질의문의 구문정보를 이용한 키워드 추출

Keyword Extraction Using Syntactic Information of Question

양수정, 서영훈
충북대학교

Yang Soo-Jeong, Seo Young-Hoon
Chungbuk National Univ.

요약

자연언어 질의문에서 추출된 키워드들은 정답추출에 미치는 비중이 다른 경우가 많지만 키워드들에 대해 상대적인 가중치를 부여하기가 어렵다. 본 논문에서는 이러한 문제점을 해결하기 위하여 질의 문장의 구문 정보를 이용하여 중심키워드와 일반키워드들로 구분하였으며 이를 기반으로 키워드들 간의 가중치 부여 방법을 제안한다. 질의문 코퍼스로부터 질문 유형을 분석하여 구문을 추출하고 추출된 구문정보를 이용하여 질의문에서 키워드들을 추출한다. 이렇게 얻어진 키워드들을 이용하여 다량의 문서들 속에서 중심키워드와 일반키워드들 간의 불린 검색을 통해 질의문의 정답이 포함되었을 가능성이 큰 단락을 추출하고, 질의문과 추출된 단락간의 유사도 측정을 통해 단락을 순위화 한다. 본 논문에서 제안하는 시스템은 질의문의 정답이 포함된 단락추출에 대한 정확도를 향상시킬 것으로 기대된다.

1. 서론

질의 응답시스템은 일반적으로 질의분석, 문서검색, 정답추출로 구성된다[1][2][3][4]. 질의 분석(question analysis)단계에서는 질문 유형(question type)의 분류를 통하여 질문으로부터 사용자가 찾고자 하는 것이 무엇인지를 파악한다. 그리고 질문의 유형이 분류되었다면 질문의 조건이 무엇인지를 파악하는 것이 중요하다. 찾고자 하는 대상이 사람이라면 그 사람이 어떤 사람인지를 설명하는 중심이 되는 키워드를 찾아야 한다. 하지만 자연언어 질의문에서 중심이 되는 키워드를 추출하는 것에는 많은 어려움이 있다. 일반적으로 정보검색에서는 키워드 질의어 방식으로 질의어의 가중치를 통계적인 방법으로 계산하여도 적합하다. 그러나 자연언어 질의문은 질문 유형을 결정하는 단어들과 질문에 대한 구체적인 조건을 제시하는 단어들로 구성되어 있으며 질의문을 구성하는 모든 단어들이 정답을 찾기 위한 정보를 제공해주는 중심 키워드가 아니다[5]. 따라서 자연언어 질의문에서 추출된 키워드들이 정답추출에 미치는 비중이 다른 경우가 많기 때문에 질의어에 대해 통계적인 가중치 계산 방식을 적용하는 것은 부적합하다. 즉, 용어의 출현빈도와 문서빈도를 이용하여 키워드에 가중치를 부여했을 경우 일시적으로 자주 사용되는 명사나 단순히 출현빈도만 높은 용어에 높은 가중치가 부여되는 오류가 발생할 수 있다. 따라서 질의문의 주제어가 일치하지 않더라도 일상적인 용어들이 일치할 경우에 임계치보다 높은 값이 계산되어 질의문과 관련된 정답이 포함되어 있는 단락으로 추출될 경우가 발생된다[6]. 본 논문에서는 이러한 문제점에 대한 해결 방법으로 질의 문장의 구문 정보를 이용하여 중심키워드와 일반키워드들로 구분하였으며 이를 기반으로 키워드들 간의 가중치 부여 방법을 제안한다. 이렇게 구문정보로 얻어진 키워드들을 이용하여 불린모델을 이용한 검색을 통해 단락을 추출하고 백터 모델을 이용하여 키워드들과 추출된 단락간의 유사도 측정을 통하여 단락을 순위화 한다.

본 논문은 2장에서 관련연구에 대한 정리, 3장은 구문정보와 키워드 추출방법에 대해, 4장에서는 문서 검색을 위한 방법으로 불린모델과 벡터모델에 대하여 설명한다. 마지막으로 5장에서 실험 평가 및 결론으로 구성한다.

2. 관련연구

2.1 형태소 분석을 통한 키워드추출

한국어와 같이 조사가 발달한 언어의 경우에는 한 어절에서 조사를 제외한 나머지를 색인어로 채택하는데, 이 방법이 가지고 있는 가장 큰 문제점은 복합명사에 대한 처리이다. 예를 들어, "정보검색"이라는 복합명사는 '정보검색에 대한.."에서와 같이 붙여 쓸 수도 있고 "정보 검색에 대한..."에서와 같이 띄어 쓸 수도 있는데, 위에서 설명한 방법에 따라 색인어를 채택한다면 전자의 경우에는 '정보검색'이 색인어로 채택되는 반면, 후자의 경우에는 '검색'이 색인어로 채택되게 되므로, 색인을 할 때에는 이에 대한 처리도 고려해야만 한다. 그 밖에 "정보의 검색"에서와 같이 조사 '-의'가 두 명사 사이에 오는 경우나 '정보에 대한 검색","정보를 검색하는" 등과 같이 다양한 표현에 대해서도 '정보 검색' 이라는 복합 명사를 찾아 색인어로 채택할 수 있어야 한다[7][8].

2.2 구문 분석을 통한 키워드 추출

문헌을 이루고 있는 각 문장에 대한 구문 분석을 통해 특정 기능을 가진 단어나 구를 식별하여 이것을 색인어로 사용한다. 자동 색인 기법을 위한 구문 분석은 수준에 따라 의미적인 처리를 포함하여 완벽한 구문 분석 기법과 의미 처리를 제외한 통사적 구문 분석 기법으로 나눠 볼 수 있는데, 자연언어처리에 있어서 완벽한 구문 분석은 대단히 어려운 일인 반면, 그러한 어려움에 비추어 볼 때 구문 분석의 효과가 그다지 크지 않다는데 구문 분석의 문제점이 있다.

2.3 통계적 기법에 의한 키워드 추출

통계적 기법에 의한 자동 색인에서는 색인어의 선정을 위한 기준으로 주어진 문헌에서 특징 단어가 얼마나 자주 사용되었는가 하는 빈도수 정보가 사용된다. 단어의 사용빈도수는 산출 방식에 따라 단순 빈도수와 상대 빈도수로 구분되며 단순 빈도수와 상대 빈도수의 산출 방식은 물론, 주제어 선택 시 사용될 빈도수의 한계치는 모두 실험적으로 결정된다. 통계적 기법에서 고려해야 할 사항은 빈도수가 지나치게 높거나 지나치게 낮은 단어는 주제어에서 제외한다. 즉, 문헌에서 빈번하게 나타나는 기능어를 수록한 불용어 리스트를 사용하여 고빈도어를 먼저 제거한 다음 나머지 단어들을 빈도수 순으로 배열하고, 임의로 정한 빈도수의 최저 한계치를 초과하는 단어들을 색인어로 선택한다.

추출된 색인어들에 대한 가중치는 단어의 출현 빈도수로부터 가중치를 계산하고 이 값이 일정한 범위 안에 드는 단어를 선택하여 가중치 없이 사용하거나, 산출된 빈도수 자체를 색인어에 부여된 가중치로 사용한다[7].

3. 구문정보를 이용한 키워드 추출

3.1.문법 형태소

문법형태소는 어휘형태소에 비해 개수가 많지 않기 때문에 미등록어를 인식하거나 형태론적 중의성을 해결하는데 유용한 정보가 된다. 문법형태소는 조사와 어말어미, 선어말어미 등 허사들로 구성되는데 그중 조사는 크게 격조사, 보조사, 접속조사로 분류된다.

문장을 이루는 성분에는 서술어, 부사어, 관형어가 있으며, 이들은 각각 용언, 부사, 관형사들로 이루어지는데 비해 체언은 여러 가지 문장성분이 가능하다. 즉 체언은 문장에서 여러 가지 격(자격)을 가질 수 있는데 이들이 각기 '격'을 나타낼 수 있는 것은 격조사가 있기 때문이다[9]. 이러한 격조사의 특성을 이용하여 질의문에서 개념적 중요도가 높은 명사를 중심으로 구문정보를 추출한다.

'보조사'는 체언을 일정한 격으로 규정하지 않고 주어, 목적어, 부사어 등 여러 격에 두루 쓰이면서, 특별한 의미를 더해 주는 조사이고, 접속조사는 두 단어를 같은 자격으로 이어 주는 구실을 하는 조사로 '-와/과'이외에 다수가 있지만 본 논문에서는 두 조사에 대해서만 고려를 하였으며 이러한 조사를 중심으로 질의문에서 체언이 지니고 있는 정보를 추출하기 위해, 질의문 코퍼스에서 자주 사용되는 고빈도 조사 정보를 얻었다. 질의문 코퍼스를 통해 추출된 조사의 종류는 약 40여 개 정도가 되었고, 그중 11개의 조사들이 전체 95% 정도를 차지하고 있다. 이러한 고빈도 조사와 체언과의 결합관

계를 이용하여 질의문의 구조를 추출하게 된다.

3.2 명사구로 인식할 구문

[표 3] 명사구로 인식할 구문

· **N1 [와(과)] N2** "-와/과" 는 'N1'과 'N2'를 같은 자격으로 이어 주는 역할 · **N1 [의] N2 J2** J2 종류: [가 /이 /을 /를 /에서 /의] 'N1 N2'는 조사 생략전과 의미상 같은 경우가 많기 때문에 복합 명사나, 명사가 명사를 수식하는 명사 수식 구성으로 인식한다.[3]

동등한 위치나 자격을 나타내 주는 조사에 의해 연결된 구문의 경우 해당 조사를 생략함으로써 구문을 간소화 시키고, 생략된 조사와 연관관계인 두 단어들에 대해서는 명사들의 나열로 인식하여 동일한 가중치를 적용한다.

관형격 조사 '-의'의 경우 N2의 위치에는 질문유형을 결정하는 의문사에 대한 보충단어들이 빈번히 출현하였는데 문제는 그러한 단어가 문서들에서 일반적으로 많이 쓰이는 단어들이기 때문에 그러한 경우를 보완하기 위해 J2의 종류를 제안한다.

3.3 전제조건을 의미하는 구문정보

"1989년에서의 노벨 평화상의 화폐의 가치는 얼마입니까?"란 질문 중 조사 '-에서'는 위치나 시점을 나타내 주는 부사격 조사이다. 질문에서의 핵심은 "노벨 평화상의 화폐의 가치"를 묻고 있지만 1989년이라는 시기를 제시함으로써 질문의 범위를 한정하고 있다. 이렇듯 질의문의 조건을 보다 구체적으로 명시하기 위해 쓰여지는 조사들이 존재하는데, 그런 특정 조사만을 따로 질의문의 전제조건을 추출하기 위한 구문정보로 구성하였다. 이는 질문에 대한 정답과 관련된 단락을 추출할 때 분명 사용자의 의도가 내포된 중심키워드를 기반으로 한 문서 검색이 이루어지지만 중심키워드로 추출된 단락들 중에서는 전제조건에 해당하는 키워드에 대한 고려가 분명 더 필요기 때문에 전제조건 키워드를 따로 추출한다.

[표 4] 전제조건의 구문정보

· N (으)로 / (으)로써	· N1 [(에)서 / 부터] N2 까지
· N (에)서	· 년/대/번
· N2 까지	

3.4 질의문의 구문정보를 이용한 키워드 추출

질의문에서 개념적 중요도가 높은 명사를 중심으로 키워들를 추출하기 위해서 체언과 조사를 중심으로 구문정보를 추출하였다. 구문을 추출하기 전에 "-와/과/의"조사에 대한 처리와 전제조건 키워드 추출 과정을 통하여 다양한 조사들의 결합관계를 일반화(간소화)시킨 후 질의문에 대한 구문정보를 추출하였다. [표3]에 추출된 구문정보는 2개 이상의 조사정보를 가지고 있는 질의문중 85%를 차지하고 있다. 이 구문의 위치정보를 이용하여 중심키워드와 일반 키워드를 추출되게 된다.

[표 5] 구문정보

구문정보	키워드 추출
N1 의 N2는	중심키워드: N1/ 일반키워드: N2
N1가 T N2는	중심키워드: N1/ 일반키워드:T N2
N1 가 N2을	중심키워드: N1/ 일반키워드: N2
N1을 T N2는	중심키워드: N1/ 일반키워드:T N2
N1가 N2을 V N3는	중심키워드: N1, N2/ 일반키워드:T N3
N1을 T N2의 N3은	중심키워드: N1/ 일반키워드:T N2, N3

T= [용언+관형형어미]

• 키워드 추출방법

- 조사의 개수가 1개인 경우.[그림 1 :: A부분]

조사의 격 정보를 이용하여 주격, 목적격, 그리고 관형격조사 '-의'에 해당 하는 조사들과 결합하여 나타나는 명사들에 한하여 중심키워드로 추출

- 조사의 개수가 2개인 경우.[그림 1:: B부분]

추출된 구문정보[표3]와 일치하는 경우는 구문정보의 위치에 따라 중심키워드와 일반키워드로 추출.

구문정보를 이용할 수 없을 경우는 A부분으로 분기된다.

- 조사의 개수가 3개 이상인 경우.[그림 1:: B부분]

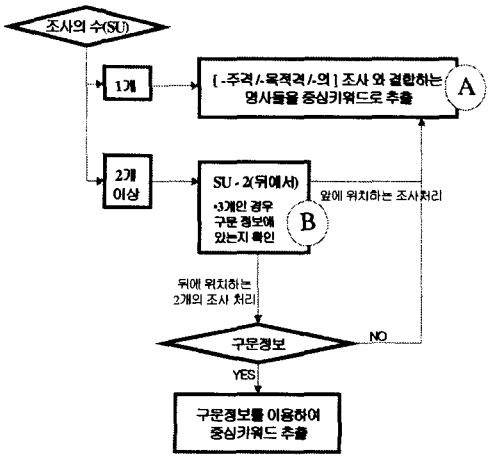
① 조사가 3개의 경우 표3 구문정보에 있는 [N1가 N2을 NP는], [N1을 NP의 N3은]의 구문과 일치

할 경우 구문정보의 위치에서 중심키워드를 추출

② 구문정보가 일치하지 않을 경우. [표4]에서 보여주는 예시의 방법으로 처리한다.

[표 6] 구문정보와 일치하지 않을 경우 처리 방법

질문	경복궁을 재건하기 위해 당백전 사용을 엄명하였으며, 쇄국정책을 실시한 사람은 누구인가?
의문사 제거	누구인가(제거)
조사 추출	을 을 을 은 → ① 을 을 / ② 을 은
② 조사 처리	• ②의 조사 → 구문구조를 이용한다. ②의 조사를 구문구조정보로 이용할 수 없을 경우 A로 분기
① 조사 처리	• ①의 조사는 주절인지, 보조절 인지 구분 [그림1: A]로 분기. 위 예문은 두 개의 서술어가 한 문장으로 연결되어 있다. 그렇기 때문에 ①에 해당하는 문장은 ② 를 부연 설명하는 보조절 이므로 주절보다는 낮은 중요도를 가진다.
중심키워드	쇄국정책, 쇄국, 정책
일반키워드	경복궁, 재건하기, 재건, 위해, 당백전....



▶▶ 그림 1 중심키워드 추출 구성도

4. 불린모델과 벡터모델의 문서검색

불린모델에서 검색식을 모든 키워드들의 OR연산 식으로 만 구성하였다. 그 이유로는 질의 응답시스템에서 사용되는 검색 시스템은 검색된 문서들 내에 질의와 관련된 문서가 얼마나 많이 분포하고 정확도가 얼마나 높은가 하는 것 보다는 검색된 결과 내에 정답이 포함되어 있는가 하는 것에 관점을 두었다[10]. 그래서 가급적 많은 양을 검색해 내어 정답이 그 안에 들어있을 확률을 높이기 위해서이다. 그 다음으로 검출된 단락들에 대해 키워드들과의 유사도 측정을 통하여 정답이 포함되었을 가능성이 큰 단락들 순으로 순위화 한다. 불린 모델은 질의나 문서의 키워드에 모두 이진(binary) 가중치를 할당하여 단락에 대한 상대적인 가중치 부여가 어려운 점을 보안하기 위해 벡터 공간 모델을 적용하였다. 질의문에서 추출한 키워드들에 대해서는 중심키워드에 대한 가중치 변수(α), 전제조건 키워드에 대한 가중치 변수(β), 일반키워드에 대한 가중치 변수(γ)의 조율에 의해 가중치에 대해 변화를 주었으며 변수크기는 α>β>γ순이다. 따라서 질의와 단락의 유사도(cosine coefficient similarity)[수식1]에 따라 랭킹을 줄 수 있다.

$$sim(p_i, q) = \frac{\vec{p_i} \cdot \vec{q}}{|\vec{p_i}| \, |\vec{q}|} \qquad [수식\ 152\ ]$$

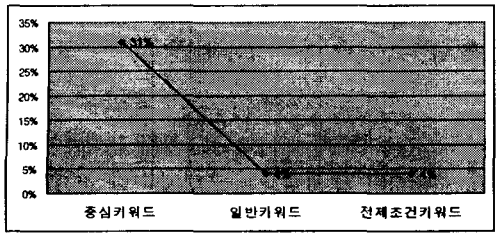
5. 실험 평가 및 결과

질의문에서 중심키워드를 얼마나 잘 추출했는지에 대한 평가 방법으로 정확률과 재현율을 사용하였다.

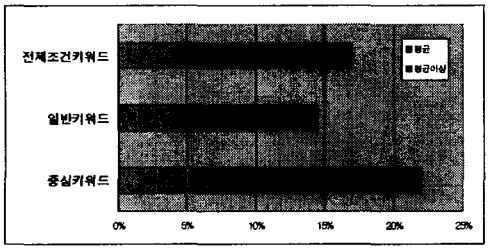
$$정확률 = \frac{키워드와\ 정답이\ 함께\ 포함된\ 단락}{정답이\ 포함된\ 단락}$$

$$재현율 = \frac{키워드와\ 정답이\ 함께\ 포함된\ 단락}{키워드가\ 포함된\ 단락}$$

[그림 2]에서 보면 중심키워드는 다른 키워드들보다 정답과 함께 출현하는 빈도가 8배 가까이 높은 것을 알 수 있다.



▶▶그림 2 정확률



▶▶ 그림 3 재현율

이것은 일반키워드로 추출된 키워드들의 경우는 여러 문서에서 자주 사용되는 일반적인 단어들이기 때문에 정답과 연관지어볼 때 중심키워드와 비교해 상대적으로 중요도가 낮게 나타난다. 이는 일반적으로 사용하는 통계적인 가중치 방법을 부여했을 경우 일시적으로 자주 사용되는 명사나 단순히 출현빈도만 높은 용어에 높은 가중치가 부여되는 오류가 발생할 수 있다. 따라서 질의문의 중심키워드가 일치하지 않더라도 일상적인 용어들이 일치할 경우에 임계치보다 높은 값이 계산되어 정답이 포함되어 있는 단락으로 추출될 경우가 발생될 가능성이 크다는 것을 말해준다. 실험 결과, 추출된 단락에 나타난 단어의 빈도로 산출한 평가이므로 단락의 수에 따라 정확률과 재현율값에 많은 변화가 있다. 실험에서는 가급적 많은 양을 검색해 내서 정답이 그 안에 들어 있을 확률을 높이기 위해 키워드들과 관련된 모든 단락들을 추출하였기 때문에 단락이 많이 추출이 되었는데 이는 키워드들의 가중치 변수값의 조율을 통한 유사도 측정으로 조정할 수 있다. 따라서 이 논문에서의 평가는 결과의 값보다는 키워드에 따른 상대적인 비교 평가로 이루어 져야 한다. 그 결과 구문정보를 이용한 중심키워드는 다른 일반키워드보다 정답과 근접한 거리에 있음을 나타내 주며 일반키워드 보다 정답을 추출할 가능성이 더 크다는 것을 보여준다.

본 논문에서는 자연언어 질의문의 핵심이 되는 키워드 추출을 위해 질의문의 구문정보를 이용하여 중심키워드와 일반키워드를 추출하는 기법을 제안하였다. 실험 결과 구문정보를 통해 추출된 중심 키워드는 일반키워드보다 정답과 함께 출현할 가능성이 8배나 높았으며 통계적 방법에 의한 키워드추출에서 발생할 수 있는 오류들에 대해 효과적으로 대처할 수 있는 모델이다. 향후 연구 과제로는 키워드들의 가중치 변수값의 변화에 따른 추출 단락 수에 대한 조절을 통하여 중심키워드에 대한 정확률과 재현율에 대한 실험과 정답관련 단락 추출에 대한 정확률과 재현율에 대한 실험이 필요하다.

▌참고문헌▐

[1] 장정선, 외 3, "질의생성 모델을 이용한 전자우편 질의응답 시스템", 제14회 한글 및 한국어 정보처리 학술 대회, pp176-183, 2002.

[2] C.L.A. Clarke, 외 3, "Question Answering by Passage Selection (MultiText Experiments for TREC-9)". TREC 2000

[3] Dan Moldovan, 외 6," LASSO: A Tool for Surfing the Answer Net", in Proceedings of the Text Retrieval Conference (TREC-8), November, 1999.

[4] Jun Suzuki, 외 2, 'SVM Answer Selection for Open-Domain Question Answering", 19th International Conference on Computational Linguistics (Coling-2002), Taipei, pp.974-980, 2002.

[5] 강승식, 외4. "자연언어 질의 문장의 용어 가중치 부여 기법". 제14회 한글 및 한국어 정보처리 학술 대회, pp.223-227, 2002.

[6] 강승식, "한국어 형태소분석과 정보검색", 홍릉과학 출판사, pp.184-185, 326-332.

[7] 김영택, 외 공저 "자연언어처리", 생능출판사, p366-375.

[8] 김지영, 외1, "한국어 정보검색에서의 복합명사 가중치 부여 방법 및 평가", 제13회 한글 및 한국어 정보처리 학술 대회, pp.157-162, 2001.

[9] 김기혁, "국어 문법 연구(형태*통어론)", 도서출판 박이정, pp.231-241, pp.507-521

[10] 이영신, 외2, "질의 응답 시스템을 위한 가변 길이 단락 검색", 제14회 한글 및 한국어 정보처리 학술 대회, pp.259-266, 2002.