \cdot idf(t,D) = \frac{tf(t,d)}{df(t,D)}$$

TF-IDF값을 벡터모델로 이루어진 정답 후보들의 지식에 곱하여 출연하는 단어들의 가중치를 매긴다. 이를 이용하여 질의어와 정답후보들의 유사도를 계산한다. 유사도 계산은 정보검색에서 전통적으로 사용되는 Cosine Similarity를 사용한다 [4].

$$similarity(Q,D) = \frac{Q \cdot D}{|Q| \times |D|}$$

2.6 정답 결정

Cosine Similarity 값은, 0에서 1사이의 값이 나온다. 이 값

3) http://ko.wikipedia.org/wiki/이순신
4) http://ko.wikipedia.org/wiki/권율

이 1에 가까울수록 두 문서의 유사도가 높다 [5]. 따라서 질의어와 정답후보간의 Cosine Similarity값이 가장 높은 정답 후보를 정답으로 결정한다.

3. 실험

3.1 실험데이터

본 논문에서 사용된 입력데이터는 장학퀴즈에 출제된 객관식 인물 문제 111개이다. 지식데이터는 한국어 위키피디아 백과사전에서 수집하였다. 수집된 지식데이터는 총 66만개이며 지식데이터 사전의 단어 개수는 130만이었다.

3.2 실험 결과 및 분석

본 논문의 실험데이터에 대한 실험 결과는 표 1과 같이 69.3%의 우수한 성능을 보여주었다.

인물문제(개)	정답률(%)
111	69.3

표 1 <실험 결과>

이러한 성능은 1장에서 기술한 문제점을 해결한 결과이다. 컴퓨터는 지능과 지식을 가지고 있지 않은 문제점을 지식데이터를 구축하고, 단어의 가중치를 매겨 유사도를 측정함으로써 해결하였다.

반면, 오답률 30.7% 중에서 지식데이터에 정답후보에 대한 정보가 없거나 부족했을 경우가 21.8%이다. 정답후보에 대한 정보가 없는 경우는 위키피디아 상에서 문서 자체를 찾을 수 없는 경우에 발생한다. 그리고 정보가 부족한 문제점은 위키피디아 문서 안에 내용이 충분하지 않아서 발생한다. 이 문제점들은 지식데이터를 위키피디아뿐만 아니라 다른 데이터를 추가하여 보완하면 해결 될 것이다.

그리고 나머지 오답률 8.9%는 다른 문서의 유사도가 더 높았을 경우이다. 이 원인은 조사해보니 동의어의 문제점이라는 것을 알게 되었다. 예를 들어, 질의어로는 “조선의 4대 국왕이며 누구나 쉽게 배울 수 있는 효율적이고 과학적인 문자 체계인 훈민정음을 창제한 사람이다. 이 사람은 누구인가?”가 입력되고, 정답후보의 키워드로 “세종대왕”이 입력된다. 이 키워드들을 바탕으로 지식을 추출한 결과 “세종대왕”은 “조선의 4대 왕으로 수월하게 학습할 수 있는 능률적이고 체계적인 한글을 만들었다.”이다. 이들의 유사도를 계산할 경우 표 2의 동의어들이 같은 단어인지를 판별할 수 없기 때문에 질의어와 “세종대왕” 키워드의 유사도가 낮아진다. 따라서 동의어를 처리하는 과정을 적용한다면 더 높은 유사도를 보일 것이다.

동의어	
국왕	왕
쉽다	수월하다
배우다	학습하다
효율적	능률적
과학적	체계적
한글	훈민정음
만들다	창제하다

표 2 <동의어>

4. 결론 및 향후 연구

본 논문에서는 자연어 처리와 정보검색을 이용해 질의응답 시스템을 구축하는 방법에 대하여 제안하였다. 우선 질의어가 들어오면 해당 질의어를 형태소 분석을 하고 질의에 대한 정답 후보가 주어져 있을 때, 지식데이터를 이용하여 정답후보에 대한 지식을 추출한다. 이 추출한 지식에 벡터 모델을 적용하고 TF-IDF와 Cosine Similarity를 사용하여 후보 순위를 정해서 정답을 결정한다. 실험 결과 본 논문의 시스템은 객관식 인물 문제에 대해 69.3%의 정답률을 보여 우수한 성능을 나타냈다.

향후 연구에서는 단답형 질의에 대해서도 응답이 가능하도록 하는 방법을 연구할 것이다. 또한 위의 실험결과를 바탕으로 향후 연구에서는 동의어를 고려하여 정답률을 높일 것이다.

Acknowledgement

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 서울어코드활성화지원사업의 연구결과로 수행되었음.
(IITP-2015-H1807-14)

참 고 문 헌

[1] Sang-Zoo Lee, Jun-ichi Tsujii and Hae-Chang Rim, Lexicalized Hidden Markov Models for Part-of-Speech Tagging

[2] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, nr. 11, pages 613 - 620 , 1975

[3] Singhal, Amit, "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35 - 43 , 2001

[4] Rada Mihalcea and Courtney Corley, Corpus-based and Knowledge-based Measures of Text Semantic Similarity, 2006

[5] Salton G. and McGill, M. J., Introduction to modern information retrieval. McGraw-Hill, ISBN 0-07-054484-0, 1983