형태소 네트웍을 이용한 한글 문헌의 자동 키워드 추출

김 철 완, 장 재 우
전북대학교 컴퓨터공학과

Automatic Keyword Extraction using Morpheme Network for Korean Texts

Chul-Wan Kim and Jaw-Woo Chang
Chonbuk National Univ. Dept. of Computer Engineering

요 약

본 논문은 한글 문헌의 자동 키워드 추출을 위한 새로운 접근 기법을 제시한다. 한글에서 나타나는 형식형태소는 어절내에서 일정한 결합규칙을 가지며 또한 명사구나 동사구에서 보여지는 것처럼 어절간의 연결에도 관계된다. 유한개의 형식형태소를 노드로 하여 구성된 형태소 네트웍은 어휘사전 및 문헌을 통해 링크를 생성하게 되며 형태소분석 과정에서 이를 이용하면 명사 추출의 정확성을 높일 수 있고 사전 탐색을 최소화하여 미등록어 추정 및 분석 속도를 향상시킬 수 있다.

1. 서론

도서 정보, 신문 기사, 특허 및 법률 정보와 같은 문헌 정보는 기하급수적으로 증가하고 있으며 이러한 정보를 관리하기 위해 자연언어 처리기술을 이용한 정보 검색 시스템이 최근 활발히 연구되고 있다. 정보 검색 시스템은 사용자로부터 정보에 대한 요구가 발생하였을 때 사용자가 필요로 하는 문서에 대한 질의표현과 저장된 문서들을 구분하는 표현간의 유사성을 계산해서 관련된 문서를 제공해야 한다. 따라서 기본적인 정보검색 시스템은 적어도 질의 표현에 대한 정의와 함께 문서로부터 문서를 대표하는 표현을 도출하는 방법 그리고 유사성 계산을 통한 검색 방법등을 구현함으로써 설계되어진다[2]. 정보 검색 시스템에서 문서를 대표하는 표현을 추출하는 색인과정은 자연언어 처리기술과 연계되어 가장 어려운 부분이라고 말할 수 있다. 또한 시스템의 성능에 적지 않은 영향을 미치기 때문에 국내에서도 색인과 관련된 많은 연구가 진행되고 있다.

자동색인이란 문헌으로부터 그 문헌을 대표할 수 있는 대표어구를 컴퓨터를 이용하여 자동으로 찾아내는 것을 말한다[1]. 자동색인을 위한 여러 가지 방법들이 제안되고 있지만 현재의 자연언어 처리기술의 한계와, 투자에 비해 얻을 수 있는 성능향상이 불확실한 이유 등으로 구문분석이나 의미분석을 이용한 방법은 실험적인 수준에 그치고 있다[3]. 본 논문은 비록 형태소 분석 방법이 구문, 의미분석 방법에 비해 정확성이 떨어지고 구단위 색인이 어렵다는 문제가 있지만, 한국어 자동색인에 있어서 형태소 분석 방법을 가장 현실적인 방법으로 판단하였다. 이러한 입장에서 본 논문은 형태소분석을 이용한 기존의 자동색인 시스템의 문제점을 해결하기 위해 새로운 접근 방법을 제안한다.

본 논문에서 제안하는 형태소 네트웍은 문법형태소가 어절내에 또는 어절간에서 특정한 결합 패턴을 보이는 것에 착안하여 이를 이용하여 품사 모호성 해결이 가능함을 보이고 또한 어휘 사전의 사용을 최소화하여 분석속도를 높이고 미등록어 추정이 가능함을 보인다. 2장에서는 형태소 네트웍이 어떻게 구성되는가를 언급하고, 형태소분석 방법을 3장에서, 구현 및 실험결과를 4장에서 보이고 결론을 맺도록 하겠다.

2. 형태소 네트웍

실질적인 의미를 갖지 못하고 조사나 어미처럼 문법적 기능을 하는 형태소를 형식형태소 또는 문법형태소(grammatical morpheme)라 한다. 문법형태소는 하나의 어절내에서 독립적으로 또는 일정한 결합규칙을 가지고 체언과 용언에 붙어 문법적 기능을 하게 되며 명사구나 동사구에서 처럼 어절간에도 특정한 연결패턴을 보이게 된다. 본 논문에서 제안하는 형태소 네트웍이란 한국어에서 문법형태소로 사용되는 음절 및 음소 151개를 노드로 갖고 어절내에서 혹은 어절간에서의 문법형태소 연결 패턴을 링크로 표현한 문법형태소 네트웍을 말한다.

2.1 내부링크(Internal Link)

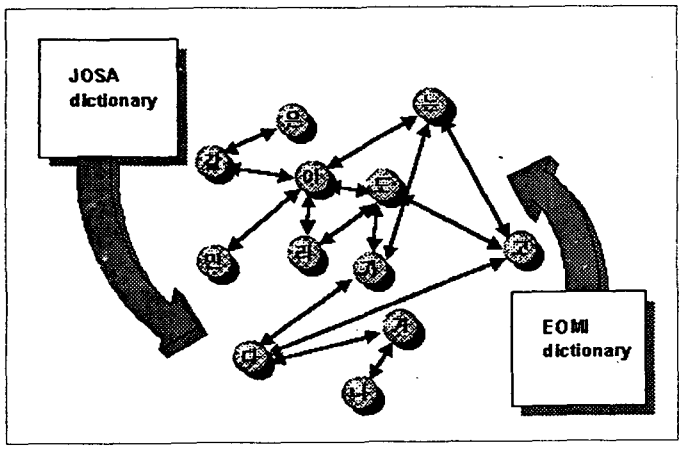

그림 1. 내부링크의 설정

내부링크는 어절내에서의 문법형태소 결합형태를 나타내는 것으로 453개의 조사와 938개의 어미 결합형을 통해 각각의 연결패턴을 노드간의 링크로 표현한 것을 말하며 이것은 형태소분석 과정에서 조사 및 어미 사전으로 사용된다. 내부링크는 문법형태소를 분리하기 위해서 오토마타나 접속정보를 사용할 때 보다 간단하면서도 결합형 단위로 분석하는 것과 동일 효과를 볼 수 있다는 장점이 있다. 그림 1은 내부링크의 설정을 나타내고 있다.

2.2 외부링크(External Link)

한국어와 같이 문장 내에서 단어의 역할이 구문의 구조보다 기능어에 의해 결정되는 언어에서는, 문장 내에서 단어의 위치나 이웃한 단어의 분석 결과에 의해 단어의 분석 결과를 예측할 수 있을 뿐 아니라, 통사분석에서 처리했던 모호성을 형태소 분석 단계에서 처리하는 것이 가능하다[4]. 외부링크는 이러한 개념을 바탕으로 Text learning을 통해 어절간의 연결이 강하게 나타나는 조사와 조사, 어미와 어미, 조사와 어미 등을 문법형태소의 링크로 표현한 것을 말하는 것으로 4가지의 링크타입을 표 1에서 보여주고 있다. 그러나 자동색인이 형태소분석에 목적이 있는 것이 아니기 때문에 외부링크를 형태소분석에 이용하지는 않으며 형태소분석 결과로부터 체언과 용언간의 품사 모호성 해결을 위해 사용하게 된다.

2.2.1 모호성 해결

형태소 분석과 관련되는 모호성은 품사 모호성과 어휘 모호성이며 자동색인의 경우 체언과 용언의 구별이 관건이므로 품사 모호성만을 처리하면 된다. 품사 모호성은 하나의 형태소가 여러가지 품사로 분석되는 경우로, 한국어에서는 단일 형태소로 이루어진 단어에서 주로 나타나고, 조사나 어미와 결합하여 나타날 때에는 품사 모호성이 중첩되게 된다[7]. 예를 들어 "본 연구는 두 개의 중요한 목적을 가지고 있다"라는 문장에서 '가지고'는 아래와 같이 분석된다.

(NOUN '가지') + (JOSA '고')
(VERB '가지') + (EOMI '고')

위와 같이 품사 모호성이 발생할 경우 '목적을'에서

표 1. 외부링크의 분류

TYPE 1	조사와 조사의 링크
TYPE 2	조사와 어미의 링크
TYPE 3	어미와 조사의 링크
TYPE 4	어미와 어미의 링크

'을'과 '가지고'의 '고'에는 조사와 어미의 관계, 즉 외부링크 TYPE 2가 설정될 수 있으며 형태소 네트웍상에 이와 같은 외부링크가 일단 설정되면 모호성 해결 단계에서 인접 단어의 분석결과와 외부링크를 통해 올바른 품사를 확정할 수 있다. 링크 설정시 주의할 점은 링크 자체에 모호성이 없어야 한다는 것이다. 예를 들어 '을'과 '고'에는 TYPE 2이외의 링크가 설정될 수 없기 때문에 이를 모호성 해결의 단서로 이용할 수 있지만 또 다른 타잎의 링크가 존재한다면 결국 링크를 설정할 수 없다.

2.3 노드의 구성

앞서 언급했듯이 형태소 네트웍은 문법형태소를 노드로 갖고 내부링크와 외부링크로 구성된다. 노드는 링크설정을 위한 Map 영역과, 이 문법형태소가 조사나 어미로 사용될 때 어떤 특성을 갖는지를 기록하기 위한 영역을 포함하게 된다. 조사가 어휘 형태소와 결합할 때 어휘 형태소의 종성과 어울릴 수 있는가, 혹은 독립적으로 사용될 수 있는가 등을 정보로 유지하게 된다. 그림 2는 형태소 네트웍을 구성하는 노드의 구조를 나타낸다.

<table>
<tr><td>Id number</td><td>Morpheme</td><td>OFFSET in the datafile</td></tr>
<tr><td colspan="3">JOSA information field</td></tr>
<tr><td colspan="3">EOMI information field</td></tr>
<tr><td colspan="3">Internal Map field: FROM</td></tr>
<tr><td colspan="3">Internal Map field: TO</td></tr>
<tr><td colspan="3">External Map field: FROM</td></tr>
<tr><td colspan="3">External Map field: TO</td></tr>
</table>

그림 2. 노드의 구성

3. 형태소 분석

한국어에서 조사가 발달된 점을 이용해 간단한 자동색인 시스템에서는 형태소분석 없이 조사 사전만을 이용해 명사를 추출하기도 했지만 정확성에 많은 문제가 있기 때문에 자동색인에 적합한 형태소분석이 요구된다. 형태소 분석의 정도는 응용분야에 따라 달라지며 정보검색시스템에 사용할 경우는 최대한 많은 단어를 분석해야 하므로, 복합명사나 사전 미등록어를 처리하는데 많은 비중을 두어야 한다[4].

자동색인에서는 단어를 구성하는 모든 형태소의 결합관계를 해석하는 정도의 분석은 필요치 않고 체언과 용언을 구분하고 각각에 포함된 명사만을 정확히 분리하면 되므로 형태소 분석도 이에 적합하도록 구성해야 한다. 본 논문은 명사 사전을 사용하는 데서 발생하는 시간적 문제와 미등록어 처리를 위해 명사 사전을 사용하지 않고 대신 어미분석을 강화하고 용언사전을 두어 정확한 용언분석을 통해 체언을 구별할 수 있도록 형태소 분석기를 구성하였다. 전체적인 형태소분석은 강승식[4,5,6]의 방법을 따르되 이를 본 시스템에 적합하도록 변형시켜 사용하였다.

3.1 용언분석

한국어 형태소분석은 용언분석이라고 말할 수 있을 정도로 한국어에서 용언은 다양한 형태론적 변형 현상과 관계되어 나타난다. 본 논문에서 용언분석은 어간과 어미를 분리한 후 어간의 원형을 복원하여 사전 탐색을 통해 용언을 확인하는 과정으로 요약된다. 앞서 언급했듯이 본 논문은 체언과 용언간의 품사적 모호성이 발생할 때 인접한 어절의 분석 결과를 이용하기 위해 외부링크를 설정하게 되는데, 형태소 분석시 어미 '은(는)/을(를)'이나 '아/어'가 축약 또는 생략될 경우를 고려해야 한다. 또한 용언의 원형만을 저장하고 있는 사전을 탐색하기 위해서는 불규칙용언의 원형 복원이 필수적이라 하겠다. 분석과정은 다음과 같다.

- 어말어미의 시작 위치 추정

음절 정보[4]를 이용하여 어말어미의 시작위치를 찾고 네트웍상에 링크가 존재하면 이를 어말어미로 추정한다.

- 4종성('ㄴ', 'ㄹ', 'ㅁ', 'ㅂ')의 분리

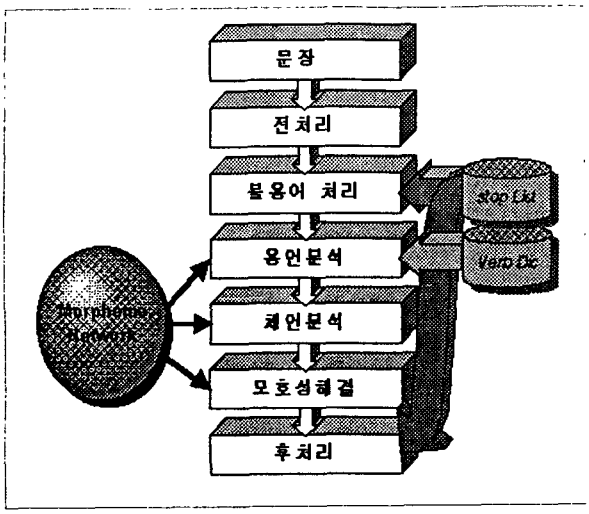

그림 3. 시스템 구성

어간의 말음이 'ㄴ/ㄹ/ㅁ/ㅂ'이고 'ㄴ/ㄹ/ㅁ/ㅂ'+어미의 내부링크가 존재하면 이를 어미로 추정한다.

- 선어말어미의 분리
어간의 말음이 'ㅆ'이나 '시'이면 선어말어미에 사용되는 음절을 찾게 되는데 이 과정을 반복하면 모든 선어말 어미를 분리할 수 있다.

- '하다'류의 접사 분리
어간에서 용언화 접미사가 발견되면 이를 접사로 분리한다.

- '아/어'의 변이체 분리
어간의 말음이 'ㅏ/ㅘ/ㅕ/ㅐ/ㅔ'이고 '아/어'+어미의 내부링크가 존재하면 이를 어미로 추정한다.

이상과 같은 과정을 거쳐 생성된 분석 후보들은 불규칙용언의 처리와 사전 탐색을 거쳐 분석을 완료하게 된다.

3.2 체언분석

본 시스템에서 체언의 분석은 음절 정보를 이용하여 조사의 시작 위치를 추정하고 이를 내부링크로 확인하는 과정이다. 분리된 조사는 체언의 종성에 따라 결합이 가능한 지를 검사하여 분석을 마친다. 서술격 조사나 명사에 용언화접미사+명사형어미+조사(ex. '검색함을')와 같은 체언에 나타나는 어미 활용 현상은 모두 어미분석 과정에서 처리하며 그 외의 조사 변형 현상은 없는 것으로 가정한다.

4. 시스템 구성

전체적인 시스템의 구성은 그림 3과 같다. 어절 단위의 분석이 지니는 한계는 그간의 연구에서 문제시 된 사항이며 앞으로는 문장 단위의 분석이 필수적이라 생각된다. 본 시스템에서는 인접 단어의 분석 결과를 모호성 해결의 단서로 사용하기 때문에 문장 단위로 분석을 한다. 전처리 과정은 한글 이외의 코드에 대해 적절한 조치를 취하여 이들을 형태소 분석에서 제외 시키는 것이 주목적이며, 아라비아 숫자, 또는 괄호 등의 특수문자를 처리한다.

대부분의 자동색인 시스템에서는 체언 분석을 용언 분석전에 수행시켜 어떤 단어가 체언 어절일 가능성이 없다고 판단되면 처리를 중지하기도 하지만, 본 시스템의 경우 문장의 모든 어절이 분석되어야 하고 체언 분석 과정에서 사용하는 어휘정보가 적기 때문에 그와 같은 판단을 하지 않고 대신에 용언 분석을 먼저 수행하게 한다. 용언 분석과정에서는 선어말 어미나 용언에만 나타나는 음소 등의 정보를 이용하여 체언일 가능성이 없는 어절을 체언 분석과정에서 제외시킬 수 있다.

4.1 불용어 사전

표 2. 불용어 사전

품 사	갯 수
의존명사	397
인칭 대명사	142
지시 대명사	35
수사	81
관형사	356
부사	6150

현재 한글 정보검색 시스템에서 불용어(Stopword)에 대해 정확히 정의된 바는 없지만 일반적으로 색인어로 사용하기에는 부적절한 단어를 지칭하는 의미로 사용

되어 왔다. 한국어의 특성상 불용어의 선택이나 사전의 이용 방법등은 영어와 상당한 차이가 있기 때문에 불용어에 대한 정확한 선정 기준과 이용 방법이 필요하다. 본 논문에서는 표 2에서 보여지는 것처럼 색인어로 사용될 수 없는 6개의 품사를 선정하여 국어사전[8]에서 7012개의 단어를 추출하였다. 이 단어들은 독립적으로, 또는 문법형태소와 결합되어 나타날 수 있기 때문에 스트링 매칭 방법만으로 이들을 찾아낼 수 없으며 문법형태소를 분리하고 필요에 따라서는 부사형 어미를 첨가시키는 과정을 거쳐 사전을 탐색하게 된다. 따라서 본 시스템에서는 불용어 사전의 탐색이 형태소 분석전과 후에 이루어지도록 구성하였다.

4.2 후처리

형태소 분석과 모호성 처리 과정이 끝나면 어절 각각의 분석 결과로부터 명사 추출 여부를 후처리 과정에서 확정하게 된다. 다음은 분석 결과에 대한 후처리 과정이다.

- 불용어도 아니고 체언이나 용언으로도 분석되지 않은 경우: 문법 형태소없이 독립적으로 사용된 명사로 판정한다.

- 체언으로 분석된 경우: 어휘 형태소에 대한 불용어 검사를 통해 추출 여부를 결정한다.

- 용언으로 분석된 경우: 용언화 접미사, 서술격 조사가 분리된 경우 어휘 형태소에 대한 불용어 검사를 통해 추출 여부를 확정한다. 단, 용언화 접미사가 '하'인 경우 부사형 어미 '이/히'를 어휘 형태소에 첨가하여 불용어 검사를 반복한다. 그외의 경우 추출 대상에서 제외한다.

- 불용어로 분석된 경우: 추출 대상에서 제외한다.

- 체언과 용언의 모호성이 해결되지 않은 경우: 체언으로 처리한다.

- 문법형태소없이 연속적으로 나타나는 단어들은 복합명사로 추출한다.

4.3 분석예

예문: "나는 파란 감을 보고 난 후 마을로 가는 길을 알 수가 있었다"

위 문장은 논문[4]에서 제시한 예문으로 많은 모호성을 가지는 단어들을 포함하고 있으며 이에 대한 본 시스템에서의 형태소 분석 결과를 표 3에서 보여주고 있다. 본 시스템은 명사사전을 사용하지 않기 때문에 어절 2, 4, 5, 6, 10에 대해 각각의 명사형으로 파란(波瀾), 보고(報告), 난(蘭), 후(後), 알(卵)과 같은 후보를 생성할 수 없는 것이 해결하기 어려운 문제로 남아 있다. 그러나 위 예문에 대한 분석에서 이러한 문

표 3. 형태소 분석 예

	어 절	형 태 소 분 석 결 과
1	나는	(NOUN'나' + JOSA'는') (VERB'나' + EOMI'는')
2	파란	(NOUN'파' + JOSA'란') (VERB'파' + EOMI'란') (VERB'파랗' + EOMI'ㄴ')
3	감을	(NOUN'감' + JOSA'을') (VERB'감' + EOMI'을') (VERB'가' + EOMI'ㅁ을')
4	보고	(NOUN'보' + JOSA'고') (VERB'보' + EOMI'고')
5	난	(VERB'나' + EOMI'ㄴ') (VERB'날' + EOMI'ㄴ') (VERB'낳' + EOMI'ㄴ')
6	후	(ADV'후')
7	마을로	(NOUN'마을' + JOSA'로')
8	가는	(NOUN'가' + JOSA'는') (VERB'가' + EOMI'는')
9	길을	(NOUN'길' + JOSA'을') (VERB'길' + EOMI'을') (VERB'긷' + EOMI'을')
10	알	(VERB'알' + EOMI'ㄹ')
11	수가	(NOUN'수' + JOSA'가')
12	있었다	(VERB'이' + PEOMI'ㅆ었' + EOMI'다')

제는 어절 6에서 조사의 생략으로 오분석된 경우 뿐이며 대부분 타당한 분석이 이루어졌다.

예문에서 명사는 '나', '감', '후', '마을', '길', '수'이며 대명사, 의존명사 등의 불용어를 제외시키면 추출 대상이 되는 명사는 3단어('감', '마을', '길')에 불과하지만 모호성 해결을 하지않고 표 3의 분석결과로 명사를 추출할 경우 불필요한 단어가 색인어로 나타나게 된다. 만약 앞에서 예로 들었던 "..목적을 가지고 있다"에서 '을'과 '고'에 TYPE 2의 외부링크가 설정된 상태라면 어절 3과 4의 모호성을 해결할 수 있으며 어절 2와 3에서 관형격 어미 'ㄴ'과 조사 '을'에 TYPE 3을 설정할 수 있다. 그러나 어절 8과 9에서는 '는'과 '을'에 두개 이상의 링크가 가능하기 때문에 외부링크가 설정될 수 없다. 본 시스템에서 외부링크 설정후 재분석하였을 때 다음과 같은 결과를 얻을 수 있었다.

추출 대상 : '감', '마을', '길'
형태소 분석후 명사: '나', '파', '감', '보', '마을', '가', '길', '수'
모호성 해결후 명사: '나', '감', '마을', '가', '길', '수'
불용어 제거후: '감', '마을', '가', '길'

5. 결론 및 향후 연구 방향

자동 색인 시스템의 성능을 평가하기 위한 가장 중요한 요소는 역시 사람에 의해 선택된 색인어와 얼마나 유사하게 특정 단어를 추출해 내느냐에 있다. 이러한 요구를 만족시키 위해 그간 자연 언어 처리와 연계하여 많은 연구가 있었으며 이는 조사 절단과 같은 단순한 방법에서부터 의미분석에 이르기 까지 이론적인 면에서도 상당한 발전이 있었다. 그러나 자연 언어 처리에 있어 현재의 기술 수준이 직면한 한계는 단시간내에 해결되기는 어려울 것으로 보이며 앞으로도 많은 노력이 필요할 것이다.

본 논문에서는 한글 문헌에서 키워드 추출을 위한 새로운 접근 기법을 제시하였다. 형태소 네트웍을 이용하여 조사나 어미사전이 링크로 표현될 수 있음을 보였으며 어절간에 나타나는 문법형태소의 연결패턴을 단서로 모호성 해결을 시도하였다. 용언에 대한 형태소 분석을 통해 체언을 판별할 수 있도록 하고 불용어 사전을 구축함으로써 불필요한 단어들의 추출을 막을 수 있었다. 또한 명사 사전을 사용하지 않기 때문에 사전 미등록어를 처리할 수 있고 속도를 향상시킬 수 는 있었다.

본 논문에서 설정한 4가지 타입의 외부링크는 어절간의 연결을 이용한 저수준의 통사분석이라 말할 수 있지만 조사 및 어미를 세분화하여 외부링크의 타입을 확장하고 대량의 문헌을 통해 한국어 문장에 나타나는 구문적 특성을 링크로 설정할 수 있다면 보다 정확한 분석이 가능할 것이다. 본 논문에서는 특정 문장에 대한 간단한 실험결과를 제시했을 뿐이지만, 앞으로 대량의 문헌을 통한 전체적인 성능 평가가 이루어져야 할 것이다.

참고 문헌

[1] 정영미, 정보 검색론, 구미 무역 출판부, 1991.

[2] 최기선, "구문 및 의미분석을 통한 한국어 자동색인," 정보관리학회지 Vol.8, No.2, 1991 pp.96-106

[3] 최기선,"한국어 정보검색," 정보과학회지 Vol.12, No.8, pp.24-32, 1994

[4] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석," 서울대학교 박사학위 논문,1993

[5] 강승식 외, "한국어 형태소 분석기에서 불규칙 용언의 분석 모형," 한국정보과학회 논문지, Vol. 19, No. 2, pp.151-163, March, 1992

[6] 강승식 외, "한극어 형태소 분석기에서 선어말어미의 분석 모형," 한국정보과학회 논문지, Vol. 18, No. 5, pp.505-513, September, 1991

[7] 김영택 외, 자연 언어 처리, 교학사, pp.495-512, 1994

[8] 금성교과서, 콘사이스 국어사전, 1994