시소러스범주정보를 이용한 질의응답시스템

김수민[U] 백대호 김상범 임해창
고 려 대 학 교 컴 퓨 터 학 과
{smkim, daeho, sbkim rim}@nlp.korea.ac.kr

A Question Answering System Using the Information of the Category Information of Thesaurus

Su-Min Kim[U] Dae-Ho Baek Sang-Beom Kim Hae-Chang Rim
Dept. of Computer Science and Engineering, Korea University

요 약

정보검색시스템은 사용자의 질의를 입력받아 사용자가 원하는 정보를 검색해주는 시스템을 의미한다. 그러나, 대부분의 정보검색시스템은 단어와 연산자의 조합으로 이루어진 질의를 입력받아 문서를 검색해 주고, 사용자는 그 문서들 중에서 원하는 정보를 다시 찾아내야 한다. 본 논문에서는 영어 자연어질의를 입력받아 사용자가 원하는 정보에 좀 더 근접한 형태의 답으로서 제한된 길이의 짧은 답을 제시하는 시스템을 구현한다.

시스템은 크게 질의분석단계, 문서검색 및 분석단계, 정보추출단계의 세 단계로 나눌 수 있다. 사용자 질의분석단계에서는 의문사 정보와 오토마타, 시소러스 범주정보를 이용하여 질의에 대한 정답이 될 수 있는 단어의 속성을 예측하였다. 문서분석단계에서는 정답이 될 수 있는 단어의 후보를 선정하기 위해서 시소러스의 범주정보를 사용하였고, 선정된 정답후보 중에서 정답을 추출하기 위해 각 후보단어의 질의어단어와의 평균거리가중치, 범주간유사도, 공기질의어비율을 사용하였다. 실험을 통해 평균거리가중치만을 이용하는 것 보다 범주간유사도와 공기질의어비율을 함께 이용한 것이 성능의 향상을 보였다.

1. 서론

일반적으로 정보검색시스템이라 하더라도 그 시스템이 검색해 주는 것은 정보가 포함되어 있을 가능성이 높은 문서이다. 때문에 사용자는 검색된 문서에서 자신이 원하는 정보를 찾기 위해 문서를 읽고 스스로 정보를 찾는 과정을 거쳐야한다. 검색된 문서의 양이 적을 때는 문제가 없지만, 방대한 양의 문서가 검색된 경우에는 사용자가 일일이 문서를 읽고 정보를 찾는다는 것은 상당한 노력이 아닐 수 없다. 또한 정보검색 시스템의 검색 대상이 되는 문서의 양이 증가함에 따라 검색결과로 사용자에게 제시되는 문서의 양도 많아졌다. 때문에 사용자의 정보요구에 보다 근접한 형태의 정보를 제공해야하는 필요성이 대두되었다.

EH한 단어들과 연산자의 조합으로 질의어를 입력받는 정보검색시스템의 경우 사용자는 자신이 원하는 바를 표현하는데 적절한 단어들과 연산자를 선택해야하는 어려움을 겪는다.

When did Nixon die?

위와 같은 질의문을 명사만으로 구성할 경우, time, Nixon, death등의 의미적으로 유사한 명사들을 조합한다 하더라도 정확한 의미를 표현하기란 쉽지 않다.

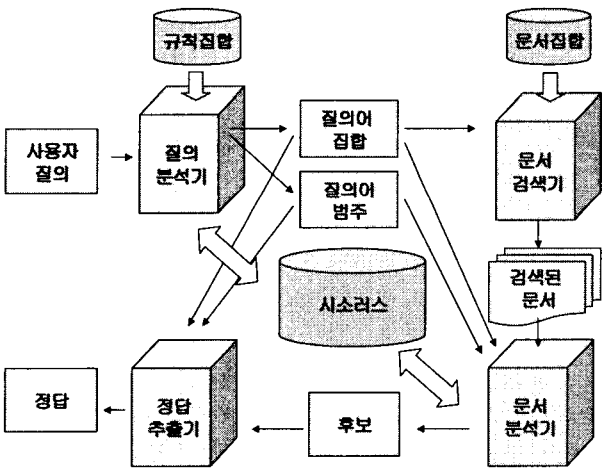
이러한 필요성에 의해 나타난 것이 자연어 질의를 처리할 수 있는 정보검색시스템인 질의응답시스템이다. 그러나 자연어질의를 대상으로 하는 정보검색시스템에서 자연어질의를 개별적인 단어들의 나열로 간주하여 When AND did AND Nixon AND die로 질의를 재구성할 경우 질의문이 요구하는 정확한 의미를 시스템에 전달할 수 없다. 즉, 자연어 질의문에 보다 면밀한 분석이 필요한 것이다.

질의응답시스템에서는 질의를 분석하는 것과 문서에 나타난 단어들 중에서 답이 될 수 있는 것을 인식하는 것이 결정적이 요소라 할 수 있다. 기존연구로 질의를 분석하기 위해서 의문사정보를 이용한 연구와 규칙기반 부

분구문분석을 이용한 연구가 있었다. 문서에 나타난 고유명사들을 인식하고 분류하기 위해서 정보추출기법을 적용한 연구와 고유명사사전을 이용한 연구, 정규표현을 이용한 연구가 있었다[1][2][3][5].
본 논문에서 구현하는 시스템은 질의를 분석하는데 수동으로 구축한 오토마타를 사용하여 주요어구를 추출하였고 주요어구의 범주를 결정하여 질의의 범주로 할당하였다. 질의의 범주를 결정하고 문서에 나타난 범주를 결정하는데는 시소러스의 범주정보를 사용하였다. 또한 각 범주는 범주 벡터로 표현을 하여 단일한 범주만을 할당하는 것이 아니라 경우에 따라 가능한 모든 범주를 할당할 수 있다. 또한 검색결과 측면에서는 사용자에게 문서를 제시하는 것이 아니라 질의문의 답으로 가장 적합한 문서의 일부분만을 제시하기 때문에 사용자가 문서 내에서 정보를 다시 찾아야 하는 부담을 줄일 수 있다.

2. 시소러스 범주정보기반 질의응답시스템

본 논문에서 구현한 시스템은 질의문을 입력받아 방대한 문서집합[1] 내에서 사용자가 원하는 답을 찾아 제시하기 때문에 시스템이 처리할 수 있는 질의문은 문서집합 내에 있는 사실에 기반한 내용을 묻는 질문이어야 한다. 즉, 질의문에 대한 답은 시스템이 검색 대상으로 하고 있는 문서 집합 내에 적어도 하나 이상의 문서에 존재해야한다.
eh한 사용자에게 제시되는 정답은 일정길이에 제한을 두기 때문에 질의문은 짧은 답을 갖는 요구하는 것이어야 한다. 여러 가지 사실들을 종합해서 도출되는 답을 요구하는 질의문이나 복잡한 과정에 대한 설명을 요구하는 질의는 시스템의 질의 대상에서 제외된다. 전체적인 시스템 구성은 [그림1]과 같다.

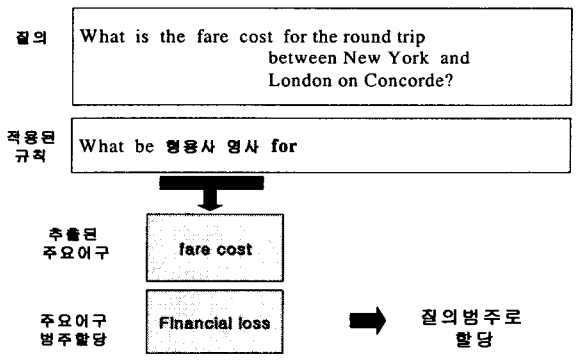


[그림 1] 전체시스템 구성도

본 시스템은 크게 세 단계로 나눌 수 있다. 사용자가 입력한 자연어 질의문을 분석하는 단계, 질의문에서 적절한 단어를 선택하여 문서를 검색하고 검색된 문서를 분석하는 단계, 분석된 문서내의 내용 중에서 답을 추출하는 단계가 그것이다.

2.1. 질의 분석 단계

단어와 연산자의 조합으로 구성된 질의어를 처리하는 시스템과 달리 자연어 질의를 처리하는 시스템이 갖추어야 하는 기능 중의 하나가 질의의 의미를 분석하는 것이다. 질의분석단계에서는 의문문의 형태로 구성되어 있는 질의문을 분석하여 질의의 답이 될 수 있는 단어의 범주를 결정한다. 그렇게 함으로써 문서 내의 단어들 중에서 답이 될 수 있는 범주와 유사한 범주를 갖는 단어를 추출할 수 있다.
질의에 범주를 할당하기 위해서는 질의문을 분석해야 한다. 질의문은 크게 How, Where, When, Who, What, Which의 의문사를 포함하는 문장과 그 외의 문장으로 나눌 수 있다. 먼저 질의문을 품사태깅하고 태깅된 질의문중 Who, Where, When등의 의문문은 의문사만으로 정답의 범주를 결정한다. How의 경우는 How다음에 나타난 형용사나 부사에 따라 범주가 결정된다. 예를 들어 How long의 경우는 시간이나 길이 범주를 할당받게 되고, How many나 How much는 수량범주를 할당받게 된다. 이와 달리 의문사가 What과 Which인 의문문과 의문사가 없는 의문문의 경우는 정답의 범주를 결정하기 어렵다.
본 논문에서 구현한 시스템은 이들 의문문의 분석을 위해 수동으로 규칙을 구축하였고, 이 규칙에 의해 의문문의 주요 어구를 추출한다. 정답의 범주가 의문사에 의해서가 아니라 주요어구에 의해 결정되는 것이다. 추출된 주요 어구는 시소러스를 통해 범주가 결정된다. 본 시스템은 시소러스의 범주 중에서 46개의 범주를 사용하여 어구의 범주를 할당한다. 결정된 주요 어구의 범주는 의문문의 범주로 할당된다. [그림2]는 이 과정을 나타낸다.



[그림 2] 질의범주할당과정

1) TIPSTER와 TREC문서 CD(1-5)

[표1]은 의문사에 따른 의문문의 분류와 의문문에 할당되는 범주를 타나낸다.
또한 질의문의 분석을 통해 질의문에 나타난 단어들을 추출하여 질의어 집합을 생성한다. 질의어 집합에는 명사, 동사, 형용사, 부사 CD(cardinal number)등이 포함된다.[2] 이 질의어 집합은 문서를 검색하는 단계와 답을 추출하는 단계에서 참조된다.

질의		질의범주
Who		PERSON
Where		COUNTRY CITY CAPITAL PENINSULA ISLAND CONTINENT PROVINCE MOUNTAIN MOUNTAIN_PEAK RIVER OCEAN
When		DAY YEAR TIME_PERIOD TIME_UNT TIME
How	far, tall	LENGTH LINEAR_UNT
	long	LENGTH LINEAR_UNT YEAR TIME TIME_PERIOD TIME_UNT TIME
	rich	MONETARY_VALUE MONETARY_UNT ECONOMIC_CONDITION FINANTIAL_LOSS
	much, many	NUMBER
What	규칙에 적용된 경우	주요어구의 범주를 할당한다.
Which		
이외의 경우	규칙에 적용되지 않은 경우	범주를 할당하지 않는다.

표 1] 질의범주 할당표

2.2. 문서 검색 및 분석단계

질의 분석을 통해 생성된 질의어 집합을 문서 검색시스템에 적용하여 문서를 검색하고, 상위의 문서들을 순차적으로 분석한다. 먼저 문서를 태깅하고, 태깅된 문서와 질의 분석 단계에서 생성한 질의 범주를 이용하여 후보를 추출한다. 후보는 문서 내의 단어들 중 후보가 될 수 있는 단어를 의미하며, 후보 단어의 품사와 질의문의 범주에 의해서 결정된다. 예를 들어, 질의 범주가 city면 후보는 고유명사들이 되고, 질의 범주가 length면 후보는 CD(cardinal number)들이 될 것이다.
추출된 후보들은 단어 범주 할당 모듈에 의해 범주를 할당한다. 명사와 고유명사의 경우는 시소러스를 이용해 범주를 할당하고, CD의 경우는 전후에 나타난 단어의 정보를 이용해 범주를 할당한다. [표2]는 단어의 범주 할당 예를 보여준다. 예를 들어, 92Km나 $600,000의 경우는 한 단어를 분리하고, Km와 $정보를 이용해 범주를 할당하게 된다. President Steven의 범주가 PERSON이므로 Steven의 범주를 PERSON으로 할당할 수 있다.

2) Penn Treebank의 품사태그집합에서 cardinal number에 부착되는 태그

President Steven	person
New York	city
Seoul	city
92m	length, number
5 may	timeperiod number
$600,000	economic condition financial loss monetary value

[표2] 단어에 할당된 범주들

2.3. 정답 추출 단계
문서 분석 단계를 통해 정답 후보가 추출되면, 정답에 가장 근접한 후보를 선택해야 한다. 정답을 선택하기 위해 사용하는 정보는 질의어 집합 내의 단어들과 평균거리가중치와 범주 유사도, 그리고 공기 질의어 비율이다.

2.3.1 평균거리가중치
질의 분석 단계에서 생성한 질의어 집합 내에는 명사, 동사, 형용사, 부사, CD등의 품사를 가진 단어들이 포함되어 있다. 후보 단어의 주변에 질의어 집합 내 단어들이 나타난 경우 그 단어와의 거리를 모두 더해 평균거리가중치를 계산한다.

$$ADW(wi) = \sum_{j=i-Win}^{i+Win} DW(I,J) \times E(Q,wi) \qquad (1)$$

$$DW(i,j) = \frac{Win - |i-j|}{Win} \qquad (2)$$

$e(Q, wi)$ 1 단어wi가 질의어집합 Q의 원소일 경우
0 원소가 아닐 경우

수식(1)에서 ADW(w_i)는 문서 내 i번째 단어인 후보w_i의 질의어 집합 내 단어들과 평균거리가중치를 Win은 후보 단어 주변의 단어들 중 거리계산 시 고려할 전후 단어의 개수를 나타낸다.$DW(i,j)$는 문서 내 거리가 가까울수록 높은 값은 갖고, 거리가 멀수록 낮은 값을 갖게 되며, 0에서 1사이의 값을 갖는다. w_i, w_j는 각각 문서내의 i,j번째 단어를 Q는 질의어 집합을 나타낸다. 즉, 후보 주변의 단어들 중 질의어 집합에 속하는 단어가 나타났을 경우, 그 단어와 후보 단어간 거리의 평균을 의미한다.

2.3.2 범주간 유사도
범주간 유사도는 질의 범주와 후보 단어 범주간의 유사도를 의미한다. 두 범주가 동일할 경우는 높은 값을 갖고, 동일하지 않더라도 유사할 경우에는 높은 값을 갖는다. 두 범주가 명백히 무관할 경우는 낮은 값을 갖고 그 외의 경우는 기본 값을 갖는다. 예를 들어, time period와 time unit은 낮은 범주간 유사도 값을 갖게 되고, length와 monetary unit은 낮은 범주간 유사도 값을 갖

게 된다. 하나의 단어에는 시소러스의 범주정보를 이용하므로 복수개의 범주가 할당될 수 있다. 두 단어의 범주가 유사도를 계산할 때 하나의 범주만이라도 일치할 경우 두 단어는 높은 유사도 값을 갖게 된다. 두 단어의 범주가 일치하는 것이 존재하지 않을 경우에는 미리 정의한 유사범주관계를 참조하여 유사하다고 판단된 경우는 높은 값을 갖게 된다. 유사범주관계는 시스템 구현에 사용한 46개 범주에 대해 미리 정의해 두었다. [표3]은 유사범주관계를 나타낸다.

MOUNTAIN	MOUNTAIN_PEAK
MONETARY_VALUE	MONETARY_UNIT
FINANTIAL_LOSS	ECONOMIC_CONDITION
TIME TIME_PERIOD	TIME_UNIT
CINEMA	MOVIE
LENGTH	LINEAR_UNIT
WORD	NAME
CAPITAL	CITY

[표3]유사 범주 관계

2.3.3 공기단어비율

공기 단어 비율은 질의문에 나타난 의미 있는 단어들의 집합인 질의어 집합에 속한 단어들이 후보 단어 주변에 공기한 비율을 나타내며 식(3)과 같다. 예를 들어, 질의어 집합의 단어 수가 6이라 하고, 후보 단어 주변에 3개의 단어가 공기한 경우 R_i의 값은 0.5가 되고, 후보단어 주변에 6개의 단어가 모두 공기한 경우 R_i의 값은 1이 될 것이다. 즉, 한 두 개의 질의어 단어와만 자주 공기한 후보 단어보다는 여러 질의어 단어와 공기한 후보 단어가 높은 값을 갖게 되고, 정답으로 추출될 가능성이 높아지게 되는 것이다.

$$Ri = \frac{\text{후보단어와 공기한 질의어집합내 단어 수}}{\text{질의어 집합의 총 단어 수}} \quad (3)$$

2.3.4 정답 가능치

최종적인 후보 단어 W_i의 정답 가능치는 다음과 같이 계산된다.

$$score_{w_i} = ADW(w_i) \times \sim(C_{w_i}, C_Q) \times R_i \quad (4)$$

$Sin(C_{Wi}, C_Q)$은 단어 범주와 질의어 범주의 범주간 유사도를 의미하고, C_{Wi}는 단어 W_i의 범주, C_Q는 질의범주를 의미한다. 정답 가능치는 질의어 단어와의 평균거리가중치가 높고 범주간 유사도가 높으며, 공기단어비율이 높은 후보가 높은 값을 갖게 됨을 알 수 있다. 이 값이 높은 후보 단어를 중심으로 일정 길이만큼의 텍스트를 잘라내었다.

3. 실험 및 평가

본 논문에서 구현한 시스템은 TREC8의 Question Answering track에서 사용한 질의를 사용하여 실험하였다. 실험에 사용한 문서집합은 [표4]와 같다.

품사 태거는 TreeTager를 사용하였고, 검색시스템은 OKAPI 알고리즘으로 구현한 시스템을 사용하였다. 시소러스는 WordNet을 사용하였다. 문서분석은 검색된 문서들 중에서 상위 10개의 문서에 대해 수행하였고, 여기서 정답을 추출하였다[4].

정답은 250바이트로 길이 제한을 두고 실험을 하였다. 즉, 문장의 구분 없이 250바이트 이내로 문서의 일부를 추출하여 답으로 제시한다. 하나의 질의에 대해서는 5개의 정답 후보를 제시하며, 각 후보는 1위에서 5위의 순위를 갖는다.

AP newswire
Wall Wtreet Journal
San Jose Mercury News
Financail Times
Los Angeles Times
Foreign Broadcast Information Service

[표4] 사용한 문서집합(TIPSTER and TREC)

단어의 범주를 할당하기 위해 사용한 범주는 시소러스 범주 중에서 46개를 선택하여 사용하였다. 결과 평가는 MRR(Mean Reciprocal answer Rank)를 사용하였고, 식(5)와 같다[1].

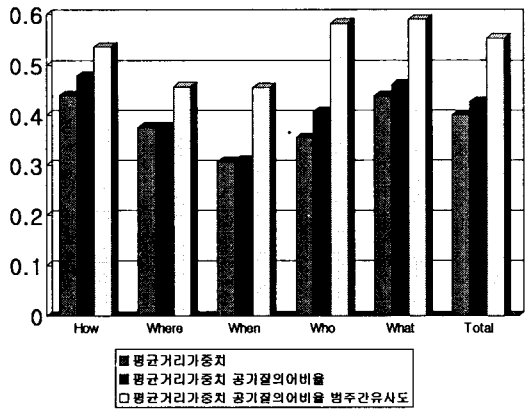
$$MRR = Mean \sum_{i=1}^{N_q} \frac{1}{\text{정답순위}} \quad (5)$$

여기서 정답 순위는 5개의 정답 후보 중에 답으로 인정된 후보의 순위를 의미한다. N_q는 실험에 사용한 전체질의어의 수를 나타낸다. 정답은 TREC8에서 사용했던 정답 패턴 인식 프로그램을 사용하여 판정하였다. 예를 들어, 정답이 1위에 있을 경우 1이 되고, 2위에 있을 경우 0.5가 되며, 5위에 있을 경우 0.2가 된다. 5위 안에 답이 존재하지 않을 경우 0이 된다. 또 5개의 정답 후보 내에 정답이 복수 개 존재할 경우, 가장 높은 값을 RR(Reciprocal Rank)값으로 채택한다.

[표5]는 정답을 추출하기 위해 평균거리 가중치만을 이용한 경우(A), 평균거리가중치와 공기질의어비율을 이용한 경우(B) 그리고 평균거리가중치와 공기질의어비율, 범주간유사도를 모두 이용한 경우(C)를 각각 실험한 결과이다. [그림3]은 이 결과를 그래프로 나타낸 것이다. [표5]에서 #는 시스템이 정답을 찾은 질의의 수를 나타낸다. 평균거리가중치만을 이용했을 경우는 전체 196질의 중 111개에 대해서만 정답을 찾았고, 그 MRR값은 0.399이다. 이에 비해 평균거리가중치와 공기질의어비율을 함께 고려한 경우는 평균거리가중치만을 고려한 경우와 동일하게 196개 질의 중 111개에 대해 정답을 찾

	A	#	B	#	C	#
How(31)	0.439	18	0.478	17	0.535	19
Where(21)	0.376	12	0.376	12	0.456	15
When(18)	0.306	9	0.310	9	0.454	13
Who(48)	0.354	26	0.406	26	0.582	32
What (78)	0.438	46	0.460	47	0.590	60
total(196)	0.399	111	0.427	111	0.552	139

[표 5] 실험결과



[그림 3] 실험결과 그래프

았지만, MRR값이 0.427로 좀 더 나은 성능을 보였다. 이는 시스템이 제시한 정답후보 1위에서 5위 중, 좀 더 상위 순위에 정답이 존재했음을 나타낸다. 마지막으로 범주간유사도까지 모두 고려한 경우는 MRR이 0.552이고, 전체 196개 질의 중 139개에 대해 정답을 제시함으로써 가장 좋은 성능을 보였다. 이것은 각각의 질의에 대해 시스템이 적절한 범주를 결정하였고, 문서에서 찾은 정답 후보에 대해서도 시소러스를 이용하여 범주를 할당하여 두 범주간의 유사도가 질의에 대한 답을 찾는데 효과적으로 작용했음을 알 수 있다.

4. 결론

본 연구에서는 영어 자연어 질의에 대한 질의 응답시스템을 구현하였다. 질의 분석 단계와 정답 후보의 선정을 위한 문서 분석 단계에서 각 단어들의 범주를 할당하기 위해 시소러스의 범주 정보와 수동으로 구축한 규칙을 사용하였다. 그리고 제한된 길이의 텍스트를 정답으로 추출하기 위해 각 후보 단어의 질의어 단어와의 평균 거리 가중치, 범주간 유사도, 공기 질의어 비율을 사용하였다.

본 연구에서는 단어의 범주를 판단하는데 시소러스의 정보에 의존하였는데, 앞으로 광범위한 고유명사분류기를 사용하는 연구와 담화 분석단계를 추가한 연구를 수행할 예정이다.

5. 참고 문헌

[1] O.Ferret, B.Grau, G.Illouz, C.Jacquemin, N.Masson, "QALC - the Question-Answering program of the Language of the Language and Cognition group at LIMSI-CNRS", *Proceedings of the Eighth Text REtrieval Conference TREC8, 2000.*

[2] D.Hull, "Xerox TREC-8 Question Answering Track Report", *Proceedings of the Eighth Text REtrieval Conference TREC8, 2000.*

[3] K.C.Litkowski, "Question-Answering Using Semantic Relation Triples", *Proceedings of the Eighth Text REtrieval Conference TREC8. NIST Special Publication, 2000.*

[4] Helmut Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees", *Proceedings of International Conference on New Methods in Language Processing. September 1994.*

[5] R.Srihari, W.Li, "Information Extraction Supported Question Answering", *Proceedings of the Eighth Text REtrieval Conference TREC8. NIST Special Publication, 2000.*

[6] E.Voorhees, D.Tice, "The TREC-8 Question Answering Track Evaluation", *Proceedings of the Eighth Text REtrieval Conference TREC8, 2000.*