(제 10 회 한글 및 한국어 정보처리 학술대회)

구문 분석에 기반한 자연어 질의로부터의 불리언 질의 생성*

박미화, 원형석, 이원일, 이근배
포항공대 전산과
자연언어 처리 연구실

Boolean Formulation of Korean Natural Language Queries Using Syntactic Analysis

Mihwa Park, Hyungsuk Won, Wonil Lee, Geunbae Lee
Natural Language Processing Laboratory
Dept. of Computer Science and Engineering, POSTECH

요 약

본 연구는 자연어 질의의 형태 및 구문 정보를 바탕으로 불리언 질의를 생성하는데 그 목적을 둔다. 일반적으로 대부분의 상용정보검색시스템은 입력형식을 검색성능이 좋은 불리언 형태로 하고 있으나, 일반 사용자는 자신이 원하는 정보를 불리언 형태로 표현하는데 익숙하지 않다. 그러므로 본 정보검색시스템은 자연어 질의를 기본 입력형태로 하여 사용자의 편의성을 높이고, 이 질의를 범주문법에 기반한 구문분석 결과에 의해 복합명사를 고려한 불리언 형태로 변환하여 검색을 수행함으로써 시스템의 검색성능의 향상을 도모하였다. 정보검색 실험용 데이터 모음인 KTSET2.0 으로 실험한 결과 본 논문에서 제안한 자연어 질의로부터 자동 생성된 불리언 질의의 검색성능이 KTSET2.0 에서 제공하는 수동으로 추출한 불리언 질의보다 8% 더 우수한 성능을 보였고, 기존 자연어질의 시스템이 수용해온 방법인 형태소 분석을 거쳐 불용어를 제거한 후 Vector 모델을 적용하여 검색을 수행한 경우보다는 23% 더 나은 성능을 보였다.

1. 서 론

이상적인 정보검색시스템은 각 문서와 질의의 내용을 완전하게 이해 하는 것이나 실제로 이것은 불가능하므로 대부분의 정보검색 시스템들은 문서들과 질의의 내용에 근접하는 어떤 구조화된 방법을 사용한다. 주로 문서들은 색인어 또는 키워드의 집합들로 표현되며, 질의는 색인어 또는 키워드들의 집합에 의해 표현 되어진다. 그리고 대부분의 상용 정보검색 시스템들은 불리언 식으로 표현되는 불리언 모델을 사용한다. 이것은 불리언 질의가 일반 사용자에게 사용하기가 어려울지 모르지만 검색 전문가들은 그들이 필요한 정보를 AND, OR, NOT 등의 연산자를 사용하여 정확하게 표현할 수 있기 때문에 좋은 검색결과를 얻을 수 있기 때문이다. 전례로 [3]의 실험결과, 불리언 모델이 가중치를 가진 색인어들의 집합으로 이루어진 질의보다 더 나은 검색성능을 나타내었다. 하지만 대부분의 일반사용자는 원하는 정보를 불리언 형태로 표현하는데 익숙하지 않다. 그러므로 정보검색시스템이 사용자의 편의성과 검색의 효율성

* 본 연구는 과기처 soft science(stepi)의 부분지원을 받은 것임.

을 동시에 만족하기 위해서는 일반사용자가 자연어로 질의를 입력했을 때 불리언 형태로 자동변환 후 검색을 수행하는 것이 타당할 것이다. 사용자가 자신이 원하는 정보를 자연어 질의의 형태로 표현할 때 그들은 특정한 언어상관관계를 선택함으로써 그들의 질의에서 단어들 사이의 다양한 의미적 연관성을 표현한다. 이러한 질의들을 단순히 가중치를 가진 색인어들의 집합으로 처리하는 것은 그들의 관계를 무시하는 것이다. 불리언 질의의 장점중의 하나는 자연어로 표현된 관계 구조 (Relational Structure)를 컴퓨터에 의해 쉽게 처리 되는 형태로 표현하는 것이다. 자연어질의로 된 입력문은 여러 단계에 걸쳐 분석되어질 수 있다. 일반적으로 형태소분석, 구문분석, 의미분석의 3 단계로 이루어진다. 자연어 질의로부터 키워드 및 연산자를 추출하고 키워드간의 수식관계 및 중요도를 도출하기 위해선 구문분석이 필수적이다. 또한 구문분석을 수행하게 되면 단어들 사이의 구문관계에 의한 복합명사 합성이 가능하다는 장점이 있다. 따라서 본 논문에서는 자연어 질의를 범주문법에 기반한 구문분석을 수행한 후 그 결과로 나오는 구문 트리에서 복합명사 합성까지 고려한 불리언 질의를 생성하고, 생성된 불리언 질의를 여러 가지 검색 모델에 적용하여 그 성능을 평가하는 실험을 수행하였다.

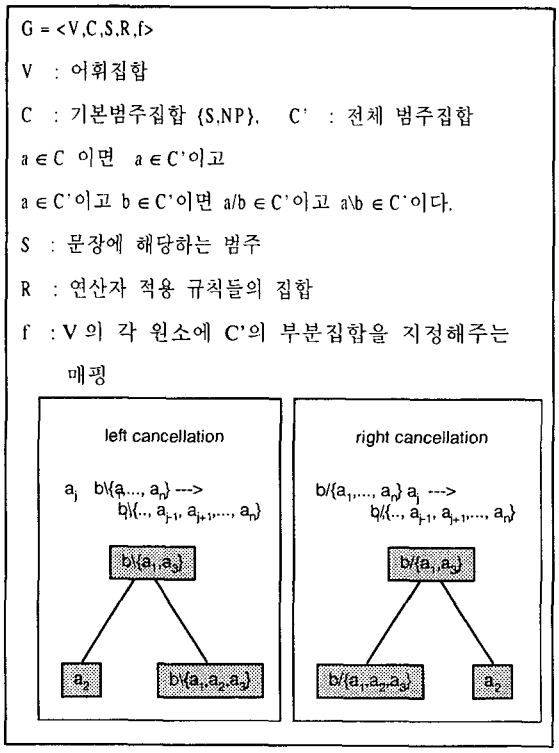
2. 관련 연구

구문분석에 의한 자연어 질의의 불리언 질의 자동생성에 관한 관련연구는 구문분석기의 성능문제로 국내에서는 구체적으로 실험한 사례가 없다. 국외의 연구 사례로는 Salton 과 Smith 의 연구가 있다. Salton [4]은 구문구조에 의한 불리언 생성이 아니라 질의에 나타난 키워드들의 빈도수에 기반한 방법론을 제안하였다. Salton 의 방법론은 원래 질의문의 의미를 잘 반영하지 않는 경향이 있고, 검색효율도 Smith 의 구문분석에 의한 방법론보다 떨어진다. Smith [5]는 자연어 질의를 구문분석기를 사용하여 불리언 형태로 변환하는 알고리즘을 제시하였다. 그녀는 구문적으로 생성된 불리언 연산자에 대해 P-norm 모델을 적용하여 3 개의 시험용 문서집합에서 수동 생성된 불리언 식 및 통계적으로 생성된 불리언 식과 비교하였다. Smith 의 불리언 식이 Salton 의 불리언 식보다 나은 성능을 보였고, 전문가에 의해서 수동으로 생성된 불리언 식과 비슷한 성능을 보였다.

3. 불리언 질의 생성 알고리즘

3.1 범주문법에 기반한 구문분석

범주 문법을 사용하여 한국어를 모델링하게 되면, 구구조 문법과는 달리 한국어가 가지는 후치사와 자유어순을 쉽게 처리할 수 있게 된다. 범주 문법은 구문규칙을 사용하지 않고, 사전에 정의된 범주에 따라 상향식으로 category cancellation 에 기초해 구문 분석이 이루어 지는 파싱 방법이다. 본 연구에서 사용하는 범주문법은 방향성(directional) 범주문법으로 [그림 1]같이 정의된다 [1].

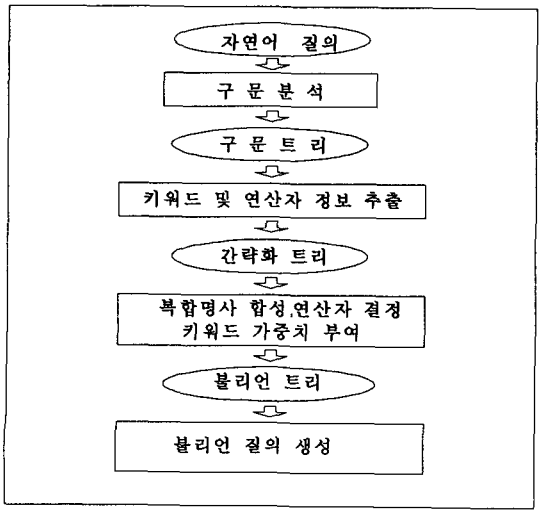


[그림 1] 범주문법의 정의

이렇게 정의된 범주문법은 한국어의 조사, 어미, 접속사등을 자연스럽게 표현할 수 있다. 한국어에 사용되는 기본범주는 "n", "np", "v", "vp", "s" 이다. "n"은 접미사와 결합되는 명사의 범주를, "np"는 그 이외의 명사의 범주를 나타낸다. "v"는 술부가 문장성분과 결합하기위한 범주이고 시제를 제외한 보조용언과 선어말어미들과 결합되는 범주를 나타낸다. "vp"는 시제가 결합된 술부를 나타낸다. 정보검색을 위한 구문분석기의 특징은 " ~에 관한 " 또는 " ~를 위한 " 등과 같이 질의어에 나타날 수 있는 관용적인 표현은 미리 구문사전에 등록하여 검색의 효율성을 높이도록 하였다. 질의어에 대한 구문분석 예는 [그림 2]와 같다.

```
예)멀티미디어 PC 나 DTP 를위한 프린터에대한 문서
                              /n---MC*<멀티미디어>
                          /np
                          /   \eng---sf<pc>
                      /np/np
                      /   \(np/np)\np---jS<나>
                  /np
                  /   \np---sf<DTP>
              /np/np
              /   \(np/np)\np---jO<를>:s<#>:DI 여<위하>:eCNMG<ㄴ>
          /np
          /   \np---MCF<프린터>
      /np/np
      /   \(np/np)\np---jO<에>:s<#>:DI 여<대하>:eCNMG<ㄴ>
/np
      \np---MCC<문서>
```

[그림 2] 구문분석 수행결과



[그림 3] 불리언 질의 생성 알고리즘

3.2 불리언 질의 생성 알고리즘

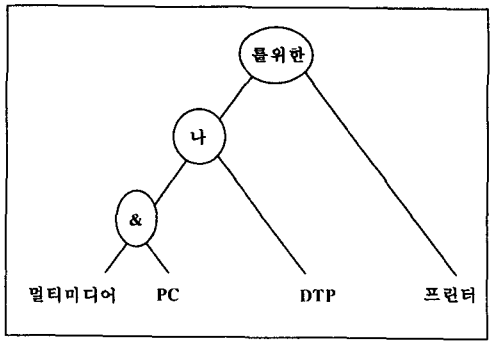
자연어 질의로부터의 불리언 질의 생성 알고리즘은 [그림 3]와 같다.

3.2.1 구문 트리의 간략화

구문분석 결과 트리를 후위 순회하여 단말노드인 경우 형태소 및 태거 정보를 검사하여 키워드와 연산자를 분류한 후, 키워드이면 단말노드에 그대로 두고 연산자이면 해당 노드의 상위 노드들 중에서 가장 처음으로 만나는 NP 범주 또는 V, VP 범주 노드에 형태소 및 태거정보를 보관한다. 그리고 해당노드와 그 노드의 부모노드를 삭제한다. 각각의 태거정보에 대한 키워드 및 연산자 분류기준은 [표 1]과 같고 구문트리를 간략화 한 예는 [그림 4]와 같다.

구 분	태 거 정 보
키워드	보통명사(MC), 고유명사(MP), 수사(S), 외국어(eng)
연산자	조사(j), 관형형 어미(eCNMG)
불용처리	의존명사(MD), 대명사(T), 관형사(G), 부사(B), 형용사(H)등

[표 1] 키워드 및 연산자 분류



[그림 4] 간략화된 구문트리

3.2.2 복합명사 합성 및 질의어 가중치

구문 트리에서 존재하는 단어와 단어들 사이의 수식관계를 이용하여 성분명사의 수가 2 또는 3 인 명사구를 생성할 수가 있다. [표 2]에서 알 수 있듯이 KTSET2.0 에서 가능한 복합명사의 길이에 대한 통계

복합명사길이	갯 수	비율
2	59,530	77.8%
3	14,547	19.0%
4 이상	2,421	3.2%
Total	76,498	100%

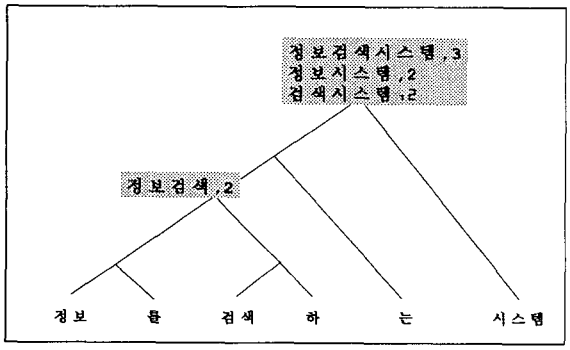
[표 2] KTSET2.0 에서의 복합명사 비율

치를 조사한 결과 길이가 3 까지인 복합명사가 전체의 96.8%를 차지하였다. 그러므로 길이가 4 이상인 복합명사의 생성은 검색에 도움을 주지 못하며, 오히려 효율을 떨어뜨릴 가능성이 크다. 길이가 2 인 복합명사의 생성규칙은 [2]의 Lexical rule 을 사용한다. 구체적인 Rule 은 [표 3]와 같다.

L1	조사가 생략된 명사/인접한 명사
L2	관형격 조사 결합명사 / 피수식 명사
L3	목적격,주격조사 결합명사 / 서술형명사
L4	관형화된 서술형 명사 / 피수식 내포문 명사
L5	조사상당 용언이 이끄는 관형화된 내포문의 명사 / 피수식 내포문의 명사

[표 3] Lexical Rule

길이가 3 인 복합명사의 합성은 구문트리의 순회과정에서 길이가 2 인 명사구의 생성규칙에 의해 다음과 같이 자동적으로 생성 가능하다. 복합명사 생성 알고리즘은 간략화된 구문트리를 후위 순회하여 단말 노드이면 해당 형태소 정보를 복합명사 후보로 등록하고 단말 노드가 아니면 그 노드의 형태소 및 태거 정보를 기반으로 양 자식 노드의 복합명사 후보가 합성이 가능한지 검사하여 합성된 명사구와 합성길이에 대한 정보를 해당 노드에 보관한다. 또한 상위 노드에서의 복합명사 합성을 위해 하위 노드의 복합명사 후보들도 가져와 보관한다.



[그림 5] 복합명사 합성 예

[그림 5]의 예에서 알 수 있듯이 합성길이가 2 인 명사구는 다시 합성길이 3 의 명사구로 합성이 되고 또한 하위노드에서 올라온 복합명사 후보들도 규칙에 의해 각각 합성된다.

3.2.3 질의어 가중치

문서에 대한 색인어의 가중치는 Fox 에 의해 제안된 방법을 적용하였다 [6]. 적용된 식은 다음과 같다.

$$\begin{cases} \left(0.5+0.5\dfrac{tf_{i,j}}{Maxtf_j}\right)\dfrac{\log(N/n)}{\log(N)} & tf_{i,j}>0 \\ 0 & tf_{i,j}=0 \end{cases}$$

Where

$tf_{i,j}$:문서 j에 나타난 색인어 i의 빈도수

$Maxtf_j$:문서 j에서 가장 많이 나타난 색인어의 빈도수

n :Term i가 나타난 문서수

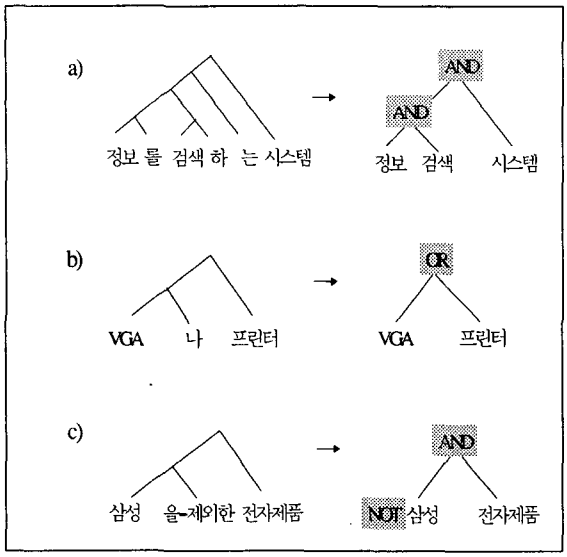
N :전체 문서수

질의어에 대한 가중치 부여 방법은 각 키워드의 기본적인 가중치는 Normalized IDF 값을 적용하였고 합성된 복합명사는 각각의 명사의 가중치의 합을 적용하여 해당 복합명사가 나타난 문서가 상위에 Rank 되도록 하였다. Normalized IDF 에 대한 식은 다음과 같다.

$$\frac{\log\left(\dfrac{N}{n}\right)}{\log(N)}$$

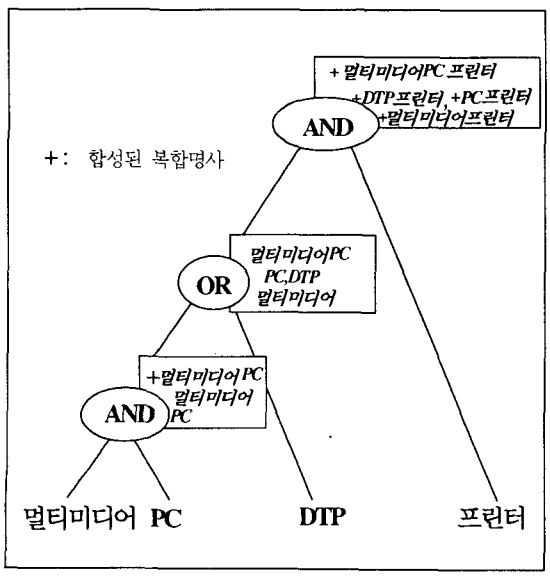
3.2.4 불리언 연산자 결정

연산자의 우선순위는 구문관계에 의해 결정이 되나 연산자 자체는 형태소 및 태거 정보에 의해 결정된다. 예를 들어 해당 노드의 형태소 및 태거 정보가 주격 또는 목적격 조사이거나 관형형 어미이거나 소유격 조사인 경우엔 AND 연산자로 결정되고, 조사 "~나" 또는 " , " 처럼 OR 의 의미가 강한 형태소는 OR 연산자로 결정된다. NOT 연산자는 "~를 제외한", "~이외의", "~를 포함하지 않는" 처럼 구문사전에 등록된 관용적인 표현들에 한하여 NOT 연산자로 결정된다. 연산자 결정에 대한 예는 [그림 6]와 같다.



[그림 6] 연산자 결정

[그림 6]의 a)는 목적격 조사와 관형형 어미가 AND 연산자로 결정되는 예를 보여 주며, b)는 접속조사 "나" 가 OR 연산자로 결정됨을 보여준다. c)는 NOT 연산자의 경우를 보여주는 예로서, "~을 제외한" 이라는 관용어구의 상위노드에는 AND 연산자를 두고 그 노드의 왼쪽 서브 트리의 루트 노드에 NOT 연산자를 표시한다.불리언 생성 알고리즘을 모두 수행한 후 생성된 불리언 트리는 [그림 7] 과 같다.



[그림 7] 불리언 Tree

3.2.5 불리언 질의 생성

불리언 질의트리를 전위 순회하면서 불리언 질의문을 생성한다. 이때 연산 우선순위를 위해 왼쪽 자식 노드를 순회하면 " ("를 생성하고 오른쪽 자식 노드를 순회하면 ") "를 생성한다. 합성된 복합명사는 AND 연산으로 해당위치에 표시한다. [그림 7]로부터 생성된 불리언 질의는 다음과 같다.

(((((멀티미디어 & PC) & 멀티미디어 PC) | DTP) & 프린터) & 멀티미디어 PC 프린터) & DTP 프린터)

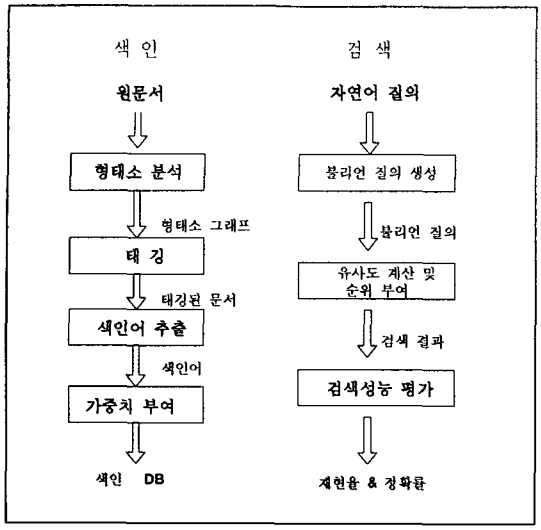
4. 실험 및 평가 방법

4.1 실험 방법

실험을 위한 대상문서는 한국통신에서 구축한 정보검색 실험용 데이터 모음인 KTSET2.0 을 이용하였다. KTSET2.0 은 약 4,414 건의 문서에 50 개의 질의문으로 구성되어 있고 질의어당 평균 적합 문헌 수는 29 문서이다. 특히 KTSET2.0 의 질의어는 자연어 질의와 불리언 질의를 동시에 제공하고 있어 이번 실험의 평가대상으로 적합하다고 판단되었다. 성능 평가 방법은 KTSET2.0 에서 제공하는 불리언 질의를 직접 입력하여 검색을 수행하였을 때의 성능과 자연어질의를 입력으로 받아 자동으로 불리언 질의를 생성하여 검색을 수행하였을 때의 검색성능을 비교 평가 하였다. 검색성능을 평가하는 척도는 일반적으로 사용하는 재현율과 정확률을 채택하였다.

4.2 시스템 구성

실험을 위해 구현한 정보검색시스템의 전체 구성은 [그림 8] 과 같다. 색인은 문서에서 필요한 정보를 추출하여 검색시에 사용토록 하는 과정이다. 입력 문서들의 구성 문장들을 입력으로 받아 형태소 분석 및 태깅 과정을 거치고 난 뒤 명사류들을 색인어로



[그림 8] 정보검색시스템 구성

추출한다. 이 과정에서 색인어로서 가치가 없는 단어들은 불용어처리 한다. 가중치가 부여된 색인어 들로 구성된 색인 DB 는 각 색인어에 대한 문서번호와 출현빈도(Term Frequency)를 가지게 된다. 색인의 결과로 만들어 지는 색인 DB 는 검색에 이용된다. 검색은 자연어 질의를 기본입력으로 하여 본 논문에서 제안한 방법으로 불리언 질의를 생성하여 검색을 수행한다. 검색모델은 P-Norm 모델로써 유사도 계산 방법은 다음과 같다.

$$Sim(D,q_{and}):1-\sqrt[p]{\frac{q_1^p(1-d_1)^p+q_2^p(1-d_2)^p}{q_1^p+q_2^p}}$$

$$Sim(D,q_{or}):\sqrt[p]{\frac{q_1^p d_1^p+q_2^p d_2^p}{q_1^p+q_2^p}}$$

$$Sim(D,q_{not}):(1-d_i)$$

Where

q_i : 질의어의 키워드 i 에 대한 가중치

d_i :문서에서 색인어 i의 가중치

$p : P-value$

4.3 실험결과 및 평가

KTSET2.0 의 50 개 질의어에 대해 구문분석을 수행한 결과 45 개 질의어에 대해 완전하게 구문분석이 완료되었고, 5 개 질의어는 구문분석이 되긴 되었으나 분석오류가 발생하여 올바른 결과가 나오지 않았다. KTSET2.0 에서 제공하는 수동 추출된 불리언 질의(MB)와 본 논문에서 제안한 자동 생성 불리언 질의(AB)에 대한 검색성능을 비교한 결과는 [표 4]와 같다. [표 4]에서 알 수 있듯이 본 논문에서 제안한 자동추출 불리언 질의(AB)의 검색성능은 전문가가 직접 추출한 수동추출 불리언질의(MB)의 검색성능보다 8.4% 더 우수한 결과를 보였고, 기존 자연어질의 시스템이 수용해온 방법인 형태소분석을 거쳐 불용어를 제거한 후 Vector 모델을 적용하여 검색을 한 경우(VEC)의 검색성능 보다 23.9% 높은 성능을 나타내었다. 각각의 질의어에 대해 자동추출 불리언이 더 나은 성능을 보이는 경우는 적합문서에 키워드가 분할되어 존재하는 경우와 합성된 복합명사를 포함하는 문서가 적합문서에 존재하는 경우로 5 개 질의문에 나타났다.

Rec.	MB	VEC	AB	MB vs AB	VEC vs AB
0	78.00	66.00	76.00	-2.56%	15.15%
10	81.45	70.93	76.57	-6.00%	7.94%
20	71.72	67.27	69.11	-3.64%	2.74%
30	62.21	55.23	65.38	5.08%	18.37%
40	52.18	46.17	58.76	12.62%	27.28%
50	45.70	40.79	56.76	24.21%	39.14%
60	36.74	28.33	46.76	27.28%	65.02%
70	28.45	21.76	36.44	28.10%	67.48%
80	19.62	16.06	28.08	43.10%	74.85%
90	10.96	12.05	13.48	23.00%	11.85%
100	6.56	7.39	7.64	16.41%	3.28%
Avg	44.87	39.27	48.63	8.38%	23.84%

[표 4] 11point 재현율에 대한 정확률

Top	MB		AB			
N	Rec.	Pre.	Rec.		Prec.	
1	6.78	78.00	6.36	(-6.19%)	76.00	(-2.56%)
5	25.46	67.50	23.65	(-7.12%)	66.25	(-1.85%)
10	33.88	55.85	35.37	(+4.41%)	57.91	(+3.68%)
15	37.74	44.71	40.92	(+8.45%)	48.55	(+8.59%)
20	43.52	38.94	46.90	(+7.77%)	43.21	(+10.95%)
25	49.24	36.12	52.00	(+5.60%)	39.49	(+9.32%)
30	52.73	33.23	56.64	(+7.42%)	36.24	(+9.05%)
Avg	35.62	50.62	37.41	(+5.01%)	52.52	(+3.75%)

[표 5] 상위 N 문서에 대한 정확률

예) NLQ: 음성의 인식 또는 생성에 관한 문서
MB: 음성인식 | 생성
AB: ((음성 & (인식 | 생성)) & 음성인식 & 음성생성)

수동추출 불리언 질의가 더 나은 성능을 보이는 경우는 구문분석 오류로 인해 자동추출 불리언 질의가 사용자 의도와 맞지않은 경우와 합성된 복합명사가 포함된 문서가 적합문서가 아니면서도 상위에 Rank 된 경우도 발생하여 모두 8 개 질의문에 이러한 현상이 나타났다.

예) NLQ: 멀티미디어를 위한 VGA나 프린터
MB: 멀티미디어 & (VGA | 프린터)
AB : ((멀티미디어 & VGA) | 프린터)

[표 5]는 자동추출 불리언과 수동추출 불리언에 대해 1, 5, 10, ..., 30 상위 등급 목록 지점들의 평균 정확률을 나타낸다. 여기서도 자동추출 불리언이 수동추출 불리언보다 높은 성능을 나타내었다. 이는 합성된 복합명사가 포함된 문서가 적절하게 상위에 Rank 되었기 때문이며, 결론적으로 구문분석이 성공적으로 수행될 때 자연어 질의로부터 자동 생성된 불리언 질의가 사용자의 의도를 잘 반영함을 보여준다.

5. 결 론

본 논문에서는 정보검색시스템의 자연어 인터페이스와 관련하여 자연어 문장으로 표현된 검색 질의문을 범주문법에 기반한 구문분석결과로부터 불리언 질의문을 생성하는 방법에 관한 문제를 다루었다. 이 과정에서 구문 트리로부터 복합명사를 합성하는 방법과 질의어의 가중치 부여 방법 및 연산자 및 연산 우선순위 결정 방법을 제안하였다. 자연어 질의로부터 불리언 질의를 생성함에 있어 구문분석 정보를 이용함으로써 첫째, 불리언질의의 연산순위가 자연스럽게 결정되고 이것은 결국 사용자의 의도를 충분히 반영한 검색결과를 얻는 데 큰 역할을 하게 되며, 둘째 구문 트리를 이용하여 복합명사를 합성함으로써 검색에 도움이 되지 않는 불필요한 복합명사의 합성을 방지할 수 있는 장점을 가진다. 본 논문에서 제안한 방법으로 생성된 불리언 질의를 KTSET2.0을 대상으로 검색성능을 평가한 결과 KTSET2.0에서 제공하는 전문가가 직접 추출한 불리언 질의문의 검색성능보다 우수했고, 또한 자연어 질의를 불리언 질의로 변환하지 않고 불용어만 제거해 Vector 모델을 적용한 검색결과보다는 훨씬 더 우수한 성능을 보였다. 앞으로 복합명사 합성시에 적절한 여과과정을 거쳐 검색에 불필요한 복합명사의 생성을 방지하는 방법론이 요구되며, 불용어도 현재의 Adhoc list보다는 좀 더 체계적인 방법론의 정립이 요구된다. 그리하여 보다 향상된 자연어 인터페이스를 제공함으로써 일반적인 정보검색에서 사용자의 편의성을 제공하면서 높은 검색성능을 지원하는 정보검색시스템을 구축하고자 한다.

감사의 글

본 연구를 수행하는데 도움을 주신 자연언어 처리 연구실의 모든 분들께 진심으로 감사 드립니다.

참고문헌

[1] 이원일, "확률범주문법에 기반한 음성 한국어 처리", 포항공대 박사학위논문, 1998.

[2] 이현아, "구문분석과 공기정보를 이용한 개념기반 명사구 색인방법", 포항공대 석사학위논문, 1996.

[3] Gerald Salton, Edward A. Fox and Harry Wu, "Extended Boolean Information Retrieval", CACM Nov , Vol. 26, p1022~1036, 1983.

[3] Gerald Salton, C.Buckley, E.A.Fox, " Automatic Query Formulation in Information Retrieval", Journal of the American Society For Information Science, Vol.34, p262~280, 1983.

[4] Smith, M.E., "Aspects of the P-NORM model of Information Retrieval : Syntactic Query Generation, Efficiency and Theoretical Properties", Ph.D. Thesis, CS, Cornell Univ., 1990.

[5] E.A. Fox, "Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types", Ph.D. Thesis, cs, Cornell Univ., 1983.