질의응답시스템을 위한 서술형 정답 추출

고병일 강유환 신승은 서영훈
충북대학교 컴퓨터공학과
{kobi, eric, seshin}@nlp.chungbuk.ac.kr yhseo@cbucc.chungbuk.ac.kr

A Extraction of Descriptive Answer for a Question-Answering System

Byeongil Ko, Yuhwan Kang, Seungeun Shin, Younghoon Seo
Dept. of Computer Engineering, Chungbuk National University.

요 약

본 논문에서는 서술형 정답을 요구하는 질의에 대해 올바른 서술형 정답을 추출하는 서술형질의응답시스템에 대해 기술한다. 질의응답시스템에서 요구되는 다양한 서술형 정답을 추출하기 위해 정답 유형을 10가지로 정의하였다. 말뭉치로부터 각 서술형 정답 유형에 대한 정답 패턴을 정의하고, 패턴별 제약 규칙 및 각 유형별 패턴 적용 순위화 등을 사용하여 정확한 서술형 정답이 추출되도록 하였다. 정답 패턴은 서술형 정답의 구문 구조 및 각 패턴 또는 정답 유형별 실마리 어휘 등으로 구성된다. 현재 학습되지 않은 일반 문서에 대해 59.2%의 서술형 정답 추출 정확도를 보이며, 시스템 성능 향상을 위해 연구가 진행중이다.

1. 서 론

전통적인 정보 검색 시스템은 사용자의 질의에 대해 정답이 포함된 문서들을 순위화하여 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 질의 응답이라는 개념이 등장하게 되었다.

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

TREC(Text REtrieval Conference)[3]에서는 Definitional Question Answering에 대한 연구가 진행되고 있으나, 다양한 서술형 정답을 추출하는 질의응답시스템에 대한 연구는 아직 초기 단계이다. 그러나 용어의 정의문을 이용하여 전문용어집을 구축하는 시소러스관련 연구들은 비교적 다양하게 진행되고 있다[4][5].

Definitional QA 연구들에 대해 살펴 보면, 초기에 FALCON System[6]같이 시스템들은 단순하고 수동으로 구축된 정의 패턴들을 적절한 문장이나 구,절로부터 추출하여 적용하는 시스템이었다.

최근 TREC-12의 시스템들은 좀더 정교해진 기술들을 사용하고 있다. [7]와 [8]들은 centroid-based 통계적 순위화 정보를 이용하여 수동 구축된 정의 패턴들을 직접화하고 이를 적용하고 있다. 이런 일련과정들은 다양한 리소스인 biography.com같이 정의관련 정보들을 갖고 있는 웹사이트나 WordNet[9]같은 시소러스 정보를 효과적으로 이용한다. [10]는 교사 학습(supervised learning) 방법을 문장들을 구별하는데 사용하였다. 최근에는 [11]과 같이 사전들이나 정의문을 지니는 코퍼스로부터 미리 수동으로 구축된 패턴들을 이용하여 정의문들을 추출하고, 이를 데이터베이스에 넣어 두었다가 정답을 추출하는데 이용하는 시스템도 있다..

대부분의 Definitional QA는 시스템들은 자동으로 패턴들을 구축해주는 다양한 기계학습 기술들을 무시하고 수동으로 구축된 정의 패턴들을 정의하는데, [12]은 표면(surface) 패턴을 학습하는 비교사 학습(unsupervised learing) 방법을 제안하였다. 그러나 이 시스템은 명확하고 정확한 질문에 의해서만 정답을 제시하는 단점이 있다.

[13]은 Instance-Based 학습을 통한 소프트 패턴 매칭 방법을 이용하여 패턴을 구축한다. 이 매칭 방법은

정규표현식을 이용하는 것이 아니라 probabilistic framework를 이용하여 테스트 문장에 대해 매칭을 실시한다.

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 4개 뿐이어서 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.

이와 같이 정의형 QA시스템들에 대한 연구가 진행되고 있지만, 아직 국내에서는 미약한 연구 분야이다. 이에 대해 본 논문은 서술형 질의응답시스템을 위한 서술형 정답 추출방법을 제시한다.

2. 서술형 정답 유형

2.1 서술형 정답 유형 및 패턴

질의응답시스템에서 요구되는 다양한 서술형 정답을 추출하기 위해 말뭉치로부터 서술형 정답 유형을 10가지로 정의하였다. 정의한 서술형 정답 유형들은 정의, 기능, 종류, 방법, 특징, 목적, 이유, 구성성분, 원리, 유래이다.

서술형 정답 패턴은 "두산세계대백과사전 엔싸이버"[14]에서 1000개의 문서를 대상으로 10개의 서술형 정답 유형에 대해 수동으로 정답 태깅을 함으로써 구축하였다. 태깅에서 X는 표제어, Y는 X를 설명하는 정답 문장으로 하여 태깅을 하였다.

이렇게 구축된 서술형 정답 패턴의 통계 정보는 표1과 같이 나타나며, 표2는 태깅하여 얻어진 서술형 정답 유형별 패턴의 예이다.

[표 1] 각 유형별 서술형정답문장 패턴 개수(개)

정의	기능	종류	방법	특성
28	64	77	14	21
목적	원인	구성성분	원리	유래
21	16	17	3	31

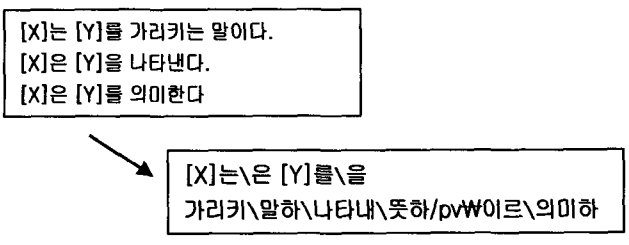
[표 2] 서술형 정답 유형별 패턴

정답유형	패 턴
정의	[Y]은 [X]이라고 한다. [Y]를 총칭하여 [X]라고 한다.
기능	[X]의 효과는 [Y]이다 [X]은 [Y]로 사용된다
종류	[X]는 [Y]등이 있다. [X]는 [Y]로 크게 구분된다.
방법	[X]의 방법으로 [Y]가 있다. [X]의 일반적인 방법은 [Y]이다.
특성	[X]의 장점은 [Y]이다. [X]의 결점으로 [Y]를 들 수 있다.
목적	[X]는 [Y]하기 위하여 [Y]가 [X]의 목적이다.
원인	[X]로 인해 [X]을 일 이들 [X]의 원인은[Y]
구성성분	[Y]가 [X]를 구성한다. [X]에는 주로 [Y]가 있다.
원리	원리는 [Y] [X]의 원리는 [Y]와 같다.
유래	[X]는 [Y]에서 비롯되었다. [X]는 [Y]에서 유래한다.

2.2 서술형 정답 문장 패턴 정규화

정의된 정답 유형으로부터 구축된 패턴들을 정답문장 추출을 위해 적용하기 전, 패턴 정제화 작업 및 순위화 작업을 실시한다. 패턴 정제화 작업은 구축된 초기 패턴들을 통합·분리하고, 의미 태깅 정보를 추가하는 것이다. 또한 각 패턴들에 대해 제약 규칙들을 정의하는 작업이고, 각 유혈별 패턴들을 순위화하는 작업이다.

패턴 정제화 작업에서 초기단계 작업으로, 패턴 통합·분리단계이다. 이것는 구축된 초기 패턴들로부터, 유사한 유형의 패턴들은 통합하고, 그러하지않은 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 의미 태그정보를 추가한다. 이것은 의미 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] 서술형 정답 유형별 패턴

정답유형	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)
기능	[X]jm nc(효과)jx [Y]co ([표제어]의) [X]jx [Y]obj_jc nc(담당)xsv(하) ([표제어]는) [Y]adv_jc nc(사용)xsv(되)|pa(쓰)
종류	[X]jx [Y]adv_jc pa+ef(게)/mag nc(구분)+xsv|pv(나뉘) [X]jx [Y1]jj [Y2]obj_jc nc(대표|일반)xsn(적)+co

패턴에 의미 태그 정보를 추가한 다음 단계 과정은 패턴에 대한 세부 규칙들을 정의하는 것이다. 여기서 규칙이란 이미 구축된 패턴들로 태깅된 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.

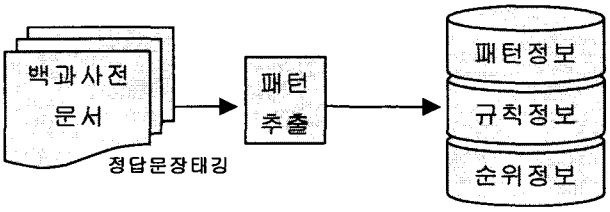
예를 들어, 정의 유형에 '[X]는 [Y]로'라는 패턴은 보통 문장에도 많이 출현하는 유형의 패턴이다. 이런 패턴의 문장이 정의 유형의 정답문장을 포함할 때에는 패턴뒤에 '용언'출현하지 않는 경우였다. 이와 같은 정보들을 모아서 규칙을 정하고 적용 하였다. 표4는 패턴의 세부 규칙의 예이다.

[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, 말, 법률, 일, 노래, 총칭, 하나

패턴별 각 세부 규칙까지 정의한 후에는 각 유형별 패턴들에 대해 순위화를 실시한다. 순위화는 패턴들 사이에 순위를 정하여 정답문장 적용의 패턴 순서를 정하는 것이다. 패턴 순서를 통하여 매칭된 문장에 대해서는 같은 유형의 다른 패턴을 적용하는 것을 방지하고, 이를 통해 시스템의 패턴 적용 시간을 줄이는 효과를 얻을 수 있다. 따라서 패턴 순위화는 다양한 문장들에 적용 되는 패턴들은 낮은 순위를 매기고, 좁은 범위의 문장들에 적용 되는 패턴들은 순위를 높게 매긴다.

이런 패턴 구축, 정제화, 의미태그 정보 추가, 패턴별 세부 규칙, 패턴 순위화의 일련의 과정은 그림 2과 같이 볼수 있다.



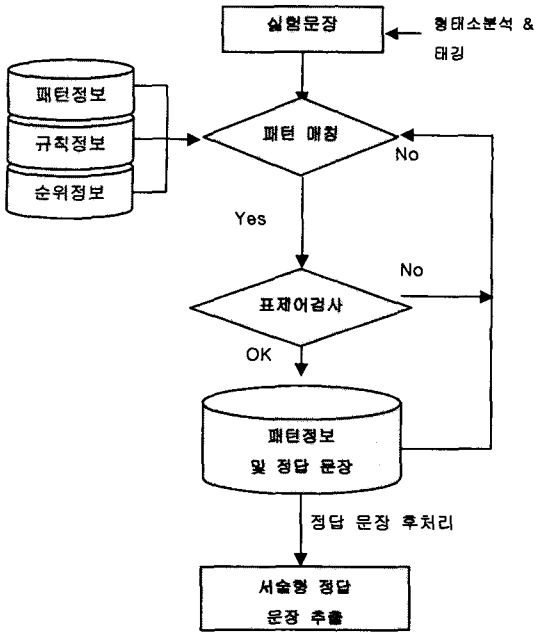
[그림 2] 패턴 추출과정

3. 서술형 정답 문장 추출 시스템

2장까지 서술형 정답 문장 추출을 위한 작업으로 서술형 정답 패턴들을 구축하였다. 구축된 패턴과 규칙정보, 순위화 정보를 이용하여 서술형 정답을 문서로부터 추출하게 된다. 그림 2는 이런 과정을 보여준다.

그림 3에서, 실험 문서의 각 문장에 대해 형태소 분석과 태깅작업을 실시하고, 이 문장에 대해 패턴정보를 적용하기 위해 정답패턴과 실험 문장간 패턴 매칭을 실시한다.

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.



[그림 3] 서술형 정답 문장 추출 과정

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

표제어 X의 경우는 연속된 명사의 나열이거나 단일 명사를 표제어 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 대명사나 관형사 "이/mm"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 표제어의 조건에 충족시키게 한다.

정답 문장 Y는 보통 한 두 어절로 이루어진 것은 제외 한다. 이것은 서술형 정답 문장을 추출하기 위해서이다. 단순 한 두 어절로 이루어진 단답형 정답은 본 연구의 목적이 아니기 때문이다.

X,Y가 조건에 만족하면, 패턴정보와 정답 문장 정보를 저장하고, 다음 유형의 패턴들을 실험 문장에 적용한다.

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

4. 실험 및 분석

본 시스템에 대한 평가는 2장의 서술형 정답 문장 태깅한 학습 문서와 다른 실험 문서 100개에 대하여 실시하였다. 100개의 문서에 대해서는 구축된 패턴정보를 이용하여 직접 수동으로 서술형 정답 문장 태깅을 실시하였다. 그리고 수동 태깅 결과와 자동 태깅을 실시하여 얻어진 자동 태깅 결과를 통하여 비교 및 평가를 실시하였고, 그 결과는 다음 표5 와 같다.

[표 5] 실험결과

수동태깅문서에서의 서술형 정답태깅 문장	136 문장
자동태깅 결과문서에서의 서술형 정답 태깅 문장	130 문장

130개의 정답문장들이 자동으로 추출되었다. 이 문장들을 수동태깅된 문장들과 비교해 본 결과, 130개의 자동 태깅 문장들 중 서술형 문장 태깅이 올바르게 된 문장은 77개의 문장으로 약 59.2%의 성능을 나타내었다.

이런 성능은 본 연구가 초기 단계의 연구이고 진행중인 연구여서 높은 성능을 나타내지 못하고 있다. 그 이유를 살펴 보면, 패턴들에 대한 정제화 작업이 자세하고 정밀하게 이루어 져야 한다. 정제화 작업을 통해 각 패턴 별 특성을 파악하고, 그 패턴들에 대한 세부 규칙들을 정하는 작업들과 함께, 각 유형 별 패턴들 순위화도 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 패턴들이 있다. 이와 같은 작업들을 통다면, 시스템의 성능 향상을 기대할 수 있을 것이다.

마지막으로 시스템에서 얻어진 실험결과 얻어진 자동 태깅 문장들을 살펴보자. 다음 표6은 정의 유형 부분에 대한 올바른 서술형 정답 문장들을 추출한 경우이다.

[표 6] 추출된 서술형 정답 문장

서 술 형 정 답 문 장
[X-정의:가드레일:X-정의]은 [Y-정의:최근에는 자동차 도로의 양쪽에 설치한 방호책:Y-정의]을 말하기도 한다.
[Y-정의:옛 이름은 지우산(智雨山)이며, 봉우리의 바위들이 마치 누룩더미로 쌓은 여러 층의 탑처럼 생겼다:Y-정의] 하여 [X-정의 : '누룩덤':Y-정의]라고도 한다.
[X-정의:염건피:X-정의]는 [Y-정의:염장피(鹽藏皮)를 건조시킨 건조시킨 것:Y-정의]으로
[X-정의:'간양(看羊)':X-정의]은 [Y-정의:흉노에 포로로 잡혀갔던 소무(蘇武)의 충절:Y-정의]을 뜻하는

오분석된 문장들은 표7를 통해 살펴보면. 첫번째 문장은 태깅이 되어야 하는데 안되는 경우이다. 'Y이 X이다.'라는 정의 유형의 규칙이지만 표제어인 '개도'에 대한 처리 중 이에 대한 형태소 분석의 오류로 일반 명사로 인식 실패하여 패턴 추출에 실패 하였다. 두번째 문장은 태깅이 잘못 된 경우이다. 이는 '[Y]이 [X]이다.' 패턴 보다는 '[X]는 [Y]이 특색이다.'라는 특징유형의 문장으로 추출이 되어야 한다.

[표 7] 서술형 정답 문장의 오류

태깅이 되어야 하는데 태깅이 안되는 경우
[Y-정의:이 개각도가 전주(全周) 360°의 몇 분의 1인가를 분수값으로 표시한 것:Y-정의]이 [X-정의:개도:X-정의]이다.
태깅이 안되어야 하는데 태깅된 경우
[Y-정의:가산관료제는 근대 관료제와는 화폐급의 뒷받침이 있는 본직(本職)으로서의 직무행위가 없는 것 : Y-정의]이 [X-정의:특색:X-정의]이다</S>.

위와 같은 오류들은 첫째, 시스템에서의 정답 문장 자동 패턴 매칭 모듈과 형태소 분석 결과의 오분석으로 인한 오류, 둘째, 적절하지 못한 패턴의 순위화를 따른 패턴 매칭으로 인한 패턴 매칭 오류를 들 수가 있다.

5. 결론 및 향후 연구

본 연구에서는 서술형 정답 유형들에 대하여 정의하고, 각 서술형 정답 유형 별 패턴들을 수집하였다. 그리고 수집된 패턴들에 대해 정제화작업을 거치고 각 패턴 별 제약 규칙들을 추가하고 각 패턴 별 순위화를 통해 패턴들을 보완하였다. 이 패턴정보들을 이용하여 실제 백과사전에서 서술형 정답들을 자동 추출하였다.

현재 높은 성능을 위해 연구가 진행 중이며, 이를 위해 다양한 정답 패턴들을 구축하고 이 패턴들에 대한 정제화 작업들을 통하여 패턴 정보를 정교화 하는 작업이 요구 된다. 또한 다양한 문장 형태에 적용 가능한 시스템의 구성도 필요하다. 추후 더 나아가서는 서술형 정답을 추출하는 서술형 질의 응답시스템의 연구가 진행될 것이다.

참고 문헌

[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.

[10] S. Blair-Goldensohn, K.R. McKeown and A. Hazen Schlaikjer, "A Hybrid Approach for QA Track Definitional Questions", *The Twelfth Text REtrieval Conference (TREC 2003) Notebook*, pp. 336-343, 2003.

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

[12] D. Ravichandran and E. Hovy, "Learning Surface Text Patternsfor a Question Answering System", *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 41-47, 2002.

[13] H. Cui, M..Y. Kan and T.-S. Chua, "Unsupervised Learning of Soft Patterns for Generating Definitions from Online News", *Proceedings of the Thirteenth World Wide Web conference (WWW 2004)*, pp.90-99, 2004.

[14] 두산세계대백과사전 엔싸이버, http://www.encyber.com