복합명사의 역방향 분해 알고리즘

이 현 민† · 박 혁 로††

요 약

본 논문에서는 단위명사 사전과 접사 사전을 이용하여 한국어 복합명사를 분해하는 새로운 알고리즘을 제안한다. 한국어 복합명사는 그 구조에 있어서 중심어가 뒤에 나타난다는 점에 착안하여 본 논문에서 제안한 분해 알고리즘은 복합명사를 끝음절에서 첫음절 방향 즉 역방향으로 분해를 시도한다. ETRI의 태깅된 코퍼스로부터 추출한 복합명사 3,230개에 대해 실험한 결과 약 96.6%의 분해 정확도를 얻었다. 미등록어를 포함한 복합명사의 경우는 77.5%의 분해 정확도를 나타냈다. 실험에 사용된 데이터중의 미등록어는 대부분 접사를 포함한 파생어로서, 제안한 복합명사 분해 알고리즘은 접사가 부착된 미등록어 분석에 있어서 보다 높은 분석 정확도를 나타냄을 알 수 있었다.

A Reverse Segmentation Algorithm of Compound Nouns

Hyun-Min Lee† · Hyuk-Ro Park††

ABSTRACT

In this paper, we propose a new segmentation algorithm for compound noun analysis in Korean. The algorithm segments a compound noun into a sequence of unit nouns and affixes using a unit noun dictionary and an affix dictionary. In most cases, the head of a compound noun appears at the end of the word, the proposed algorithm tries to segment the given compound noun from the end of the word to the beginning of the word. To evaluate the accuracy of the proposed algorithm, an experiment was conducted with 3,230 compound nouns which is extracted from ETRI tagged corpus. Experimental results shows that the accuracy of the proposed method is 96.6% on the average. In case of compound nouns with unknown words, the accuracy drops to 77.5%. From the experiment, it become clear that the proposed algorithm outperformed other methods in case of compound nouns with unknown words.

키워드 : 복합명사 분해(compound noun segmentation), 미등록어(unknown word)

1. 서 론

복합명사란 둘 이상의 명사가 결합된 형태를 말한다. 한글에서 명사와 명사 사이는 띄어쓰기 자유롭고 그 결합형태도 다양하다. 복합명사의 표기법이 실생활에서는 문제가 되지 않지만 정보 검색 시스템이나 기계 번역 시스템과 같은 정보 처리 시스템에서는 심각한 문제를 야기시킨다. 복합명사에 대한 처리 문제는 복합명사를 사전에 수록하는 방법이 있으나 수록해야할 복합명사의 수가 너무 많기 때문에 단위명사만을 사전에 등록한 후 주어진 복합명사를 단위명사로 분리하는 방법에 대한 연구가 활발히 진행되고 있다[4, 6-9].

복합명사를 분리할 때 가장 어려운 문제점은 중의적 분해 문제이다. 예를 들어, 복합명사 '정보통신망'에서 '정', '보', '통', '신', '망', '정보', '보통', '통신', '신망', '정보통' 등의 명사가 추출될 수 있다. 추출 가능한 명사, 즉 단위명사(unit noun)로 '정보통신망'을 분해했을 때, '정 + 보 + 통 + 신 + 망', '정 + 보통 + 신 + 망', '정 + 보 + 통신 + 망', '정보 + 통신 + 망' 등으로 여러 개의 분해가 가능하게 되어 올바른 후보를 찾는데 어려움이 발생하게 된다.

중의적 분해 문제와 더불어, 미등록어를 포함한 복합명사의 분해 문제도 있다. 미등록어(unknown word)는 사전에 등록되지 않은 단위명사로서 대표적으로 고유명사, 외래어, 신조어, 접사 파생어[1] 등을 들 수 있다. 예를 들어, 복합명사 '거장니콜라스루빈스타인'의 경우, '거장+니콜라스+루빈스타인'으로 분리되어야 하지만, '니콜라스'와 '루빈스타인'이 고유명사인 까닭에 사전 탐색에 실패

† 정 회 원 : (주)에이랩 전임연구원
†† 종신회원 : 전남대학교 전산학과 교수
논문접수 : 2001년 2월 20일, 심사완료 : 2001년 7월 19일

1) 접두사, 접미사가 결합된 명사로서 예를 들어, '명료화', '대공사' 등을 들 수 있다.

하게 되어 올바른 분해 결과를 얻기가 어렵다. 따라서, 복합명사 분해 문제는 중의적 분해 문제와 미등록어를 포함한 복합명사의 분해 문제를 해결함으로써 정확도를 높일 수 있다.

본 논문에서는 한국어 복합명사는 그 구조에 있어서 중심어가 뒤에 나타나는 점에 착안하여 중심어를 먼저 분해하기 위해 복합명사를 끝음절에서 첫음절 방향 즉 역방향으로 분해해 가는 역방향 분해 알고리즘을 제시한다. 제시한 역방향 분해 알고리즘은 미등록어를 포함한 복합명사 역시 분해할 수 있으며, 특히 접사를 포함한 파생어에 대한 비교적 정확한 처리의 기능도 제공한다.

본 논문의 구성은 다음과 같다. 2장에서는 복합명사 분해와 관련된 기존 연구들을 살펴보고, 3장에서는 복합명사 분해를 위한 역방향 분해 알고리즘을 제시하며, 4장에서는 제시한 알고리즘을 이용하여 실험하고 그 결과를 분석하며, 마지막으로 5장에서는 결론을 맺는다.

2. 관련 연구

기존 복합명사 분해에 대한 접근 방법은 크게 사전에 기초한 방법[4, 8]과 통계에 기초한 방법[2, 6, 7]으로 구분할 수 있다. 사전에 기초한 방법은 일단 일반 전자사전을 이용하여 복합명사를 분해한 후 여러 가지 휴리스틱을 이용하여 모호성을 제거하는 방법이다. 그러나, 이 방법은 각종 휴리스틱이 경험에 의해 만들어지므로 그 처리 범위가 모호하고, 미등록어를 포함하는 복합명사에 대해서는 분석 방법이 없다는 단점이 있다. 통계에 기초한 방법에서는 코퍼스를 이용하여 각 단어의 출현 확률을 구한 다음 이들 단어 출현 확률을 이용하여 복합명사를 분해하는 방법을 적용한다.

강승식[4]은 복합명사를 단위명사들로 분해하는 방법으로 네 가지 분해규칙과 두 가지 예외규칙을 사용하여 가능한 분해 후보들을 생성하고, 분해 후보들에 대해 가중치를 부여함으로써 최적 후보를 선택하는 알고리즘을 제안하였다. 이 알고리즘은 미등록 단위명사가 포함되어 있는 복합명사의 분해뿐만 아니라, 복합명사의 길이에 상관없이 적용된다. 또한, 분해 후보가 하나 이상 생성되면 각 분해 후보마다 가중치를 부여하고 가중치가 가장 높은 후보를 선택한다.

심광섭[6]은 합성된 상호 정보(Mutual Information)를 이용하여 띄어쓰기가 되어 있지 않은 한국어 복합명사를 단위명사로 분리하는 알고리즘을 제시하였다. 합성된 상호 정보는 네 가지 유형의 음절간 상호 정보를 합성한 것으로서 주어진 복합명사에서 단위명사로 분리 가능한 지점을 선택하는데 사용된다. 분리된 각 부분이 모두 단위명사가 될 때까지 분리 과정을 반복하며, 만약 분리된 명사가 사전에 등록된 단위명사가 아닐 경우에는 분리된 명사에 인접한 음절을 하나씩 추가하면서 사전을 탐색하는 방법으로 분리 가능한 명사를 찾게 하였다. 또한, 효율적인 최장 명사를 발견하기 위해 2음절 사전 형식을 사용하였다.

윤보현[7]은 통계 정보와 선호 규칙을 이용하여 한국어 복합명사를 단위명사로 분해하는 방법을 제안하였다. 통계 정보로는 1음절 접사 빈도, 그리고 2음절 또는 3음절 단위명사가 복합명사에서 사용된 위치정보와 빈도정보를 이용하였다. 선호 규칙은 중의적 분해를 일으키는 단위명사의 개수가 다를 때, 단위명사의 개수가 적은 복합명사의 분해 패턴을 올바른 분해 패턴으로 선호하는 규칙이다.

기존 연구들의 대부분은 일반적인 복합명사의 분해에 대해서는 비교적 높은 정확도를 보이고 있다. 그러나, 복합명사의 중의적 분해 문제나, 복합명사에 미등록어가 포함되어 있을 경우에는 비록 처리 방안이 제안되어 있기는 하지만, 만족할 만한 수준이 되지는 못한다. 또한, 기존의 복합명사 분해 방법들은 1음절 단위명사에 대한 처리를 고려하지 않았고, 3음절 복합명사의 분해 문제도 배제하고 있다.

3. 복합명사 분해 알고리즘

복합명사 분해시 고려해야 할 사항으로는 중의적 분해 문제와 미등록어를 포함한 복합명사의 분해 문제를 들 수 있다.

복합명사의 중의적 분해 문제는 복합명사 내에 다양한 단위명사가 존재하여 여러 가지 형태로 분해되는 결과가 산출되므로 이 중에서 한가지를 선택해야 하는 문제이다. 예를 들어, 어절 '대학생선교회'에서 복합명사를 분해한 결과, 두 가지 형태 '대학+생선+교회'와 '대학생+선교회'로 분해되어 올바른 분해인 '대학생+선교회'를 선택할 수 있어야 한다.

또한, 미등록 명사가 포함된 복합명사 분해 문제는 복합명사를 이루는 단위명사가 사전에 등록되지 않아서 복합명사를 분해하는 데 실패하는 문제이다. 예를 들어, 어절 '건축사협회'에서 단위명사 '건축'과 '협회'는 사전에 등록되어 있으나 3음절 단위명사 '건축사'가 없어서 복합명사의 분해를 실패하게 된다. 복합명사를 정확히 분해하기 위해서는 '건축사'와 같은 미등록어가 사전에 등록되어야 하지만, 모든 단위명사를 사전에 등록하기는 불가능하므

로, 미등록 단위명사가 포함된 복합명사도 분해 할 수 있어야 한다.

이들 문제점을 해결하기 위해서 본 논문에서는 다음의 네 가지 방법을 제안한다.

3.1 복합명사 분해 전략

3.1.1 단위명사 사전을 기반으로 한 분해

복합명사를 분해하기 위해 사전을 이용한다. 단위명사 사전은 2음절 이상의 명사로만 구성한다. 단위명사 사전에서 1음절 명사를 제외시킨 이유는 중의적 분해 문제를 완화하기 위해서이다.

3.1.2 최장일치 단위명사 우선 분해

사전탐색은 최장일치 되는 단위명사를 우선으로 분해하도록 처리한다. 예를 들어, '민주주의'가 단위명사 사전에 있을 경우, '민주'와 '주의'로 다시 분리하지 않는다. 단위명사 사전에서 분리 가능한 후보를 찾지 못할 경우는 1음절을 건너뛰고 다시 분해 가능한 후보를 탐색한다.

3.1.3 역방향 분해

한국어 복합명사의 형태는 '밤낮'과 같은 대등 구조와 '거리질서'처럼 수식어와 중심어를 갖는 심층 구조(deep structure)로 나눌 수 있다. 복합명사가 여러 개의 단위명사로 이루어져 있을 경우 수식어와 중심어 중간에 나오는 단위명사는 부-수식어나 부-중심어가 된다

본 논문에서는 복합명사의 중심어를 우선 분해하기 위해 분해하는 방향을 끝음절에서 첫음절로의 역방향 분해 방법을 적용한다.

3.1.4 접사 사전을 이용한 분해

접사 사전을 이용하여 접사에 의한 파생어를 포함한 복합명사를 분해 가능하도록 한다. 접사사전은 접두사 18개, 접미사 97개, 그리고 접두 접미사 9개로 구성하여 기존의 연구에서보다 더 많은 접사를 고려하여 처리한다. 또한, 접두-접미사의 처리를 따로 두었는데, 이는 한국어에서 사용되는 접사 중에는 접미사와 접두사 두 가지 모두로 사용되는 접사가 많은 비중을 차지하기 때문이다[10].

3.2 단위명사 사전

본 논문에서는 KORDIC의 형태소 분석에서 사용된 명사 사전으로부터 1음절 명사를 제외한 56,210개의 단위명사를 추출하여 단위명사 사전으로 이용하였다. 1음절 명사를 단위명사 사전에 수록하지 않은 이유는 한국어에서는 1음절 명사의 경우 대부분의 음절들이 1음절 명사로 사용되기 때문에 1음절 명사를 단위명사 사전에 등록하는 것은 무의미하다. 1음절 명사는 모든 복합명사에 대해 중의적 분해 문제를 갖게 하는 원인이 되기도 한다. 예를 들어, '정보통신망'의 경우 '정', '보', '통', '신', '망'이 모두 1음절 명사이기 때문에, '정 + 보 + 통 + 신 + 망'이라 분해할 수도 있게 된다. 또한, '볼가강유역'의 경우 '볼가'가 미등록어이기 때문에 2음절 단위명사로 분해되지 못하고 '볼'이라는 1음절 명사와 '가'라는 1음절 명사로 인식되어, '볼 + 가 + 강 + 유역'으로 인식되는 오류를 범하게 된다.

3.3 접사 사전

복합명사를 단위명사로 분해하는 과정에서 그 단위명사에 대해 구체적으로 분석해 보면 '책상', '하늘' 등의 단순명사로부터 '경쟁력', '재취업' 등처럼 일반명사의 앞뒤에 접사가 첨가된 명사까지 내부적 구성이 다르게 나타남을 알 수 있다. 단위명사에 첨가되어 사용되는 접미사는 거의 모든 명사에 제약없이 붙는 성질을 지닌 것이 많아 일일이 단위명사 사전에 기록하는 것은 모든 명사를 기록하는 것처럼 불가능할 수밖에 없다. 따라서 접미사를 수반한 단위명사들에 대해선 따로 처리를 해야 한다.

접사 사전은 접두사로 빈번히 사용되는 접사와 접미사로 빈번히 사용되는 접사, 그리고 접두사 및 접미사로 동시에 사용되는 접사를 각각 구분하여 사전을 구성하였다. 실제, 한국어에서 사용되는 접사 중에는 접미사와 접두사 두 가지 모두로 사용되는 접사가 많은 비중을 차지하므로, 접두사로 쓸 것인지 아니면 접미사로 사용할지를 결정해야 하는 문제가 발생하게 된다. 예를 들어, 접사 '소'는 접두사와 접미사로 각각 쓰일 수 있는 접사이다. 접사 '소'를 포함하는 복합명사의 분해 예는 다음과 같다.

```
처  음 : 소시민애환 -> 소접두사시민 + 애환
가운데 : 분향소설치 -> 분향소접미사 + 설치
끝     : 직업소개소 -> 직업 + 소개소접미사
```

본 연구에서는 이런 접사에 대해서는 복합명사 내에서 접사의 사용위치를 보고 판단하게 하였다. 즉, 접두사와 접미사로 모두 쓰이는 접사일 경우, 복합명사의 첫머리에 오면 접두사로 사용하고, 맨 마지막에 오면 접미사로 사용하며, 가운데에 오면 접미사로 인식하도록 일관성을 부여하였다. 복합명사의 중간에 나타나는 접사 중에서 접두사와 접미사로 동시에 쓰일 수 있는 접사를 단순히 접미사로 처리한 이유는 한국어의 접사 중에서 접미사의 비중이 접두사보다 높기 때문이다[10].

그리고, 숫자의 경우 복합명사 내에서 사용되는 대부분의 숫자 뒤에는 단위를 나타내는 단위명사를 동반하게 됨으로 0~9는 모두 접두사로 등록하였다. 예를 들어, 복합명사 '100만시민연대모임'의 경우 숫자 '100' 뒤에 오는 '만'은 단위를 지칭하는 의존명사임으로 '100'과 '만'은 서로 분리될 수 없다. 따라서, '100만 + 시민연대 + 모임'으로 분리되는 것이 마땅하다.

한국어와 숫자를 제외한 기타 문자(영문자, 특수 문자 등)는 복합명사 내의 위치에 따라서 접두사와 접미사로 구분될 수 있다. 또한, 영문자의 경우 파생어가 아닌 미등록어로 처리해서 분리할 수도 있다.

<표 1>과 <표 2>는 본 논문에서 사용한 접사와 다른 연구에서 사용된 것을 비교하여 정리하였다. 제시된 표에서 보는 바와 같이, 본 논문에서는 다른 논문에서 사용된 접사 목록보다 많은 양의 접두사와 접미사를 고려함으로써 접사를 포함한 미등록어에 대한 분해에 좀더 좋은 성능을 볼 수 있게 하였다.

〈표 1〉 접두사 대조표

	체 언 접 두 사
태깅분과	제
국 민 대	대, 소, 고, 저, 과, 비, 미, 불, 첫, 끝, 앞, 뒤
고 려 대	가, 개, 고, 날, 내, 대, 맏, 맨, 무, 반, 부, 불, 비, 생, 소, 수, 숫, 시, 양, 왕, 왜, 외, 웃, 재, 제, 초, 총, 최, 풋
본 논 문	가, 고, 과, 대, 명, 무, 미, 반, 부, 불, 비, 생, 소, 신, 역, 재, 저, 전, 정, 주, 준, 초, 총, 최, 타, 탈, 피, 항

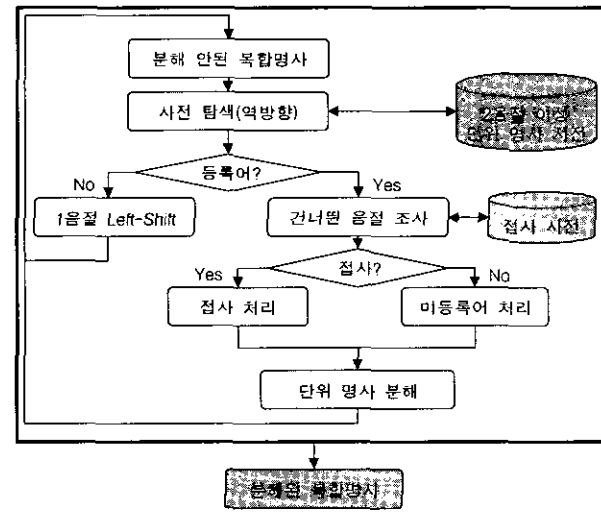
〈표 2〉 접미사 대조표

	체 언 접 미 사
태깅분과	짜리, 어치, 님, 네, 쯤, 씨, 끼리, 씩, 째, 께, 여, 간, 당, 적, 별, 이, 들, 네들, 네들끼리, 들간, 들끼리, 들쯤, 이네, 식, 가, 경, 군, 양
국 민 대	등, 적, 중, 등, 량, 뿐, 식, 상, 성, 형, 화, 별, 용, 측, 간, 째, 쪽, 쯤, 시, 내, 하, 껏, 씩, 속, 뒤, 씨, 및, 후, 경, 당 들뿐, 끼리, 들끼리, 짜리, 면, 네, 외, 때문
고 려 대	가량, 간, 군, 기, 께, 꼴, 꾸러기, 꾼, 끼리, 나기, 내기, 네, 노릇님, 단, 대, 데기, 들, 론, 붙이, 뻘, 새, 생, 설, 성, 세, 식, 씨, 씩, 양, 어치, 여, 옹, 용, 장이, 쟁이, 적, 지기, 질, 짜리, 째, 쭝, 쯤, 차, 체, 치, 토록, 퉁이, 포, 학, 화
본 논 문	가, 각, 감, 계, 고, 곡, 공, 관, 광, 구, 국, 권, 금, 기, 군, 꾼, 님, 단, 당, 대, 댁, 도, 력, 령, 로, 록, 론, 료, 류, 률, 율, 립, 만, 방, 물, 미, 민, 방, 배, 범, 보, 복, 비, 사, 산, 상, 서, 석, 선, 성, 소, 식, 수, 술, 실, 애, 액, 어, 풍, 학, 해, 행, 형, 호, 화

3.4 분해 알고리즘

본 논문에서 제시하는 복합명사의 역방향 분해의 전체 흐름은 (그림 1)과 같고, 알고리즘은 (그림 2)와 같다.

복합명사를 분해하는 방법을 살펴보면, 길이가 N인 분리해야 할 복합명사가 들어오면, 끝음절에서 첫음절 방향으로 단위명사 사전을 이용하여 최장일치 되는 명사를 추출한다. 만약 추출에 실패하면, 끝음절을 건너뛴 음절열에 추가하고, 나머지 (N-1)개의 복합명사열을 가지고 다시 사전탐색을 시작한다. 사전탐색에서 최장일치 되는 단위명사를발견하면, 기존의 건너뛴 음절열에 대해 접사여부를 판별한다. 만약 건너뛴 음절열이 접두사이면 이전 분해열 sp[i-1]에 건너뛴 음절열을 추가하고, 접미사이면 분해열 sp[i]에 사전 탐색된 단위명사와 건너뛴 음절열을 더해서 저장한다. 건너뛴 음절열이 접사가 아닐 경우에는 건너뛴 음절을 미등록어로 간주하여 분해열 sp[i]에 저장하고, 사전 탐색된 단위명사를 분해열 sp[i+1]에 저장한다. 분해된 길이만큼을 복합명사에서 제거하고 건너뛴 음절열을 초기화한다. 그리고 다시 재귀호출 방식으로 복합명사를 분해해 간다. 최종으로 얻어진 분해열 sp[]는 복합명사의 끝음절에서 첫음절 방향으로 분해해야 할 음절 정보를 갖게 된다.



(그림 1) 복합명사 분해 흐름도

복합명사 '게르만민족대이동'에 대해 본 논문에서 제안한 방법으로 분해를 해보면, 먼저 '동'으로 끝나는 최장길이의 단위명사를 탐색한다. 최장길이의 단위명사 '대이동'이 탐색되어 분할 리스트에 기록하고 복합명사로부터 '대이동'을 제거한다. 다시 복합명사 '게르만민족'의 '족'으로 끝나는 최장 길이의 단위명사를 탐색한다. 최장길이의 단위명사 '민족'이 탐색되어 분할 리스트에 기록하고 복합

명사로부터 '민족'을 제거한다. 다시 복합명사 '게르만'의 끝음절 '만'으로 끝나는 단위명사를 단위명사 사전에서 탐색하지만 탐색에 실패하게 되어, '만'이 건너뛴 음절열에 추가되고 복합명사로부터 제거된다. 나머지 '게르'의 '르'와 '게' 역시 사전 탐색에 실패하여 건너뛴 음절열에 추가되고 복합명사에서 제거된다. 더 이상 분해할 복합명사가 없으므로 마지막으로 건너뛴 음절열에 있는 '게르만'이 접사인지를 확인한다. 접사가 아니므로 미등록어로 간주하여 하나의 단위명사로서 분리를 해낸다. 이렇게 해서 최종적으로 '게르만민족대이동'은 '게르만 + 민족 + 대이동'으로 분해된다.

```
char cn[N] ;    /*분리해야 할 복합명사*/
int sp[N] ;     /*역방향으로 저장된 음절길이 정보*/
char *skipSyl ; /*건너뛴 음절열*/
int i ;         /*역방향으로 i번째 분리 위치*/

void segmentCnoun()
{
    char *ptrcn = cn ;
    int cutLen ; /*cn으로부터 제거할 음절길이*/
    cutLen = lookupDictionary(ptrcn) ; /*사전탐색*/

    if (cutLen == 0) { /*건너뛸 음절의 확장*/
       skipSyl = strcat(ptrcn+strlen(cn)-1,skipSyl) ;
       cutLen = 1 ;
    }
    else
       if (*skipSyl != NULL) {
          if (isSuffix(skipSyl))
              cutLen = cutLen+strlen(skipSyl)
          else if (isPrefix(skipSyl))
              segPos[i-1] = segPos[i-1] + strlen(skipSyl)
          else
              segPos[i++] = strlen(skipSyl) ;
          segPos[i++] = cutLen
          *skipSyl = '\0'
       }

    *(ptrcn+strlen(cn)-cutLen) = '\0' ;

    if ((cn == NULL) || (cn[0] == '\0'))
       if ( (skipSyl != NULL) &&
          (skipSyl[0] != '\0')){
          if (isPrefix(skipSyl))
             segPos[i-1] = segPos[i-1] + strlen(skipSyl)
          else
             segPos[i++] = strlen(skipSyl) ;
       }
    else
       segmentCnoun() ;
}
```

(그림 2) 복합명사 분해 알고리즘

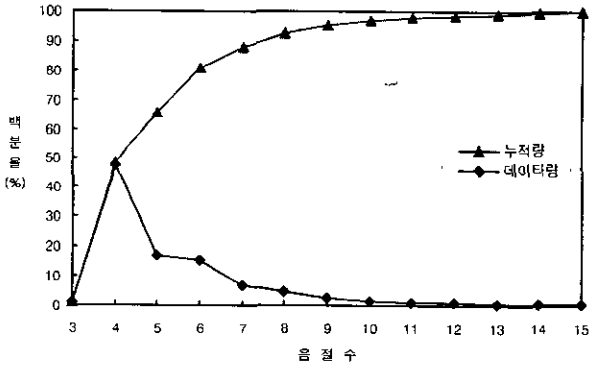
4. 실험 및 분석

4.1 실험 자료

본 논문에서 제안한 분해 알고리즘의 성능 평가를 위해 ETRI의 태깅된 말뭉치로부터 3,230개의 복합명사를 추출하였다. 사전에 포함되어 있지 않은 미등록어를 포함하는 실험 데이터는 444개로 전체의 13.7%를 차지하고 있으며, 미등록어 중에서 특히 접사 파생어는 151개로 전체 데이터 의 4.8%에 해당한다. 다양한 분야의 내용으로 구성된 말뭉치로부터 수동으로 추출된 복합명사는 <표 3>과 같은 형태로 구성되어 있으며 음절수별 전체 비율을 (그림 3)과 같다.

〈표 3〉 음절수별 실험 데이터의 구성

음 절 수	복 합 명 사 개 수
3 음절	32 (1.0%)
4 음절	1,545 (47.8%)
5 음절	542 (16.8%)
6 음절	505 (15.6%)
7 음절	215 (6.7%)
8 음절	150 (4.6%)
9 음절	87 (2.7%)
10 음절	52 (1.6%)
11 음절	30 (0.9%)
12 음절	27 (0.8%)
13 음절	14 (0.4%)
14 음절	13 (0.4%)
15 이상	18 (0.6%)
합 계	3,230 (100%)



(그림 3) 음절수별 데이터 비율

4.2 실 험

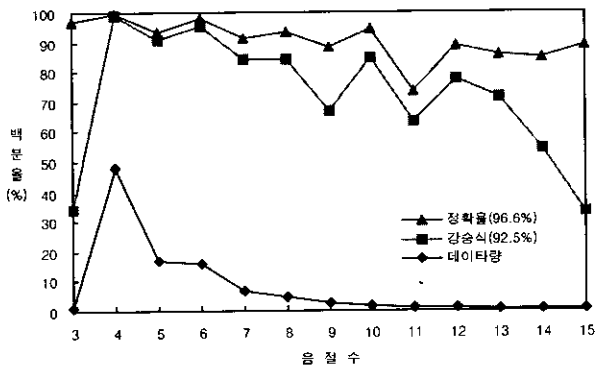
추출한 복합명사를 가지고 역방향 분해 알고리즘을 이용

하여 실험한 결과 96.6%의 분해 성공률을 얻었다. 실험한 복합명사의 음절수별 정확율은 <표 4>에서 표시한 바와 같다. 성능비교를 위해 같은 실험데이터에 대해 강승식[4]의 분해 알고리즘[2]으로 실험한 결과 92.5%의 분해 정확율을 얻었다.

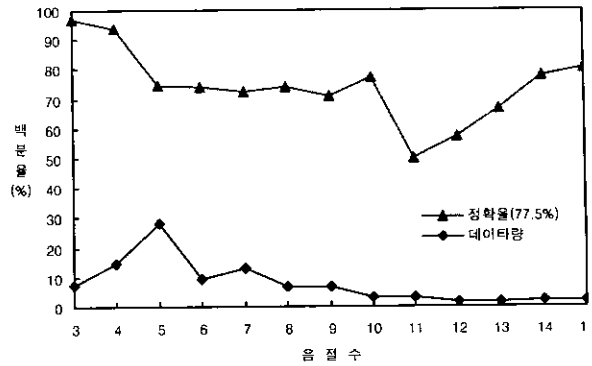
(그림 4)는 본 논문에서 제시한 역방향 분해 알고리즘과 강승식[4]의 분해 알고리즘의 음절수에 따른 분해 성공율을 도식한 것이다. (그림 5)는 미등록어를 포함한 복합명사 444개에 대한 역방향 분해 알고리즘의 분해 정확율을 도식한 것이다. 그리고, 접사 파생어를 포함한 실험데이터 151개에 대한 분해 정확율은 (그림 6)과 같다[3].

〈표 4〉 복합명사의 분해 정확율

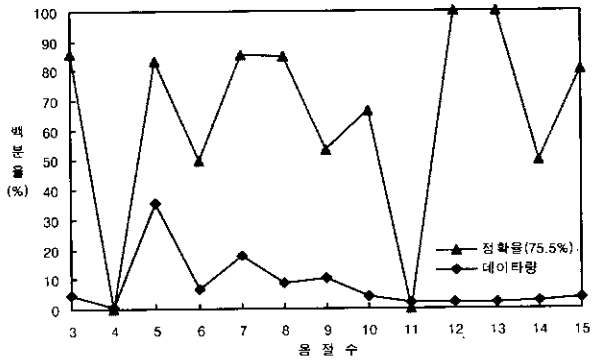
음절수	복합명사 개수	분해성공	정확율(%)
3 음절	32 (1.0%)	31	96.8
4 음절	1,545 (47.8%)	1,540	99.7
5 음절	542 (16.8%)	507	93.5
6 음절	505 (15.6%)	494	97.8
7 음절	215 (6.7%)	196	91.2
8 음절	150 (4.6%)	140	93.3
9 음절	87 (2.7%)	77	88.5
10 음절	52 (1.6%)	49	94.2
11 음절	30 (0.9%)	22	73.3
12 음절	27 (0.8%)	24	88.9
13 음절	14 (0.4%)	12	85.7
14 음절	13 (0.4%)	11	84.6
15 이상	18 (0.6%)	16	88.9
합 계	3,230 (100%)	2,014	96.6



(그림 4) 전체 분해 정확율



(그림 5) 미등록어 분해 정확율



(그림 6) 파생어 분해 정확율

4.3 분 석

(그림 4)에서 보는 바와 같이, 본 논문에서 제시한 역방향[2] 분해 알고리즘의 분해 정확율이 강승식[4]의 분해 알고리즘의 분해 정확율보다 높게 나타남을 알 수 있다. 특히, 3음절 복합명사의 분해 정확도는 상당히 앞서 있으며, 음절길이에 상관없이 비교적 고른 정확율을 보였다. 그러나, 본 논문에서 제시한 알고리즘의 분해 정확율은 음절 길이가 길어지면서 감소해 가는 추세이며, 음절수가 홀수인 경우가 짝수의 경우보다 현저히 낮아짐이 발견되었다. 특히, 음절길이가 11인 경우에는 다른 음절길이보다 정확도가 73.3%로 현저히 낮았는데, 이는 '거장니콜라이루빈스타인'이나 '바실리예프스키국방장관' 처럼 미등록어로 분리되어야할 명사에, 사전에 등록된 단위명사가 포함되어 있어, 사전에 등록된 단위명사의 앞뒤를 분리해 버린 경우가 많았기 때문이었다. 이것은 사전에 등록된 단위명사를 포함한 미등록어를 고려하지 않았기 때문이다. 이런 오류는 비단 11음절 복합명사뿐만 아니라 다른 길이의 복합명사에서도 발생하였다. 복합명사의 분해에서 미등록어에 포함된 등록어로 분해하는 오류의 예는 다음과 같다.

알타이미등록어지역	: 알 + 타이등록어 + 지역
투르키스탄미등록어지역	: 투르 + 키스등록어 + 탄 + 지역
거장니콜라이루빈스타인미등록어	: 거장 + 니 + 콜라등록어 + 이루빈 + 스타등록어인

2) 강승식 교수의 프로그램을 그대로 다시 구현하기 어렵기 때문에 실험에는 홈페이지에서 제공하는 프로그램(HAM version 4.70c)을 이용하였다.

3) 본 논문에서 제시한 알고리즘과 강승식[4]의 알고리즘이 사용하는 단위명사 사전과 접사가 다르므로, 미등록어와 파생어 분해 정확율에 대한 비교는 무의미하다.

본 논문에서 사용한 복합명사의 실험자료에는 444개의 미등록어가 포함되어 있다. 이들 미등록어가 포함된 복합명사를 역방향 분해 알고리즘을 이용하여 분해를 수행한 결과 복합명사 344개가 정확히 분해되었고, 분해 성공률은 약 77.5%였다. 미등록어를 포함한 복합명사의 분해 오류는 모두 등록어로만 구성된 복합명사의 분해 성공률 99.6%[4]보다는 매우 낮아서, 미등록어를 포함한 복합명사의 분해 문제가 복합명사 분해 시스템에 중요한 변수로 작용함을 알 수 있었다. 미등록어를 포함한 복합명사의 분해 오류로는 앞서 설명한 미등록어 속에 등록어가 포함되어 발생하는 오류와, 다음 예처럼 접사의 잘못된 결합에 의해서 발생하는 경우가 가장 많았다.

다국적기업직영화미등록어진출 : 다국적 + 기업직접미사 + 영화 + 진출

또한, 본 논문에서 제시한 알고리즘의 복합명사의 분해 순서 때문에 분해 오류가 발생하기도 하였다. 예를 들어 '환자의식'의 경우, 정방향 분해를 했을 경우 '환자 + 의식'으로 바르게 분해될 수 있으나, 역방향 분해인 까닭에 '자의식'이 '의식'보다 사전에서 먼저 탐색되어 '환 + 자의식'으로 분해되는 오류가 발생하였다.

그리고, 접사의 처리에서도 오류가 발생하였다. 접두사 혹은 접미사로 쓰일 수 있는 접사에 대해서 복합명사의 처음에 위치할 때만 접두사로 인식하고, 나머지는 항상 접미사로 인식하도록 하는 일방적인 방법을 사용한 탓에 '민족대화합'의 접사 '대'를 처리하면서, '민족대접미사 + 화합'으로 분해되는 오류가 발생하였다. '대'는 접두사와 접미사로서 각각 쓰일 수 있으나, 본 논문에서 제시한 알고리즘에서는 접두사보다는 접미사로서의 비중을 높게 평가하다보니 이런 문제가 발생하였다.

5. 결 론

정보검색분야, 기계번역분야 등의 자연어 처리 시스템에서 복합명사를 얼마나 잘 처리하느냐에 따라 시스템의 성능에 커다란 영향을 미친다. 한국어에서 복합명사는 명사간 결합이 자유롭고, 단위명사로 띄어쓰는 것을 원칙으로 하나 붙여써도 무방하기 때문에 처리가 어렵고 복잡하다.

본 논문에서는 복합명사의 역방향 분해 알고리즘을 제안하고 실험하였다. 분해후보는 사전탐색을 이용하였으며, 1음절 명사로 인해서 너무 잘게 분해되는 것을 막기 위해, 2음절 이상의 56,210개의 명사로 구성된 단위명사 사전을 이용하였다. 또한 접사의 처리를 위해 134개의 접사로 구성된 접사 사전을 사용하였다. 분해되지 않은 3,230개의 복합명사에 대해 ETRI 코퍼스로부터 추출된 복합명사를 대상으로 실험한 결과 약 96.6%의 정확도를 얻었다. 또한, 미등록어를 포함한 복합명사 444개에 대해 77.5%의 분해 성공률을 얻었다. 실험에 사용된 데이터중의 미등록어는 대부분 접사를 포함한 파생어로서, 제안한 복합명사 분해 알고리즘에서 접사를 포함한 미등록어에 대한 비교적 높은 분해 정확도를 얻을 수 있었다.

그러나, 역방향 최장일치 분해를 적용하기 때문에 발생하는 분해 오류, 접두 - 접미사가 복합명사의 가운데에 나타났을 때 일괄적으로 접미사로만 취급해서 생기는 분해 오류, 그리고 긴 미등록어 속에 작은 길이의 등록어가 포함되어 있어 등록어의 앞뒤에서 분해되는 오류 등은, 제안한 알고리즘의 성능을 저하시키는 중대한 원인으로 작용하였으며, 이러한 문제는 좀더 신중한 검토가 필요할 것으로 사료된다.

4) 전체 실험 데이터 3230개 중에서 미등록어가 포함된 444개를 제외한 2786개의 실험 결과이다.

참 고 문 헌

[1] JoonHo Lee, HyunYang Cho, HyukRo Park, "N-Gram based Indexing for Korean Text Retrieval," Information Processing & Management, 35(4), 1999.

[2] Bo-Hyun Yun, Ho Lee, Hae-Chang Rim, "Analysis of Korean Compound Nouns Using Statistical Information," Proc. of the 1995 International Conference on Computer Processing of Oriental Languages, pp.76-79, 1995.

[3] Eugene Charniak, "Statistical Language Learning," The MIT Press, 1993

[4] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회논문지(B), 25권 1호, pp.172-182, 1998.

[5] 심광섭, "음절간 상호정보를 이용한 한국어 자동 띄어쓰기", 정보과학회논문지(B), 23권 9호, pp.991-1000, 1996.

[6] 심광섭, "합성된 상호 정보를 이용한 복합명사 분리", 정보과학회논문지(B), 24권 11호, pp.1307-1317, 1997.

[7] 윤보현, 조민정, 임해창, "통계 정보와 선호 규칙을 이용한 한국어 복합명사의 분해", 정보과학회논문지(B), 24권 8호, pp. 925-928, 1995.

[8] 최재혁, "음절수에 따른 한국어 복합명사 분리 방안", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.262-267, 1996.

[9] 박혁로, 신중호, "비터비 학습 알고리즘을 이용한 한국어 복합명사 분석", 한국정보과학회 학술발표논문집, 1997.

[10] 한국전자통신연구원, "전자사전 표제어 선정 지침서", 1999.

이 현 민

e-mail : hyunmini@a-lab.co.kr
1994년 전남대학교 전산학과 졸업(학사)
2001년 전남대학교 전산학과 졸업(석사)
2001년~현재 (주)에이랩 전임연구원
관심분야 : 정보검색, 자연어처리

박 혁 로

e-mail : hyukro@chonnam.ac.kr
1987년 서울대학교 전산학과 졸업(학사)
1989년 한국과학기술원 전산학과 졸업 (석사)
1997년 한국과학기술원 전산학과 졸업 (박사)
1994년~1996년 연구개발정보센터 연구원
1997년~1998년 연구개발정보센터 선임연구원
1999년~현재 전남대학교 전산학과 조교수
관심분야 : 정보검색, 자연어처리, 데이터베이스