전문(全文) 분석을 통한 파생명사 및 합성명사의 분석

박봉래, 황영숙, 임해창
고려대학교 컴퓨터학과 자연어처리연구실

Analysis of Derived Nouns and Compound Nouns by Examining Full Text

Bong-Rae Park, Young-Sook Hwang, Hae-Chang Rim
pbr@nlp.korea.ac.kr, yshwang@nlp.korea.ac.kr, rim@nlp.korea.ac.kr

대부분의 한국어 형태소 분석기는 파생명사나 합성명사가 포함된 어절을 오분석 또는 과분석하는 경향이 있다. 이는 하나의 어절에서 오분석이나 과분석을 방지하기 위하여 획득할 수 있는 정보가 제한적이기 때문이다. 이에 본 논문은 파생명사나 합성명사 후보가 포함된 어절뿐만 아니라 주변 및 전문에서 분석에 필요한 정보를 수집하여 이용하는 방법을 제시한다. 제안한 방법은 오분석된 파생명사나 합성명사에만 나타나는 저빈도 단어를 제거하고, 파생명사나 합성명사 후보의 주변 어휘들을 실마리로 이용하며, 문서 전역에서 동일한 파생명사나 합성명사 후보가 포함된 둘 이상의 어절을 비교분석하여 파생명사 및 합성명사 후보가 포함된 어절을 처리한다. 실험 결과 제안한 방법은 99.8%의 정확도와 95.3%의 재현율로 파생명사나 합성명사 후보가 포함된 어절을 올바르게 분석할 수 있었다.

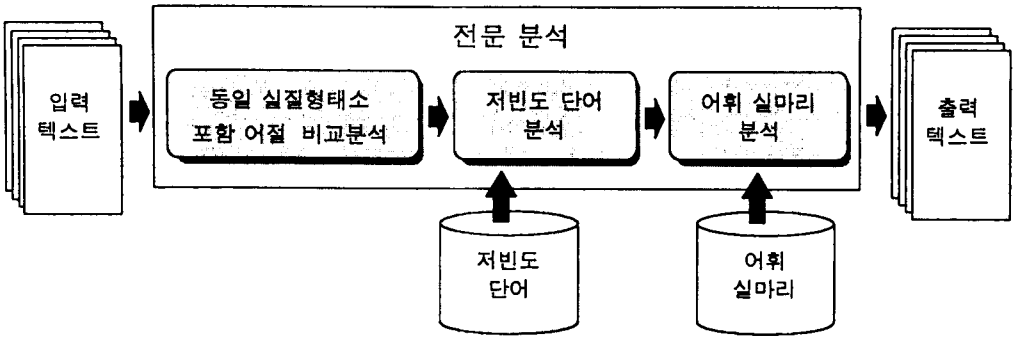
1. 서론

한국어에서 어절은 띄어쓰기의 단위로서 보통 두개 이상의 형태소들로 구성되어 있고 형태소들은 일반적으로 하나 이상의 음절들로 표현된다[4,7]. 따라서 어절을 구성하고 있는 음절들을 분리하여 구성 형태소들을 밝혀내는 한국어 형태소 분석기는 분리 위치에 따라서 다양한 분석 결과를 내놓을 수 있다[6]. 특히, 1음절 접사에 의한 파생명사나 1음절 명사에 의한 합성명사들을 모두 분석할 경우 종종 백여개 이상의 분석결과를 내놓는다. 이러한 과분석은 자연어 처리의 상위 단계인 구문분석 및 의미분석을 어렵게 만들 뿐 아니라 이들의 응용 시스템인 정보 검색 시스템, 기계 번역 시스템 및 각종 인식기의 후처리 시스템의 성능 저하의 원인이 될 수 있다.

지금까지 보고된 합성명사 및 파생명사 처리 방법은 크게 사전 의존적인 방법[5,6]과 결합제약정보를 이용하는 방법으로 분류할 수 있다[2]. 사전 의존적인 방법이란 모든 가능한 합성명사 및 파생명사를 사전에 등록하고 생산성 높은 접사[1]에서 파생된 몇 종류의 파생명사들만을 분석함으로써 과분석을 방지하는 방법이다. 이 방법은 크게 두 가지의 문제점을 가지고 있다. 첫번째는 실세계의 모든 파생명사와 합성명사를 모두 사전에 등록할 수 없기 때문에 결국에 미등록 파생명사 및 합성명사가 포함된 어절을 분석하지 못하거나 오분석하게 된다는 점이다[3]. 이 문제점은 과분석만큼이나 자연어 처리 상위 시스템 및 응용 시스템의 성능 저하의 큰 원인이 되고 있다. 두번째 문제점은 사전에 등록된 파생명사 및 합성명사가 포함된 어절도 여전히 과분석 및 오분석의 가능성이 존재한다는 점이다. 예를 들어, 어절 '물리학과'의 경우에 사전에 '물리학'이 포함되어 있어도 '물리학_명사+과_조사'와 '물리_명사+학과_조사'로 분석될 수 있다.

결합제약정보를 이용하는 방법은 기본적으로는 사전

1) 생산성 높은 접사란 다른 접사에 비해 상대적으로 다양한 실질형태소와 결합할 수 있는 접사이다[8].



[그림 1] 시스템 구성도

의존적인 방법과 비슷하지만 생산성이 낮은 접사에서 파생된 파생명사들 중에서도 접사의 결합 가능 여부를 평가하여 과분석을 방지하는 방법이다[2]. 이 방법은 대량의 코퍼스를 분석하여 접사의 결합 특징을 추출하고 모든 명사들을 적절하게 분류하여 서로 결합 가능한 접사와 명사의 결합만을 인정하는 방법인데, 이 방법도 모든 명사들을 올바르게 분류하고 접사와의 결합 가능 여부를 모두 추출하는 것이 쉽지 않기 때문에 결국 과분석을 피할 수 없다. 그리고, 현재로서는 접미파생명사에 대해서만 이러한 연구결과가 보고되어 있다[2]. 물론, 복합명사 제약 정보를 이용하는 방법으로 합성명사 분석 제약도 가능하겠지만 아직까지 구체적으로 보고된 것은 없다[2]. 그리고 이 또한 모든 명사들이 미리 분류되어 있어야 하고, 분류되어 있다고 해도 모호한 경우들이 많기 때문에 상당한 과분석이 야기될 수밖에 없다.

2. 제안한 파생명사 및 합성명사 분석 방법

본 논문에서 제안하는 전문(全文) 분석을 통한 파생명사 및 합성명사의 분석 방법이란 파생명사 및 합성명사 후보가 포함된 어절과 주변 어휘 및 전문에서 파생명사 및 합성명사 분석에 필요한 정보를 수집하여 파생명사 및 합성명사 분석에 이용하는 것이다. 따라서 제안한 방법은 세 가지 방법으로 이루어져 있다. 첫번째 방법은 파생명사 및 합성명사 후보를 포함한 어절에 오분석된 파생명사 및 합성명사 포함 어절에만 나타나는 저빈도 단어[3]가 존재하는지 조사하고 존재할 경우에 저빈도 단어를 포함한 분석 결과를 배제한다. 그리고, 두번째 방법은 특정 파생명사 및 합성명사 후보를 포함한 어절 주변에 이 파생명사 및 합성명사와 함께 자주 나타나는 어휘들이 존재할 경우에 이들을 어휘 실마리로 이용하여 해당 파생명사 및 합성명사 후보가 포함된 분석 결과를 올바른 분석결과로 선택한다[4]. 끝으로 세 번째 방법은 정보 추출 영역을 확대하여 입력 문서 전역에서 동일한 파생명사 및 합성명사 후보를 포함한 어절들을 수집하고 각 어절에 나타난 해당 파생명사 및 합성명사 후보가 결합한 어휘에 근거하여 해당 파생명사 및 합성명사 후보를 포함한 분석 결과를 올바른 분석결과로 선택한다.

[그림 1]은 제안한 방법의 시스템 구성도로서 파생명사 및 합성명사 분석을 위한 세 가지 방법들을 상호보완관계와 정확도를 고려하여 효과적으로 통합시킨 상태를 표현하고 있다. 즉, 동일 실질형태소 포함 어절의 비교분석 방법(이하, '동일 실질형태소 분석 방법'이라 약술함)이 제일 먼저 입력 텍스트 내의 모든 파생명사 및 합성명사 후보를 포함한 어절에 대해 적용되고 이 방법에 의해 처리되지 못한 어절들에 대해서 저빈도 단어 분석 방법이 적용되며 이 방법에서도 처리되지 못한 어절들에 대해서 어휘 실마리 분석 방법이 적용된다[5]. 동일 실질형태소 분석 방법이 가장 먼저 적용된 이유는 나머지 두 방법이 오분석하기 쉬운 어절들[6] 중 두번

2) 다만, 관련 연구로서 명사간 의미 관계를 고려한 합성명사의 의미 관계 분석 방법이 보고되어 있다[1].

3) 오분석된 파생명사 및 합성명사 포함 어절에만 나타나는 저빈도 단어들은 미리 학습 코퍼스로부터 추출되어 저빈도 사전에 존재함.

4) 특정 파생명사 및 합성명사와 함께 자주 나타나는 어휘들도 미리 학습 코퍼스로부터 추출되어 어휘 실마리 사전 내에 존재함.

5) 실제로는 형태소 분석기가 입력 텍스트 내의 각 어절들을 분석하고 분석된 어절들 중 파생명사 및 합성명사 후보를 포함한 어절들을 대상으로 제안한 방법들이 적용되지만 제안한 방법을 위주로 설명하기 위해 시스템 구성도에서 형태소분석기는 제외하였다.

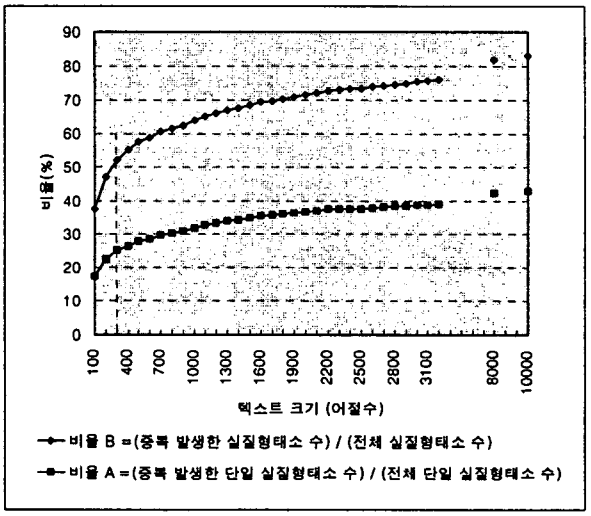
6) 저빈도 단어 분석 방법이 오분석하기 쉬운 어절은 해당 문맥에서 실제로 의미있게 사용된 저빈도 단어가 포함된 어

이상 발생한 실질형태소를 포함한 어절들을 동일 실질형태소 분석 방법이 미리 분석함으로써 나머지 두 방법의 오분석을 최소화할 수 있기 때문이다. 또한, 저빈도 단어 분석 방법이 어휘 실마리 분석 방법에 앞서 적용된 이유는 각각의 실험 결과 저빈도 단어 분석 방법이 높은 재현율과 정확도로 파생명사 및 합성명사 분석을 수행할 수 있는 반면에 어휘 실마리 분석 방법은 훨씬 낮은 재현율과 정확도로 파생명사 및 합성명사 분석을 수행하는 데에다 주변에 존재하는 어휘 실마리를 찾기 위해 주변 어절들을 분석해야 하는 부담이 있기 때문이다.

2.1. 동일 실질형태소 포함 어절의 비교분석에 의한 파생명사 및 합성명사 분석

태깅된 약 100만 어절의 코퍼스를 분석한 결과에 따르면 동일한 실질형태소가 텍스트의 크기에 따라 두 번 이상 발생할 확률은 [그림 2]와 같다. 그림에 따르면 300어절 크기의 텍스트 내에 존재하는 단일한 실질형태소들 중 약 25%(비율 A)가 두번 이상 발생하고 이들이 전체 발생한 실질형태소의 52%(비율 B)를 차지한다. 그리고 텍스트 크기가 커질수록 두번 이상 발생하는 단일 실질형태소의 비율이 점점 높아져서 텍스트 크기가 만 어절이 되면 약 42%(비율 A)의 단일 실질형태소가 두번 이상 발생하고 이들이 전체의 83%(비율 B)를 차지하게 된다. 이러한 분석 결과는 동일한 실질형태소에 대한 전문 분석 방법이 작은 크기의 텍스트인 경우도 전체 어휘의 약 50%를 처리할 수 있는 효용성이 있고 텍스트의 크기가 커질수록 효용성은 더욱 증가함을 의미한다. 제안한 방법은 이러한 효용성에 근거하여 동일한 파생명사 및 합성명사 후보를 포함한 둘 이상의 어절들을 비교분석하여 해당 파생명사 및 합성명사 후보의 실제 어휘 여부를 결정한다.

한국어에서 동일한 3음절 이상의 형태소가 서로 다른 의미로 동일한 텍스트에 나타날 가능성은 거의 없다. 따라서 중의성이 발생한 어절에 존재하는 3음절 이상의 실질형태소 후보가 동일한 텍스트의 다른 어절에도 나타난다면 이들은 같은 의미로 사용된 실제 어휘일 가능성이 높다. 물론 두 어절이 동일한 형태라면 중의적으로 나타난 각 실질형태소 후보들이 똑같이 두 어절에 나타난 것이므로 어느 쪽이 실제 어휘라고 결정할 수는 없다. 하지만 3음절 이상의 동일한 실질형태소가 서로 다른 형태소와 결합하여 서로 다른 어절에 존재한다면 대부분의 경우에 이 실질형태소 후보를 포함한 분석 결과가 올바른 분석일 가능성이 높다. 제안한 방법은 이러한 점을 이용한 분석 방법으로 <알고리즘 1>에 제시한 바와 같이 동일한 텍스트 내에서 3음절 이상의 동일한 형태소 분석 후보를 갖는 모든 어절들을 수집하고 이들이 결합한 서로 다른 형태소들에 근거하여 올바른 분석 결과를 선택한다.



[그림 2] 전문 분석 방법의 효용성

<알고리즘 1> 동일 실질형태소 포함 어절의 비교분석에 의한 파생명사 및 합성명사 분석

① 3음절 이상의 동일한 실질형태소 후보를 갖는 어절 수집

예) 고려대가 ==> 고려_명사+대가_명사
or 고려대_명사+가_조사
고려대를 ==> 고려대_명사+를_조사

② 서로 다른 어휘와 결합한 3음절 이상의 동일한 실질형태소 후보를 올바른 후보로 선택

예) 고려대가 ==> 고려대_명사+가_조사

그러나, 이 방법은 한가지 문제점을 가지고 있다. 예를 들어, 어절 '합의문서'와 '합의문에'가 동일한 텍스트에 존재할 경우에 어절 '합의문서'가 '합의문_명사+서_조사'로 오분석될 수 있다는 점이다[3]. 이러한 현상은 접미사 '문'과 명사 '문서'가 비슷한 의미를 가지고 있기 때문에 나타나는 현상이다. 따라서 '문서'와 같이 접미사와 비슷한 의미를 갖고 접미사의 음절로 시작하는 어휘들을 미리 구분할 수 있다면 이러한 문제는 해소될

절이고, 어휘 실마리 분석 방법이 오분석하기 쉬운 어절은 주위에 학습 코퍼스로부터 잘못 선정된 어휘 실마리들이 존재하는, 파생명사 및 합성명사 후보를 가진 어절이다.

수 있다. 이러한 어휘는 대량의 코퍼스에서 동일 실질형태소 분석 방법을 수행할 때 두 가지 이상의 분석 결과가 유도될 수 있는 어절들에서 자동으로 추출할 수 있을 것이다. 예를 들어, 어절 '합의문서'에서 동일 실질형태소 분석 결과로 '합의문_명사+서_조사' 분석도 나오고 '합의_명사+문서_명사' 분석도 나올 경우에 '문서'를 따로 관리하는 것이다[7].

2.2. 저빈도 단어 배제에 의한 파생명사 및 합성명사의 오분석 방지

파생명사 및 합성명사를 포함하고 오분석된 어절들 중 가장 문제가 되는 경우는 2음절 명사들로 구성된 복합명사로 오분석되는 어절들이다. 다른 경우들은 오분석될 때 나타나는 어휘들이 몇 안되기 때문에 이들을 배제하면 높은 정확도로 분석이 가능하지만 복합명사로 오분석될 때에는 다양한 어휘들이 오분석된 어절에 나타날 수 있기 때문이다. 그러나 복합명사로 오분석될 때에 나타나는 어휘들도 분석해 보면 현재는 거의 사용되지 않는 저빈도 단어들이 많이 존재한다. 따라서 이러한 어휘들을 미리 선정하여 분석을 통해 배제하면 그만큼의 오분석을 방지할 수 있다.

저빈도 단어를 선정하는 가장 간단한 방법은 태깅된 코퍼스에서 단어들의 발생빈도를 측정하여 전자사전에 존재하면서 발생 빈도가 0인 어휘들을 수집하는 것이다. 그러나 이 방법을 신뢰하기 위해서는 대용량의 태깅된 코퍼스가 필요한 단점이 있다. 따라서 제안한 방법은 오분석된 파생명사 및 합성명사 포함 어절에서 저빈도 단어가 발생하는 위치에 근거하여 저빈도 단어를 추출한다. 다음은 오분석된 파생명사 및 합성명사 포함 어절에서 저빈도 단어가 발생하는 위치들이다.

<저빈도 단어 발생 위치>

■ 접사와 실질형태소의 경계

예) 난방기-설치의
==오분석 결과==> 난방_명사+기설_명사+치의_명사

■ 접사와 형식형태소의 경계

예) 기준일-로 ==오분석 결과==> 기준_명사+일로_명사

■ 접사와 접사의 경계

예) 고려대-등이
==오분석 결과==> 고려_명사+대등_명사+이_조사

■ 실질형태소와 형식형태소의 경계

예) 군대표-로 ==오분석 결과==> 군대_명사+표로_명사

■ 실질형태소와 실질형태소의 경계

예) 군대표-주장도
==오분석 결과==> 군대_명사+표주_명사+장도_명사

제안한 방법은 <알고리즘 2>에 제시한 바와 같이 먼저 저빈도 단어 발생 위치에 자주 나타나는 어휘들을 수집하고 이들이 실제로 현재는 거의 사용되지 않는 저빈도 단어인지 조사하기 위하여 대량의 학습 코퍼스에서 실제 어휘로 사용된 적이 있는지를 검증한다.

<알고리즘 2> 저빈도 단어 선정

① 대량의 학습 코퍼스에서 사전에 등록된 파생명사 및 합성명사를 포함한 어절을 수집한다.

예) 난방기설치의, 기준일로, 고려대등이, 군대표로, 군대표주장도

② 수집된 어절들 중 복합명사 분석이 가능한 어절들을 추출한다.

예) 난방기설치의, 기준일로, 고려대등이, 군대표로, 군대표주장도

③ <저빈도 단어 발생 위치>에 일정 빈도이상 자주 발생한 저빈도 단어 후보들을 추출한다.

예) 기설, 일로, 대등, 표로, 표주

④ 대량의 학습 코퍼스에 존재하는 중의성 없이 분석된 어절들에 한번도 발생하지 않은 저빈도 단어 후보들을 최종 저빈도 단어로 선정한다.

예) 기설, 일로, 표로, 표주

학습 코퍼스로부터 파생명사 및 합성명사를 포함한 어절에 대해 오분석을 야기하는 저빈도 단어들이 선정되었으면, 실제 응용 과정에서 저빈도 단어를 포함한 복합명사로의 오분석을 배제하는 것은 간단하다. 예를 들어, 어절 '기준일로'는 '기준_명사+일로_명사'와 '기준일_명사+로_조사'로의 분석이 가능한데, '일로'가 저빈도 단어이면, '기준_명사+일로_명사' 분석이 배제됨으로서 올바른 분석 결과인 '기준일_명사+로_조사' 분석만이 남게 된다.

그러나, 이 방법은 저빈도 단어가 실제로 의미있게 사용된 경우와 데이터 부족으로 잘못 저빈도 단어로 선

7) 이 문제는 어휘 실마리 분석 단계에서도 다시 언급된다.

정된 어휘도 함께 배제하게 되는 문제점이 있다. 예를 들어, 어휘 '대가'가 저빈도 단어로 잘못 선정되면[8], 어절 '미술대가'의 올바른 분석인 '미술_명사+대가_명사'를 배제할 가능성이 존재한다는 점이다. 그러나, 어절 '미술대가'와 함께 어절 '미술대가로서'와 같은, 동일한 실질형태소 후보 '미술대가'를 포함한 어절이 존재한다면 선행 과정인 동일 실질형태소 분석 방법에 의해 '미술_명사+대가_명사'가 미리 선택되어 저빈도 단어 분석에서 처리되지 않을 수 있다.

2.3. 어휘 실마리를 이용한 파생명사 및 합성명사의 분석

어휘 실마리를 이용한 분석 방법이란 특정 파생명사나 합성명사 주변에 자주 나타나는 어휘들을 실마리로 이용하여 파생명사 및 합성명사 후보의 실제 어휘 여부를 결정하는 것이다. 예를 들어, 어절 '고려대가'에서 '고려대_명사+가_조사'와 '고려_명사+대가_명사'의 분석이 가능할 때, 주변에 '학생', '학교' 등이 존재하면 '고려대'가 포함된 분석 결과를 선호하는 방법이다. 이를 위해서는 각 파생명사 및 합성명사에 대해서 어느 어휘들이 실마리로 사용될 수 있는지 미리 선정되어 있어야 한다.

어휘 실마리 선정은 원칙적으로 대량의 학습 코퍼스로부터 각각의 파생명사 및 합성명사와 주변에 나타나는 어휘와의 관계를 상호 정보(mutual information)로 측정하여 상호 정보가 높은 어휘들을 해당 파생명사 및 합성명사의 어휘 실마리로 선정함으로써 수행될 수 있다. 그러나 모든 파생명사 및 합성명사에 대해서 이를 수행하는 것은 데이터 부족의 문제가 있다. 따라서 제안한 방법은 동일한 접미사나 1음절 명사로 끝나는 파생명사 및 합성명사를 동일 부류의 어휘들로 간주하고 이들에 대해서 어휘 실마리를 선정함으로써 데이터 부족 문제를 최소화한다. 예를 들어, '고려대'의 어휘 실마리들을 '고려대'는 물론이고 '서울대', '연세대', '한양대' 등의 접미사 '대'로 끝나는 모든 3음절 명사들 주변에서 추출한다. <알고리즘 3>은 이러한 어휘 실마리 추출 과정을 구체적으로 서술한 것이다.

<알고리즘 3> 어휘 실마리 선정

① 대량의 코퍼스에서 동일한 접미사나 1음절 명사로 끝나는 파생명사 및 합성명사와 주변 어휘들을 발생 빈도와 함께 수집한다.

예) 고려대: 입시, 학생, 편의, ...
연세대: 학생, 교수, 시설, ...

② 동일한 접미사나 1음절 명사에 대해 주변 발생 어휘들을 발생 빈도와 함께 수집한다.

예) **대: 입시, 학생, 편의, 교수, 시설, ...

③ 접미사나 1음절 명사와 주변 어휘들 사이의 관련 정도를 발생빈도에 근거한 상호 정보 값(MI)으로 계산한다.

예) MI(**대,입시), MI(**대,학생), MI(**대,편의), MI(**대,교수), MI(**대,시설), ...

④ 일정 크기 이상의 상호 정보 값을 갖는 주변 어휘를 해당 접미사나 1음절 명사의 어휘 실마리로 선정한다.

예) '**대'의 어휘 실마리: 입시, 학생, 교수, ...

이렇게 코퍼스로부터 어휘 실마리들이 선정되었으면, 실제 응용 과정에서는 파생명사 및 합성명사 후보를 포함하고 중의성이 존재하는 어절 주변에 해당 파생명사 및 합성명사 후보의 끝음절에 대한 어휘 실마리가 존재할 경우에 이들에 근거하여 파생명사 및 합성명사 후보를 실제 어휘로 인식한다. 이때 우연히 어휘 실마리가 나타날 경우를 고려하여 어휘 실마리는 둘 이상일 때에만 해당 파생명사 및 합성명사 후보를 실제 어휘로 인식한다. 예를 들어, 어절 '고려대가'의 주변에 어휘 실마리 '학생' 및 '학교'가 나타난 경우에 '고려대가'의 분석 결과 '고려대_명사+가_조사'와 '고려_명사+대가_명사' 중에서 '고려대_명사+가_조사'를 올바른 분석 결과로 결정한다.

텍스트내에 동일한 파생명사 및 합성명사 후보가 포함된 동일한 어절이 둘 이상 존재하는 경우에는 각각의 주변에 존재하는 어휘들을 모두 고려하여 파생명사 및 합성명사 가능성을 평가한다. 예를 들어, 어절 '고려대가'가 텍스트에 두번 발생하고 각각 주변에 서로 다른 어휘 실마리가 하나씩 존재하면 어절 '고려대가' 주변에 두개의 어휘 실마리가 존재하는 것으로 간주하여 이 경우도 '고려대_명사+가_조사'를 올바른 분석 결과로 결정한다. 동일한 파생명사 및 합성명사 후보가 포함된 서로 다른 어절이 둘 이상 존재할 경우도 마찬가지이지만 이 경우의 대부분은 동일 실질형태소 분석 방법이 미리 처리할 가능성이 높다.

이러한 어휘 실마리 분석 방법에도 동일 실질형태소

8) 실제의 실험 결과에 따르면 '대가'는 저빈도 단어가 아니다. 학습 코퍼스에서 저빈도 단어 후보로 추출되지만 검증단계에서 배제된다.

분석 방법에서와 같은 문제점이 존재한다. 예를 들어, 어절 '종합주가'의 경우에 주변에 어휘 실마리 '주가', '증시', '매수' 등이 존재하면 '종합주_명사+가_조사'로 오분석할 가능성이 있다. 이러한 문제는 2.1절에서 제시한 바와 같이 접미사 '주'와 명사 '주가'가 비슷한 의미를 가지고 있기 때문에 나타나는 현상이다. 따라서 이런 류의 명사들 '학과', '회의', '주가', '부서', '제도' 등을 미리 추출하여 이들이 나타난 어절에 대해서는 어휘 실마리 분석 방법을 적용하지 않아야 한다. 이들 어휘를 추출하는 방법은, 예를 들어, 대량의 코퍼스에서 동일 실질형태소 분석 방법에 의해 복합명사 '종합주가'가 선택되었을 때 어휘 실마리 분석 방법에서도 파생명사로서 '종합주'가 선택된 경우에 두 방법이 충돌한 것이므로 이때 '주가'와 같은 어휘들을 추출하면 된다.

3. 실험 및 평가

저빈도 단어와 어휘 실마리를 추출하기 위하여 100만 어절 코퍼스를 이용하였고, 이 코퍼스로부터 저빈도 단어를 720개 추출하고 어휘 실마리는 다음과 같은 접미사 또는 합성명사의 오른쪽 끝음절 명사에 대해서 평균 23개의 어휘 실마리를 추출하였다.

가, 감, 계, 관, 국, 군, 금, 기, 난, 단, 당, 대, 량, 력,
령, 로, 론, 료, 률, 문, 물, 법, 병, 부, 비, 사, 산, 생,
선, 성, 세, 소, 시, 식, 실, 쌀, 안, 업, 원, 인, 장, 재,
점, 주, 지, 차, 채, 철, 청, 체, 촌, 통, 품, 학, 회

제안한 방법을 테스트하기 위해서 20만 테스트 코퍼스에서 다음과 같은 중의적 분석을 갖는 1,000개의 어절들과 주변에 존재하는 좌·우 5어절 이내의 어절들을 함께 수집하여 이용하였다.

<3음절 파생명사 및 합성명사 후보>+<기타 형태소>*
or <2음절 명사>+<2음절 명사>+<기타 형태소>*

이 1,000개의 어절들 중에 실제로 861개는 파생명사 및 합성명사를 포함하고 있고 나머지 139개는 2음절명사들의 복합명사를 포함하고 있다. 이는 1,000개의 어절을 모두 파생명사 및 합성명사를 포함한 어절로 분석해도 86.1%의 정확도를 얻을 수 있다는 의미이다.

[표 1]은 세 가지 방법의 공헌 정도를 비교하기 위하여 개별적으로 각 방법을 파생명사 및 합성명사 분석에 적용한 실험결과이다. 이 표에 따르면 동일 실질형태소 분석 방법이 정확도 99.8%로 646개를 분석할 수 있다[9]. 그리고, 저빈도 단어 분석 방법은 정확도 100%로 803개의 어절을 처리할 수 있어 가장 우수하다[10]. 끝으로 좌·우 5어절씩 10어절 이내의 어휘를 분석한 어휘 실마리 분석의 경우에는 정확도 90.1%로 81개만을 처리할 수 있다[11].

[표 1] 개별 방법의 비교 평가

동일 실질형태소 분석			저빈도 단어 분석			어휘 실마리 분석		
입력 어절	처리 어절	정확도	입력 어절	처리 어절	정확도	입력 어절	처리 어절	정확도
1000	646	99.8%	1000	803	100%	1000	81	90.1%

[표 2]는 세 가지 방법을 통합적으로 적용한 실험결과이다. 이 표에 따르면 동일 실질형태소 분석 방법이 1,000개의 어절 중 646개를 정확도 99.8%로 처리하고 나머지 354개 중 291개를 저빈도 단어 분석 방법이 100%의 정확도로 처리하고 있다. 끝으로 나머지 63개의 어절 중 18개를 어휘 실마리 분석 방법이 정확도 94%로 처리하고 1,000개 중 45개의 어절이 처리되지 못하였다. 따라서 제안한 방법은 정확도 99.8% 및 재현율 95.3%로 파생명사 및 합성명사를 처리할 수 있다.

[표 2] 통합 방법의 평가 결과

동일 실질형태소 분석 =>			저빈도 단어 분석 =>			어휘 실마리 분석			미처리 어절
입력 어절	처리 어절	정확도	입력 어절	처리 어절	정확도	입력 어절	처리 어절	정확도	
1000	646	99.8%	354	291	100%	63	18	94%	45

9) 동일한 실질형태소 분석 방법이 분석한 646개 중 567개는 파생명사 및 합성명사 후보를 선택한 것이고 79개는 2음절 명사들의 복합명사 후보를 선택한 것이다. 그리고 유일하게 오분석된 어절은 '합의문서'이다.

10) 저빈도 단어를 추출한 코퍼스와 테스트 데이터를 추출한 코퍼스는 서로 다르지만 둘 다 신문 내용이기 때문에 100%의 정확도가 획득된 것으로 판단된다. 따라서 이 정확도를 완전히 신뢰하기는 어렵다.

11) 어휘 실마리 추출 영역인 윈도우의 크기를 10이상으로 확장할 경우에 이보다는 높은 재현율을 얻을 수 있을 것이다. 그리고, 오분석된 어절들이 '종합주가', '국장회의', '정책부서의', '검사제도', '실무회의', '각료회의', '인기학과' 등인데, 이들은 2.3절에서 설명한 바와 같이 '주가', '회의', '부서', '제도', '학과' 등의 어휘가 해당 접미사와 의미가 비슷하여 발생한 오류이므로 이들을 별도로 처리할 경우에는 정확도도 향상될 수 있을 것이다.

[표 1]에 따르면 각각의 방법은 높은 정확도로 파생명사 및 합성명사 분석을 수행할 수 있지만 재현율은 높지 못하다. 그러나 이들 방법이 통합된 경우에는 [표 2]에 제시된 바와 같이 높은 정확도와 재현율로 파생명사 및 합성명사 분석을 수행할 수 있다.

4. 결론 및 향후 연구

본 논문은 파생명사 및 합성명사 포함 어절의 오분석 및 과분석을 최소화하기 위하여 파생명사 및 합성명사 후보를 포함한 어절과 주변 어휘 및 전문(全文)으로부터 정보를 획득하여 이용하는 방법으로 저빈도 단어 분석 방법, 어휘 실마리 분석 방법 및 동일 실질 형태소 포함 어절의 비교분석 방법을 제시하였다. 저빈도 단어 분석 방법은 오분석된 파생명사 및 합성명사 포함 어절에만 나타나는 저빈도 단어들을 배제함으로써 저빈도 단어들로 인한 오분석을 방지하고, 어휘 실마리 분석 방법은 주변 어휘들을 어휘 실마리로 이용하여 파생명사 및 합성명사를 인식하며, 동일 실질형태소 분석 방법은 전문에서 3음절 이상의 동일한 실질형태소들을 포함한 어절들을 추출하여 이들이 결합한 어휘들에 근거하여 올바른 분석 결과를 선택하는 방법이다. 실험결과는 이들 세 방법의 통합 시스템이 99.8%의 정확도와 95.3%의 재현율로 파생명사 및 합성명사 후보를 포함한 어절들을 올바르게 분석할 수 있음을 보였다.

앞으로 동일 실질형태소 분석과 어휘 실마리 분석에서 문제로 드러난 접미사와 의미가 비슷하면서 접미사의 음절로 시작하는 2음절 명사들을 자동으로 선정하여 이들이 포함된 어절에 대해서는 동일 실질형태소 분석 방법과 어휘 실마리 분석 방법을 적용하지 않음으로써 정확도를 향상시킬 계획이다. 그리고 어휘 실마리를 대량의 코퍼스에서 추출함으로써 더 많은 어휘 실마리를 선정하고 윈도우 크기를 10어절 이상으로 적용하여 어휘 실마리 적용 방법의 재현율을 개선할 계획이다.

참고 문헌

[1] 김지영, 권혁철, "합성명사의 의미 관계 분석 시스템을 위한 지식베이스 구축 기법," *한국정보과학회 가을 학술발표논문집*, pp. 985-988, 1992.

[2] 남윤진, 옥철영, "말뭉치 분석에 기반한 명사파생접미사의 사전정보 구축," *정보과학회논문지*, 제23권, 제4호, pp. 389-401, 1996.

[3] 박봉래, 황영숙, 임해창, "확장 정의된 유사어절의 분석에 근거한 실시간 미등록어 인식," *한글 및 한국어 정보처리 학술발표논문집*, pp. 222-228, 1996.

[4] 시정곤, *국어의 단어 형성 원리*, 국학자료원, 1994.

[5] 윤보현, "복합명사 구성패턴과 통계정보를 이용한 한국어 복합명사 분석," *고려대학교 전산과학과 석사학위 논문*, 1995.

[6] 임희석, "어절의 중의성 유형 분류에 근거한 한국어 형태소 분석기," *고려대학교 전산과학과 석사학위 논문*, 1994.

[7] 조규빈, *고교문법*, 지학사, 1993.

[8] 차준경, "한국어 파생어의 생산성에 대한 계량적 접근," *고려대학교 철학과 석사학위논문*, 1955.