정의형 질의응답시스템을 위한 정의형 정답 문장 추출

A Extraction of Definitional Answer Sentence for a Definitional Question-Answering System

고병일, 강유환, 신승은, 서영훈
충북대학교 컴퓨터공학과

Byeong Il Ko, Yu Hwan Kang, Seung Eun Shin, Young Hoon S
Dept. of Computer Engineering, Chungbuk National University

요약

본 논문에서는 정의형 정답 문장을 요구하는 질의에 대하여 올바른 정답 문장을 추출하는 방법에 대해 기술한다. 말뭉치로부터 정의형 정답 문장 패턴을 정의하고, 패턴별 제약 규칙 및 패턴 순위화 같은 방법들을 이용하여 정확한 정의형 정답 문장이 추출되도록 하였다. 정답 패턴은 정의형 정답 문장의 구문 구조 및 각 패턴 또는 정답 패턴 별 실마리 어휘 등으로 구성된다. 현재 학습되지 않은 일반 문서에 대해 약 83%의 정의형 정답 문장 추출 정확도를 보이고 있다.

Abstract

In this paper, we propose a method to extract a definitional answer sentence for a Definitional Question-Answering System. definitional answer sentence patterns are manually constructed with restriction rules to patterns, and a ranking information of the pattern using its frequency from the corpus. answer sentence pattern consists of the syntactic structure of a definitional answer sentence, and clue words. this system show 83% accuracy for untrained corpus.

Ⅰ. 서 론

전통적인 정보 검색 시스템은 사용자의 질의에 대해 정답이 포함된 문서들을 유사도에 따라 사용자에게 제공한다. 이는 시스템이 제시하는 문서 내에서, 사용자 자신이 원하는 정보를 찾아야 하는 별도의 과정을 필요로 한다. 그러나 대다수의 사용자들은 다량의 문서보다는 구체적인 대답을 요구하는 경우가 많다[1]. 이러한 사용자들의 요구로 인해 질의응답이라는 개념이 등장하게 되었다.

질의응답 시스템이란 사용자가 제시하는 질의를 분석하여 거대한 문서 집합으로부터 제한된 길이의 정답을 추출해 내는 시스템이다[2]. 일반적인 질의응답 시스템들은 단답형 정답만을 제공하기 때문에 정의형 문장의 정답을 요구하는 질의에 대한 정확한 정답을 제공하는데 어려움이 발생하게 된다.

TREC(Text REtrieval Conference)[3]에서는 Definitional Question Answering System에 대한 연구가 최근 진행되기 시작하였으나 아직 국내에서는 정의형 정답 문장을 추출하는 질의응답시스템에 대한 연구는 아직 미미하다. 그러나 정의문에 대한 기존 연구로서 용어의 정의문을 이용하여 전문용어 집들을 구축하는 시소러스 관련 연구들은 비교적 다양하게 진행되고 있다[4][5].

정의형 질의응답시스템 연구들에 대해 살펴보면, 초기에 FALCON System[6] 시스템은 단순하고 수

동으로 구축된 정의 패턴들을 적절한 문장이나 구,절로 추출하여 실제 질의응답시스템에 적용하였다. 그리고 최근 TREC 12(2003)에서의 시스템들은 좀더 정교해진 기술들을 사용하고 있다. 이런 기술들은 수동으로 구축된 패턴들을 효과적으로 적용하는 방법들과 다양한 기반지식들을 이용하는 기술이다. [7]와 [8]에서는 centroid-based 통계적 순위화 정보를 이용하여 수동 구축된 정의 패턴들을 직접화하고 이를 정답 추출에 적용하고 있다. 이와 함께 다양한 기반 리소스로서 biography.com같은 정의관련 정보들이 기구축 된 웹사이트나 WordNet[9]같은 시소러스 정보를 효과적으로 이용하고 있다.

국내에서는 정의문 추출을 통한 용어의 전문용어 정의문 구축에 관한 연구가 있었다.

[4]은 훈련 코퍼스를 분석하여 만들어낸 정의문 패턴를 통해 정보과학 분야 뉴스 기사에서 정의문을 추출하였다. 이 연구는 정의문 패턴의 수가 4개라서 다양한 정의문들을 추출하는데 한계가 있다.

[5]은 전문 용어 사전을 구축하기 위해 의학분야 코퍼스를 이용한 연구이다. 여기서는 정의문 자동 추출을 위한 텍스트 코퍼스로부터 용어 정의문 관련 정보를 사전의 정의문을 통해 정의문의 패턴을 자동으로 추출하는 방법을 제시하였고, 단순 구문적 패턴 뿐만 아니라, 용어의 어휘 구성 패턴, 정의문의 의미적 패턴까지 고려한 정의문 추출을 하였다. 이러한 연구들은 그 패턴이 너무 일반적이고, 그 수가 작아서 패턴의 적용범위가 작은 단점을 지닌다.

외국에서는 정의형 질의응답시스템들에 대한 연구가 진행되고는 있지만, 국내에서는 정의형 정답 문장에 대한 연구는 미약한 것이 현실이다. 이에 대해 본 논문은 질의응답시스템을 위한 정의형 정답 문장 추출방법을 제시한다.

II. 정의형 정답 유형

1. 정의형 정답과 패턴

정의형 질의응답 시스템에서의 정의형 정답이란 질문에 대한 정답이 단답형이 아닌 서술형이면서 '용어'를 정의하는 정답을 정의형 정답이라 한다.

"동의보감의 저자는?"라는 질문은 "허준"이라는 단답형 정답을 요구하는 질문이고 "허준"은 그 질문에 대한 정답이다. 하지만 " '브라우저'란 무엇인가요?" 라는 질문에 대한 정의형 정답을 요구하는 질문에 대한 정답은 단답형 정답일 수 없다. 이 질문에 대한 정답은 다음과 같은 정의형 정답이 될 것이다.

질문	브라우저란 무엇인가요?
정답	'훑어보다'라는 의미를 가지며 단순히 문서의 내용만을 보여주는 것이 아니라, 하이퍼텍스트 문서를 검색하는 것을 도와주는 도구이다

질의응답시스템에서 요구되는 정의형 정답 문장을 추출하기 위한 패턴 구축을 위해 1000문서의 학습문서를 이용하였고, 수동으로 패턴구축을 하였다.

수동으로 패턴 구축 시, 정의형 정답을 요구하는 용어는 X라 하였고, 정의형 정답에 해당하는 문장에 대해서 Y라 하여 패턴으로 구축하였다.

이렇게 구축된 정의형 정답 문장 패턴의 수는 표 1과 같이 나타내며, 표 2는 태깅하여 얻어진 패턴 예를 나타 낸다.

[표 1] 정의형 정답 문장 패턴 개수

정의형 정답 문장 패턴 수	339 개

[표 2] 구축되어진 정답 문장 패턴 예

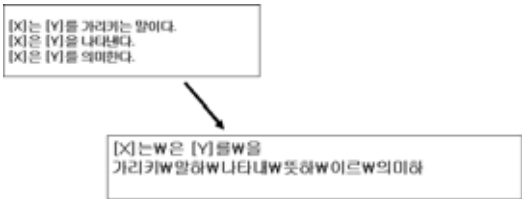
정의형 정답 문장 패턴 예
[X]는 [Y]라는 뜻을 가진다. [X]는 [Y]로서, [X]는 [Y]을 말하지만,

2. 정답 문장 패턴 정제화

정의된 정답 유형으로부터 구축된 패턴들에 대해 정답문장추출을 위해 시스템에 적용하기 전에 패턴

정제화 작업 및 순위화 작업을 실시한다. 패턴 정제화 작업은 구축된 초기 패턴들을 통합·분리하고, 형태소 태그 정보를 추가하는 것이다. 또한 각 패턴들에 대해 제약 규칙들을 정의하는 작업이고, 각 유형별 패턴들에 대해 순위화 작업을 실시하는 것이다.

패턴 정제화 작업에서 초기단계 작업으로, 패턴 통합·분리단계이다. 이것은 구축된 초기 패턴들로부터, 유사한 유형의 패턴들은 통합하고, 그러하지 않은 패턴들은 분리, 구분하는 과정이다.



▶▶ 그림 1. 패턴 통합 과정

그림 1에서 "X는/은 Y를/을"부분의 비슷한 유형에 동사부분이 다른 패턴들로서 같은 유형의 패턴으로 통합을 실시할 수가 있다.

통합된 패턴들에 대해서는 정확한 정답 문장 추출을 위해 형태소 태그 정보를 추가한다. 이것은 형태소 태그 정보를 이용하여 단순 패턴 매칭에서 발생할 수 있는 조사부분 관련 오분석을 줄일 수 있기 때문이다. 의미 태그 정보가 추가된 패턴의 예는 표 3이다.

[표 3] 정의형 정답 유형별 패턴

정답 유형	패 턴
정의	[X]jx [Y]obj_jc pv(가리키|말하|나타내|뜻하|이르|의미하) [X]jx [Y]obj_jc mag pv(가리키|말하|나타내|뜻하|이르)

패턴에 형태소 태그 정보를 추가한 다음 단계 과정은 패턴에 대한 세부 규칙들을 정의하는 것이다. 여기서 규칙이란 이미 구축된 패턴들로 태깅 된 문장에서 출현하는 문장들의 공통적인 특징들을 규칙이라 하고, 이 규칙들을 통합하여 패턴 별 규칙에 정의하는 것이다. 이런 규칙들은 세부적인 정보를 가지고 있는 패턴들에게서는 나타나지 않는다. 그러나 다양한 문장에 걸쳐 출현되는 패턴들에게서는 이런 규칙들이 출현하고, 이런 규칙들을 규칙화 함으로써 정답 문장 추출에 불필요한 결과들을 줄일 수 있게 된다. 예를 들어, 정의 유형에 '[X]는 [Y]로'라는 패턴은 보통 문장에도 많이 출현하는 유형의 패턴이다. 이런 패턴의 문장이 정의 유형의 정답문장을 포함할 때에는 패턴 뒤에 '용언'출현하지 않는 경우에 정의형 문장에 해당 하였다. 이와 같은 정보들을 수집하여 규칙을 정하고 이를 패턴정보에 적용 하였다. 표4는 패턴의 세부 규칙의 예이다.

[표 4] 패턴 별 세부 규칙

정의	[X]jx [Y]adv_jc
Rule	패턴 뒤에 '용언' 출현하는 경우 정의 아님
정의	[X]jx [Y]co
Rule	Y의 끝에 오는 단어 : 것, 개념, 뜻, 등

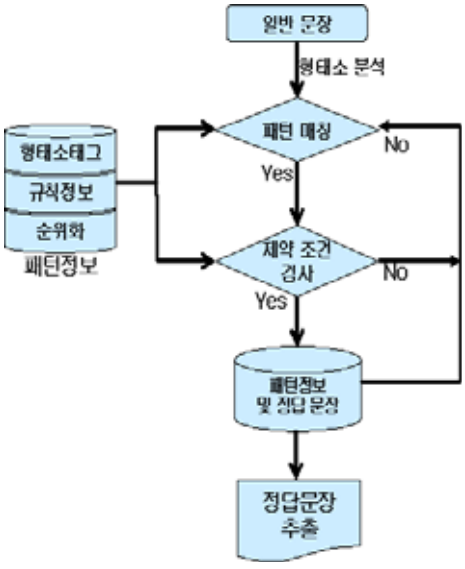
패턴별 각 세부 규칙까지 정의한 후에는 패턴들에 대해 순위화를 실시한다. 순위화는 패턴들 사이에 순위를 정하여 정답문장 적용의 패턴 순서를 정하는 것이다. 패턴 순서화에 따라 매칭 된 문장에 대해서는 다른 패턴을 적용하는 것을 방지하고, 이를 통해 시스템의 패턴 적용 시간을 줄이는 효과를 얻을 수 있다. 따라서 패턴 순위화는 패턴의 출현 빈도를 이용하여 고빈도의 패턴들은 상위 순위에, 저빈도의 패턴들은 하위 순위로 하여 순위화를 하였다.

이와같은 패턴 구축, 정제화, 의미태그 정보 추가, 패턴 별 세부 규칙, 패턴 순위화의 일련의 과정은 그림 2과 같다.

▶▶ 그림 2. 패턴 정제화 과정

III. 정답 문장 추출 시스템

구축된 패턴과 규칙정보, 순위화 정보를 이용하여 정의형 정답을 문서로부터 추출하게 된다. 그림 3는 이런 과정을 보인다. 그림 3에서, 실험 문서의 각 문장에 대해 형태소 분석과 태깅작업을 실시하고, 이 문장에 대해 패턴정보를 적용하기 위해 정답패턴과 실험 문장간 패턴 매칭을 실시한다.



▶▶ 그림 3. 정의형 정답 문장 추출 과정

정답 패턴과의 패턴 매칭을 통해 실험문장에 적용하다 보면, 부사 같은 패턴 매칭에 필요 없는 것들이 출현한다. 이와 같은 경우, 패턴 적용 할 때에는 부사를 무시하고, 정답 문장 추출부분에서는 부사부분을 유지하게 한다. 이와 같은 부사처리를 통해 정답 문장 패턴 적용에 융통성이 생기게 된다.

정답 패턴과 일치하는 문장에 대해서 용어부분에 해당하는 X와 그 X를 설명하는 정답 문장 Y부분을 검사한다. 이 검사를 통해 X는 용어에 만족하는 지를 검사하고, Y는 정답 문장에 만족하는 지를 검사한다.

용어 X의 경우는 연속된 명사의 나열이거나 단일 명사를 용어 X로 정의한다. X에 그 외의 것들이 등장한다면 X에서 제외를 한다. 실제 예로, X에 "이것은 ~", "이 모델은 ~", "이 별당 건축은 ~"과 같이 관형사"이/mm"나 대명사 "이것/np"같은 것들이 나오면 X에 만족되는 조건이 아니므로 정답 문장에서 제외한다. 또한 연속한 명사나, 명사가 출현한 후 X를 수식거나 한정하는 부분이 나오게 되면 이부분만을 X에서 제외를 하여 X를 용어의 조건에 충족시키게 한다.

정답 문장 Y는 보통 한 두 어절로 이루어진 것은 제외 한다. 이것은 정의형 정답 문장을 추출하기 위해서이다. X와 Y가 조건에 만족하면, 패턴정보와 정답 문장 정보를 저장하게 된다.

모든 패턴 적용 작업이 완료된 후에 저장된 패턴 정보와 정답 문장들을 정답 문장 후처리를 통하여 정의형 정답 문장을 추출한다.

IV. 실험 및 분석

본 시스템에 대한 평가는 실험 문서 100개에 대하여 자동 정답 문장 추출을 실시하였다. 수동 태깅 결과와 자동 태깅을 실시하여 얻어진 자동 태깅 결과를 통하여 비교 및 평가를 실시하였고, 그 결과는 다음 표 5와 같다.

[표 5] 실험결과

수동태깅문서에서의 정답 태깅 문장	78 문장
자동태깅 결과문서에서의 정답 태깅 문장	86 문장

실험을 실시한 결과, 86개의 정답문장들이 자동으로 추출되었다. 이 문장들 중 수동태깅된 문장들과 비교해 본 결과, 86개의 자동 태깅 문장들 중 정의형 정답 문장 태깅이 올바르게 된 문장은 72개의 문장으로 약 83.7%의 정확도를 나타내었다.

정확도는 높은 성능을 보였지만, 아직도 오분석된 문장이 나오는 부분과, 문장에 알맞은 패턴이 적용되

지 못하는 경우가 나타났다. 이러한 결과의 이유를 살펴 보면, 패턴들에 대한 정제화 작업이 조금 더 자세하고 정밀하게 이루어 져야 할것이라 본다. 정제화 작업을 통해 각 패턴의 특성을 파악하고, 그 패턴들에 대한 세부 규칙들을 정하는 작업들과 함께, 패턴의 순위화작업도 현재의 빈수도의 따른 패턴 순위화가 아닌 문장 형태의 따른 패턴 순위화를 통하여 순휘화를 재조정을 해야 한다. 또한, 패턴 정제화작업과 함께 다양한 문장 길이에 따른 패턴 적용이 같이 이루어 져야 한다. 현 시스템에서는 한 문장 단위의 자동 문장 태깅 만이 이루어지고 있지만, 패턴에서는 두문장, 세문장, 더 나가서는 문장 전체적으로 적용을 해줘야 하는 패턴들도 출현하고 있다. 이런 패턴의 형태에 대한 시스템의 튜닝도 필요하다.

마지막으로 시스템에서 얻어진 실험결과 얻어진 자동 태깅 문장들을 살펴보자. 다음 표6은 올바른 정의형 정답 문장들을 추출한 경우이다.

[표 6] 추출된 정답 문장

정 답 문 장
{X:약건피}는 {Y:동남아시아·아프리카 등의 열대지방의 원피(原皮)에 많고, 한번 가볍게 건조시킨 것을 5 % 정도의 아비산(亞砒酸)나트륨에 담갔다가 다시 건조·방부(防腐) 처리를 한 것}이다.
{Y:규장각도서.국왕이나 왕세자가 결혼할 때는 임시로 설치한 가례도감에서 의식 전반을 관장하고, 그 절차를 일일이 기록}하여 {X:《가례의궤(嘉禮儀軌)》}라 하였다
{Y:본래 물질의 3태(態) 중 하나인 기체를 지칭하나, 일반적으로는 화산이나 온천에서 분출하는 가스, 산이나 해상에서 발생하는 연무(煙霧), 신체의 소화기 내에서 발효하는 가스} 등도 {X:가스}라고 말한다.

오분석된 문장들은 표 7을 통해 살펴보면. 첫번째는 'Y이 X이다.'패턴 보다는 'X는 Y이다.'라는 문장으로 추출이 되어야 한다. 즉 '가산관료제'라는 용어의 정의형 문장으로 추출이 되어야 하는데, 그렇지 못한 경우 이다. 두번째 문장은 태깅이 안된 경우이다. 이 문장은 'Y이 X이다.'라는 유형의 규칙이지만 용어인 '개도'에 대한 처리 중 이에 대한 형태소 분석 오류로 일반 명사로 인식 실패하여 패턴 추출에 실패하였다.

[표 7] 정의형 정답 문장의 오류

태깅이 잘못된 경우
{Y:가산관료제는 근대 관료제와는 화폐급의 뒷받침이 있는 본직(本職)으로서의 직무행위가 없는 것:}이 {X:특색}이다
태깅이 안된 경우
{Y:이 개각도가 전주(全周) 360°의 몇 분의 1인가를 분수값으로 표시한 것}이 {X:개도}이다.

V. 결론 및 향후 연구

본 연구에서는 정의형 정답 유형들에 대하여 정의하고, 정답 문장들의 패턴들을 수집하였고, 그 수집된 패턴들에 대해 정제화 작업으로 각 패턴 별 제약 규칙들을 추가 및 형태소 태그 부착, 패턴의 순위화를 통해 수동으로 구축된 패턴들을 보완하였다. 이 패턴들을 이용하여 실제 실험문서에서 정의형 질의 응답 시스템을 위한 정의형 정답 문장들을 자동 추출하였다.

현재 높은 성능 높이기 위해 연구가 계속 진행 중이며, 이를 위해 다양한 정답 패턴들을 구축 및 이 패턴들에 대한 정제화 작업들을 통하여 패턴 정보를 정교화 하는 작업들이 요구 된다. 또한 다양한 문장 형태에 적용 가능한 시스템의 구성도 필요하다. 추후 더 나아가서는 현제 제한된 영역의 문서와 함께 일반 웹 문서에 적용하는 연구가 진행될 것이다.

▌참고문헌▐

[1] Ellen M. Voorhees, Dawn M. Tice, "Building a Question Answering Test Collection", In Proceeding of SIGIR 2000, pp.200-207, 2000

[2] Daisuke Kawahara, Nobuhiro Kaji, Sadao Kurohashi, "Question and Answering System based on Predicate-Argument Matching", In Proceedings of the Third NTCIR Workshop, 2002.

[3] TREC (Text Retrieval Conference) Overview, http://trec.nist.gov/overview.html

[4] 신효식, 김재호, 이해윤, 최기선 "텍스트로부터 정의문의 자동추출", 제 14회 한글 및 한국어 정보처리 학술대회, pp.292-299. 2002.
[5] 김재호, 배선미, 신효식, 최기선 "의학 전문용어의 정의문 자동추출", 한국정보과학회 2004 봄 학술발표논문집(B), pp.922-924. 2004.
[6] S. Harabagiu, D. Moldovan, R. Mihalcea M. Pasca, R.Bunescu, M. Surdeanu, R. G irju, V. Rus, and P. Morarescu,"Falcon: Boosting knowledge for answer engines", Proc. Of Ninth Text Retrieval Conference (TREC 9), pp. 479-488, 2000.
[7] 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.
[8] A. Echihabi, U. Hermjakob, E. Hovy, D. Marcu, E. Melz and D. Ravichandran, "Multiple-Engine Question Answering in TextMap", The Twelfth Text REtrieval Conference (TREC 2003) Notebook, pp. 713-722, 2003.
[9] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.