정의형 질의응답 시스템을 위한 정답 패턴

Answer Pattern for Definitional Question-Answering System

신승은*, 서영훈**
충북대학교 전기전자컴퓨터공학부

Seung-Eun Shin(seshin@nlp.chungbuk.ac.kr)*, Young-Hoon Seo(yhseo@chungbuk.ac.kr)**

요약

본 논문에서는 정의형 질의응답 시스템을 위한 정답 패턴에 대하여 기술한다. 정의형 질의응답 시스템은 정의형 질의에 대한 정답으로 단답형 정답이 아닌 서술형 정답을 제공하기 때문에, 정답 추출 방법이 일반적인 단답형 정답 추출 방법과 다르다. 정의형 정답 패턴을 이용한 정의형 정답 추출은 의미 분석 없이 정확한 정의형 정답을 추출할 수 있다. 정의형 정답 패턴은 정확한 정답 추출을 위해 정답 패턴과 패턴별 제약 규칙, 우선순위로 구성된다. 정의형 정답 학습 코퍼스로부터 정답 패턴을 추출하고, 각각의 정답 패턴에 대한 F-measure에 따라 최적화하여 패턴별 제약 규칙을 구성한다. 마지막으로 정확률과 정답 패턴 구문 구조를 이용하여 우선순위를 결정한다. 제안한 정의형 정답 패턴을 이용한 정의형 정답 추출은 실험 코퍼스에 대해 정확률 0.8207, 재현율 0.9268, F-measure 0.8705를 보였다. 이것은 제안한 방법이 정의형 질의응답 시스템에 효율적으로 사용될 수 있음을 의미한다.

■ 중심어 : | 질의응답 시스템 | 정의형 정답 추출 | 정답 패턴 | 제약 규칙 |

Abstract

In this paper, we describe the answer pattern for definitional question-answering system. The answer extraction method of a definitional question-answering system is different from the general answer extraction method because it presents the descriptive answer for a definitional question. The definitional answer extraction using the definitional answer pattern can extract the definitional answer correctly without the semantic analysis. The definitional answer pattern is consist of answer pattern, conditional rule and priority to extract the correct definitional answer. We extract the answer pattern from the definitional training corpus and determine the optimum conditional rule using F-measure. Next, we determine the priority of answer patterns using precision and syntactic structure. Our experiments show that our approach results in the precision(0.8207), the recall(0.9268) and the F-measure(0.8705). It means that our approach can be used efficiently for a definitional question-answering system.

■ Keyword : | Question-Answering System | Definitional Answer Extraction | Answer Pattern | Conditional Rule |

* 이 논문은 2004년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음(This work was supported by Chungbuk National University Grant in 2004)

접수번호 : #050321-003
접수일자 : 2005년 03월 21일
심사완료일 : 2005년 04월 26일
교신저자 : 신승은, e-mail : seshin@nlp.chungbuk.ac.kr

I. 서 론

인터넷 기술의 발전과 함께 정보의 양이 급격히 증가함에 따라 사용자는 더욱 정확하고 빠르게, 그리고 원하는 정보를 더욱 편리하게 얻을 수 있는 새로운 정보검색 기술을 요구하고 있다. 최근 사용자의 질의 의도를 파악한 후, 검색 대상 문서로부터 답을 찾아 제공하는 질의응답 시스템에 대해 많은 관심이 집중되고 있다.

질의응답 시스템은 사용자의 자연어 질문과 검색 대상 문서의 의미를 파악하기 위한 고정밀 자연어 처리 기술과 대상 문서로부터 답을 추출하기 위한 정보추출 기술을 필요로 하며, 많은 후보 문서들로부터 답을 포함하는 문서를 걸러주는 역할을 위해 기존의 문서검색 기술도 활용한다[1].

질의응답 시스템과 정보검색 시스템의 큰 차이점 중 하나는 자연어를 입력하여 문서를 검색하는 것이 아니라 정답을 찾는 것에 있다. 이를 위해 질의의 처리과정에서 사용자가 원하는 정답이 무엇인지 질의의도를 파악할 수 있는 질의유형이나 키워드 등의 정보를 질의로부터 추출한다. 또한, 기존의 정보검색 방법에 의해 질의와 유사한 문서를 추출하고, 문서에서 다시 정답을 포함할 가능성이 있는 단락을 추출한 후, 단락에서 질의유형과 동일한 개체를 찾아내어 사용자에게 정답으로 제시한다.

질의응답 시스템을 평가하는 대표적인 것이 미국 NIST(National Institute of Standards and Technology)가 주관하는 질의응답 평가대회 TREC(Text REtrieval Conference)의 QA Track이다. 첫 대회인 TREC-8에서는 약 53만 문서의 코퍼스로부터 수작업으로 작성한 200질문에 대해서 50바이트 및 250바이트 크기의 정답을 포함한 단락을 제시하도록 하였다. TREC 2001의 경우, 약 25% 정도가 정의(definition)에 관련된 질문이었고, TREC 2002에서는 정답을 포함한 문자열을 제시하는 것이 아니라 실제 정답을 제시하도록 요구하였다. TREC QA 로드맵[2]에 따르면 정답이 단순히 문장의 일부분을 제시하는 것이 아니라 문장들 사이의 추론이 필요한 질문, 새로운 문장을 생성하여 이를 정답으로 제시하거나, 주어진 정답에 대한 배경 설명, 정답의 정당성 검증, 정답의 모호성 해결, 전문가 수준의 의견 제시가 필요한 질문, 이질적 정보의 통합을 통한 정답의 제시 등 점점 질의응답 시스템의 난이도를 높여 갈 계획이다.

일반적인 질의응답 시스템들은 단답형 정답만을 제공하기 때문에 정의형 정답을 요구하는 질의에 대한 정확한 정답을 제공하는데 어려움이 있다. 따라서 정확한 정답을 제공하기 위해 "디버그(Debug)란 무엇인가?" 혹은 "장보고는 누구인가?"와 같은 정의형 질의들은 최근에 많은 관심을 받고 있고, 정의형 질의응답에 대한 연구가 진행되고 있다[3].

정의형 질의응답에 대한 연구는 초기에는 단순하고 수동으로 구축된 정의 패턴들을 적절한 문장이나 구, 절로부터 추출하여 적용하였고, 최근 시스템들은 통계 정보나 WordNet[4]과 같은 언어 자원을 이용하는 좀 더 정교해진 기술들을 사용하고 있다[5, 6, 7]. 또한 사전이나 코퍼스로부터 미리 수동으로 구축된 패턴들을 이용하여 정의문을 추출하고, 이를 데이터베이스에 넣어 두었다가 정답을 추출하는 데 이용하는 시스템도 있다[8]. 일반적으로 정의형 정답을 추출할 때 패턴을 구축하여 사용한다. 그러나 패턴을 구축할 때에 많은 노력과 비용이 요구되며, 다양한 형태의 정답 패턴을 효율적으로 구축하는데 어려움이 있다.

본 논문에서는 정의형 질의에 대한 정확한 정답을 제공하는 정의형 질의응답 시스템을 위한 정답 패턴에 대하여 기술한다. 정의형 정답 패턴은 정확한 정답 추출을 위해 정답 패턴과 패턴별 제약 규칙, 우선순위로 구성된다. 정의형 정답 학습 코퍼스로부터 정답 패턴을 추출하고, 각각의 정답 패턴에 대한 F-measure에 따라 최적화하여 패턴별 제약 규칙을 구성된다. 마지막으로 정확률과 정답 패턴 구문 구조를 이용하여 우선순위를 결정한다. 정의형 정답 패턴을 이용하여 효율적으로 정의형 정답을 추출할 수 있다.

II. 정의형 정답 학습 코퍼스

1. 정의형 정답

정의형 질의에 대한 정답으로 단답 형식이 아닌 구나

하나의 문장으로 이루어진 경우를 정의형 정답이라 한다. 이러한 정의형 정답은 단답형으로 대답할 수 없는 형식이며, 일반적으로 패턴에 의해 추출할 수 있다. [표 1]은 정의형 질의와 정답을 포함하고 있는 문장, 정의형 정답을 나타낸다.

표 1. 정의형 질의와 정답

정의형 질의	디버그(Debug)란 무엇인가?
정답 문장	버그(bug)는 벌레를 뜻하며, 디버그(debug)는 원래 '해충을 잡다'라는 뜻이며, 프로그램의 오류를 벌레에 비유하여 오류를 찾아 수정하는 일이라는 의미로 쓰인다.
정의형 정답	원래 '해충을 잡다'라는 뜻이며, 프로그램의 오류를 벌레에 비유하여 오류를 찾아 수정하는 일

정의형 질의응답 시스템은 [표 1]에서의 정의형 질의에 대한 정답으로 정답 포함 문장에서 정의형 정답을 추출해야 한다. 정답 포함 문장에서 정의형 정답을 추출하기 위해 "X는 V이라는 의미로 쓰인다."라는 패턴이 필요하며, 이러한 패턴을 구축하기 위해 학습 코퍼스를 먼저 구축한다.

2. 학습 코퍼스

정의형 정답 패턴의 구축을 위해 정의형 정답 학습 코퍼스를 구축한다. 학습 코퍼스는 백과사전의 각 범주에서 균형적으로 선택한 500문서의 본문을 대상으로 정의형 정답을 수동으로 태깅하여 구축하였다. 백과사전은 각 항목(표제어)에 대한 일체의 지식을 잘 표현한 문서이기 때문에 다수의 정의형 문장과 패턴을 포함한다. 따라서 학습 코퍼스의 대상 문서를 백과사전으로 선택하였다. 정의형 정답 학습 코퍼스는 백과사전 표제어의 도메인에 관련 없이 패턴화 될 수 있는 정의형 정답에 태깅하였다. 전체 500문서에 대해 384개의 정의형 정답이 태깅되었고, 그 결과는 [표 1]과 같다.

표 2. 정의형 정답 학습 코퍼스 구축 결과

백과사전 문서 수	500
정의형 정답 수	405
정의형 정답 패턴 수	22

[표 2]에서 표제어 생략은 백과사전의 표제어에 대한 정의 문장에서 표제어가 주어인 경우 빈번하게 생략됨을 나타낸다. 백과사전과 같이 표제어가 있고, 해당 표제어에 대한 내용으로 이루어진 문서에서의 특징이므로 표제어가 생략된 패턴은 표제어가 생략되지 않은 패턴으로 통합한다.

III. 정의형 정답 패턴

정의형 정답 패턴은 정확한 정답 추출을 위해 정답 패턴과 패턴별 제약 규칙, 우선순위로 구성된다. 정의형 정답 학습 코퍼스로부터 정답 패턴을 추출하고, 각각의 정답 패턴에 대한 F-measure에 따라 최적화하여 패턴별 제약 규칙을 구성된다. 마지막으로 정확률과 정답 패턴 구문 구조를 이용하여 우선순위를 결정한다.

1. 정답 패턴

정의형 정답 학습 코퍼스에서의 각 패턴별 출현 수와 정확률을 이용하여 정답 패턴을 정의한다. [표 3]은 학습 코퍼스를 이용한 정답 패턴과 각 패턴별 출현 빈도수, 정확률을 나타낸다.

표 3. 정답 패턴 및 통계 정보(상위 10개 패턴)

순위	정답 패턴	출현 빈도수	정확률
1	[X]+l은|는|도|란|이란/jx [V]+l이/co	159	0.309
2	[X]+l은|는|란|이란/jx [V]+l로|으로|로서|으로서|으로써/jc+l,/s	54	0.740
3	[X]+l은|는|란|이란/jx [V]+l이라|이라고|이라고도/j_ l하|생각되/pv	50	0.595
4	[V]+l은|는|도|를|을/j_ [X]+l이라|이라고|이라고도/j_ l하/pv	42	0.429
5	[V]+l은|는|란|이란|개이/j_ [X]+l이/co	34	0.139
6	[X]+l은|는|란|이란/jx [V]+l을|를/jc 〈정의〉+l*/*	33	0.767
7	[V]+l에|에서/e_ [X]+l이라|이라고|이라고도/j_ l하/pv	10	0.556
8	[X]+l이라는|란|이란/j_ 〈용어〉+l은|는/jx [V]+l이/co+l*/*	5	0.833
9	[V]+l를|을/jc 〈정의〉+l에|에서/ec [X]+l이라|이라고|이라고도/j_ l하/pv	3	1.000
10	[X]+l의/jm 〈정의〉+l은|는/jx [V]+l*/*	3	1.000

각 패턴별 출현 빈도수는 학습 코퍼스에서 각 패턴이 출현한 문장 수를 의미하며, 정확률은 학습 코퍼스에 정답 패턴을 적용하였을 때의 정확률을 의미한다. [표 3]은 출현 빈도수 상위 10개의 패턴에 대한 정보를 나타낸다. [표 3]에서 출현 빈도수가 높은 패턴들이 낮은 정확률을 보인다. 이것은 정의를 표현하기 위해 자주 사용되는 패턴이 정의 문장이 아닌 일반 문장을 표현하는 경우도 많기 때문이다. 출현 빈도수가 높은 패턴이 일반적으로 정확률이 낮기 때문에, 정확률이 낮은 패턴에 대해 정확률을 높이는 것이 중요하다.

2. 패턴별 제약 규칙

정답 패턴을 이용하여 정확한 정의형 정답을 추출하기 위해 패턴별 제약 규칙이 필요하다. 패턴별 제약 규칙을 사용하여 정확률이 낮은 패턴에 대해 정확률을 높일 수 있다. 패턴별 제약 규칙은 NotXWord, 어절 수, Vend, !etm이 있다. 각각의 패턴에 제약 규칙을 다양하게 적용함으로써 최적의 패턴별 제약 규칙을 구성한다. 최적의 패턴별 제약 규칙을 결정하는 요소로써 검색 시스템의 성능을 평가하는 기준인 F-measure를 사용하였다[9]. F-measure는 정확률과 재현율을 이용하여 시스템의 성능을 하나의 척도로 평가하는 방법으로 식 (1)과 같다.

$$F = \frac{(r^2+1)PR}{r^2P+R} \qquad (1)$$

여기서 r 의 의미는 정확률(P)과 재현율(R)의 비중을 선택할 수 있게 하는 변수로써 $r > 1$이면 정확률의 비중을, $r < 1$이면 재현율의 비중을 높게 둔다는 의미이다. 즉, r=0.5일 때 0.5배의 가중치를 적용하여 정확률 값의 중요도를 높여서 계산한다. 실험에서는 패턴별 최적의 제약 규칙을 결정하기 위해 r의 값으로 1을 사용한다.

2.1 NotXWord

NotXWord는 전체 패턴에 대해 적용하는 제약 규칙으로 X 항목으로 올 수 없는 단어들이다. 패턴에서 X는 정의형 정답의 대상으로, "[X]는 [V]이다."라는 패턴에 의해 "폐곡면은 경계가 없는 연속곡면이다."라는 정의형 정답 문장이 추출된 경우 X는 "폐곡면"이며, 앞문장은 "폐곡면"에 대한 정의형 정답이 된다.

표 4. NotXWord 제약 조건 예

NotXWord	개념, 것, 곳, 그, 그곳, 그들, 대부분, 데, 둘째, 등, 때, 뜻, 말, 명칭, 수, 원인, 유래, 의미, 이, 이것, 이곳, 이들, 이름, 이유, 정의, 종류, 크기, 특색, 특성, 특징, 하나, 후자, 등	
적용 예	예문 1 (정의:O)	간정기사는 초서로 쓰여 있고, 당쟁사 연구의 좋은 자료이다.
	예문 2 (정의:X)	이것은 초서로 쓰여 있고, 당쟁사 연구의 좋은 자료이다.

[표 4]는 NotXWord의 예이다. [표 4]의 예문 1에서 해당 문장은 "간정기사"의 정의형 정답이 될 수 있으나, 예문 2에서의 문장은 "이것"의 정의형 정답이 될 수 없다. 이런한 NotXWord는 전체 패턴에 적용하는 제약 조건이다.

2.2 어절 수

정의형 정답 패턴에서 V의 어절 수를 제한함으로써 정확한 정의형 정답을 추출한다. [표 5]는 V 어절 수 제약 조건에 대한 예이다.

표 5. V 어절 수 제약 조건 예

	정의형 정답 패턴	정답 문장	V 어절 수	정의
1	X는 V이다.	코란은 아랍어로 '읽혀야 할 것'이라는 뜻이다.	5	O
		기업은 개인소유이다.	1	X
2	X는 V를 말한다.	가드레일은 자동차 도로의 양쪽에 설치한 방호책을 말한다.	5	O
		공군은 좁은 뜻으로 독립 공군을 말한다.	2	X
3	X는 V로,	국사는 왕사(王師) 위의 최고의 승직으로,	4	O
		갑각은 육각형으로,	1	X

각각의 정의형 정답 패턴에 V의 어절 수를 다양하게 적용함으로써 정확한 정의형 정답을 추출한다. 일반적으로 패턴이 정의 문장을 표현하는 어휘를 포함하면 적용되는 V 어절 수는 작아지며, 패턴이 정의 문장을 표현하는 어휘를 포함하지 않으면 V 어절 수는 커진다.

후자의 경우 V 안에서 정의를 표현하기 위한 어휘들이 필요하기 때문이다. 각각의 패턴들은 자신의 적용 V 어절 수를 포함한다. 각각의 패턴들은 가장 높은 F-measure의 값을 갖는 V의 어절 수를 선택한다.

2.3 Vend

각 패턴에서 V의 끝 어휘를 확인하는 제약 조건으로 X와 V의 뒷 단어가 일치하는 경우(Vend 1)와 V의 단서 어휘를 사용하는 경우(Vend 2)가 있다. [표 6]은 Vend에 대한 예이다.

각각의 패턴은 Vend의 적용 여부를 포함하며, 적용 여부는 가장 높은 F-measure의 값을 갖도록 설정한다.

표 6. Vend 제약 조건 예

단서어휘	개념, 것, 뜻, 말, 명칭, 암석, 약물, 악칭, 영화, 요소, 일기, 일부, 일종, 자료, 작품, 제도, 조직, 총칭, 프로그램, 하나, 학교, 학문, 행위, 현상, 회사, 등		
적용 예	예문 1 (정의:O)	폐곡면은 경계가 없는 연속곡면이다.	Vend 1 조건 만족
	예문 2 (정의:O)	고려장은 늙은 어머니를 깊은 산골에 버리는 몽건적인 비정의 시대상을 파헤친 작품이다.	Vend 2 조건 만족
	예문 3 (정의:X)	가우스의 곡률은 곡면의 등장변환(BW0)으로 불변이다.	Vend 1, 2 조건 불만족

Vend 제약 조건은 Vend 1과 Vend 2 중 하나의 조건만 만족하면 해당 패턴으로 추출된 정답을 정의형 정답으로 결정할 수 있다.

2.4 !etm

패턴이 적용된 문장이 뒤에 나오는 명사(구)를 수식하는 경우 정의형 정답으로 추출하지 않는 제약 조건이다. [표 7]은 !etm 제약 조건에 대한 예이다.

표 7. !etm 제약 조건 예

	정의형 정답 패턴	정답 문장	!etm	정의
1	X는 V이다.	코란은 아랍어로 '읽혀야 할 것'이라는 뜻이다.	O	O
		경전은 아랍어로 '읽혀야 할 것'이라는 뜻인 코란도 포함한다.	X	X
2	X는 V를 말한다.	가드레일은 자동차 도로의 양쪽에 설치한 방호책을 말한다.	O	O
		교통 안전 장치는 자동차 도로의 양쪽에 설치한 방호책을 말하는 가드레일도 있다.	X	X

!etm 제약 조건은 용언으로 끝나는 패턴에 적용되는 제약 조건이다. !etm 제약 조건을 적용하는 패턴에 의해 추출된 정답 문장은 !etm 제약 조건을 만족하는 경우에 정의형 정답으로 결정할 수 있다.

3. 패턴별 우선순위

패턴별 우선순위는 각 패턴에 대한 가중치를 의미한다. 하나의 문장에 여러 패턴들을 적용하여 여러 개의 정답 문장을 추출한 경우, 우선순위가 높은 패턴에 의해 추출한 정답 문장을 정의형 정답으로 결정한다. 정답 문장패턴별 우선순위는 정확률과 정답 패턴 구문 구조를 이용하여 결정한다. 다음의 2가지 패턴을 보자.

패턴 1) X는 V이다.

패턴 2) X는 V로,

예문 : 가스압입법은 천연가스의 저장법으로도 널리 이용되고 있는 채유법(採油法)으로, 노후(老朽)유전을 다시 소생시키기 위해 고안된 것이다.

패턴 2가 패턴 1보다 정확률이 높고, 패턴 2에 의해 추출한 "가스압입법은 천연가스의 저장법으로도 널리 이용되고 있는 채유법(採油法)으로,"가 가스압입법에 대한 정의형 정답이지만, 패턴 1에 의해 추출한 "가스압입법은 천연가스의 저장법으로도 널리 이용되고 있는 채유법(採油法)으로, 노후(老朽)유전을 다시 소생시키기 위해 고안된 것"이 더 정확한 정의형 정답이라 할 수 있다. 따라서 우선순위는 패턴별 정확률과 정답 패턴 구문 구조를 이용하여 결정한다. [표 8]은 각 패턴별 우선순위를 나타낸다.

표 8. 정의형 정답 패턴별 우선순위(상위 10개 패턴)

우선순위	정의형 정답 패턴
1	[X]+I이라는I란I이란/j_ 〈용어〉+I은I는/jx [V]+I이/co+I*/*
2	[V]+I를I을/jc (정의)+I에어서/ec [X]+I이라I이라고I이라고도/j_ I하/pv
3	[X]+I이라는I란I이란/j_ 〈용어〉+I은I는/jx [V]+I을I를/jc (정의)+I*/*
4	[X]+I의/jm 〈정의〉+I은I는/jx [V]+I*/*
5	[X]+I은I는/jx [V]+I*/* I의미/nc+I로/jc (사용)+I*/*
6	[X]+I이라는I란I이란/j_ 〈용어〉+I은I는/jx [V]+I의/jm 〈정의〉+I*/*
7	[X]+I은I는I란I이란/jx [V]+I을I를/jc (정의)+I*/*
8	[X]+I은I는I도I란I이란/jx [V]+I이/co
9	[X]+I은I는I란I이란/jx [V]+I로I으로I로서I으로서/jc 〈정의〉+I*/*
10	[X]+I은I는/jx 〈정의〉+I로는I으로는/jc [V]+I도/jx (포함)+I*/*

IV. 실험 및 평가

기본 정의형 정답 패턴에 패턴별 제약 규칙인 어절수, Vend, !etm을 다양하게 적용하여 최적의 패턴 규칙을 실험하였다. NotXWord는 전체 패턴에 적용하는 규칙이므로, 모든 패턴에 적용하였다. 백과사전 500개 문서를 대상으로 구축한 정의형 정답 학습 코퍼스에 대해 패턴별 제약 규칙 적용 실험을 하여 [표 9]와 같은 정확률과 재현율, F-measure를 보였으며, 제약 규칙을 적용하기 전과 비교하여 F-measure +0.2938의 성능 향상을 보였다. [표 10]은 실험 코퍼스에 대한 실험 결과를 나타낸다. 실험 코퍼스는 백과사전을 대상으로 전체 분야에서 균형적으로 추출한 500 문서를 사용하였다.

표 9. 제약 규칙 적용에 따른 F-measure (학습 코퍼스)

학습 코퍼스	정확률	재현율	F-measure (r=1)
제약 규칙 적용 전	0.4167	1.0000	0.5883
제약 규칙 적용 후	0.8359	0.9432	0.8863

표 10. 제약 규칙 적용에 따른 F-measure (실험 코퍼스)

실험 코퍼스	정확률	재현율	F-measure (r=1)
제약 규칙 적용 전	0.4093	0.9683	0.5754
제약 규칙 적용 후	0.8207	0.9268	0.8705

표 11. 제약 규칙을 적용한 정의형 정답 패턴 (상위 10개 패턴)

우선순위	정의형 정답 패턴	어절수	Vend	!etm
1	[X]+ㅣ이라는ㅣ란ㅣ이란/j_ 〈용어〉+ㅣ은는/jx [V]+ㅣ이/co+ㅣ*/*	2	X	X
2	[V]+ㅣ를ㅣ을/jc (정의)+ㅣ에어서/ec [X]+ㅣ이라ㅣ이라고ㅣ이라고도/j_ ㅣ하/pv	2	X	X
3	[X]+ㅣ이라는ㅣ란ㅣ이란/j_ 〈용어〉+ㅣ은는/jx [V]+ㅣ을ㅣ를/jc (정의)+ㅣ*/*	2	X	X
4	[X]+ㅣ의/jm 〈정의〉+ㅣ은는/jx [V]+ㅣ*/*	2	X	X
5	[X]+ㅣ은는/jx [V]+ㅣ*/* ㅣ의미/nc+ㅣ로/jc (사용)+ㅣ*/*	2	X	O
6	[X]+ㅣ이라는ㅣ란ㅣ이란/j_ 〈용어〉+ㅣ은는/jx [V]+ㅣ의/jm 〈정의〉+ㅣ*/*	2	X	O
7	[X]+ㅣ은는ㅣ란ㅣ이란/jx [V]+ㅣ을ㅣ를/jc (정의)+ㅣ*/*	2	X	O
8	[X]+ㅣ은는ㅣ도ㅣ란ㅣ이란/jx [V]+ㅣ이/co	3	O	O
9	[X]+ㅣ은는ㅣ란ㅣ이란/jx [V]+ㅣ로ㅣ으로ㅣ로서ㅣ으로서/jc (정의)+ㅣ*/*	2	X	O
10	[X]+ㅣ은는/jx 〈정의〉+ㅣ로는ㅣ으로는/jc [V]+ㅣ도/jx (포함)+ㅣ*/*	2	X	O

실험 코퍼스에 대한 실험에서도 제약 규칙을 적용했을 때, F-measure +0.2951의 성능 향상을 보였다. 또한 정의형 정답 패턴의 결과가 정확률 0.8207, 재현율 0.9268, F-measure 0.8705를 보였다. 이는 정의형 질의응답 시트템을 위한 정의형 패턴으로써 효율적으로 사용할 수 있음을 의미한다. [표 11]은 최적의 제약 규칙을 적용한 정의형 정답 패턴을 나타낸다.

V. 결론 및 향후 연구

본 논문에서는 정의형 질의에 대한 정확한 정답을 제공하는 정의형 질의응답 시스템을 위한 정의형 정답 패턴을 제안하였다. 정의형 정답 패턴은 백과사전 문서 집합을 이용하여 정의형 정답 학습 코퍼스를 구축하고, 이를 이용하여 정답 패턴을 구축한다. 정의형 정답 패턴은 정확한 정답 추출을 위해 정답 패턴과 패턴별 제약 규칙, 우선순위로 구성된다. 정의형 정답 학습 코퍼스로부터 정답 패턴을 추출하고, 각각의 정답 패턴에 대한 F-measure에 따라 최적화하여 패턴별 제약 규칙을 구성된다. 마지막으로 정확률과 정답 패턴 구문 구조를 이용하여 우선순위를 결정한다.

제안한 정의형 정답 패턴을 이용한 정의형 정답 추출은 실험을 통해 백과사전 실험 코퍼스에 대해 재현율 0.9268, 정확률 0.8207, F-measure 0.8705를 보였다. 또한 패턴별 제약 규칙과 우선순위 적용 전과 비교하여 정확률 +0.4114, F-measure +0.2951의 성능 향상을 보였다. 이것은 제안한 정의형 정답 패턴을 통해 효율적으로 정의형 정답을 추출할 수 있음을 의미한다.

향후 연구로는 백과사전과 같은 정형화된 문서 이외의 일반 문서에 대한 실험 및 패턴 구축 방법에 대한 연구와 학습을 통한 자동 패턴 구축 방법에 대한 연구가 필요하다. 또한 정의형 질의 이외의 다양한 형태의 서술형 질의에 대한 연구도 필요하다.

참 고 문 헌

[1] 황이규, 김현진, 장명길, "질의응답 기술 개발", 정보처리학회지, 제11권 제2호, pp.48-56, 2004.

[2] J. Burger, C. Cardie. et.al,"Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)," NIST DUC Vision and Roadmap Documents, http://www-nlpir.nist.gov/projects/duc/roadmapping.html, 2001.

[3] E.M. Voorhees, "Overview of the TREC 2001 question answering track," Proceedings of the Eleventh Text REtrieval Conference (TREC 2001), 2001.

[4] C.Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[5] S. Harabagiu, D. Moldovan, R. Mihalcea M. Pasca, R. Bunescu, M. Surdeanu, R. Girju, V. Rus, and P. Morarescu, "Falcon: Boosting knowledge for answer engines," Proc. Of Ninth Text REtrieval Conference (TREC 9), pp.479-488, 2000.

[6] J. Xu, A. Licuanan and R. Weischedel, "TREC 2003 QA at BBN : Answering Definitional Questions," The Twelfth Text REtrieval Conference(TREC 2003) Notebook, pp.28-35, 2003.

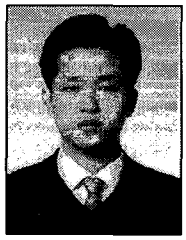
[7] A. Echihabi, U. Hermjakob, E. Hovy, D. March, E. Melz and D. Ravichandran, "Multiple-Engine Question Answering in TextMap," The Twelfth Text REtrieval Conference(TREC 2003) Notebook, pp.713-722, 2003.

[8] W.Hildebrandt, B.Katz and J.Lin, "Answering definition questions using multiple knowledge sources," Proceedings of HLT/NAACL 2004, Boston, MA, pp.49-56, 2004.

[9] D.D. Lewis, "Evaluating and optimizing autonomous text classification systems," Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.246-254, 1995.

저 자 소 개

신 승 은(Seung-Eun Shin) 정회원



- 1999년 : 충북대학교 컴퓨터공학과(공학사)
- 2001년 : 충북대학교 컴퓨터공학과(공학석사)
- 2001년 3월~현재 : 충북대학교 컴퓨터공학과 박사과정

<관심분야> : 정보검색, 자연언어처리

서 영 훈(Young-Hun Seo) 정회원

- 1983년 : 서울대학교 컴퓨터공학과(공학사)
- 1985년 : 서울대학교 컴퓨터공학과(공학석사)
- 1991년 : 서울대학교 컴퓨터공학과(공학박사)
- 1994년~1995년 : 미국 Carnegie-Mellon 대학 기계번역센터 객원교수
- 1988~현재 : 충북대학교 전기전자컴퓨터공학부 교수

<관심분야> : 정보검색, 자연언어처리, 기계번역