구문형태소를 이용한 색인어 추출

황이규, 이근용, 김남수, 이용석

전북대학교 컴퓨터과학과 언어정보공학실

{yghwang, keylee, nskim}@cs.chonbuk.ac.kr, yslee@moak.chonbuk.ac.kr

Index Extraction Using Syntactic Morpheme

Y. G. Hwang, K. Y. Lee, N. S. Kim, Y. S Lee

Dept. of Computer Science, Chonbuk National University

요 약

문서를 대표하는 단어를 추출하는 색인어 추출은 정보검색 시스템의 질을 좌우한다. 대부분의 색인어 추출 시스템은 명사를 추출하고 있으며, 가능한 모든 명사를 추출하고 있다. 이러한 방법은 불필요한 단어가 그 문장을 대표하는 색인어로 추출될 가능성이 높으며, 이는 정보 검색 시스템의 효율을 저하시킨다. 이를 해결하기 위해 품사 태깅이나 구문 해석 단계 등을 통해 불필요한 후보를 제거할 수 있지만, 태거를 구축하거나 구문 해석을 위해서는 많은 비용과 시간이 필요하다. 본 논문에서는 구문 형태소 단위의 형태소 해석에 기반한 색인어 추출 방법을 제안한다. 구문 형태소는 통사적/의미적으로 강한 공기 관계를 가지면서 문장에서 하나의 통사적 단위나 자질의 단위로 표현되기 때문에 구문 형태소내에 포함된 단어열들은 대두분 색인어가 될 수 없다. 이러한 방법을 이용하여, 형태소 해석 결과를 이용한 색인어 추출에서 발생하는 색인 오류를 제거함으로써 색인기의 성능을 높이는 방법을 제안한다.

1. 서론

색인이란 어떤 문헌에 대해 그 문헌의 전체적 내용을 나타내거나, 그 문헌을 다른 문서들로부터 구별할 수 있도록 그 문서의 선택 단서가 되는 단어 또는 단어구 등을 추출하는 것을 말한다[1]. 즉, 각 문헌을 변별할 수 있는 대표어구를 각 문헌에 부여하는 것을 의미하며, 동시에 검색시 사용되는 유용한 어구를 추출하는 것을 목적으로 한다. 색인 방법은 크게 자동 색인과 수동 색인으로 나눌 수 있는데, 자동 색인 방법에는 크게 단어의 빈도를 계산하여 출현빈도가 많은 순으로 색인어를 정하는 통계적인 방법과 형태소 해석, 구문 해석, 의미 해석 등의 다양한 기법을 이용하는 언어학적인 기법으로 나눌 수 있다.

이중 언어학적 기법중의 하나인 의미 해석을 이용한 기법은 가장 정확한 색인어 추출이 되는 반면 현실적으로 각종 사전의 구성과 문장의 완전한 이해가 불가능하므로 구현의 어려움이 있다. 또한 구문 해석을 이용한 기법은 구단위의 색인어와 보다 정확한 색인어가 추출된다는 장점이 있는 반면, 분석 결과의 모호성과 구문 해석기의 구현이 복잡하다는 어려움이 있다. 색인어 추출에 가장 많이 이용되는 형태소 해석을 이용한 기법

은 한국어에 적용이 쉽고 구현이 간단하다는 장점이 있다[2].

그러나 이러한 방법은 한 단어가 문장에서 어떠한 품사를 가지는지 정확하게 파악하지 않고 단순한 사전 지식만을 가지고 색인어를 추출하기 때문에 불필요한 색인어를 포함하는 경우가 많다. 한국어는 특히 첨가어 특성을 가지고 있으며, 많은 단어가 두 개 이상의 명사로 분해될 수 있으며, 다품사 모호성과 어휘 모호성을 가지고 있다. 이러한 문제로 인해, 순수하게 형태소 해석만을 이용하여 색인어를 추출하면 색인 오류를 포함하게 된다.

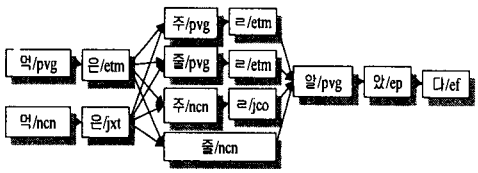
우리는 이 논문에서 색인어 추출에서 자주 발생하는 문제점을 살펴보고, 이런 문제중 몇가지를 특별한 지식없이 해결하여 불필요한 색인어를 배제하는 방법으로 구문 형태소 단위의 형태소 해석[4]을 통한 색인 방법을 제안한다.

구문 형태소는 통사적/의미적으로 강한 공기관계를 가지면서 문장에서 하나의 통사적 단위나 자질의 단위로 표현된다. 따라서 이러한 구문 형태소 내에 포함된 단어열들은 비록 형태소 분석에서 품사가 명사로 분석되더라도 색인어가 될 수 없다. 이러한 방법을 이용하여, 형태소 분석시에 불필요하게 색인어로 추정되는 명사들을 제거함으로써 명사 색인기의 성능을 높이는 방법을 제안한다.

2. 구문형태소를 이용한 색인

2.1 형태소 해석 기반의 색인

형태소 해석을 통한 색인어 추출은 한 어절에서 기능 형태소를 제외한 실질 형태소가 명사임이 밝혀지면 이를 하나의 색인어로 가정하는 방법이다. 한국어는 기능 형태소인 어미나 조사가 매우 발달한 언어이기 때문에 다양한 어미와 조사를 가지고 있다. 따라서 한 어절을 가능한 모든 방법을 통해 분리하면 다양한 형태소 분석 결과가 발생한다. 예를 들어, "먹은 줄 알았다"란 문장을 형태소 분석한 후, 명사만을 추출하는 과정을 살펴보자.



[그림 1] "먹은 줄 알았다"의 형태소 분석 결과

형태소 해석 후에 태깅 단계를 거치면 [그림 1]의 결과에서 추출될 수 있는 명사는 없다는 것을 쉽게 알 수 있지만, 형태소 해석 후에 명사를 추출한다면, "먹/ncn", "주/ncn", "줄/ncn"이 색인어로 선택될 수 있다.

이렇게, 색인어를 자동으로 추출할 때 나타날 수 있는 오류들을 분류해 보면 다음과 같다[3].

1) 품사 중의성을 가지는 단어
   예) '가장'

2) 명사와 조사, 명사와 접미사, 또는 용언과 어미로 분석될 가능성이 있으면서 하나의 명사로도 분석되는 단어
   예) '벨기에', '오페라는', '보고서'

3) 띄어쓰기의 오류로 인해 미지어로 인식되는 단어
   예) '이에앞서', '먹을수'

4) 용언과 어미로 분석되면서 단일 명사로도 분석되는 단어

예) '대해', '하고', '위해', '대한'

5) 복합명사에서 발생하는 오류

예) '거주지도', '색인이란', '공직자가', '문서내의'

이중 구문 분석을 통해서 해결할 수 있거나, 태깅을 통해서 해결할 수 있는 경우도 있지만, 3)과 4)에 의한 색인 오류는 구문 형태소에 기반한 형태소 해석을 통해 해결할 수 있다.

2.2 구문 형태소

구문형태소란 여러 기능형태소들이 결합하여 하나의 구문/의미적 단위를 형성하는 형태소 열을 말한다[4]. 한국어에서 구문형태소로 정의할 만한 기능 형태소의 결합으로 크게 두 가지가 있다. 기능형태소들이 결합하여 하나의 양상 자질을 나타내는 경우와 기능형태소와 용언이 결합하여 하나의 심층격 조사 역할을 하는 형태소열이 그것이다. 양상이란 어떤 사건이나 행동, 상태에 대한 화자의 태도를 표현한다. 예를 들면 겸양, 추측, 피동, 소망, 가능, 부정, 진행, 시도, 완료 등이다. 이 양상은 구문 해석 단계에서 구구조 규칙에 의해 용언에 대한 부가 자질의 형태로 표현될 수 있고, 의미 해석 단계에서 여러 지식을 이용하여 인식된 구구조를 하나의 의미 자질로 나타낼 수 있다. 또한 심층격 조사 상당 형태소 열은 구구조 관점에서 볼 때, 주어진 문장의 정확한 문법적 구조를 파악하는데 도움이 되지만 의미 해석을 어렵게 한다. 이런 양상 자질이나 조사 상당 어구를 구문 형태소 단위로 구문 해석 전에 인식하면 형태소의 모호성 축소와 구문 해석의 과정을 단순화시킬 수 있다.

또한 이들은 분리될 수 없는 하나의 의미적 단위를 형성하기 때문에 표층적으로 나타나는 용언이나 명사를 독립적인 단어로 파악하지 않아도 된다. 따라서 이러한 구문 형태소를 포함하는 어절내에 나타나는 체언들은 문장을 대표하는 색인어가 되기에는 적합하지 않다. 이러한 지식을 바탕으로 문장을 색인할 경우, 불필요한 색인어 후보가 많이 배제될 수 있다.

아래는 구문 형태소의 구조와 예를 보인 것이다.[1]

Type 1 : <v> {<ecx> [jxc] <px>}+ {<e>}+

예) "먹/pvg+고/ecx 싶/px+다/ef"

"먹/pvg+어/ecx+는/jxc 보/px+다/ef"

"먹/pvg+어/ecx 보/px+고/ecx+도/jxc 싶/px+다/ef"

(-아/-어) 지다, (-게) 되다, (-게) 하다, (-게) 만들다, (-아/-어) 가다, (-아/-어) 오다, (-고) 있다, (-고) 계시다, (-아/-어) 내다, (-아/-어) 버리다, (-고) 나다, (-고) 말다, (-아/-어) 주다, (-아/-어) 드리다, (-아/-어) 두다, (-아/-어) 놓다, (-아/-어) 가지다, (-아/-어) 대다, (-지) 말다, (-지) 못하다, (-지) 않다, (-아/-어) 보다, (-아/-어) 보이다, (-어야) 하다, (-고) 싶다, (-아/-어) 있다, (-아/-어) 계시다, (-는가/-ㄴ가/-나) 보다, ...

[표 1] type 1형 구문형태소

Type 2 : <v1> <etm> {<nbn>}+ [<j>]* <v> {<e>}+

예) "먹/pvg+을/etm 수/nbn 있/paa+다/ef"

"먹/pvg+을/etm 리/nbn+도/jxc 없/paa+다/ef"

Type 2' : <v1> <etm> <nbn> <jp> {<e>}+

예) "먹/pvg+을/etm 터/nbn 이/jp+다/ef"

"먹/pvg+을/etm 모양/nbn+이/jp+다/ef"

(-ㄴ/-는) 경우가 많다/있다/흔하다, (-ㄴ/-는) 바(가/도) 있다/없다, (-ㄴ/-는) 바에 따르다, (-ㄴ/-는) 셈이다, (-ㄴ/-는) 수가 많다/있다, (-ㄴ/-는) 적(도/은/이) 있다/없다, (-ㄴ/-는) 줄(도/은) 몰랐다/알았다, (-ㄴ/-은) 편이다, (-ㄹ/-을) 리(가/는) 없다, (-ㄹ/-을) 만(은) 하다, (-ㄹ/-을) 모양이다, (-ㄹ/-을) 바(는/를) 모릅니다/없습니다, (-ㄹ/-을) 뿐(만) 아니다, (-ㄹ/-을) 뿐이다, (-ㄹ/-을) 줄 모르다/알다, (-ㄹ/-을) 지 모르다/알다, (-ㄹ/-을) 지경에 이르다, (-ㄹ/-을) 지경이다, (-ㄹ/-을) 터이다, ...

[표 2] type 2, 2'형 구문형태소

1) 여기에서 사용되는 품사태그는 모두 국어정보베이스[5] 품사태그를 따랐다.

Type 3 : <n> <j> <v> <ecs>

예) "컴퓨터/ncn+에/jca 대하/pvg+어/ecs"

Type 3' : <n> <j> <n> <j>

예) "한국/ncn+과/jcj 마찬가지/ncn+로/jca"

(-에) 달해, (-에) 따라, (-에) 비해, (-에) 의해, (-에) 처해, (-에) 한해, (-에) 반해, (-와) 같이, (-와) 견주어, (-와) 관련하여, (-와) 달리, (-와) 함께, (-와) 더불어, (-를) 비롯해, (-를) 통해, (-를) 향해, (-를) 두고, (-를) 맞아, (-을) 가지고 (-로) 말미암아, (-로) 미루어, (-와) 마찬가지로, (-와) 반대로, (-와) 벌도로, (-에) 있어,

[표 3] type 3, 3'형 구문형태소

3. 구문 형태소를 이용한 색인 오류 제거

2.2절에서 설명한 구문형태소를 이용할 경우, 불필요한 형태소 모호성이 해소되기 때문에 불필요한 색인어의 축소에도 많은 도움이 된다. 이것을 크게 두가지로 분류 할 수 있다.

첫째, 구문 형태소는 의미적 최장일치를 바탕으로 하기 때문에, 문서에서 발생하는 띄어쓰기 오류에 대해 정확한 분석을 할 수 있다. 따라서, 2.1절의 3) 유형의 색인 오류에 의해 발생하는 문제를 조기에 제거할 수 있다.

둘째, 구문 형태소는 서로 관련있으며 공기하는 형태소열을 하나의 구문적 요소 또는 자질정보로 간주하기 때문에 이러한 형태소열이 비록 형태소 분석 후보로 명사를 포함하더라도 이를 하나의 단위로 간주한다.

예를 들어, "사랑할 경우가 많다"와 같은 문장에서 "경우"는 "-할 경우가 많다"라는 구문 형태소에 포함된 하나의 단위로 인식되기 때문에 그 단어 자체의 의미를 가지지 않는다. 또한 "-ㄹ" 다음에 나타나는 "경우"는 반드시 앞에서와 같은 방법으로 해석된다. 따라서 "경우"는 명사 색인어가 될 수 없다.

이러한 종류의 명사가 문장에서 많이 발생하는데, type2형 구조에 속하는 구문 형태소 중 "경우", "바", "바다", "셈", "수가", "수", "수도", "적", "적도", "줄", "편", "만", "모양", "지경", "터"등이 이에 해당된다. 또한 type3형에 포함된 명사들 중에 "대해", "인해", "대한", "반대", "별도", "의해", "달리", "반해", "마찬가지" 등과 같은 명사는 type3 환경에서는 단순히 의사 조사의 역할을 수행하고 있기 때문에 명사로 볼 수 없는 것이다. 이러한 구문 형태소가 실제 문장에서의 출현 빈도와 이에 따르는 색인 오류를 감소시킬 수 있다.

4. 실험 및 평가

실험을 위하여 한글 테스트 컬렉션(HANTEC) ver 2.0[6]에서 과학 분야 문서 3000건을 대상으로 기존의 색인 방법인 형태소 해석에서의 명사를 추출하여 색인하는 방법과 구문형태소를 이용하여 명사를 추출하여 색인하는 방법을 비교하였다. 문서의 특징을 살펴 보면 한 문서당 약 1040여 글자에 295단어를 포함하고 있는 일본 공업신문 기사를 한국어로 번역한 것이다.

하래는 순수 형태소 분석만을 이용한 색인어 추출과 구문 형태소 단위를 이용한 색인어 추출의 결과를 보여주고 있다. 실험 결과, 약 8%의 색인어 감소를 볼 수 있었다.[2)]

2) 모든 색인어를 빈도수 누적 없이 추출하였으며, 품사 수준의 불용어는 제거하였지만 단어 수준의 불용어는 고려하지 않았다.

	형태소 분석	구문 형태소 단위 분석
색인어 수	564,462	519,382

[표 4] 구문형태소 단위를 이용한 색인 결과

전체 913,000여 어절 중, 구문 형태소와 관련이 있는 어절은 약 49,000여 어절이며, 이중 type1과 type2에 속하는 어절이 약 36500여 어절, type3형이 약 12500여 어절이었다.

type1과 type2형 어절중 명사를 포함하는 예는 "확보할 수 있" , "채용할 수가 있", "주목될 것 같다"와 같은 어절에 나타나는 한음절이나 두음절이 대부분이다. 이들 중 대부분은 불용어 리스트를 이용해 처리할 수도 있으나, 그럴 경우, 불용어가 아님에도 기계적으로 불용어로 간주되는 단어도 많이 존재함을 알 수 있었다.

"수"와 같은 색인어는 형태소 분석을 이용한 색인어 추출에서 약 6300여 번 출현하였는데, 이중 대부분은 구문 형태소 단위의 분석에서는 나타나지 않는 경우였다.

또한, type3형 어절중 명사를 포함하는 예로는 "가부에 대한", "예방을 위해", "개에 대해", "원인에 대해서는", "신회사와는 별도로"등과 같은 어절이 나타나고 있다.

이렇게 구문 형태소 분석 방법에 의해 제거되는 색인어들은 기존 방법에서는 불용어 리스트를 통해 해결해 왔다. 그러나, 이들 단어중 상당수가 실제로 문장에서 의미 있는 역할을 수행하는 경우도 존재하기 때문에 무조건적으로 불용어로 처리해서는 안된다. 구문형태소 단위에서는 주위 형태소열들과의 관계를 고려하기 때문에, 몇몇 색인 오류는 쉽게 제거할 수 있었다.

5. 결론 및 향후 연구과제

우리는 본 논문에서 구문 형태소 단위를 이용한 색인어 추출기에 대해 기술하였다. 구문 형태소 단위의 형태소 해석을 이용할 경우, 실제 문장에서 자주 나타나는 색인 오류를 제거할 수 있다. 대부분의 정보검색 엔진이 형태소 분석만을 통해서 색인어를 추출함으로써 불필요한 색인어가 문장을 대표하는 색인어로 추출되는 경우가 많으며, 이를 해결하는 방법으로 불용어 리스트를 이용하는데, 구문 형태소 단위의 형태소 분석과 품사 수준의 불용어 처리를 통해 색인 오류의 감소와 정확한 색인어 추출의 효과를 얻을 수 있었다.

참고문헌

[1] Willaim B. Frakes, Richard Baeza-Yates, Information Retrieval : Data Structures & Algorithms, Prentice-Hall, 1992.

[2] 김영택, 자연언어처리, 교학사, 1994.

[3] 강승식, "형태소 분석과 자동색인", 제2회 자연언어처리 튜토리얼, 2000.

[4] 황이규, 이현영, 이용석, "형태소 및 구문 모호성 축소를 위한 구문단위 형태소의 이용", 정보과학회논문지:소프트웨어 및 응용, pp. 784-793, 제27권 7호, 2000.

[5] KIBS : Korean Information Base System, http://kibs.kaist.ac.kr/

[6] Hantec : http://www.kordic.re.kr/~giis/hantec