미등록어 처리가 강화된 복합명사 분해

충북대학교 컴퓨터공학과
김 응 균† · 서 영 훈

Compound Noun Analysis Strengthened Unknown Noun Processing

Eung-Gyun Kim, Young-Hoon Seo
Department of Computer Engineering, Chungbuk National University, Chungbuk, Korea

요 약

본 논문에서는 분해 패턴을 이용한 재사용 분해 알고리즘과 외래어 인식, 이름 명사 인식, 지명 인식에 의한 미등록어 추정을 이용한 복합명사 분해 방법을 제안한다. 재사용 분해 알고리즘은 현재 분해되는 음절보다 짧은 길이의 음절에서 사용된 분해 방법을 재사용하여 분해가 이루어짐을 의미한다. 외래어 인식에서는 한국어 음절에서 비교적 사용 빈도가 낮은 음절들로 외래어가 구성이 됨을 이용한다. 이름 명사는 한국인의 이름 특성에서 한자 독음을 차용하여 작명이 이루어지기 때문에 일정한 수의 음절이 반복적으로 사용되는 점을 이용하여 인식한다. 지명 인식 방법은 지명이 출현하는 패턴을 분석하여 지명 사전의 검색으로 인식한다. 이와 같이 지명 사전에 의한 지명 인식과 알고리즘에 의한 외래어 및 이름 명사 인식 방법을 사용함으로써 미등록어 추정에 정확성을 높이고 분해 정확율 향상에 기여한다. 실험 결과 미등록어가 포함된 약 1,500어절에 대해 약 98%의 정확율이 나타났고, 미등록어가 사전에 모두 등재된 후의 실험에서는 약 99%의 정확율을 보였다.

서 론

일반적으로 한국어 복합명사는 띄어 쓰기를 원칙으로 하나 붙여 써도 무방하기 때문에 기계 번역 분야, 정보 검색 분야, 그리고 맞춤법 검사 분야에서 분해 필요성이 발생하였을 경우 많은 어려움이 따른다. 이를 해결하기 위한 가장 단순한 방법은 모든 복합명사를 사전에 등재하는 것인데 한국어는 단위 명사의 조합이 자유롭기 때문에 조합된 단어의 양이 많아져 모든 복합명사를 사전에 등재하는 것은 매우 비효율적이다. 그래서 모든 복합명사의 사전 등재 대신 복합명사 분해를 위한 여러 방법들이 연구되고 있다.

기존의 연구 내용을 살펴보면 최재혁(1996)은 음절 수에 따른 분해 패턴을 미리 정의해 놓은 상태에서의 복합명사 분리 방법[1]을 제안했다. 이 연구는 이미 구축해 놓은 유형을 이용하기 때문에 분석 우선 순위 상의 오류가 존재하고, 분해 패턴과 사전에 의존적인 성능을 보인다. 윤보현(1997)은 분해 패턴과 접사의 위치에 대한 통계 정보와 단위 명사의 중심어와 수식어로써의 사용 빈도를 이용한 복합명사 분해를 제안하였고,[2] 심광섭(1997)은 대형 코퍼스에서 추출한 음절간 상호 정보를 이용한 분해 방법[3]으로 음절간의 연관도를 계산한 통계적 기법에 기반한 분석 방법을 제안하였다. 강승식(1998)은 네 가지 분해 규칙과 두 가지 예외 규칙을 사용하여 분해 후보들을 생성하고, 분해 후보들에 가중치를 부여함으로써 최적의 분해 후보를 선택하는 분해 알고리즘을 제안하였다.[4] 정래정(1996)은 고유 명사의 출현 패턴을 조사하여 고유 명사가 존재하는 부근에 나타나는 실마리 단어로 고유 명사를 인식하고, 성씨 사전과 이름 사전으로 이름 명사를 인식[5]방법을 제안하였다. 이 연구에서는 사전을 성씨와 이름 사전으로 구성하여 한국인의 이름에서 나타나는 두 번째, 세 번째 음절간의 빈도차 즉 돌림자 사용으로 인한 특성을 고려하지 않았다. 이현민(2000)은 복합명사 분해에 있어 일반적인 좌에서 우 방향 분해 대신 우에서 좌 방향으로의 분해를 제안하였다.[6] 일반적인 좌에서 우로의 분해에서 발생하는 문제는 어느 정도 해결하였으나, 우에서 좌로의 분해에 따른 또 다

†E-mail : jchern@white.chungbuk.ac.kr
E-mail : yhseo@cubcc.chungbuk.ac.kr

른 문제가 발생하였다. "환자의식"의 예를 들면 순방향 분해에서는 "환자+의식"으로 정확히 분해되지만 역방향 분해시 "환+자의식"으로 오분석이 일어난다. "이재성(2001)은 번역문에서 외래어 표기 용례를 자동 구축하기 위한 외래어 인식에 관한 연구[7]에서 음절 정보를 이용한 외래어 인식을 제안하였다.

대부분의 복합명사 분해에서 음절 분해 패턴을 이용하고 있다.[1,2,4,6] 미등록어가 포함된 어절에 대한 분해에서는 단지 미등록어를 제외한 음절이 등록어일 경우, 미등록어와 등록어의 분리 위치를 결정하는 수준에서 이루어지기 때문에 미등록어 어절 중 등록어가 포함된 경우 잘못된 분석을 하는 것으로 나타났다. 따라서 복합명사의 분해 정확도를 향상시키고 미등록어에 대한 의미적 범주를 결정하기 위해 본 논문에서는 외래어 및 이름 명사 그리고 지명 인식을 통하여 미등록어 추정이 포함된 분해 방법을 제안한다.

복합명사에서 미등록어로 인한 분해 오류

복합명사 분해에서 미등록어는 정확율을 떨어뜨리는 주요한 원인 중에 하나이다. 그 중 이름 명사, 외래어, 지명 등이 미등록어로 나타날 때 분해 효율의 하락이 발생할 수 있다. 예를 들면 다음과 같다. 이름 명사가 미등록어일 경우 발생하는 문제로써 사전에만 의존해 분해하고 별도의 이름 명사 처리를 하지 않았을 경우 "김대중대통령"에서 "김대중"이 미등록어이고 "대중"과 "대통령"이 등록어일 경우 "김+대중+대통령"으로 분해가 된다. 그리고 외래어가 미등록어일 경우 "브루나이국왕"에서 "브루+나이+국왕"으로 분해가 된다. "브루나이"라는 미등록 외래어에서 "나이"라는 단위 명사가 포함되어 있음으로 해서 발생하는 오분석이다. 지명 역시 "사창동사거리"에서 분해 빈도가 가장 높은 2|2|2로 분해 패턴으로 분해가 이루어졌을 때 "사창", "동사", "거리"모두 사전에 등재돼 있는 단어들이기 때문에 "사창+동사+거리"로 오분석될 가능성이 높다. 이와 같이 미등록어 사이에 등록어가 포함되어 있을 때 별도의 미등록어 처리 방법이 없다면 오분석이 일어날 가능성이 높다.

재사용 분해 알고리즘

한국어에서 복합명사는 3음절을 제외했을 때 4, 5, 6음절이 전체 복합명사의 97%을 차지한다.[11] 따라서 4, 5, 6음절에 대한 복합명사 처리만으로 대부분의 복합명사를 처리할 수 있다. 본 논문에서는 복합명사 분해시 분해 알고리즘의 재사용성에 기반한 6음절 까지의 복합명사 분해를 시도한다. 먼저 기분석된 복합명사 패턴을 검토하여 해당 음절에 나타나는 모든 분해 패턴을 조사하고 가장 높은 확률값을 가지는 분해 패턴 순으로 분해를 실시한다.

분해 패턴 적용의 예를 들면 "컴퓨터공학"은 5음절이므로 5음절 분해 패턴 중 우선 순위가 가장 높은 2|3으로 분해를 시도한다. "컴퓨|터공학"이라고 분해하고자 할 때 사전에는 "컴퓨"라는 단어가 없기 때문에 분해에 실패하고 다음 패턴인 3|2패턴으로 진행한다. "컴퓨터|공학"으로 분해되어 시스템 사전의 검색을 통해 두 단어 모두 사전에 등재되어 있기 때문에 "컴퓨터+공학"으로 결과를 출력한다.

다음은 각 음절 별 분해 패턴이다.

3음절 : [1|2] [2|1]

4음절 : [2|2] [1|3] [3|1]

5음절 : [2|3] [3|2]

6음절 : [2|2|2] [3|3] [1|2|3] [5|1] [1|5] [2|4]

한국어에서 단위 명사는 95%이상이 4음절을 초과하지 않는다. 이런 단위 명사들의 조합으로 복합명사가 만들어지는데 단위 명사들 중 대부분은 2음절 또는 3음절로 이루어지고 여기에 1음절 접사가 붙는 형태가 대부분이다. 따라서 접사 처리와 3음절 분해만 할 수 있다면 재사용 분해 알고리즘과 사전 탐색으로 5음절 이상의 분해도 가능함을 알 수 있다. 예를 들면 5음절 복합명사를 분해하고자 할 경우 5음절은 3음절과 2음절 또는 1음절과 4음절로 분해가 이루어진다. 그런데 여기서 4음절 역시 두개의 2음절의 조합으로 구성되는 경우가 대부분이기 때문에 1음절과 두개의 2음절로 구성되었다고 볼 수 있다. 즉 앞의 1음절을 접두사로 보면 접두사 1음절과 뒤에 오는 2음절이 하나의 단어로 구성되고 다시 뒤에 오는 2음절에 하나의 단위로 구성되면 5음절은 3음절과 2음절 단위 명사의 조합으로 이루어졌음을 알 수 있다. 따라서 2음절 단위 명사에 대한 사전 검색과 3음절 단위 명사 또는 접사를 포함한 복합명사에 대한 검색과 분해만 이루어 진다면 분해가 가능하다. 5음절 이외에 다른 음절도 접사 처리가 포함된 3음절 분해 방법과 분해 패턴이 포함된 각 음절별 분해 알고리즘만으로 분해가 가능하다고 할 수 있다.

외래어 인식

외래어라 함은 원래 외국어였던 것이 국어의 체계에 동화되어 사회적으로 그 사용이 허용된 단어로써 한국어의

음절 중에 비교적 사용 빈도가 낮은 음절들로 구성되어있다. 외래어는 대부분 인명 또는 지명이기 때문에 색인어로써 중요한 역할을 한다. 하지만 사전 검색으로 외래어를 인식하기 위해 모든 외래어를 사전에 등재하는 것은 현실적으로 어려운 일이다. 외래어의 특성상 수백여개국의 말들이 계속적인 생성과 소멸이 발생하고 국어 체계로의 계속적인 유입 때문에 사전 등재는 불가능에 가깝다. 따라서 외래어 인식을 위한 별도의 외래어 인식 알고리즘을 요구한다. 본 논문에서는 외래어 음절의 출현 특성과 음소 결합 특성을 이용한다.

1. 음절 출현 특성

음절 출현 특성이란 한국어에서 쓰이는 음절의 빈도수와 외래어에서 쓰이는 음절의 빈도수를 계산하여 얻어진 특성으로 외래어인지를 판단하는 기준으로 사용한다. 본 연구실에서 소유한 형태소 분석용 사전에서 외래어를 수작업으로 삭제한 후 남은 어절들을 한국어 어절이라고 가정하고, 이를 기반으로 출현하는 모든 음절에 대한 빈도수를 계산하여 얻어진 결과를 한국어 음절에 대한 통계값으로 사용한다. 그리고 웹사이트 encyber[8]에서 "외래어 표기 용례"에서 추출한 외래어 어절 약 3,000여개와 웹사이트에서 수작업으로 얻은 2,000여개 엔트리를 추가하여 전체 5,000여 엔트리에서 각 음절의 출현 빈도수를 계산하여 외래어 음절에 대한 통계값으로 사용한다. 또한 한국어 음절과 외래어 음절을 비교하여 한국어 음절에는 나타나지만 외래어 음절에는 나타나지 않는 음절을 Onlykorea(앞으로는 OK로 표시한다)라고 명명하고 이와는 반대로 외래어 음절에는 나타나지만 한국어 음절에는 나타나지 않는 음절을 Onlyforeign(앞으로는 OF으로 표시한다)으로 명명한다. OK와 OF은 외래어로 의심되는 어절이 입력되었을 경우 해당 어절에 하나의 음절이라도 OK 또는 OF이 포함되었을 경우 즉각적인 외래어 여부를 판단할 수 있다. 해당 어절에 OK리스트 중 하나의 음절이라도 포함되어 있다면 즉시 한국어로 판별하고 OF리스트 중 하나의 음절이라도 포함되어 있다면 외래어로 판별한다. 외래어 인식에 있어 음절 정보 사용에 대한 예를 들면 "월드컵예선"에서 "월드컵"이 미등록어일 경우 분해 패턴의 실패로 추정을 하기 위해 "월드컵"을 외래어 인식을 시도한다. 월드컵에서 "드"와 "컵"이 한국어에서는 드물게 출현하는 음절이고 외래어에서는 비교적 출현 빈도가 높은 음절이므로 외래어로 인식된다. 다음은 "월드컵" 각 음절의 통계값이다. 이 통계값은(외래어에서의 출현 음절 수/총외래어 음절 수)-(한국어에서의 출현 음절 수/총한국어 음절 수)로써 얻어진다.

월 : -15.2

드 : 183.7

컵 : 1.6

따라서 "월드컵"의 외래어 출현 빈도는 세 음절에 대한 통계값의 합(-15.7+183.7+1.6=170.1)으로써 실험에 의해 얻어진 외래어 임계치와 비교하여 임계치 보다 크다면 외래어일 가능성이 높다고 판단하고 그렇지 않을 경우에는 한국어로 판단한다. "월드컵"이 외래어로 판단이 되면 "월드컵예선"은 3|2로 분해가 이루어 진다.

2. 음소 결합 특성

음소 결합 특성도 음절 특성과 마찬가지로 음소의 결합 특성이 외래어에서만 나타나는 음절이 포함된 어절에 외래어의 가능성을 부여하는 방법이다. 음소 결합 특성이란 외래어 음절에만 나타나는 고유한 결합 특성을 의미한다. 예를 들면 "커피숍"이라는 어절에서 마지막 음절의 "숍"의 예를 들면 중성의 'ㅛ'과 종성의 'ㅂ'만으로 초성에는 관계없이 외래어라는 특성을 얻을 수 있다. 이와 같은 음소 결합 특성은 한국어 어절과 외래어 어절에 대한 모든 조합을 비교하여 얻어진다.

Table 1은 세 가지 형태의 음소 결합 특성을 보여 준다. 타입 1은 초성과 중성의 결합 특성을 의미하고 2는 중성과 종성, 3는 초성과 종성의 결합 특성을 의미한다. 위와 같이 구축된 외래어 음소 결합 특성을 이용하여 보다 높은 정확도의 외래어 인식을 시도 할 수 있다. 타입 1에 해당하는 조합 규칙 19개와 타입 2에 해당 하는 조합 규칙 14개를 발견했으나 타입 3은 조합 규칙을 발견할 수 없었다(Fig. 1)

이름 명사 인식

한국어의 이름은 95%이상이 3음절로 이루어져 있고 거의 대부분의 이름 명사가 작명시에 제한된 범위 내의 음절(한자 독음)을 사용하고 있다. 본 연구에서는 이와 같은 배경으로 이름 명사 인식을 3음절로 제한시키고 순수 한글 이름을 제외한 한자 독음으로 구성된 이름 명사 인식을 시도한다.

한국 통신에서 제공하는 전화번호부에서 추출한 이름 명사 리스트로부터 외국인 이름을 제거한다. 외래어 이름을

Table 1. 음소 결합 특성 예

초성	중성	종성	타입
ㅂ	ㅟ	All	1
All	ㅛ	ㅂ	2
-	-	-	3

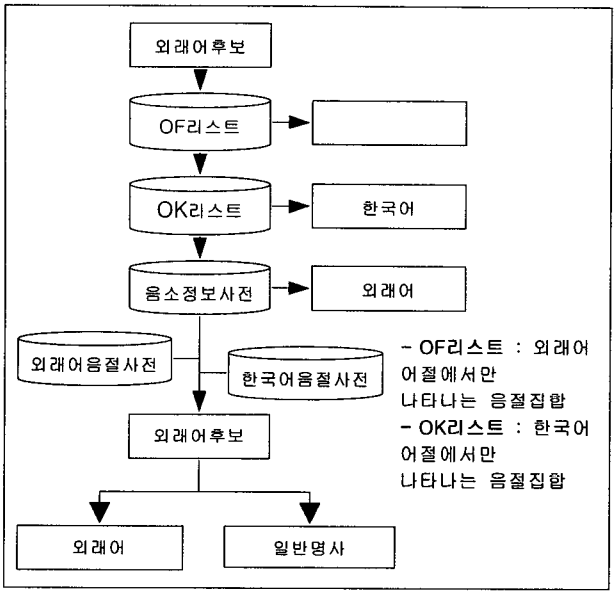

Fig. 1. 외래어 인식 순서도.

제거한 후 모든 성씨를 조사한 결과 400여개의 성씨를 확인할 수 있다. 그러나 400여개 성씨를 모두 사용함써 과분석이 심화되기 때문에 시스템상에서는 상위 50개의 성씨만을 선택하여 사용한다. 그리고 2번째 이름 명사의 음절과 3번째 이름 명사의 음절에 대한 빈도수를 계산하여 첫번째 음절 사전(성씨), 두 번째 음절 사전, 세 번째 음절 사전을 구축한다. 2번째 음절과 3번째 음절을 분리하여 통계 정보를 구축한 것은 한국 이름에는 아직까지 돌림자라는 것이 존재하기 때문에 이런 음절 별 위치 특성을 반영하기 위함이다. 또한 보다 정확한 이름 명사 인식을 위해 2, 3음절에 나타나는 Bigram 음절 정보를 이용하여 이름 사전을 구축하였다. 그리고 이름 명사 인식의 효율성과 복합명사 분해의 정확율 향상을 위해 1, 2, 3음절 실마리 단어 리스트를 구축하였다(Fig. 2).

1. 실마리 단어 구축

이름 명사 인식의 효율을 높이기 위해 이름 뒤에 올 수 있는 약 50여개의 2음절 실마리 단어를 구축하였다.

실마리 단어가 검색되면 실마리 단어 앞의 3음절은 이름 명사일 가능성이 높기는 하지만 실마리 단어 앞의 모든 3음절 단어가 이름 명사는 아니므로 3음절 명사에 대해 이름 명사 인식을 시도한다. 실마리 단어 앞의 3음절이 이름 명사가 아닐 경우의 예를 들면

현대차[사장] :

사장 : [실마리 단어]

에서 "사장"이라는 실마리 단어를 인식하였지만 실마리 단어 앞 단에 위치한 "현대차"가 사람 이름인지의 여부를 이

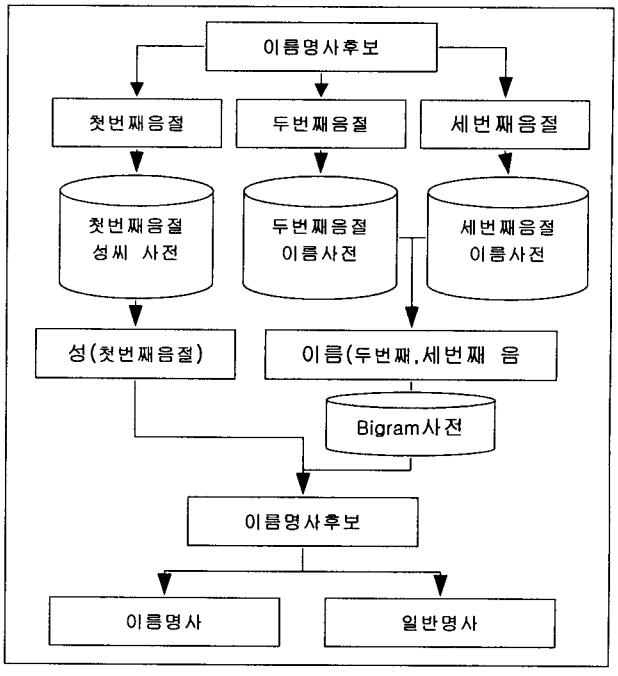

Fig. 2. 이름명사 인식 순서도.

Table 2. 1, 2음절 실마리 단어의 예외 상황의 예

실마리 단어	예 외	예외 상황	예외 어절
1음절	군	육	육군
	님	형	형님
2음절	판사	출	출판사
	사장	이	이사장

름 명사 인식을 통하여 확인할 수 있다.

이외에도 1음절과 3음절 실마리 단어를 수집하였고, 1음절 실마리 단어로는 "씨", "군", "옹", "님", "양", "전"이 있다. "님"과 같은 경우 "성철스님"과 같이 마지막 음절에 "님"이라는 실마리 단어가 있지만 앞의 3음절이 이름 명사가 아니기 때문에 이렇게 실마리 단어와 결합하여 2음절 명사의 형태를 갖추는 어절을 수집하여 실마리 단어에 대한 예외 처리를 한다. 수집한 대부분의 3음절 실마리 단어는 거의 대부분이 2음절 실마리 단어에 접미사가 붙은 형태이기 때문에 이들을 제외한 "대변인"과 같이 2음절 실마리 단어에 포함되지 않은 5개의 단어만이 실마리 단어로 등록되었다. 그래서 실마리 단어는 1음절 6개 2음절 51개 3음절 5개로 총 62개를 사용한다. 실마리 단어는 복합명사 분해의 정확도를 개선하고 수행 시간을 단축할 수 있고 실마리 단어 앞 단에 이름 명사가 오는 경우가 대부분이므로 부수적 효과로서 이름 명사 인식을 위한 좋은 위치를 제공한다.

Table 2는 1, 2음절 실마리 단어에 대한 예외 처리 부분

이다. 실마리 단어가 해당 어절의 마지막 부분에 위치했을 때 바로 앞 음절을 검사하여 예외 상황 인지를 판단한다. 예를 들면 "영진출판사"의 경우 마지막에 오는 "판사"라는 실마리 단어만을 보고 "영진출"을 이름 명사 또는 일반 명사로 추정하는 경우가 발생하기 때문에 "판사"라는 실마리 단어가 검색되었을 때 앞 단의 1음절을 추가로 검색하여 "출"인지의 여부를 확인하여 "영진+출판사"로 분해가 이루어진다.

지명 사전 구축

지명은 이름이나 외래어와는 달리 엔트리 수가 제한적이므로 구축이 가능하다. 우리 나라의 경우 8개의 단위로 세분화되어 있어 마지막 음절의 체크만으로 어느 정도 지명 여부를 가늠해 볼 수 있다. 지명 사전 구축을 위해 사이트[9]에서 우편 번호를 다운 받아 마지막 음절이 "도, 시, 구, 군, 동, 면, 읍, 리"로 끝나는 어절만을 이용하여 15,860개의 엔트리의 지명 사전을 구축한다. 지명 인식은 별도의 알고리즘보다는 일정 수의 지명이 존재하기 때문에 지명 사전을 별도로 구축하고, 지명이 나타나는 위치를 고려하여 인식하는 방법을 제안한다. 지명이 나타나는 위치에 대한 하나의 예를 들면 "수곡동|지점"과 같이 5음절 분해에서 3|2분해의 앞단의 3을 의미한다.

실험 및 분석

복합명사 실험 전 이름 명사와 외래어 인식에 따라 복합명사의 성능에 영향을 미칠 수 있기 때문에 이름 명사와 외래어 인식에 대한 성능 평가를 먼저 실시하였다.

외래어 인식은 태깅된 코퍼스에서 추출한 일반 명사에 대해서 테스트를 실시하였다. 외래어와 이름 명사의 성능 측정을 위해 정보 검색에 쓰이는 재현율과 정확율을 사용하였다(Table 3).

* 재현율(Recall) :
  시스템이 올바르게 인식한 외래어 어절 수 /
  총 외래어 어절 수
* 정확율(Precision) :
  시스템이 올바르게 인식한 외래어 /
  시스템이 외래어로 인식한 총 어절 수
* 총 외래어 어절 수 : 551 어절
  시스템이 올바르게 인식한 외래어 어절 수 : 478
  시스템이 외래어로 인식한 총 어절 수 : 507

MATEC99의 태깅된 말뭉치[12]로부터 6,828개의 단위 명사를 추출, 수작업으로 한국어와 외래어를 구분한 후 실험을 진행한다. 전체 어절 수 6,828개의 어절 중 외래어인 어절 수는 551개의 어절, 그 중 시스템은 478개의 어절을 인식 87%의 재현율 보였고, 시스템이 올바르게 인식한 외래어 어절을 총 외래어 어절수로 나누어 계산하였을 때 94%의 정확율이 나타났다. 외래어의 재현율과 정확율은 trade-off 관계에 있어 재현율을 높인다면 정확율이 떨어지고 반대의 경우는 재현율을 낮춘다면 정확율을 높일 수 있다. 전체 6,828어절 중 2,534어절이 OF과OK 리스트에 의해 판별되어 전체 어절 수에 대한 포함 비율은 37%정도로 나타났다. 포함 비율이란 전체 어절 중 OK와 OF리스트에 의해 판단되어진 어절의 비율을 의미한다.

외래어 이외에 이름 명사에 대한 정확도 테스트를 별도로 실시한다. 이름 명사는 보통 명사와 중의적 의미를 가지는 것들이 많이 존재하기 때문에 성능평가에 있어서 유동적인 부분이 존재한다. 예를 들면 "현미경"은 보통 명사이지만 이름 명사로도 인식할 수 있기 때문에 본 시스템의 이름 명사 인식에서는 잘못된 인식으로 판단하지 않는다. 이름 명사 인식을 테스트하기 위해 3음절 단위 명사를 웹상에서 추출 전체 2,190어절에 대해서 외래어 인식과 동일한 방식으로 평가하였다. 전체 2,190어절에 대해서 수작업으로 이름 명사로 판단할 수 있는 어절들을 수집하고 이를 바탕으로 하여 시스템의 성능을 평가한다. 이름 명사는 3음절로 제한하였기 때문에 이름 명사 외에 테스트에 사용된 모든 단위 명사는 3음절만을 추출한다(Table 4).

* 재현율(Recall) :
  시스템이 올바르게 인식한 이름 명사 어절 수 /
  총 이름 명사 어절 수
* 정확율(Precision) :
  시스템이 올바르게 인식한 이름 명사 어절 수 /
  시스템이 인식한 전체 이름 명사 어절 수
  전체 어절 수 : 2,190
* 총 이름 명사 어절 수 : 712개
  시스템이 이름 명사로 인식한 어절 수 : 729
  시스템이 올바르게 이름 명사로 인식한 어절수 : 671

보통 명사 중에도 이름 명사로 인식될 수 있는 것들이 상당수 존재 하였다. 예를 들면 "현미경, 최고수, 전성기" 등

Table 3. 외래어 인식의 정확율과 재현율

구 분	정확율	재현율
성 능	94%	87%

Table 4. 이름 명사 인식의 정확율과 재현율

구 분	정확율	재현율
성 능	92%	94%

Table 5. 복합명사 테스트 결과

분 석 / 음 절	어절 수	오분석 어절 수	정확율
4음절	755	10	98.6%
5음절	437	14	96.8%
6음절	369	5	98.6%
전 체	1561	29	98.1%

Table 6. 복합명사 테스트 결과 중 음절 별 출현 비율

미등록어 / 음절 수	이름 명사	외래어	지명
4음절	8	2	0
5음절	30	7	2
6음절	8	4	2

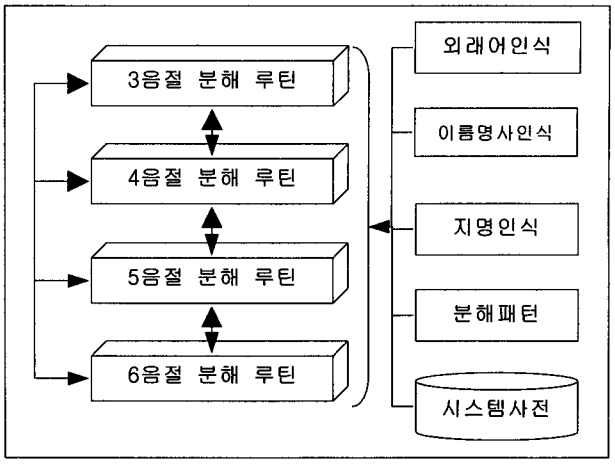

Fig. 3. 전체 시스템 구성도.

이 있다. 그리고 전체 이름 명사 712개 중에 23개 정도가 보통 명사로도 존재하는 이름들이었기 때문에 이름 명사의 정확도는 유동적일 수 있다.

외래어 인식과 이름 명사 인식에서 90%의 이상의 성능을 보였고 이를 바탕으로 하여 복합명사 실험을 실시하였다. 복합명사 분석을 위해서 웹에서 수집한 4음절에서 6음절까지의 복합명사 1,561어절을 대상으로 한다(Table 5).

각 음절별로 정확도를 조사했을 때 4음절일 경우가 근소한 차이로 가장 높은 정확율을 보였다.

Table 6은 각 음절별 이름 명사, 외래어, 지명에 대한 출현 빈도수이다. 모든 음절에서 외래어와 이름 명사가 출현한 반면 지명은 4음절에서는 출현하지 않았다. 상대적으로 5음절에서 이름 명사 출현 비율이 다른 음절보다 월등히 높은 이유는 이름 명사 뒤에 2음절 실마리 단어가 오는 경우가 많았기 때문이다.

오류 유형을 살펴보면 접사의 분리 유무를 철저히 검사하지 않았을 경우 시스템 정확도는 98.1%로 나타났다. 접사의 분리 유무란 "건축사협회"에서 "건축+사+협회"로 분해되지 않고 "건축사+협회"로 분해되었을 경우를 말한다. 이름 명사와 외래어 그리고 지명 등이 정확율 상승에 소폭 작용했음을 Table 6를 통해 확인할 수 있다. 오류 유형을 살펴보면 4음절에서 발생한 오류는 4음절에 나타난 전체 오류 어절 10개 가운데 8개가 접사로 인한 분리 오류였다. "탈당자수"의 경우 "탈당+자수"로 분리 되었다. "탈당자+수"가 옳은 분석이지만, 두 가지 분해패턴에 해당하는 2|2와 3|1에 해당하는 단어들이 모두 사전에 등재된 단어들로 구성이 되어 있지만 2|2 분해 패턴이 3|1 분해 패턴에 우선하기 때문에 발생하는 오류이다. 통계적 방법을 사용하더라도 "탈당자"의 통계치 보다 "탈당"과 "자수"의 각각의 통계치 합이 더 높기 때문에 통계적 방법으로는 해결할 수 없는 분석 오류가 존재한다. 또한 5음절에 나타난 오류 어절을 살펴보면 2|3과 3|2 사이의 중의적 분해에 의한 오류가 주를 이루고 있다. 5음절에 나타난 전체 오류 어절 14어절 중 10어절이 중의적 분해에 의한 오류로 나타났다. "진료비실사"에서 보면 "진료+비+실사"로 분해가 이루어져야 하지만 분해 우선 순위에 의해 2|3 분해가 먼저 이루어진고, 또한 "비실"이라는 것이 사전에 존재하는 단어이고 "사"가 접미사로 사용빈도가 높기 때문에 결과는 "진료+비실+사"의 형태로 잘못된 결과를 출력한다. 그리고 마지막으로 6음절에 나타난 오류 어절을 살펴보면 미등록어 포함에 따른 오류로 미등록어와 접사에 의한 과분석이 오류 유형으로 나타났다. 또한 외래어로 인한 오류도 발생하였는데 "대림레미+콘"에서 "대림"이 미등록어이기 때문에 "대림"과 "레미콘"사이가 분리되지 못하고 "콘"만 접사로 분리되어 잘못된 분석으로 나타났다. "레미콘"이 외래어로 추정되지 못한 이유는 외래어 인식에 있어 2음절 단위 명사+3음절 외래어로 출현하는 비율이 낮아 2|3분해에서 3음절에 대한 외래어 인식을 시도하지 않았기 때문이다. 앞에서 설명한 오류 유형 이외에도 축약어들이 포함되어 잘못된 분석으로 나타난다. "대민경협차관"에서 "경협"이 "경제협력"이라는 단어의 축약어인데 축약어 대부분이 미등록어일 가능성이 높기 때문에 오분석의 가능성을 내포하고 있다(Fig. 3).

결 론

본 논문에서는 재사용 분해 알고리즘과 외래어와 이름 명사, 지명 인식을 통한 미등록어 추정이 포함된 복합명사 분해 방법을 제안하였다. 실험의 분석 정확도를 관찰하기 위해 중복 데이터의 허용은 배제하였다. 그 결과 약 98.1%

의 분해 정확도를 보였고, 미등록어를 모두 사전에 등재했을 때의 정확도는 98.9%였다. 실험 결과 홀수 음절일 때 접사의 출현이 더욱 빈번하기 때문에 짝수 음절보다는 홀수 음절에서 분해 정확도가 떨어지는 것으로 나타났다. 실험에서는 4음절에서 6음절까지의 어절을 분석하였지만 한국어 단위 명사의 길이를 고려하면 전체 단위 명사 중 95% 이상의 단위 명사는 4음절을 초과하지 않기 때문에 이후의 연구에서 3, 4, 5, 6음절의 분해 알고리즘을 조합하여 7음절 이상의 분해에서도 이용 가능하다. 7음절의 예를 들면 3|4 또는 4|3의 분해 알고리즘을 이용하여 출력된 결과를 조합함으로써 7음절의 분해가 이러질 수 있고 이러한 방식의 확장으로 7음절 이상의 복합명사에서도 분해가 가능하다. 각 음절별 분해에 있어 미리 만들어진 분해 패턴을 이용함써 사전상에 등재된 단어들의 조합으로 구성된 복합명사의 경우, 큰 어려움 없이 분해가 가능하기 때문에 7음절 이상의 분해에서도 최소한의 분해 효율이 보장된다.

향후 연구로는 외래어 인식에 있어 bigram 정보의 추가적 이용과 지명 인식에 있어서 마지막 음절에 오는 접미사 없이 지명 검색이 가능한 방법에 대한 추가 연구가 진행중이다. 또한 이름 명사, 외래어, 지명 이외의 개체명 인식 기법을 도입한 복합명사의 미등록어 추정에 관한 연구가 필요하다.

REFERENCES

1) 최재혁(1996) : "음절수에 따른 한국어 복합 명사 분리 방안", 한글 및 한국어정보처리
2) 윤보현, 조민정, 임해창(1997) : "통계정보와 선호규칙을 이용한 한국어 복합명사의 분해", 정보과학회논문지(*B*), 제 *24*권, 제*8*호
3) 심광섭(1997) : "합성된 상호 정보를 이용한 복합명사 분리", 정보과학회 논문지(*B*), 제*24*권, 제*11*호, *pp1307-1317*
4) 강승식(1998) : "한국어 복합명사 분해 알고리즘", 정보과학회논문지(*B*), 제*25*권, 제*1*호, *pp172-182*
5) 정래정, 김준태(1996) : "고유명사의 출현 패턴을 이용한 색인의 성능 향상에 관한 연구" 한글 및 한국어정보처리
6) 이현민, 박혁로(2000) : "복합명사의 역방향 분해 알고리즘", 한글 및 한국어정보처리
7) 이재성(2001) : "번역문에서의 외래어 표기용례 자동 구축" *Journal of Reach Institute for Computer and Information Commmunication*
8) www.encyber.com.
9) www.koreapost.go.kr.
10) 충북대 자연어처리연구실(2002) : "형태소 분석용 시스템 사전"
11) 강승식(1993) : "한국어 형태소 분석을 위한 복합 명사의 인식 방법", 한국 인지과학회 춘계 학술 발표대회 논문집, *pp175-189*
12) ETRI(1999) : *"Tagged corpus Matec99", ETRI*