한국어정보검색에서 구문적 용어불일치 완화방안

윤보현, 김상범, 임해창
고려대학교 컴퓨터학과 자연어처리연구실
{ybh, bewise, rim}@nlp.korea.ac.kr

Alleviating Syntactic Term Mismatches in Korean Information Retrieval

Bo-Hyun, Yun, Sang-bum, Kim, Han, Hae-Chang, Rim
NLP Lab., Dept. of Computer Science and Engineering, Korea University

요약

한국어 정보검색에서 복합명사와 명사구로 발생하는 색인어와 질의어간의 구문적 용어 불일치는 많은 문제를 일으켜왔다. 본 논문에서는 복합명사 분해와 명사구 정규화를 함께 수행하여 유사도 측정값을 적당히 유지함으로써 재현율을 저하시키지 않고서 정확률을 향상시킬 수 있는 구문적 용어불일치 완화방안을 제시하고자 한다. 색인모듈에서는 통계정보를 이용하여 복합명사를 분해하고, 의존관계를 이용하여 명사구를 정규화한다. 분해되고 정규화된 키워드에 경계정보 '/'가 할당되고, 가중치가 계산된다. 검색모듈에서는 경계정보를 이용하여 부분일치를 고려하는 유사도 계산을 수행한다. KTSET 2.0으로 실험한 결과, 제안한 방법은 구문적 용어불일치를 완화할 수 있으며, 재현율을 저하시키지 않고서 정확률을 향상시킬 수 있음을 보인다.

1. 서론

한국어에서 색인어와 질의어간의 구문적 용어불일치는 검색성능을 향상하는데 있어서 심각한 장애요소로 작용하여 왔다. 이러한 구문적 용어불일치는 띄어쓰기가 자유로운 복합명사와 다양한 형태의 명사구로 인하여 발생한다. 먼저 복합명사는 그것을 구성하는 단일명사들을 띄어쓰기도 하고 붙여쓰기도 한다. 예를 들어, '정보 검색'과 '정보검색'은 같은 의미를 가지는 복합명사이다. 한편, 하나의 명사구는 다양한 형태로 존재한다. 예를 들어, 색인어구(phrasal term)[1]인 '정보/검색/'은 복합명사 '정보 검색', 수식어와 중심어 관계인 '정보의 검색', 주어와 동사의 관계인 '정보가 검색되다', 목적어와 서술어 관계인 '정보를 검색하다', 그리고 관형어형 동사가 존재하는 관계인 '정보에 대한 검색' 등으로 다양하게 문서에서 나타난다.

그러나 구문적 용어불일치 완화에 대한 기존의 연구는 복합명사 분해 또는 명사구 정규화중 어느 한가지만을 수행하여 왔다. 그러나 복합명사 분해만을 수행한 경우 표 1과 같은 문제가 발생한다. 표 1에서 보면, 색인어와 질의어가 서로 의미적으로 다르지만 복합명사 분해를 통해 같은 키워드 '시스템'과 '평가'가 추출된다. 이러한 현상은 분해된 단일명사가 복합명사 '평가시스템'과 '시스템평가'를 구분하지 못하기 때문에 유사도 계산과정에서 유사도 측정값을 불필요하게 증가시킨다. 따라서 재현율은 증가하지만 정확률은 감소한다.

표 1. 복합명사 분해의 예

	색인어	질의어
복합명사	평가시스템	시스템평가
분해된 복합명사	시스템, 평가	시스템, 평가

한편, 명사구의 정규화만을 수행할 경우 표 2와 같은 문제가 발생한다. 색인어와 질의어가 같은 단일명사 '시스템'과 '평가'를 가지고 있지만 정규화된 명사구에 의해서는 매치가 되지 않는다. 이러한 현상은 정규화된

1) 본 논문에서는 명사구를 정규화하여 색인어구를 만든다. 예를 들어, '정보의 검색'은 명사구이고, '정보/검색/'은 색인어구이다.

명사구의 특정성으로 인해 구문적 용어불일치 완화에 한계가 있다. 그리하여 유사도 측정값을 감소시키는 경향이 있어서 정확률은 증가하지만 재현율은 감소한다.

표 2. 명사구 정규화의 예

	색인어	질의어
명사구	시스템 평가	시스템을 평가하는 방법
	시스템의 평가	시스템의 평가 방법
정규화된 명사구	시스템평가	시스템평가방법

복합명사와 명사구의 이러한 특성은 복합명사와 명사구의 부분패턴을 처리할 수 있는 색인 및 검색 방법을 요구한다. 즉, 구문적으로 다른 복합명사나 명사구가 질의로 입력되어도 의미가 동일한 복합명사와 명사구를 포함한 문서들을 검색할 수 있어야 한다. 따라서 본 논문에서는 복합명사 분해와 명사구 정규화를 함께 수행하여 유사도 측정값을 적당히 조절함으로써 검색성능을 향상시키는 구문적 용어불일치 완화 방안을 제시하고자 한다.

2. 관련연구

구문적 용어 불일치를 완화하기 위한 기존의 연구는 복합명사와 명사구를 여러 방법으로 다루어 왔다. 복합명사 분해를 통하여 단일명사를 색인어로 사용하는 연구는 복합명사를 분해하기 위해 형태소분석기의 분석결과를 이용하는 방법[17], 형태소 분석기를 이용하지 않고 복합명사를 N-gram으로 분해하는 방법[2], 단일명사의 분포정보를 이용하여 분석하는 방법[3], 통계정보와 선호규칙을 이용하여 분해하는 방법[5], 명사의 연어정보와 서술성 명사의 공기정보를 이용한 복합명사를 분석하는 방법[10]으로 나뉘어 진다. 그러나 위와 같은 방법은 단일명사가 여러 개의 다른 복합명사의 형태를 구별해내지 못하기 때문에 재현율은 증가되나 정확률은 향상되기 어렵다.

한편, 한국어에 있어서 명사구 색인에 관한 기존의 연구는 부분구문분석(partial parsing)을 수행하는데 언어적 휴리스틱을 이용하는 방법[7]과 의존관계를 이용하는 방법[12]이 있다. 그러나 이러한 방법들은 어느 정도 명사구를 정규화함으로써 구문적 용어 불일치를 완화할 수 있으나, 정규화된 명사구의 특정성으로 인하여 유사도 측정값을 감소시켜 정확도는 향상되나 재현율은 저하되는 문제점이 있다.

기존 연구의 문제점인 복합명사와 명사구를 다루는데 무조건적으로 결합 또는 분해하는 방법에서 벗어나 복합명사 개개의 특성을 고려하는 방법[9]도 있다. 이 방법은 복합명사의 어휘적 정보를 단일명사들의 통계적 행태에 기반하여 자동으로 획득하고, 이러한 어휘적 정보를 검색에 적용하는 방법이다.

3. 시스템 구성도

본 논문에서 색인어와 질의어간에 형태상 불일치를 완화하기 위해 제안한 시스템 구성도는 그림 1에서 보이고 있다. 색인 모듈은 크게 품사태깅 및 미등록어 처리모듈, 단문분할모듈, 복합명사 분해모듈, 명사구 정규화 모듈, 색인어구 필터링모듈 그리고 가중치 계산모듈로 나뉘어 진다.

품사태깅 및 미등록어 처리모듈은 먼저 입력 문서를 본 연구실에서 개발된 형태소 분석기[13]를 통해 형태소 분석하고, 품사 태거[8,11,14]에 의해 품사 태깅한다.

본 논문에서 이용하는 품사 태거는 이중 은닉마르코프 모델(twoply hidden Markov model)에 기반한다. 이 품사태깅시스템은 미등록어가 포함된 어절 전체를 명사로 태깅할 수 있으므로 미등록어에 조사가 붙어있는 결과를 산출하기도 한다. 따라서 미등록어 처리모듈에서는 미등록어에 조사가 붙어있는 어절에서 최장조사를 제거한 다음 미등록어를 명사로 추출한다.

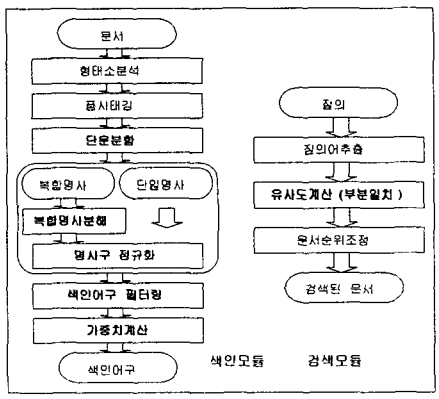

그림 1. 시스템구성도

단문분할모듈에서는 하나의 절내의 단어들이 다른 절에 있는 단어들보다 의미적으로 보다 밀접한 연관관계를 갖는다고 가정하고 복문을 단문으로 분할한다. 본 논문에서는 단문분할하기 위해 연결어미(예: '고')를 이용

하여 분할한다. 하나의 절은 하나의 서술어와 여러 개의 명사상당어구를 가질 수 있다. 이와 같이 단문분할하는 이유는 명사구 정규화 단계에서 의미있는 색인어구후보를 생성하기 위함이다.

복합명사 분해모듈은 품사 태깅된 문서에서 단일명사는 색인어로 추출하고, 복합명사는 통계 정보와 선호규칙에 의해 단일명사로 분해하여 단일명사 사이에 경계정보인 분리기호 '/'를 첨가하여 색인어구로 추출한다. 명사구 정규화모듈은 의존관계에 의해서 많은 색인어구 후보를 추출하고, 단일명사 사이에 경계정보인 분리 기호 '/'를 첨가하여 색인어구로 정규화한다. 색인어구 필터링모듈은 많은 색인어구 후보 중에서 상호정보(mutual information)와 명사 범주쌍의 상대빈도(relative frequency of noun category pair)를 이용하여 의미없는 색인어구를 제거한다. 마지막으로 가중치 계산모듈에서는 추출된 모든 색인어구에 대해 가중치가 계산되어 색인어 화일에 저장된다.

검색 모듈은 질의어 추출모듈, 유사도 계산모듈, 그리고 문서순위조정 모듈로 구성된다. 질의가 입력되면, 자동색인에서와 같은 기술을 사용하여 복합명사를 분해하고 명사구를 정규화하여 검색어구를 추출한다. 그리고 유사도 계산 모듈에서는 부분일치(partial matching)를 고려하여 색인어와 질의어간의 유사도를 계산한다. 또한 문서순위조정 모듈에서는 계산된 유사도에 근거하여 P-norm[1] 모델을 이용해서 질의에 적합한 순서로 문서를 정렬하여 출력한다.

4. 구문적 용어불일치 완화방안

4.1 복합명사의 분해

한국어 문서에서 복합명사를 단일명사로 올바르게 분해하는 것은 어려운 작업이다. 그 이유는 단일명사들의 다양한 조합으로 중의적 분해가 발생하며, 미등록어로 인해 분해가 되지 않는 복합명사가 많기 때문이다.

본 논문에서는 단문에서 존재하는 복합명사를 분해하기 위해서 본 연구실에서 개발된 복합명사 분해 시스템[15]을 이용한다. 분해 시스템은 통계 정보와 선호 규칙을 이용하여 한국어 복합명사를 단일명사로 분해한다. 통계 정보란 1음절 접사 빈도, 그리고 2음절 또는 3음절 단일명사가 복합명사 내에서 사용된 위치 정보와 빈도 정보를 이용하는 CNFP(Compound Noun Formation Probability)를 말한다. 선호 규칙이란 분해된 명사의 개수가 최소인 분해 패턴을 선호하는 규칙인 MNPR(Minimal Noun Preference Rule)이다. 또한 미등록어가 포함된 복합명사를 분해하기 위해 네가지의 휴리스틱을 사용하여 분해한다. 이렇게 분해된 복합명사에서 경계정보를 추가하기위해 분리기호 '/'를 첨가한다. 경계정보를 첨가하는 이유는 검색시 질의어와 부분일치시킬 때 단일명사들이 부분적으로 일치되는지 쉽게 판단하기 위해서이다.

4.2 명사구 정규화

한국어 명사구 정규화는 문장중에 같은 의미의 내용이 구문적으로 다르게 표현될 때 하나의 일관된 형태, 즉 색인어구로 바꾸는 작업을 의미한다. 다양한 형태의 명사구를 인식하기 위해서는 복잡한 구문 분석을 수행해야 한다. 그러나 완벽한 구문 분석을 수행하여 명사구를 인식하는 것은 현실적으로 어려운 일이다. 따라서 본 논문에서는 복잡한 구문 분석을 수행하지 않더라도 명사구를 인식할 수 있는 의존관계를 이용하여 정규화한다. 의존관계는 색인어구의 수식어와 중심어간의 구문관계를 나타낸다. 본 논문에서는 하나의 절내에 있는 명사들간에 교차(crossing branch)를 허용하여 보다 많은 색인어구를 추출한다. 본 논문에서 사용하는 의존관계는 표 3과 같다.

표 3. 의존 관계

Types	색인어구 수식어	색인어구 중심어
Type(1)	명사	뒤에 오는 명사
Type(2)	명사+소유격조사	뒤에 오는 명사
Type(3)	명사+주격조사 or 목적격조사 or 부사격조사 or 관형사격조사 or 대격조사 or 보조격조사	뒤의 동작성 명사 or 상태성 명사
Type(4)	명사+관형사형 어미	뒤에 오는 명사
Type(5)	명사+조사 관형어형동사	뒤에 오는 명사

표 3의 의존 관계를 이용하여 명사구를 인식하고 색인어구로 정규화하는 과정은 다음과 같다.

(1) 하나의 절내에 있는 단어에 색인번호를 할당한다.
(2) 절을 오른쪽에서 왼쪽으로 스캔해 가면서 만약 현재 단어가 명사상당어구이면, 이 명사를 색인어구의 중심어로 간주하고 단계(3)으로 간다. 그렇지 않으면 명사상당어구를

찾을 때까지 반복한다.
(3) 단계 (2)에서 발견된 중심어에서 왼쪽으로 절을 스캔한다. 만약 현재 단어가 명사상당어구이면, 이 명사를 색인어구의 수식어로 간주한다.
(4) 단계 (2)에서 발견된 중심어와 단계 (3)에서 수식어로 이루어진 색인어구가 의존관계에 속하는지 판단한다. 속하면 색인어구후보로 추출하고, 그렇지 않으면 의미없는 색인어구로 간주하여 제거한다.

4.3 색인어구 필터링

의존관계에 의해 정규화된 색인어구 후보 중에서 의미없는 색인어구를 제거하기 위해 본 논문에서는 상호정보와 명사의 범주쌍간의 상대빈도를 이용한다. 먼저 상호정보를 이용하여 수식어와 중심어간의 연관성이 높은 색인어구만을 추출한다. 상호정보는 단일명사들간의 연관 정도를 측정하는 어휘 연관율(lexical association)이다. 색인어구를 제거하기 위한 상호정보의 수식은 (1)과 같다.

$$MI(x,y) = \log \frac{P(x,y)}{P(x)*P(y)} = \log \frac{N\times f(x,y)}{f(x)\times f(y)} \quad (1)$$

여기서 x와 y는 단일명사를 의미하며, N은 총 발생빈도를 나타낸다. $f(x)$와 $f(y)$는 각각 x와 y가 발생할 빈도를 의미하고, $f(x,y)$는 x와 y가 동시에 발생할 빈도를 나타낸다. 또한 임계치를 0[2]로 정하고 $MI(x,y)$가 0 이상이면 색인어구로 받아들이고 그렇지 않으면 배제한다.

그러나 이 수식은 동시발생 빈도가 0이 되는 자료 부족 문제가 심각하게 발생한다. 따라서 이를 보완할 수 있는 평탄화(smoothing) 기법을 필요로 한다. 본 논문에서는 명사의 범주쌍간의 상대빈도를 175,468어절의 학습 코퍼스에서 조사하여 평탄화하는데 사용한다. 본 논문에서는 색인어구를 구성할 수 있는 품사를 다음과 같이 분류한다.

- 보통명사(NNCG)
- 동작성 보통명사(NNCV)
- 상태성 보통명사(NNCJ)
- 고유명사(NNP)
- 인칭대명사(NPP)

조사한 명사의 범주쌍간의 상대빈도는 표 4와 같다. 이러한 상대빈도는 단일명사가 수식어와 중심어 관계를 이루며, 한국어 색인어구를 이루는데 특징적인 형태가 있음을 보여준다.

표 4. 명사의 범주쌍 사이의 상대빈도

수식어 \ 중심어	NNCG	NNCV	NNCJ	NNP	NPP
NNCG	1949	865	42	0	0
NNCV	748	299	9	0	0
NNCJ	77	120	0	0	0
NNP	183	56	0	3	0
NPP	12	0	0	0	0

본 논문에서는 명사의 범주쌍의 빈도가 0이 아닌 품사가 색인어구를 구성할 수 있다고 가정한다. 어휘 연관률을 계산하는 상호정보의 수식이 0이 되는 경우, 표 4에서 나타난 명사의 범주쌍간의 상대빈도가 0이 아닌 명사의 범주쌍이면 이를 색인어구로 간주하고 그렇지 않으면 배제한다.

4.4 가중치 계산

일반적으로 사용하는 전통적인 가중치 계산식은 식 (2)와 같다. 식 (2)의 왼쪽 부분은 색인어빈도(term frequency)로서 한 문서에서의 색인어의 중요성을 의미하고, 오른쪽 부분은 역문서빈도(inverse document frequency)로서 전체 문서들 중에서 색인어의 분별력을 표현한다.

$$w = \log(tf+1.0)\times \log(N/n) \quad (2)$$

여기에서 tf 는 문서 D에서 색인어 t_i의 발생 회수이고, $\log(N/n)$는 색인어 t_i가 나타나는 문서 D의 개수(n)의 역이고, N은 전체 문서에서의 문서의 수를 의미한다.

그러나, 색인어빈도와 역문서빈도를 함께 사용하는 것은 몇 가지 이유로 문서집합의 특성에 따라 성능이 달라질 수 있다. 첫째, 고빈도 색인어나 문서길이를 보상하기 위해 색인어 빈도의 정규화가 필요하다. 그러나 아직까지 정규화의 실험은 뚜렷한 성능향상을 보여주지 못하고 있다. 둘째, 짧은 문서가 많은 문서집합에서는 색인어빈도가 큰 역할을 하지 못한다.

본 논문에서는 실험대상으로 문서길이가 비교적 짧은 논문초록으로 구성된 KTSET

2) 두 명사 x와 y가 밀접한 관계가 있으면, 상호정보 $MI(x,y)\gg 0$ 이다.

2.0[9]을 사용한다. 따라서 전통적인 가중치 계산식의 변형이 필요하다. 본 논문에서는 색인어빈도에 따라 다른 가중치 계산식을 적용하는 식 (3)을 이용한다. 색인어빈도에 따른 실험결과에 의하면, 색인어빈도가 1 이상인 경우에 역문서빈도만을 사용하고, 색인어빈도가 1 이하인 경우에 색인어빈도와 역문서빈도를 함께 사용하는 것이 가장 좋은 성능결과를 보였다.

$$\begin{aligned} &If\ tf > 1,\ w = \log(N/n) \\ &Otherwise,\ w = \log(tf+1.0) \times \log(N/n) \end{aligned} \qquad (3)$$

4.5 부분일치

앞의 절에서 설명하였듯이 문서와 질의어에 대해 복합명사를 분해하고 명사구를 정규화한 후 문서와 질의어의 유사도를 계산하는 과정에서 binary Tanimoto measure[4]에 의해 부분일치를 수행한다. binary Tanimoto measure는 개념간의 유사도를 거리로서 표현하는 것으로 값이 작으면 개념들이 유사하지 않다는 것을 의미하며, 값이 크면 개념들이 유사하다는 것을 의미한다.

부분일치는 색인어와 질의어가 완전히 일치하지 않더라도 복합명사를 이루는 단일명사가 일치하면 의미적으로 연관성이 있다는 가정에 기반한다. 부분일치는 색인어와 질의어의 가중치를 이용하고, 색인어와 질의어에서 일치하는 단일명사의 개수에 비례하여 유사도 값을 부여한다. 부분일치에 의해 가중치 w_i를 갖는 질의어 q_i와 색인어의 가중치 w_j를 가지는 문서 d_j 간에 유사도를 계산하는 식은 (4)과 같다.

$$Sim(q_i, d_j) = w_i \times w_j \times \alpha_{ij} \qquad (4)$$

위의 식에서 $\alpha_{ij}(0 < \alpha \leq 1) = \frac{|\text{색인어와 질의어에서 일치하는 단일명사}|}{|\text{색인어와 질의어에서 유일한 단일명사}|}$ 이다.

예를 들어, 질의어 q_1은 '정보/검색/'이고 가중치는 0.3525이다. 문서 d_1은 가중치가 0.2517인 색인어 '정보/검색/', 문서 d_2는 가중치가 0.4817인 색인어 '정보/검색/시스템/', 그리고 문서 d_3는 가중치가 0.6942인 색인어 '정보/시스템/'을 갖는다고 가정하자. 부분일치를 수행하는 예는 다음과 같다.

Sim(정보/검색/, 정보/검색/) = 0.3525×0.2517×2/2

Sim(정보/검색/, 정보/검색/시스템) = 0.3525×0.4817×2/3

Sim(정보/검색/, 정보/시스템/) = 0.3525×0.6942×1/3

위의 예에서 질의어 q_1인 '정보/검색/'에 대해 문서 d_1은 완전일치하는 색인어 '정보/검색/'을 가지고 있기 때문에 α_{11}는 2/2이다. 또한 문서 d_2와 질의어 q_1은 유일한 단일명사로 '정보/', '검색/', 그리고 '시스템/'을 가지며, 일치하는 단일명사로 '정보/' 그리고 '검색/'을 갖는다. 따라서 α_{12}는 2/3이다. 아울러 문서 d_3와 질의어 q_1은 유일한 단일명사로 '정보/', '검색/', 그리고 '시스템/'을 가지며, 일치하는 단일명사로 '정보/'만을 갖는다. 따라서 α_{13}은 1/3이다.

5. 실험 및 평가

본 논문에서는 평가자료로 한국어 정보검색 실험용 데이터 모음인 KTSET 2.0을 사용하고, 평가인자로서 재현율과 정확률을 이용하여 색인어구의 효율성과 제안한 방법의 성능을 평가한다.

실험 방법은 다음과 같이 네가지 방법을 수행하였다. 먼저 단일명사와 붙여쓴 복합명사를 추출하는 방법을 Baseline 방법으로 정한다. 두 번째 방법인 "분해"는 Baseline방법에 덧붙여서 복합명사를 분해하여 단일명사까지 색인어로 추출하는 방법이다. 세 번째 방법인 "분해+정규화"는 Baseline방법에 추가해서 복합명사를 분해하고 명사구를 정규화하여 단일명사와 복합명사를 추출하는 방법이다. 마지막으로 본 논문에서 제안한 방법은 Baseline방법에 덧붙여서 복합명사를 분해하고, 명사구를 정규화하여 색인어구를 추출하고, 변형된 가중치 계산식에 의해 가중치를 계산한다. 아울러 검색시에는 부분일치를 고려하여 유사도를 계산하는 방법이다.

그림 2에서는 네가지 실험 방법의 결과를 보여준다. Baseline방법보다 복합명사를 분해하여 단일명사를 추출하는 방법이 정확률을 평균 11.6% 향상시켰고, 복합명사를 분해와 명사구를 정규화하는 방법은 정확률을 평균 19.3% 향상시켰다. 마지막으로 제안하는 방법인 복합명사를 분해하고 명사구를 정규화하며 검색에서 부분일치를 시도한 방법은 Baseline 방법보다 평균 정확률을 26.5% 향상시켰다. 이 실험결과에서 복합명사와 명사구로 인한 색인어와 질의어의 구문적 불일치를 완화하고, 정확률을 향상시키는데 복합명사의 분해 및 명사구의 정규화 그리고 부분

표 5. 다양한 문서순위에 따른 정확률과 재현율

방법 / 문서순위	Baseline		분해		분해+정규화		제안한 방법	
	정확률	재현율	정확률	재현율	정확률	재현율	정확률	재현율
5	0.8583	0.8000	0.8066	0.8066	0.9433	0.9333	0.9600	0.9333
10	0.4345	0.4204	0.4061	0.4316	0.4856	0.4916	0.5169	0.4916
15	0.3039	0.3080	0.2739	0.3306	0.3430	0.3630	0.3796	0.3637
20	0.2667	0.2580	0.2078	0.2787	0.2832	0.3078	0.3203	0.3075
30	0.2095	0.2164	0.1439	0.2549	0.2327	0.2634	0.2665	0.2619
100	0.1534	0.1872	0.0617	0.2319	0.1852	0.2352	0.2124	0.2396
200	0.1041	0.1862	0.0457	0.2311	0.1782	0.2344	0.2038	0.2386
500	0.0821	0.1862	0.0413	0.2311	0.1768	0.2344	0.2023	0.2386

일치 기법이 유용함을 알 수 있다. 따라서 제안한 방법은 색인어와 질의어간의 구문적 용어불일치를 완화할 수 있다.

표 5에서는 검색된 문서의 다양한 순위에 따른 네가지 방법의 검색결과를 보여준다. 복합명사 분해만을 수행한 연구는 재현율을 증가시키나 정확률을 감소시킨다. 한편 복합명사 분해와 명사구 정규화를 수행한 방법은 Baseline방법보다 어느 정도 재현율과 정확률을 개선시킨다. 또한 제안한 방법은 Baseline방법에 비해 정확률을 상당히 증가시키고, 재현율을 감소시키지 않음을 알 수 있다. 이러한 결과는 제안한 방법이 재현율을 저하시키지 않고서 정확률을 증가시키는 방법임을 제시한다.

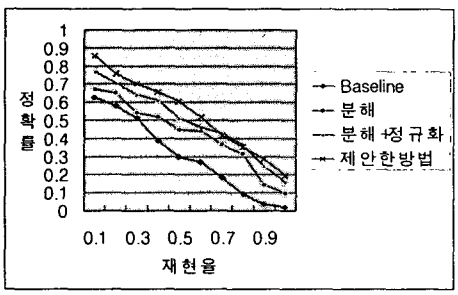

그림 2. 재현율에 따른 정확률

그림 3에서는 제안한 방법과 같은 자료 KTSET 2.0을 사용하는 [9]의 방법과 [16]의 방법을 비교하고, 서로 다른 실험 자료 KTSET 1.0을 사용하는 [2]의 방법과 [3]의 방법을 비교한 결과를 보여준다. 다른 실험 자료를 사용한 방법도 있으므로 객관적으로 비교하기 어렵지만 그림에서 알 수 있듯이 제안한 방법이 가장 나은 정확도를 보이고 있다.

6. 결론

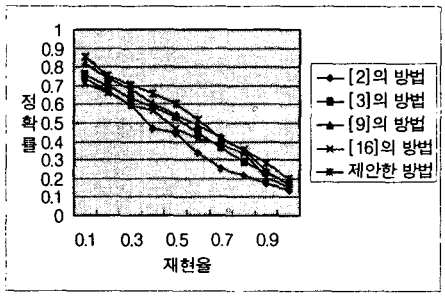

그림 3. 다른 연구와의 비교

본 논문에서는 구문적 용어불일치를 완화함으로써 검색의 성능을 향상시키는 모델을 제안했다. 구문적 용어불일치를 완화하기 위해 통계정보와 선호규칙을 이용하는 복합명사 분해와 의존관계를 이용하는 명사구를 정규화한다. 정규화된 명사구에서 상호정보와 명사의 범주쌍간의 상대빈도를 이용하여 의미없는 복합 명사를 제거하고, 경계정보인 '/'를 단일명사 사이에 추가하여 색인어구를 추출한다. 또한 문서집합 KTSET 2.0에 대해 적합한 가중계산식을 적용하였다. 아울러 문서와 질의어의 유사도 계산과정에서 부분일치를 수행하는 방안을 제시하였다.

제안한 방법이 구문적 용어불일치를 완화함으로써 재현율을 저하시키지 않고 정확률을 향상시킬 수 있고, 색인어와 질의어의 형태상의 차이를 충분히 극복할 수 있음을 실험을 통해 보였다. 향후에는 철자오류나 외래어로 인해 발생하는 어휘적 용어불일치와 동의어 및 다의어에 의해 발생하는 의미적 용어불일치의 완화에 관한 연구를 수행하고자 한다.

참고문헌

[1] Frakes, W. B., Baeza-Yates, R., Information Retrieval: Data Structures & Algorithm, Prentice-Hall, 1992.

[2] Lee, J. H., Shin, J. H., Ahn, J. S., "An

Effective Indexing Method for Korean Text Retrieval," IROL-96, pp. 79-84, 1996.
[3] Park, H. R., Han, Y. S., Lee, K. H., "A Probabilistic Approach to Compound Noun Indexing in Korean Texts," Proc. of the 16th International Conference on Computational Linguistics (COLING-96), pp.514- 518.
[4] Tanimoto, T. T., "An elementary mathematical theory of classification", IBM Technical Report, 1958.
[5] Yun, B. H., Cho, M. J., Rim, H. C., "Korean Compound Noun Indexing based on Lexical Association and Conceptual Association", IRAL-97, pp. 31-42, 1997.
[6] Yun, B. H., Cho, M. J., Rim, H. C., "A Korean Information Retrieval Model Alleviating Syntactic Term Mismatches", NLPRS-97, pp. 107-112, 1997.
[7] 김판구, "한국어 정보 검색을 위한 상호 정보량에 기반한 복합어 자동색인", 서울대학교 박사학위논문, 1994.
[8] 김진동, "어절 문맥을 고려하는 형태소 단위의 한국어 품사태깅 모델", 고려대학교 석사학위 논문, 1996.
[9] 박영찬, 최기선, "통계적 명사패턴 분류를 이용한 복합명사 검색모델", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp. 21-31, 1996.
[10] 양성현, 정의석, 윤준태, 송만석, "명사의 연어정보와 서술성 명사의 공기정보를 활용한 복합명사 분석 및 자동색인", 제9회 한글 및 한국어 정보처리 학술발표논문집, pp. 59-64, 199.
[11] 이상주, "은닉 마르코프 모델을 이용한 두 단계 한국어 품사 태깅", 고려대학교 석사학위 논문, 1994.
[12] 이현아, 이종혁, 이근배, "구문분석과 공기정보를 이용한 개념기반 명사구 색인 방법", 제 7 회 한글 및 한국어 정보처리 학술발표논문집, pp. 3-7, 1995.
[13] 임희석, "어절의 중의성 유형 분류에 근거한 한국어 형태소 분석기", 고려대학교 석사학위 논문, 1993.
[14] 임희석, 김진동, 임해창, "변형 규칙 기반 한국어 품사 태거의 개선", 제 8 회 한글 및 한국어 정보처리 학술발표논문집, pp.216-221, 1996.
[15] 윤보현, 조민정, 임해창, '통계정보와 선호규칙을 이용한 한국어 복합명사의 분해", 정보과학회논문지, 24권, 8호, pp. 900-909, 1997.
[16] 조민정, 윤보현, 임해창, "단어 형성 원리에 기반한 한국어 정보 검색 시스템", 고려대학교 석사학위논문, 1997.