# 자연어 질의응답 시스템을 위한 is-a 관계 패턴의 구축과 활용[1)]

심보준
서강대학교 컴퓨터학과
simbj@nlpzodiac.sogang.ac.kr

고영중
동아대학교
전기전자컴퓨터공학부
컴퓨터공학전공
yjko@daunet.donga.ac.kr

김학수
The CIIR in UMass, Amherst
Massachusetts, USA
hskim@cs.umass.edu

서정연
서강대학교 컴퓨터학과
seojy@ccs.sogang.ac.kr

## Extracting and Utilizing is-a Relation Patterns for Question Answering System

Bojun Shim
Dept. of Computer Science, Sogang University

Yungjoong Ko
School of Electrical, Electronics & Computer Engineering,Dong-A University

Harksoo Kim
The CIIR in UMass, Amherst
Massachusetts, USA

Jungyun Seo
Dept. of Computer Science, Sogang University

## 요 약

대다수의 개방영역 자연어 질의응답 시스템은 답을 선택할 수 있는 개념영역을 미리 정의 하고 있기 때문에 시스템이 준비하지 못한 범주의 개념을 묻는 질의문에 대해서는 올바른 응답을 생성하지 못하거나 예외 처리 방식으로 응답을 생성해 낸다. 본 논문에서는 전형적인 범주에 속하지 않는 명사 개념에 관한 질의문에 대해 범용적으로 대응할 수 있는 개방영역 자연어 질의응답 시스템을 제안한다.

제안하는 시스템은 상위 개념 명사구(Hypernym)에 포함되는 하위 개념의 명사구(Hyponym)들을 추출할 수 있는 일반적인 패턴들을 그 신뢰도와 함께 가지고 있다. 따라서 질의문이 임의의 명사구 개념을 요청할 때 정답의 후보들을 동적으로 생성되는 가상의 is-a 의미관계 사전으로부터 신뢰 순위로 정렬하여 추출해 낼 수가 있다.

제안하는 시스템은 "What 명사구 동사구" 형태의 질의문들 중에서 개체명 인식기나 시소러스를 이용하여 정답 후보를 손쉽게 생성할 수 있는 질의문을 배제한 실험용 질의문 집합을 이용한 실험에서 42%의 재현율을 보였다.

## 1. 서 론

TREC(Text REtrieval Conference)이 주도하고 있는 개방영역 자연어 질의응답 시스템(이하 질의응답 시스템)은 일반적인 정보검색 시스템과 다른 두 가지의 큰 특징을 가진다. 첫번째는 사용자가 요청하는 질의가 단어의 나열이 아닌 자연어 문장이라는 점이고, 둘째는 시스템의 출력이 문서의 집합 또는 문장들의 집합이 아니라, 불필요한 정보를 포함하지 않는 단어 또는 짧은 구로 제한된다는 점이다. 그러나, 이처럼 사용자에게 높은 효율과 편의성을 제공할 수 있는 질의응답 시스템이 아직은 실용화 되고 있지는 못한 상황이다. 낮은 정확률과 재현율, 느린 응답속도 등 여러 가지 이유가 있겠지만, 본 연구에서는 기존의 시스템이 주로 전형적이고 제한적인 개념들을 묻는 질의문들을 처리하는 능력들을 향상시켜 왔다는 점에 주목하였다.

아래의 예와 같이 TREC 2003의 질의문 집합에서 city와 country를 묻는 질의문이 각각 21개와 27개씩으로 상당히 많은 비율을 차지한다.

"What city is disneyland in?"
"What country did Catherine the Great rule?"

이전의 많은 시스템들은 이처럼 제한적인 대상 개념들을 처리하기 위해서 지명, 인명, 날짜 등을 인식 할 수 있는 개체명 인식기를 중요한 모듈로 활용하였으며, 질의의 범주를 정답의 개념 범주로 분류하는 방식을 사용하였다. 이러한 시스템들은 일반적으로 정답의 개념 범주를 개체명 인식기의 개체명 집합과 동일하게 사용하거나 이를 확장한 유한개의 개념 범주를 가지고, 이

1) 본 연구는 한국과학재단 목적기초연구(R01-2003-000-11588-0) 지원으로 수행되었음.

범주 안에 속하는 개념만을 정답으로 선택하여 출력할 수 있다. 아래의 예에서 만약 'book'이란 개념 범주를 시스템이 가지고 있지 못하다면, 정답의 후보를 선정하지 못하게 된다.

"What book did Rachel Carson write in 1962?"

따라서, 새로운 범주를 만날 때 마다 이를 처리하기 위해서 새로운 개념 범주와 이를 인식하는 규칙을 추가하여야 하는 문제점이 있으며, 이는 모든 종류의 명사구가 사용자가 궁금해 하는 개념이 될 수 있다는 측면에서 현실적으로 불가능한 작업이다.

TREC에서도 다양한 형태의 명사구를 묻는 질의문이 증가하고 있다. TREC 2003의 질의문 집합에서 'What 명사구 동사구' 형태의 질의문의 경우, 135개중 89개가 부가적인 추론과정이 없다면 우리가 사용하고 있는 개체명 인식기(MINIPAR[6])로 인식할 수 없는 개념을 질의하고 있다.

본 연구에서 제안하는 방식은 시스템이 준비하고 있지 않은 개념 범주에 대해 정답의 후보 집합을 동적으로 생성하는 방식이다. 이를 위해 자주 출현하는 is-a 관계 패턴 집합을 만들고 각 패턴 집합은 신뢰점수를 갖도록 하였다. 한편, 작은 수의 표본들로부터 일반적인 is-a 패턴들을 추출하기 위하여 기계학습에서 사용하는 표본 확장방법을 사용하였다. 이렇게 구성된 is-a 패턴은 임의의 명사구에 대해서 그 명사구에 의미적으로 포함되는 후보 집합들을 생성해 낼 수가 있게 한다.

## 2. 관련 연구

현재 TREC에서 평가 되는 시스템 중 많은 시스템들이 예상되는 정답의 범주를 정의해 놓고 각 종류별로 미리 준비된 패턴 집합을 적용하는 방법을 사용한다. 이러한 방식은 예상하지 못한 범주에 대해서는 답을 구할 수 있는 패턴을 갖고 있지 않다는 단점이 있다.

2003년 TREC의 사실 질의문 집합에 대해 가장좋은 성능을 보여준 Harabagiu[2]의 시스템은 이 방식을 사용하는 대표적인 예이다. Harabagiu의 시스템에서 정답을 추출하기 위해 패턴을 적용하는 한 가지 예를 [그림 1]에 나타내었다.

"How did Virginia Wollf die?"
예상 정답 범주 : Manner-Of-Death
적용 패턴 : X {DIE, be killed} in ACCIDENT
ACCIDENT : train accident, car wreck

[그림 1] Harabagiu 시스템의 패턴 적용 예

Ravichandran[7]은 정답을 찾는데 필요한 패턴 집합을 각각의 정답 유형별로 적은 양의 표본 집합으로부터 자동으로 생성해 내는 기법을 제안하였다. 이 방법은 기계학습에서 범주가 부착된 학습 말뭉치가 많지 않을 때에 이를 대량의 말뭉치로 자동 확장 하기위한 Riloff[10]의 연구를 응용한 것이다. 이 연구의 특징은 품사나 의미 정보를 사용하지 않고 어휘와 문장부호만을 포함하고 있는 패턴을 얻어낸다는 것이다. 따라서, 매우 높은 정확성을 보인다는 장점이 있지만, 전형적인 어휘 및 문장부호 패턴이 나타나지 않는 정답 범주에는 사용되기 어렵다는 점에서 범용성이 떨어진다고 말할 수 있다.

본 연구에서는 소량의 is-a 의미관계 예제로부터 높은 범용성을 가진 품사, 어휘 및 부호 패턴을 추출하기 위하여 Ravichandran의 연구와 유사한 알고리즘을 사용하였다.

다음은 질의응답 시스템에 어휘들간의 is-a 관계를 이용하고자 했던 연구들을 살펴보겠다.

Mann[5]은 영문 Wordnet에 등록되어 있지 않은 고유명사들의 is-a 관계를 구하기 위하여, '명사 고유명사' 패턴을 '고유명사 is-a 명사'로 간주하고 약 3 기가바이트의 뉴스 기사 문서로부터 is-a 관계 쌍을 구축하여 Wordnet의 보조 온톨로지로 사용하였다. Hearst[3]는 경험적으로 명백한 어휘 및 품사 패턴을 수동으로 정의하여 대량의 말뭉치로부터 is-a 관계쌍을 추출하였다. Mann과 Hearst는 모두 정적인 지식자원으로 활용되는 온톨로지를 구축하기 위하여 경험과 관찰에 의해 수동으로 정의된 품사 패턴을 사용하였다.

본 연구에서는 사용자에 의해 요청되는 임의의 개념에 대해서 동적으로 반응할 수 있도록 품사 패턴 집합을 구축하였다. 패턴 집합의 범용성을 보장하기 위하여 대량과 다양함을 특징으로 하는 웹 문서 집합으로부터 패턴을 추출하였다.

## 3. 질의문 유형의 분류

### 3.1 정답 개념 범주에 의한 분류 방법의 문제점

1장에서 서술 하였듯이 많은 시스템들은 질의문의 유형을 정답의 범주로 분류를 해 놓은 후 정보검색의 결과로 얻어진 후보 문장들로부터 결정된 정답 범주를 탐색하는 방법을 사용한다. 이러한 시스템들에서는 정답의 범주를 잘 정의 하는 것이 시스템의 성능을 크게 좌우한다. 더 다양한 형태의 질의문들을 처리하기 위해 새로운 정답의 범주는 계속 추가 되고 있다. Pasca[8]는 새로운 개념이 등장할 때 마다 이를 정답 계층에 추

가하고, 이를 Wordnet의 노드에 연결시키는 연구를 진행하기도 하였다. 대표적인 질의문 범주 분류 연구인 Li[4]의 연구에서 보여 지듯이 대부분의 정답 범주는 'others' 범주를 포함하고 있으며 이것은 정답 범주의 사각 지역으로 작용하게 된다.

본 연구에서는 상위 개념어에 포함되는 하위 명사구를 묻는 질의문들을 포괄적인 하나의 질의문 유형로 간주함으로써 이 문제를 해결하려 한다.

### 3.2 'What 명사구 동사구'형태의 질의문의 분류

본 연구에서는 모든 종류의 질의문을 연구 범위에 포함시키는 것이 아니라, 'What 명사구 동사구'형태를 갖는 문제들을 대상으로 하였다. 이 형태의 질의문들은 'What 동사구 명사구'형태의 질문과 등가의 질의문으로 상호 변환 될 수 있지만, 대상 개념 명사구의 범위 설정은 본 연구의 범위에 포함되지 않으므로 개념 명사구가 명확히 드러나는, 의문사 What에 명사구가 뒤따르는 형태의 질의문들로 한정짓게 되었다. 즉, 아래의 형태를 갖는 질의문들이다.

"What X T1 T2 ... Tn"

여기서 X는 명사구이고, T1은 동사이다. 그렇다면 질의응답 처리과정은 T1,...,Tn의 제약조건과 Y is-a X 관계를 동시에 만족하는 명사구 Y로 탐색공간이 한정 된다. 'What 명사구 동사구' 형태의 질의문을 본 연구에서는 세 가지의 형태로 분류 하였다.

형태 1 : "What city is Disneyland in?"
형태 2 : "What color belt is first in karate?"
형태 3 : "What book did Rachel Carson write in 1962?"

각각 city, color, book을 개념 명사구로 갖는 질의문들이다.

형태 1은 정답의 후보 명사구들이 개체명 인식기로 명확히 인식 가능한 부류의 질의문들이다. 'What city', 'What country', 등의 질의문들이 이 부류에 속한다.

형태 2는 개체명 인식기로는 인식이 불가능 하지만, Wordnet과 같은 시소러스나 사전을 이용하여 후보 집합을 찾아낼 수 있는 부류들이다. 이 형태의 특징은 질의문의 다른 구성요소에 상관없이 같은 개념은 같은 후보 집합을 가지며 후보 집합의 원소는 새로 생성되거나 확장 소멸되지 않는 정적인 집합이라는 것이다.

형태 3은 개념 명사구에 의미적으로 포함되는 명사들이 개체명 인식기로 인식 가능한 어휘 패턴을 가지고 있지도 않고, 사전이나 시소러스등에 미리 저장되어 있지도 않은 열린 집합의 원소들이다. 형태 3은 준비되지 않은 개념에 포함된 대상을 묻는 질의문이 된다.

[표 1]은 TREC 2003의 질의와 수행 성능들을 본 연구의 기준대로 분류하고 분석한 결과이다.

[표 1] 질의문 유형별 평균 정답 개수

| 구 분 | 사실 질의 전체 | 형태 1 | 형태 2 | 형태 3 |
|---|---|---|---|---|
| 질의 개수 | 413 | 46 | 17 | 72 |
| 평균 정답 수 | 3.67 | 5.02 | 2.59 | 3.42 |

표1 에서 평균 정답 수란 TREC에 정답을 제출한 시스템 중 답을 맞춘 것으로 평가된 시스템의 개수를 이야기 한다. 표를 보면, 형태 3의 평균 정답 수가 형태 1 이나 전체 평균 정답 개수에 비해서 낮은 성능을 보이고 있는 것을 알 수 있다.

형태 2가 낮은 성능을 보이고 있는 것은 정답 후보 집합을 선정하는 데에 있어서의 어려움 보다는 후보 집단 간의 유사성(예. animal : dog, cat, cow)과 후보단어를 포함하는 문장이 매우 많다는 사실(예. color : red, blue, yellow)에 의한 것으로 판단된다. 이에 관한 분석은 향후 연구 과제로 남겨두기로 하겠다.

본 연구에서는 형태 3 의 낮은 성능이 후보 집합을 선정하는 데에 있어서의 어려움에 기인한다고 판단하고, 이와 같이 열린 집합을 정답 후보 집합으로 가질 수 있는 개념 명사구와 is-a 관계로 연결되는 명사구를 탐색할 수 있는 방법을 모색하였다.

## 4. is-a 관계 패턴의 구축과 사용

### 4.1 is-a 관계 패턴 표현에 사용된 태그 집합

X is-a Y 관계가 문장 안에서 출현하는 패턴을 표현하기 위한 태그는 질의응답 시스템의 특성을 고려하여 3가지 종류로 구성된다.

첫째는 품사 태그이다. 태그 집합으로는 펜트리뱅크(PennTree Bank[9]) 품사 태그 집합을 사용한다.

둘째는 개체명 태그를 사용한다. 질의응답 시스템에서 사용가능한 X is-a Y 패턴을 관찰한 결과 많은 수의 X가 고유명사인 것을 발견할 수가 있다.

셋째는 어휘 자체 또는 문장부호, 인용구문을 태그로 사용한 경우이다. 아래의 예에서 보듯이 일반적인 정보검색 시스템과 다르게 질의응답시스템에서는 기능 어휘가 정답을 찾는 중요한 단서가 되는 경우가 많이 있다.

"Korea is one of Asian country."
"Korea, an Asian country,"

첫 번째 예에서는 is와 of가, 두 번째 예에서는 문장부호','가 is-a 관계를 표현하는 데에 중요한 역할을 한다.

그 외에 작품명이나 속담, 발언 등 많은 경우에 인용부호 ""로 둘러싸인 부분이 정답이 될 수 있다는 점을 고려하여 인용부호로 둘러싸인 부분에'quoted' 태그를 부착한다. 태그의 우선순위는 어휘 -> 개체명 -> 품사 태그 순서이고, 최종적으로 태깅이 수행된 예를 [그림 2]에 나타내었다.

```
The publication of her 1962 book, "Silent Spring" Carson, a ...
=> The/the publication/NNP of/of her/PRP$ 1962/date book/NN ,/, Silent_Spring/quoted Carson/person ,/, a/a...
```

[그림 2] 태거의 수행 예

[그림 2]의 예에서 Silent Spring is-a book 관계가 성립하므로, 우리는 Y, X/quoted 라는 하나의 패턴을 얻을 수 있다.

### 4.2 is-a 관계 패턴의 구축과 활용을 위한 웹 문서의 활용

질의응답 시스템의 많은 이전 연구들에서 웹 문서와 웹 문서집합 수집을 위한 웹 검색 엔진은 다양한 도구로 활용되어 왔다. Brill[1]의 AskMSR 시스템은 질의문으로부터 확장된 질의문을 웹 검색엔진에 질의 하여 얻어진 문서에서 가장 자주 출현한 N-Gram을 정련과정을 거친 후 답으로 반환하는 전략을 사용하였다.

본 연구에서는 첫째, 패턴의 수집과정에서 둘째, 정답 후보의 탐색과정에서 모두 구글 웹 검색엔진을 사용하여 얻은 스니핏[2]을 이용한다.

패턴을 수집하기 위한 또 하나의 대안은 질의응답 시스템의 검색 대상 말뭉치인 AQUAINT 말뭉치에서 패턴을 수집하는 방법이다. 하지만, 우리가 AQUAINT를 배제하고, 웹 문서들을 패턴 수집 대상으로 정한 것은 AQUAINT의 문서 성격을 고려하였기 때문이다. AQUAINT는 뉴스 기사만으로 이루어진 약 3 기가바이트의 데이터로, 성격이 편중된 이 말뭉치 보다는 웹 문서에 더 많은 is-a 표현들이 있다고 가정한다.

### 4.3 is-a 관계 패턴의 구축

X is-a Y 관계 패턴을 구축하기 위한 첫 단계는 명백히 is-a 관계인 표본 단어쌍을 구축하는 것이다. 우리는 TREC1999~TREC2002의 질의문 - 대답 쌍과 영문 Wordnet의 상위어(Hypernym) - 하위어(Hyponym) 관계중에서 하위어가 고유명사인 것만을 선정하여 60개의 X, Y 표본 단어쌍을 선정하였다. 선정 기준은 질의문에서 물어볼 수 있는 다양한 개념들을 포괄할 수 있는 다양성과 표본의 파급효과가 높은 범용성 두 가지로, 수동으로 단어쌍을 선정하였다. 선정된 단어의 일부분을 [표 2]에 소개하였다.

[표 2] X is-a Y 패턴 추출을 위한 표본 예

| X | Y | 출처 |
|---|---|---|
| Jolt | drink | TREC |
| Conservatives | party | TREC |
| Bonaparte | ruler | TREC |
| Bertillon system | procedure | Wordnet |
| Mexican standoff | situation | Wordnet |

앞으로 소개되는 알고리즘에 사용하는 패턴을 저장하는 자료구조는 다음과 같다.

```
data_structure pattern
{
        string //패턴 문자열
        pos_x //X의 위치
        pos_y //Y의 위치
        confidence_score //신뢰도
}
```

구해진 단어쌍 집합으로부터 is-a 관계 패턴을 추출하는 알고리즘은 다음과 같다.

```
1 For each ⟨xi, yi⟩ in PAIRS
2 {
3 Bi = boolean_query(xi and yi)
4 Di = get_web_documents(Bi, DOC_NUM)
5 Di_tagged = do_tagging(Di)
6 for each sentence si in Di_tagged
7 {
8       pattern.X = xi
9       pattern.Y = yi
10      pattern.string = xi(.*)yi or yi(.*)xi
11      pattern.pos_x = pos_of(pattern.string, xi)
12      pattern.pos_y = pos_of(pattern.string, yi)
13      remain_only_tags(pattern)
14      put_into(pattern, PATTERNS)
15 }
16 }
```

[알고리즘 1] is-a 관계 패턴의 추출

2) 스니핏(snippet) 이란 웹 검색 결과에 표시되는 서너줄의 요약문을 이야기한다.

PAIRS는 [표 2]에 표현된 X, Y 단어쌍의 집합을 의미하고, 구해진 패턴들은 패턴집합 PATTERNS에 저장된다.

8, 9번 줄에서 X와 Y를 저장하는 이유는 다음에 소개하는 패턴의 신뢰도를 구하는 알고리즘에서 활용하기 위해서 이다.

10번 줄에서 패턴에 저장되는 문자열은 X로 시작해서 Y로 끝나는 부분 문자열(substring) 또는 Y로 시작해서 X로 끝나는 부분 문자열 이다.

11, 12번 줄에서 패턴내의 X 와 Y의 위치를 저장한다.

13번 줄에서 패턴은 태그만으로 이루어진 문자열로 변환된다.

14번 줄에서 최종적으로 저장되는 패턴은 태그의 나열로만 이루어진 패턴 문자열과 X 와 Y의 패턴 내에서의 위치정보, 패턴 추출에 사용된 X와 Y 이다.

이렇게 생성된 패턴들의 신뢰도를 구하는 알고리즘은 다음과 같다.

```
1 For each patterni in PATTERNS
2 {
3 right_cnt = wrong_cnt = 0
4 x_pattern
5 = replace(patterni, patterni.pos_x, patterni.x)
6 Bi = boolean_query(patterni.X)
7 Di = get_web_documents(Bi, DOC_NUM)
8 for each sentence sj in Di
9 {
        sj_tagged = do_tagging_except_x(sj, x)
10      candidate_pattern
11      = matched_substr(sj_tagged, x_pattern)
12      if (exist(candidate_pattern))
13      {
14          if (get_word_or_phrase
15      (candidate_pattern, patterni.pos_y)
16              == patterni.y)
17              right_cnt++;
18          else
19              wrong_cnt++;
20      }
21 patterni.confidence_score
22    = right_cnt / (right_cnt + wrong_cnt)
23 }
```

[알고리즘 2] is-a 관계 패턴의 신뢰도 구하기

예를 들어 X = tungsten Y = mineral 쌍을 이용하여 [알고리즘 1]에 의해 추출된 패턴 'X is the JJ Y'을 평가한다고 할 때,

4번 줄에서 패턴을 'tungsten is the JJ Y'로 치환한다.

6, 7번 줄에서 'tungsten'을 불리언 질의로 웹 검색을 실행한다.

10~12번 줄에서 반환된 문서의 각 문장에 대해서 만일 'X is the JJ Y'의 패턴이 발견된다면,

14~17번 줄에서 Y가 mineral인 경우 right_cnt를 증가시키고,

18, 19번 줄에서 Y가 mineral이 아니라면 wrong_cnt를 증가 시킨다.

21, 22에서 최종적으로 패턴의 신뢰도를 구한다.

두 알고리즘 모두 문서개수 DOC_NUM은 1000개를 사용하고, 추출된 패턴 중 출현 빈도수가 3이상인 것들만을 유효한 패턴으로 간주한다. 최종적으로 추출된 신뢰도가 높은 상위 10개 의 패턴과 신뢰도를 [표 3]에 나열하였다.

[표 3] 신뢰도 상위 10개의 is-a 관계 패턴. X is-a Y에서 밑줄로 표시된 부분이 X, 기울임체로 표시된 부분이 Y이다.

| 패 턴 | 신뢰도 |
|---|---|
| NN is the RBS JJ NN | 1.00 |
| NNP is the JJ NN | 1.00 |
| NN in the NNP NNP NNP is VBG | 1.00 |
| quoted is the RBS JJ NN | 1.00 |
| quoted is the JJ NN | 1.00 |
| NN in the NNP NNP NNP is quoted | 1.00 |
| NNS of NNP | 0.86 |
| NNS of location | 0.86 |
| NNS of quoted | 0.86 |
| NN is DT JJ NN | 0.50 |

[표 3]에서 X 부분이 quoted로 표시된 부분은 자동으로 추출된 패턴에서 X 부분의 태그를 quoted로 치환하여 치환이전의 패턴과 같은 신뢰도를 할당한 것으로, quoted 패턴은 문장 안에서 명사구와 동일한 기능으로 작용하며, 질의응답 시스템에서 정답이 될 확률이 높다는 경험적 관찰에 근거한 것이다.

### 4.4 is-a 관계 패턴의 사용

준비된 is-a 관계 패턴을 이용하여 정답 후보 집합을 선정하는 과정 역시 웹 검색 엔진을 이용한다. 사용자의 질의문에서 명시한 대상 개념인 Y를 필수 질의어로 하고, 그 외 선택된 질의어 들을 질의어의 중요도에 따

라서 필수 질의어 또는 선택 질의어로 구성한다.

정답 후보 선정 알고리즘은 패턴 신뢰도를 구하는 알고리즘과 기본적으로 유사하지만 한 가지 큰 차이를 가진다. 패턴 신뢰도를 구하는 알고리즘은 X를 알고 있는 상황에서 Y의 정, 오 여부를 판단하는 것이고, 패턴을 사용할 때에는 Y를 알고 있는 상황으로 바뀌게 된다. [알고리즘 2]의 12번째 줄에서 부합되는 패턴을 만난다면 X 위치의 단어 또는 구를 정답의 후보로 반환하게 된다.

답을 선택하는 단계에서 정답후보로 선택된 단어의 앞 또는 뒤 단어의 품사가 명사일 때는 명사가 아닌 품사가 나올 때 까지 정답 후보로 포함을 시킨다. 이것은 사용된 표본이 대부분 한 단어 명사구를 사용했기 때문에 두 단어 이상으로 구성된 명사구가 정답 후보에서 배제되는 현상을 방지하기 위한 것이다. 또한, 반복해서 나타나는 정답후보에 대해서는 신뢰도를 누적해서 가산한다. 최종적으로 정답 후보의 누적된 신뢰도 수준으로 정렬하여 정답 후보의 순위를 결정한다.

구축된 패턴을 실제로 사용하는 예를 [그림 3] 에서 표시하였다.

**question** : What book did Rachel Carson write in 1962?
target concept : book
**query** : book AND Rachel-Carson AND 1962 OR write

**sentence_from_web_search_engine** : But/CC in/in the/the 60s/CD ,/, US/country environmental/JJ awareness/NN picked/VBD up/NN ,/, spurred/VBN in/in part/NN by/by the/the publication/NN of/of Rachel/NNP Carsons/NNS 1962/CD book/NN Silent/NNP Spring/NNP ./.

**pattern** : NN is the JJ NN
**pattern** : NN is the RBS JJ NN
.....
**pattern** : NN NNP NNP
answer candidate detected : Silent Spring
confidence score : 0.6000.

[그림 3] is-a 관계 패턴을 이용한 정답 후보의 탐색

## 5. 실험 및 평가

### 5.1 실험을 위한 질의문의 선정

앞서 밝혔듯이 본 연구의 목적은 개체명 인식기 또는 정적으로 구축된 외부자원인 Wordnet과 같은 시소러스로는 찾아 낼 수 없는 개념 범주에 포함된 정답 후보들을 찾는 것을 목적으로 한다. 따라서, 정답의 개념 범주는 시스템이 정확히 인지할 수 있다는 가정을 하였다. 이 가정에 의해 실험용 질의문 "What 명사구 동사구" 패턴의 질의문으로 한정하였다. 여기서 명사구 부분이 사용자가 원하는 개념의 범주가 된다. 또한, 연구의 목적에 부합되는 테스트용 질의문들을 선정하기 위해서 다음과 같은 기준을 마련하였다.

첫째, 개념 범주가 명시적으로 개체명 인식기 태그이름을 지칭하거나, 어렵지 않은 추론으로 개체명 태그로 인식될 수 있는 경우를 배제하였다.

예)

"What country is Aswan High Dam located in?"
"What state was Amelia Earhart born in?"
(LOCATION)

둘째, 개념 범주가 Wordnet에 등록된 단어로 정답 후보가 개념범주의 하위어(hyponym) 이외에는 극히 적은 경우를 배제하였다.

예) color, animal

이와 같은 기준으로 TREC 2003의 사실 질의문 집합 중에서 72개의 실험용 질의문 집합을 만들었다.

### 5.2 간단한 질의응답 시스템의 구축

본 연구의 목적을 정답의 후보를 생성하는 문제로 한정하였지만, 단순한 형태의 정보검색 시스템과 결합하여 TREC 2003의 작업 정의에 의거하여 실험을 진행하고, 성능을 평가하였다.

이 시스템에서는 불리언 질의를 지원하는 공개용 정보검색 엔진을 이용하여 TREC의 검색대상 말뭉치인 AQUAINT 문서집합으로부터 후보 문장들을 추출하였다. 시스템에서는 일반적인 질의응답 시스템에서 사용하는 어근 추출이나 질의어의 확장 등을 일체 사용하지 않고, 질의에 포함된 단어들에서 불용어를 제외한 단어들의 포함 여부만을 문장 추출의 근거로 사용하였다. 최종적으로 가장 높은 점수를 가진 정답을 선정하기 위해 사용된 점수의 계산방법은 아래와 같다.

$$Ss = \frac{Nq}{Dmax} \qquad [식\ 1]$$

$$As = Ss \times Cs \qquad [식\ 2]$$

[식 1]에서 Ss 는 문장의 신뢰점수, Nq는 문장에 포함된 질의어의 개수, Dmax는 질의어와 정답 후보를 포

함한 단어집합의 후보 문장 안에서의 최장거리를 나타내고, [식 2]에서 Cs는 정답 후보의 누적 가산된 신뢰점수, As는 문장 안에 포함된 정답 후보의 신뢰점수를 나타낸다. [식 2]에서 가장 높은 As를 받은 정답이 최종적인 정답으로 출력된다.

### 5.3 평가 및 분석

72개의 선정된 실험용 질의문 집합에 대해서 수행된 정답 후보의 수집 결과와 정보검색 모듈이 결합된 질의응답 시스템의 수행 결과를 [표 4] 에 나타내었다.

[표 4] 실험결과. 재현율과 MRR은 정답후보 집합에 대한 평가이고, 간단한 정보검색 모듈과 결합된 질의응답 시스템에 대한 평가이다.

| 질의문 개수 | 재현율 | MRR | Accuracy |
|---|---|---|---|
| 72 | 0.42 | 0.23 | 0.24 |

정답 후보 집합의 수집 결과의 재현율은 후보 개수의 제한 없이 수집된 후보 중에서 올바른 정답이 있는지 없는지를 판단한 수치이다. 실험결과 약 42%의 질문에 대해서 정답을 후보로 포함하고 있는 것으로 나타났다. 이 수치는 또한 이상적인 정보검색 모듈과 결합했을 경우 Accuracy의 최대치이기도 하다.

MRR(Mean reciprocal rank) 신뢰도 상위 5개의 정답 후보에 대한 평가로 정답후보의 신뢰도를 구하는 정책의 적합성을 판단하는 지표로 생각할 수 있다. 0.23의 수치는 평균적으로 정답이 약 4.35 번째에 위치하는 것을 의미한다.

마지막으로 Accuracy는 TREC 2003 사실 질의문 집합에 사용된 평가 방법으로 5.2 절에서 설명한 간단한 질의응답 시스템에서 출력한 최상위 1개의 답만으로 정답 비율을 계산한 것이다.

0.24 의 Accuracy는 평가 결과가 공개된 TREC 2003 의 상위 15개 시스템의 평균정도에 해당되는 성능으로 선정된 질의문들의 평가 결과의 평균치가 전체 보다 낮은 것과 정보검색 모듈의 낮은 성능을 감안할 때 의미있는 결과로 평가할 수 있을 것이다.

시스템의 출력을 분석한 결과 정답 후보를 제출하지 못한 가장 큰 이유는 후보 수집을 위한 웹 검색 단계에서 충분한 양의 웹 문서를 수집하지 못한 것으로 밝혀졌다. 따라서, 질의응답 시스템의 정보검색 모듈에서 사용되는 질의의 확장이나 다양한 질의 구성 전략이 웹 검색에도 사용되어야 할 것이다.

## 6. 결 론

본 논문에서는 시스템이 미리 준비하고 있기 어려운 또는 불가능한 정답 범주에 대해서 준비된 패턴 집합을 사용하여 동적으로 is-a 의미관계를 가지는 정답 후보를 추출해 내는 연구를 수행하였다.

본 연구의 성과는 크게 세 가지로 요약할 수가 있다.

첫째로 질의응답 시스템에서 질의문이 요구하는 개념에 대해 범용적으로 대응할 수 있는 수행 방식을 제안했다.

둘째로 그동안 경험과 직관에 의해 만들어져 왔던 is-a 의미관계를 가지는 단어들의 출현 패턴을 대량의 웹 데이터와 다양한 의미 영역의 표본들을 이용하여 그 신뢰도와 함께 추출해 냈다.

셋째로 범용적으로 패턴 추출에 사용되던 알고리즘을 의미관계 패턴 영역에 적용하여 그 이용 가능성을 확인하였다. 즉, is-a 의미관계 이외의 의미관계에서도 의미관계 구성단어들 간의 출현 패턴을 추출하는 데에 본 연구의 방법론이 응용될 수 있을 것으로 기대된다.

## 참고 문헌

[1] E. Brill, J. Lin, M. Banko, S. Dumais and A. Ng, "Data-intensive question answering", In *Proceedings of the Tenth Text Retrieval Conference (TREC 2001)*, 2001.

[2] S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams and J. Bensley, "Answer Mining by Combining Extraction Techniques with Abductive Reasoning", In *Proceedings of the Tenth Text Retrieval Conference (TREC 2003)*, 2003.

[3] M. Hearst, "Automatic acquisition of hyponyms from large text corpora", In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-92)*, 1992.

[4] X. Li and D. Roth, "Learning question classifiers", In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-2002)*, 2002.

[5] G. S. Mann, "Fine-Grained Proper Noun Ontologies for Question Answering", *SemaNet-2002 : Building and Using Semantic Networks*, 2002.

[6] minipar : www.cs.ualberta.ca/~lindek/

[7] D. Ravichandran and E. Hovy, "Learning surface text patterns for a question answering system", In *Proceeding fo ACL 2002*, 2002.

[8] M. Pasca and S. Harabagiu, "The informative role of wordnet in open-domain question answering", In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources : Applications, Extensions and Customizations*, 2001

[9] PennTree Bank Project : www.cis.upenn.edu/

[10] E. Riloff, "Automatically Generating Extraction Patterns from untagged text", In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996.