효율적인 색인을 위한 복합 명사의 분해

박수준, 이현아, 장명길, 박재득, 박동인
시스템공학연구소 자연어정보처리연구부

Breaking Compound Nouns for Better Indexing

Soojun Park, Hyun-A Lee, Myung-Gil Jang, Jae Deuk Park, Dong-In Park
Dept. of Natural Language Information Processing, SERI

요약

정보검색에서의 효율적인 복합명사의 분석은 정확도와 재현율의 향상을 통해 색인의 질을 높여준다. 복합명사의 분석은 많은 노력이 요구되는 작업이다. 본 논문은 간단한 분해규칙을 이용하여 복합명사의 의미해석을 대신하였다. 실험을 위해 동아일보 사설을 대상으로 복합명사를 추출하고 이를 도출된 분해규칙을 이용하여 분해하였다. 실험을 통해 평균 96.2%의 분해 성공률을 보였다.

1. 서론

정보 검색의 두 가지 중요 기능은 색인(indexing)과 검색(retrieval)이다. 색인은 문서의 내용을 표현하는 것이고 검색은 색인된 질의와 문서의 비교이다. 정보 검색에서 기본적인 가정은 문서의 겉으로 드러난 표현만을 가지고는 그 내용을 충분히 표현하기 어렵다는 것이다[Gay, 1990]. 이를 해결하기 위해 문서의 내용을 대표하고 의미적으로 모호하지 않은 표현 방법이 필요하다. 중의성은 자연언어 처리 전반에 걸친 문제로 복합명사를 단위 명사로 분해하는데도 나타난다. 복합명사의 분석은 정보 검색 시스템의 성능에 영향을 미치며 효율적인 복합 명사의 분석은 다음과 같은 잇점을 제공한다. 첫째, 복합 명사가 갖는 의미를 명확히 하여 정확도(precision)를 향상 시키고 복합 명사의 적절한 분해로 같은 의미의 여러 표현을 갖는 색인어로 재현율(recall)을 높일 수 있다. 본 연구는 자동 색인에서 복합 명사의 분해를 통한 검색 성능의 향상을 목표로 하고 있다. 이를 위해 복합 명사 분해 규칙을 도출하고 말뭉치를 이용한 통계적 기법을 적용한다. 본 연구의 대상 코퍼스는 신문(동아일보)사설이다. 코퍼스를 이용하여 복합명사를 추출하고 이에 복합명사 분해 규칙을 적용한다. 실험 결과를 토대로 분해 규칙의 타당성을 알아본다.

2. 본론

2.1 복합명사 분석의 필요성

색인 시스템의 효율을 결정하는 두 가지 요인은 재현율 (recall rate)과 정확률(precision rate)로 정의될 수 있다. 넓은 의미의 색인어는 많은 수의 문서가 검색되게 하지만 검색된 문서 중에는 부적합한 문서가 포함될 확률이 높다 반면 의미가 좁은 색인어는 상대적으로 적은 수의 문서가 검색되게 하지만 적합한 문서가 검색되지 않을 수 있다[이현아, 1996]. 한편 색인어의 가장 중요한 특징 중 하나는 색인어가 갖는 변별력이다. 색인어가 갖는 변별력이 크면 클수록 그 색인어로 인해 불확실성이 줄어들며 더 좋은 색인어가 된다. 일반적으로 단일어는 복합어에 비해 변별력이 떨어진다, 다시 말해 광의의 뜻을 갖는다고 할 수 있다. 예를 들어 '정보'와 '검색'이라는 색인어가 있을 때 각각의 색인어가 갖는 의미보다는 '정보검색'이 갖는 색인어가 더 좁은 의미를 갖는다. 그리고 두 개 이상의 명사가 결합하여 복합명사를 이룰 때 전혀 새로운 의미의 단어를 생성하기도 한다[김민정,1992].

복합 명사는 두개 이상의 명사가 결합하여 하나의 의미를 가지는 단어 또는 구를 이루는 것을 말한다. 복합 명사의 분석은 자연언어처리에 있어서 또 하나의 어려운 문제이다. 복합 명사를 모두 사전에 등록하는 것은 실용적이지 못한 방법이다. 왜냐하면 기존의 명사 조합으로 우리는 새로운 복합명사를 만들 수 있기 때문이다. 복합 명사의 분석을 위해서는 먼저 복합명사를 확인하는 작업이 필요하며 경계가 확인된 명사에 대해 의미적으로 분석을 하는 작업이 필요하다. 본 논문은 형태소 해석기를 통해 경계가 확인된 복합명사에 대해 의미적인 분석만을 한다. 의미적으로 올바르게 복합명사를 분해하는 일은 World Knowledge를 필요로 할 정도로 어려운 일이다. 따라서 색인에서의 이러한 처리는 큰 overhead라 할 수 있다. 정보검색에서의 복합명사 분석은 색인어의 철저성과 특정성을 높일 수 있다. 이 경우 복합어 전체의 뜻 뿐만 아니라 이를 분해하여 부분적인 개념을 표현하도록 하는 것이 필요하다. 결국 복합 명사를 분해하여 또 다른 복합 명사로 표현하는 것은 내포된 복합어의 개념이 색인어가 될 수 있어 색인의 질을 향상 시킬 수 있다. 따라서 세 개 이상의 명사로 구성된 복합어는 여러 가지 개념을 포함하고 있기 때문에 올바르게 복합어를 분해해서 여러 개념을 나타낼 수 있도록 해야 한다[김판구,1994]. 예를 들어 '정보검색시스템'이란 색인어를 '정보검색'과 '검색시스템'으로 나눌 수 있다. 이렇듯 복합명사에 내포된 의미도 색인어로 추출함으로써 문서 색인어와 질의어 색인어의 불일치를 줄이고 어휘의 특정성을 유지하여 결국 색인의 질을 높일 수 있다.

이상에서 언급한 복합명사의 효율적 처리와 실시간에 이용할 수 있는 방법으로 분해규칙을 이용한다. 완벽한 구문 분석과 의미해석을 이용한 복합명사의 처리를 통한 색인어 추출은 실제 정보검색 시스템의 속도를 늦춘다. 본 논문이 제시하는 분해 규칙의 이용은 완벽한 구문분석과 의미관계를 이용한 색인어 추출에서 오는 속도 저하를 간단한 분해 규칙으로 복합명사에 대한 구문분석과 의미해석을 대신할 수 있는 잇점이 있다.

2.2 복합명사의 분석에 대한 기존의 연구

복합 명사는 ①하나의 어절 안에 묶여서 나타나는 경우와 ②두 개 이상의 어절 사이에 걸쳐 연달아 나타나는 경우, ③조사나 동사를 사이에 두고 두개의 명사가 복합 명사의 의미를 가지는 경우가 있다[채영숙,1996]. 이 중에서 하나의 어절 안에 나타나는 복합 명사는 하나의 단어로 취급할 수 있

으나 복합 명사의 의미를 분석하는 일은 쉽지 않다. 색인 과정에서 두개 이상의 어절에서 연달아 나타나는 경우는 띄어쓰기를 기준으로 복합 명사를 분해, 분석할 수 있다. 또, 조사나 동사를 사이에 두고 두개의 명사가 복합 명사를 이루는 경우 구문 분석 단계에서 이들에 대한 분석이 가능 하다.

복합 명사의 분석에 대한 기존의 연구는 통계적 방법을 이용한 복합어 구성 및 분해 규칙의 연구[김판구,1994]가 있으나 획일적인 패턴의 적용이라는 단점이 있다. 또 다른 연구로, 의미 정보를 이용하여 복합명사를 이루는 명사들의 의미적 연관성을 조사, 복합 명사의 해석 및 합성에 관한 연구가 있다[최기선,1993]. 외국의 경우 [Kobayasi,1994]는 collocational information 과 thesaurus 를 이용하여 복합어 분해를 시도하기도 하였다. 복합명사 분석의 어려운 점은 복합 명사 분해 시 발생할 수 있는 중의성이다. 중의성에는 구조적 중의성과 의미적 중의성이 있는데 특히 의미적 중의성은 해결이 어렵다. 이를 해결하기 위해 시소러스, Knowledge Base 등을 이용하고있다. 정보 검색 시스템에서 복합 명사의 적절한 분해로 같은 의미를 갖는 여러 표현을 통해 재현율 향상에 목표를 두고 있으므로 의미적인 중의성 해결은 중요한 의미를 갖는다. 본 연구에서는 하나의 어절 안에 나타나는 복합 명사의 의미 분석을 통하여 사용자의 질의어에 대한 재현율과 정확율을 높일 수 있는 방법을 모색 하고자 한다.

2.3 복합 명사의 분해

한국어에서는 명사들이 연속적으로 결합을 하여 복합명사를 이루는 경우가 많으며 복합명사의 분석을 위해 이진 트리 구조를 가정할 수 있다[정원수, 1994]. 복합명사는 트리 구조로 표현될 수 있고 트리의 오른쪽은 중심어로 복합명사 전체의 개념을 나타내고 왼쪽의 단어는 중심어를 수식하는 구조를 이룬다. 다시 말해 두 개의 단위 명사로 이루어진 복합 명사는 수식명사(Modifier)와 핵심명사(Modifiee/Head)의 구조를 갖는다. 일반적으로 복합명사에는 여러 개념이 포함될 수 있다. 이런 복합명사의 분해는 복합명사 내의 정확한 수식어와 피수식어와의 관계를 파악하는 것이 필요하다. 하지만 복합명사를 구성하는 명사들의 의미에 따라 수식 관계가 달라지므로 일정한 분해 규칙을 찾아내기가 힘들다.

2.4 복합 명사 분해 규칙

본 연구에서는 복잡한 의미 정보를 이용하는 복합 명사의 분해 보다는 간단한 의미 정보를 이용하여 복합 명사를 분해하는 방법을 제시한다. 대부분의 복합 명사의 유형은 수식명사(Modifier)와 핵심명사(Head)의 형태로 분리 될 수 있다. 이 때, 핵심 명사는 복합 명사의 의미적 형태적 중심이 된다. 결국 핵심명사(Head)를 찾아내고 이를 수식하는 수식명사(Modifier)와의 관계를 밝히는 일이 중요하다.

복합 명사의 결합 관계를 크게 둘로 나누면 '수식 명사' + '핵심 명사'의 형태로 나타나는 유속합성명사와 서로 대등한 관계로 나타나는 병렬 합성 명사가 있다. 복합명사의 대부분은 유속합성명사의 형태를 띄고 이 경우 형태적, 의미적으로 중요한 핵심 명사는 복합 명사의 의미적 분석에 중요한 역할을 한다. 이 때 핵심 명사에 대한 결합 의미 관계는 핵심명사의 성질에 따라 다양하게 나타난다

[김지영, 1992]. 핵심 명사의 성질을 분류해 보면 크게 '-하다', '-되다' 등과 결합하여 동작 및 상태를 나타내는 경우(서술형 명사)와 그렇지 않은 경우로 나눌 수 있다. 본 연구에서는 복합 명사에서 나타날 수 있는 서술형 명사를 중심으로 복합 명사를 분해하는 방법을 제시하고 실험을 통해 분해 방법의 타당성을 검토한다.

한국어의 경우에도 명사의 종류를 앞에서 언급한 것과 같이 크게 두 가지로 분류할 수 있다. 본 연구에서 나눈 핵심 명사의 유형은 다음과 같다. 명사의 종류를 서술형과 비서술형으로 나누어 서술형(-하다/-되다) 명사를 복합 명사의 핵심 명사로 정한다. 복합어 내에서 서술형 명사는 대부분의 경우 동작의 주체 등 핵심적인 역할을 하는 경우가 많다. 이를 근거로 서술형 명사를 중심으로 복합 명사를 분해한다.

2.5 실험

(1) 복합 명사 추출

실험을 위해 신문 사설[동아일보 사설선집 CD-ROM] 약 5년 치의 분량을 수집, 이용하였다. 신문사설 코퍼스의 전체 크기는 약 8.6MB 정도이며 각 사설 평균 323 단어로 총 3,343 개의 사설을 대상으로 하였다. 코퍼스인 신문 사설을 대상으로 포항 공대에서 개발한 형태소 해석기를 이용 형태소 해석을 거쳐 3 단어 이상의 복합명사를 가려내고, 마지막으로 3 단어로 이루어진 복합명사를 추출 복합명사 코퍼스를 만든다. 3 단어 복합명사 코퍼스는 총 4,787 개로 약 0.17MB 정도의 크기를 갖는다. 3 단어 복합명사는 연속된 명사로 $Noun_1$-$Noun_2$-$Noun_3$ 의 형태를 갖는다. 본 연구에서 다루고자 하는 복합명사는 색인기의 처리 결과로 나타나는 연속되는 세단어 이상의 복합명사를 대상으로 한다.

(2) 분해 규칙

복합 명사를 구성하는 명사의 의미적 관계를 고려하여 복합 명사를 분해하는 방법으로 서술형 명사를 찾아내어 서술형 명사를 중심으로 복합 명사를 분해한다.

분해 규칙의 기본은 수식명사 + 핵심명사의 유속 복합 명사에서 서술형 명사가 수식 명사냐 핵심 명사냐를 가려 분해를 하고 그 다음으로는 지배소 후위의 원리에 의해 앞에 오는 명사가 뒤의 명사를 수식하는 관계로 분해를 한다. 한국어의 수식 구조에서는 이 때 바로 앞에 오는 명사가 바로 뒤에 오는 명사를 수식할 확률이 높다. 결국 서술형 명사를 중심으로 분해를 하고 서술형 명사가 없을 때는 지배소 후위의 원리를 적용 분해를 한다. 각 복합명사의 형태별 분해 규칙은 다음과 같다.

- $N_1N_2N_3$ 형 : 서술형 명사가 복합 명사 내에 없는 형태로 규칙의 적용이 어렵다 이 경우 N_1N_2 와 N_2N_3 로 분해한다.
  예) 공공-요금-체계 : [공공-요금], [요금-체계]
- $N_1N_2P_1$ 형 : 먼저 서술형 명사 P_1 이 중심어가 되어 N_2P_1 로 분해하고 N_1N_2 로 묶는다.
  예) 경제-위기-타개 : [경제-위기], [위기-타개]

- $N_1P_1N_2$ 형 : 서술형 명사 P_1 을 중심으로 N_1P_1 과 P_1N_2 로 분해한다.
  예) 정보-검색-시스템 : [정보-검색], [검색-시스템]
- $N_1P_1P_2$ 형 : 서술형 명사 P_1 을 중심으로 N_1P_1 과 P_1P_2 로 분해한다.
  예) 수해-복구-지원 : [수해-복구], [복구-지원]
- $P_1N_1N_2$ 형 : 서술형 명사 P_1 을 중심으로 P_1N_1 과 N_1N_2 로 분해한다.
  예) 무장-공비-만행 : [무장-공비], [공비-만행]
- $P_1N_1P_2$ 형 : 서술형 명사 P_1 을 중심으로 P_1N_1 으로 분해하고 서술형 명사 P_2 를 중심으로 N_1P_2 로 분해한다.
  예) 정치-자금-마련 : [정치-자금], [자금-마련]
- $P_1P_2N_1$ 형 : 서술형 명사 P_2 를 중심으로 P_1P_2 와 P_2N_1 으로 분해한다.
  예) 시위-진압-방법 : [시위-진압], [진압-방법]
- $P_1P_2P_3$ 형 : 서술형 명사 P_2 를 중심으로 P_1P_2 와 P_2P_3 로 분해한다.
  예) 탈법-선거-운동 : [탈법-선거], [선거-운동]

(3) 결과 측정

규칙을 적용하여 분해한 결과 적합한 것과 적합하지 않은 것의 비율을 조사하고 본 연구에서 제시한 방법론의 타당성을 확인하였다. 먼저 분해된 결과에 대해서 수작업을 통해 오분석 및 명사가 아닌 경우 등을 제거하였다. 각 형태별 분해 규칙에 따른 복합명사의 분해가 올바른지를 수작업을 통해 확인하였다.

3. 결과

복합명사 타입	복합명사 수	올바른 분해	올바르지 않은 분해	분해 성공률
$N_1N_2N_3$ 형	824	713	111	86.5%
$N_1N_2P_1$ 형	748	719	29	96.1%
$N_1P_1N_2$ 형	1310	1273	37	97.2%
$N_1P_1P_2$ 형	752	746	6	99.2%
$P_1N_1N_2$ 형	195	183	12	93.9%
$P_1N_1P_2$ 형	274	272	2	99.3%
$P_1P_2N_1$ 형	468	457	11	97.7%
$P_1P_2P_3$ 형	208	208	0	100%

위의 결과에서 보듯이 복합명사 평균 분해 성공률이 96.2%를 나타내었다. $N_1N_2N_3$ 형을 제외한 나머지의 경우는 97.6%의 평균 분해 성공률을 보였다.

4. 결론

본 논문은 한국어 복합명사 분석 시 서술형 명사를 중심으로 복합명사를 분해하는 방법을 제안하였다. 실험을 통해 간단한 의미 정보인 명사의 서술형/비서술형을 이용하여 복합 명사를 분해한 결과 96.2%의 분해 성공률을 보였다. 본 실험의 목적은 복합명사를 단위명사로 분해함으로써 색인의 효율을 높이는데 있다. 복합명사의 분석은 의미 해석의 일부로 여겨질 수 있으나 그 자체로도 중요한 의미를 가진다. 복합명사의 분석은 구문 해석이나 형태소 해석과 복합 명사의 인식을 필요로 하는 정보 검색에 사용될 수 있다. 정보 검색에서는 문서 내의 단어에만 관심이 있으므로 시간이 많이 걸리고 overhead가 큰 구문분석에 의존하기 보다는 간단한 방법으로 복합명사의 의미해석을 대신하였다.

정보 검색 시스템에서는 복합 명사의 적절한 분해로 같은 의미를 갖는 부분적인 표현을 통해 재현율을 향상시킬 수 있다. 이를 위해 명사의 의미 정보와 이들간의 결합 관계를 고려해야 한다. 복합명사를 의미적으로 분석하기 위해 필요한 것으로는 첫째 복합 명사 내의 명사들 간의 관계 설정과 둘째, 복합어 구문 구조 파악, 마지막으로 복합어 내의 명사들의 의미적 중의성을 해결해야 한다. 이상의 세가지 Issue는 서로 독립적이지 않고 연관되어있다. 한편 복합명사의 결합 관계를 설정하기 위해서는 구문 정보 만으로는 부족하고 world knowledge가 필요하며 올바른 구문 구조 파악을 위해서는 완벽한 구문 분석이 필요하며 의미적인 중의성 해결을 위해서는 복합명사와 그것의 개념에 대한 의미적 개념 연결(mapping)이 필요하다. 이렇듯 의미 정보와 이들간의 결합 관계를 고려한 복합 명사의 분석은 많은 노력이 요구된다고 할 수 있다.

본 논문은 많은 노력이 요구되는 복합명사의 의미적 분석을 간단한 규칙을 통해 정보검색에 이용해 보았다. 비록 완벽하지는 않지만 실시간 이용과 overhead 감소라는 의미를 갖는다. 향후 과제로는 4단어 이상의 복합 명사에 대한 분해 및 규칙에 적용되지 않는 복합명사를 대상으로 유형 분류 및 각 단어의 특성에 따른 휴리스틱 유추가 있다.

참고문헌

[1] L.S. Gay and W.B. Croft, "Interpreting Nominal Compounds for Information Retrieval," Information Processing & Management Vol. 26, No.1, pp.21-38, 1990

[2] 이현아, "구문분석과 공기정보를 이용한 개념 기반 명사구 색인 방법," 포항공과대학교 대학원 석사논문, 1996

[3] 김민정, 권혁철, "한국어 특성을 이용한 자동 색인 기법," 한국정보과학회 가을 학술발표논문집, Vol.20, No.1, pp809-812, 1992

[4] 채영숙, "신문에 나타나는 한국어 복합명사의 결합구조 분석," 한글 및 한국어정보처리 학술발표논문집, 1996

[5] 김판구, 조유근, "상호 정보에 기반한 한국어 텍스트의 복합어 자동 색인," 한국정보과학회 논문지, Vol.21, No.7, pp1333-1340, July 1994

[6] 최기선 et. al, "한국어에서의 복합명사 인식에 대한 연구," 한국전자통신연구소 최종 보고서, 1993

[7] K. Yosiyuki, T. Takenobu, T. Hozumi, "Analysis of Japanese Compound Nouns using Collocation Information," Proc. Of the 14th Conference on Computational Linguistics (COLING-94), pp.865-869, 1994

[8] 정원수, "국어의 단어 형성론," 한신문화사, 1994

[9] 김지영, 권혁철, "지식 베이스에 의한 합성명사의 의미 관계 분석 시스템," 제 6 회 한국인지과학회 춘계 학술발표논문집, pp113-126, 1992