키워드 패턴을 이용한 질의유형 분류 시스템 구현

안혁주*, 김민경, 김학수
강원대학교 컴퓨터정보통신공학과
e-mail : zingiskan12@kangwon.ac.kr, kmink0817@kangwon.ac.kr,
nlpdrkim@kangwon.ac.kr

Implementation of a Question Type Classification System using Keyword Patterns

HyeokJu Ahn*, MinKyoung Kim, Harksoo Kim
Dept of Computer and Communication Engineering,
Kangwon National University

요 약

질의응답 시스템에서 정답선택의 정확률을 향상시키기 위해 본 논문은 패턴과 휴리스틱을 기반으로 하는 질의유형 추출 시스템을 구현하는 방법을 제안한다. 질의유형은 DBPedia에서 사용하는 클래스타입을 기반으로 추출되며 질의유형에 포함하는 키워드패턴들을 수집하여 키워드패턴 데이터를 생성한다. 그 후 한국어 질의에서 많이 발생하는 유형을 분석하여 휴리스틱을 이용해 사용자가 의도한 질의유형을 출력한다. 제안시스템은 기존 연구에 비해 구축과 수정이 쉽다는 장점이 있다.

1. 서론

IRQA(Information Retrieval Question and Anwering)는 주어진 질의에 대해 응답을 구하는 시스템으로 사용자가 질의를 입력하면 이를 바탕으로 정답을 출력한다. 현재 질의응답 시스템과 관련된 연구들은 AAAI[1]와 TREC[2]을 중심으로 수행되고 있다. 그 중 질의에 대해 정확한 정보를 단답식으로 제공하는 질의응답 서비스는 사용자의 검색 편의 향상을 위한 방법으로 네이버, 다음과 같은 대형 검색 포털사이트에서 두드러지게 사용되고 있다. 이러한 서비스를 구현하기 위해서는 먼저 사용자가 입력한 질의에 대한 질의유형을 정확히 파악해야 한다.

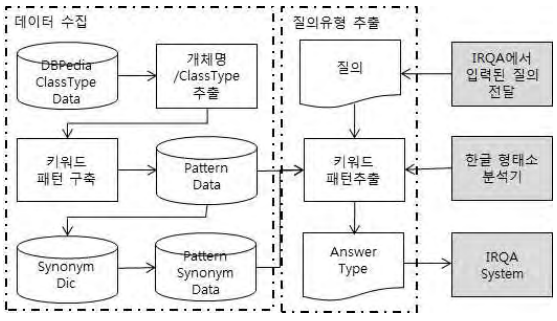
질의에서 질의유형을 추출하는 연구는 보통 패턴을 이용하거나 기계학습을 이용하여 진행되어 왔다. MURAX(1993)[3]는 대표적인 명사구 추출시스템으로 품사태거(POS tagger), 어휘구문패턴(Lexico-Syntactic Pattern) 매칭을 위한 유한 상태 인식기와 같은 비교적 저급의 언어지식을 이용하여 정답을 추천한다. SiteQ(2001)[4]는 어휘의미패턴(Lexico-Semantic Pattern)을 이용하여 질의유형을 출력하는 규칙문법을 구축하였다. Vijay(2005)[5]는 CRFs를 이용하여 질문에 대한 트리를 추출한 뒤 이를 SVM모델을 이용하여 질문에 대한 질의유형을 분류한다. Heo(2012)[6]는 sSVM과 어휘-구문 패턴을 기반으로 한 규칙모델을 이용하여 질의유형을 측정하는 방법을 제시하였다. 그러나 어휘의미패턴과 어휘구문 패턴에 관한 연구는 복잡한 규칙이 필요하고 수정이 어렵다는 단점이 있다. 기계학습을 이용한 연구의 경우에는 대용량의 학습데이터를 필요로 한다. 본 논문에서는 문장에 포함된 키워드패턴을 매칭하고 휴리스틱을 이용하여 질의유형을 추출하는 방법을 제안한다. 본 연구는 기존 연구에 비해 매우 단순하지만 질의유형 추출 부분에 있어서는 빠르고 쉽게 구현할 수 있으며, 문제점 발생 시 유연성이 돋보이는 장점이 있다.

본 논문의 구성은 다음과 같다. 2장에서는 시스템 구성 및 구현 과정에 대해 설명하고, 3장에서는 실험에서 사용한 데이터와 제안 연구방법의 성능을 측정, 마지막으로 4장에서 결론 및 향후연구를 제시한다.

2. 키워드 패턴 기반 질의유형 분류

2.1 시스템 구조도

(그림 1)은 본 논문에서 제안하는 시스템 구조도이다.



(그림 1) 시스템 구조도

제안 시스템은 데이터 수집단계와 질의유형 추출단계로 구성된다. 데이터 수집단계는 질의유형 추출에 필요한 데이터들을 생성하기 위한 단계이다. 먼저 DBPedia 클래스타입 데이터[7]에서 질의유형 목록을 추출한 뒤, Q&A 시스템에 전달한다. 그 후, 질의유형별로 키워드패턴을 구축하고 유의어사전을 이용하여 키워드패턴 데이터와 키워드패턴에 관한 유의어데이터를 생성한다. 질의유형 추출단계는 수집된 데이터를 바탕으로 질의를 받아 형태소분하여 패턴을 분석한 뒤 질의유형을 추출하여 다시 Q&A 시스템으로 전달한다.

2.2 키워드패턴 데이터 구축 및 유의어 데이터 생성

키워드패턴을 구축하기에 앞서 질의유형의 범위를 지정하는 것은 중요한 문제이다. IRQA시스템과 질의유형 추출 시스템이 결과로 출력할 수 있는 질의유형의 범위가 같아야 정확한 결과를 출력할 수 있기 때문이다. 질의유형은 DBPedia 클래스타입을 바탕으로 추출되며 (그림 2)와 같이 계단형으로 범위가 구성되어 있다.

(그림 2) DBPedia Activity 클래스타입의 하위 계층

본 논문에서는 <표 1>과 같이 각 클래스타입들의 깊이를 정의하였다.

<표 1> (그림 2) 클래스타입 깊이 정의

ClassType	Depth	ClassType	Depth
Activity	1	Athletics	3
Game	2	Boxing	3
BoardGame	3	BoxingCategory	x
CardGame	3	BoxingStyle	x
Sports	2	HorseRiding	3

[표 1]과 같이 많은 수의 클래스타입을 고려하지 않기 위해 깊이가 3을 초과하게 될 경우 패턴으로서 추출해야할 질의유형의 목록에서 제거하였다. 클래스타입의 깊이는 추출단계에서 하나의 질의에 여러 질의유형이 생성될 경우 선택지로서의 역할을 한다.

추출된 클래스타입은 질의유형으로서의 역할을 수행하기 때문에 각각에 대한 키워드패턴을 구축해야 한다. 키워드패턴은 각 질의유형을 결과로 출력할 수 있는 질의들을 분석하여 수작업으로 구축하였다. 패턴은 2~3개의 체언, 용언 그리고 부사로 이루어져 있다. <표 2>는 질의유형에 포함되는 키워드패턴들에 대한 예를 보여준다.

<표 2> 질의유형에 포함되는 키워드패턴 예

질의유형	키워드패턴
Politician	{출신,정치인}, {당선,정치인}, {정당,정치인}, {정책,정치인}…
Animal	{먹이,동물}, {수명,동물}, {서식,동물}, {무게,동물} {가장,동물}…
Monarch	{출생,왕}, {사망,왕}, {업적,왕} {즉위,왕} {재위,왕} {현재,왕}…

그러나 질의유형에 대한 키워드패턴을 수작업만으로 계속 진행하는 것은 시간이 오래 걸리고 효율적이지 못하다는 단점이 존재한다. 이를 극복하기 위해 본 논문에서는 유의어사전과 키워드패턴 데이터에서 사용한 체언, 용언, 부사를 이용하여 키워드패턴에 대한 유의어 데이터를 생성한다. 키워드패턴 유의어 데이터를 이용하면 많은 키워드패턴 데이터를 추출하지 않아도 되며, 질의유형 추출단계에서 보다 유연하게 키워드패턴매칭을 수행할 수 있다.

2.3 질의유형 추출

Q&A 시스템에서 받은 질의를 분석하여 질의유형을 추출하는 과정은 기본적으로 다음과 같은 단계를 거친다.

- 질의문장 형태소분석 및 체언, 용언, 부사 추출
- 패턴 유의어 사전을 이용한 단어 변경
- 휴리스틱을 이용한 질의유형 추출

기본적으로 패턴은 체언과 용언 그리고 부사의 구성으로 이루어져 있기 때문에 질의문장을 형태소분석한 뒤, 체언과 용언, 부사를 따로 추출한다. 그 후 각 단어별로 유

의어를 탐색하여 패턴에 매칭될 수 있도록 단어를 수정한다. 마지막으로 휴리스틱을 이용해 질의유형을 추출한다. 질의유형 추출과정에서 대표적으로 발생하는 휴리스틱은 다음과 같다.

- 질의유형이 맨 앞에 위치할 경우
- 후보 질의유형이 2개 이상 발생할 경우
- 후보 질의유형의 깊이가 모두 같을 경우

'~중에서', '순위'와 같은 표현은 앞의 문장중에서 질의유형이 될 가능성이 있는 단어가 발생할 수 있기 때문에 키워드 패턴매칭 될 수 있도록 문장을 바꿔준다. 후보 질의유형이 2개 이상 발생할 경우 우선 <표 1>과 같이 설정한 질의유형의 깊이를 판별하여 가장 깊은 질의유형을 정답으로 가진다. 마지막으로 후보 질의유형의 깊이가 모두 같을 경우, 한국어 질의는 대부분 문장에서 가장 뒤에 있을 단어가 질의유형일 가능성이 높다는 점을 이용하여 질의유형을 추출한다. 위와 같은 휴리스틱을 이용 또는 조합하여 원하는 질의유형을 추출한다.

3. 실험 및 분석

실험에 사용된 데이터는 랜덤으로 추출한 20개의 질의유형에서 4명의 대학생으로부터 각 질의유형 당 3개의 질의를 수집하였다. 결과적으로 총 240개의 질의 목록을 수집하여 실험데이터로 사용하였다. 성능은 수집한 실험데이터가 시스템을 통해 정확한 질의유형을 결과로 출력하는지를 기준으로 측정하였다. <표 3>의 실험 결과는 구현한 시스템의 성능과 기존 연구의[5,6] 평균 성능이다.

<표 3> 실험 결과

시스템	성능	질의유형 수
제안 시스템	91.67	20
기존 연구[5]	94.20	7
기존 연구[6]	83.88	명시되지않음

제안 시스템과 기존 연구의 시스템은 사용한 데이터와 질의유형 그리고 질의유형의 개수가 다르기 때문에 직접적인 비교는 불가능하지만 본 논문에서 제안하는 시스템이 쉽고 간편하게 구축할 수 있으며 다른 시스템과도 견줄만한 성능이 나옴을 확인할 수 있다.

4. 결론 및 향후연구

본 논문에서는 Q&A 시스템을 구축하는데 필요한 질의유형을 추출하는 시스템을 구축하였고, 이를 위해 키워드 패턴과 휴리스틱 기법을 사용하였다. 그 결과, 시스템을 통해 출력된 결과가 신뢰할만한 수준의 질의유형을 추출한다는 것을 확인할 수 있었다. 본 시스템은 질의유형을 추출하는 부분에 있어서 기존에 사용한 기계 학습 기법, 다른 패턴추출 기법보다 쉽게 구축 가능하며, 새로운 질의유형이 발생할 경우 빠르게 피드백할 수 있다는 장점이 있다. 그러나 패턴을 수동으로 수정한다는 점은 본 연구에 있어 확실한 약점이 된다는 것은 부정할 수 없는 사실이다. 따라서 추후에는 새로운 질의유형이 발생할 경우 자동으로 패턴을 생성하여 추가하는 방법을 모색할 예정이다.

감사의 글

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2013R1A1A4A 01005074). 또한 본 연구는 LG전자 산학연구용역 과제의 지원을 받아 수행되었음.

참고문헌

[1] AAAI Fall Symposium on Question Answering, *http://www.aaai.org/Press/Reports/Symposia/Fall/*

[2] TREC (Text REtrieval) Overview, *http://trec.nist.gov/overview.html/*

[3] Kupiec J., "Murax : A Robust Linguistic Approach for Question Answering Using an On-line Encyclopedia", In Proceedings of SIGIR'93, 1993.

[4] Gary Geunbae Lee, Jungyun Seo, Seungwoo Lee, Hanmin Jung, Bong-Hyun Cho, Changki Lee, ByungKwan Kwak, Jeongwon Cha, Dongseok Kim, JooHui An, Harksoo Kim, Kyungsun Kim, "SiteQ : Engineering High Performance QA system Using Lexico-Semantic Pattern Matching and Shallow NLP", In Proceedings of TREC, 2001.

[5] Vijay Krishnan, Sujatha Das, Soumen Chakrabarti, "Enhanced Answer Type Inference from Questions using Sequential Models", In Proceedings of HLT'05, pp.315-322, 2005.

[6] 허정, 류법모, 장명길, 김현기, "오픈 도메인 질의응답을 위한 검색문서 제약 및 정답유형 분류기술", 정보과학회논문지 : 소프트웨어 및 응용 제39권 제2호, pp.118-132, 2012.

[7] DBPedia Ontology Classes, *http://http://mappings.dbpedia.org/server/ontology/classes/*