Journal of Korea Multimedia Society Vol. 9, No. 2, February 2006(pp. 151-159)

선호 음절 정보를 이용한 복합명사의 분해 방법

박찬이[†], 류 방[††], 김상복[†††]

요 약

정보검색이나 언어번역에서의 복합명사는 사전 미등록 명사에 대한 처리에 크게 영향을 받는다. 한국어 복합명사는 그 구조가 한자어에 의해 파생한 것이 대부분으로 단위명사간 조합시 선호하는 음절이 존재한다. 이를 복합명사의 분해규칙으로 이용할 수 있다. 본 논문에서는 음절간 선호도를 이용하여 사전 미등록 복합명사에 대한 분해 방법을 제안한다. 사전 미등록 복합명사를 제안한 방법으로 분해한 결과 분해 정확률은 88.49%로서 기존의 방법보다 우수한 분해율을 보이고 있다.

A Segmentation Method of Compound Nouns Using Syllable Preference

Chan-Ee Park[†], Bang Ryu[††], Sang-Bok Kim[†††]

ABSTRACT

The ratio of a segmentation algorithm of compound nouns causes an effect a lot in nouns which are not in the dictionary. The structure of Korean compound nouns are mostly derived from the Chinese characters and it includes some preference ratio. So it will be able to use segmentation rule of compound nouns. This paper suggests a segmentation algorithm using some preference ratio of Korean compound nouns which are not in the dictionary. The experiment resulted in getting 88.49% of correct segmentation and showed effective result from the comparative experimentation with other algorithm.

Key words: Compound Nouns(복합명사), Segmentation(분해), Unknown Nouns(미등록어)

1. 서 론

한국어에서의 복합명사란 두개 이상의 단위 명사가 띄어씀 없이 연결된 명사를 말한다. 한글 맞춤법에서는 연속되는 명사들 사이는 띄어서 쓰는 것이 원칙이나 붙여서 쓰는 것도 허용하므로 붙여 쓰는 것이 때론 자연스러워서 글을 쓰는 사람에 따라 다양한 형태로 사용한다. 복합명사의 경우 실생활에서는 정보의 전달이나 사용상 문제가 되지 않지만 사전 의존적인 정보 검색 시스템이나 기계 번역 시스템에서는 심각한 문제를 발생시킨다. 그래서 복합명사의 적절한 처리가 반드시 이루어져야 한다. 이 때 고려할 수 있는 방법은 복합명사 자체를 일일이 사전에 등록하는 것과, 명사사전에 존재하는 복합명사를 단위명사(unit noun) 수준으로 분해하는 방법이 있다. 전자의 경우는 단위 명사들이 서로 결합하여 발생하는 복합명사들이 수 없이 많기 때문에 사전 등록방법은 현실적으로 불가능하다. 따라서 단위명사 사전을 이용하여 주어진 복합명사를 단위명사로 분리하는 방법에 대해 여러 가지 알고리즘이 제시되었다.[1]

복합명사를 분해할 때 고려해야 할 상황은 중의성과 미등록어 문제이다. 이 중에서 가장 어려운 점은

※ 교신저자(Corresponding Author) : 김상복, 주소 : 경남 진주시 가좌동 900번지(660-701), 전화 : 055)751-5994, FAX : 055)762-1944, E-mail : sbkim@nongae.gsnu.ac.kr
접수일 : 2005년 6월 23일, 완료일 : 2005년 10월 24일
[†] 준회원, 경상대학교 컴퓨터과학과 박사과정 (E-mail : sally0115@naver.com)
[††] 정회원, 진주 보건전문대학 사무정보계열 교수 (E-mail : bianc@chc.ac.kr)
[†††] 정회원, 경상대학교 컴퓨터과학과 교수, 경상대학교 컴퓨터정보통신연구소 연구원

중의성 문제이다. 예를 들어, 복합명사인 '특기적성교육'을 분해하는 경우 '특', '기', '적', '성', '교', '육', '특기', '기적', '적성', '교육', '특기적', '성교육' 등의 명사들이 추출될 수 있다. 추출 가능한 명사들인 단위명사로 '특기적성교육'을 분해했을 때, 단음절의 명사를 제외한 '특기+적성+교육', '특기적+성교육' 등으로 분해가 이루어지게 되어 올바른 분해결과를 추출하는데 어려움이 발생하게 된다. 이러한 문제점은 비단 한국어에만 존재하는 것이 아니라 한국어와 문법적으로 유사한 구조를 가진 일본어나 한자문화권 국가들에서도 나타나는 현상이다. 일본어 경우는 복합명사의 구조를 규명하는 쪽으로 연구가 진행되고 있다.[2]

또 하나는 단어사전에 존재하지 않는 미등록어를 포함한 복합명사의 분해 문제가 있다. 대표적인 미등록어로 수치값, 고유명사, 외국어의 한글표기, 새로 생성된 용어 등을 들 수 있다. 예로서 복합명사 '피아노천재볼프강'의 경우, '피아노+천재+볼프강'로 분리되어야 한다. 이 경우 '볼프강'은 고유명사이므로 단위명사 사전에 검색되지 않으므로, 단위사전에 나타나는 단위명사인 '강'을 강제분리하고 '볼프'를 미등록어로 처리한다. 이처럼 복합명사 분해 과정에서 중의적 분해가 발생하는 경우 올바른 단어를 선택하는 문제와 더불어 미등록어를 효율적으로 처리하는 문제를 해결해야 한다.

본 논문에서는 한국어 복합명사의 기반이 대부분 한자어에서 유래한 것들이 많고, 한자는 음절마다 의미를 가지고 있기 때문에 같은 의미를 지닌 한자어가 대부분 여러 개 존재하지만 선호하는 음절의 조합으로 단어가 구성된 점을 착안하여 음절간 선호도를 활용한 복합명사 분리 방법을 제시한다. 이 방법을 사용하기위해 기존의 복합명사 분리 알고리즘 중 분해율이 높은 역방향 분해 알고리즘[3]을 기반으로 하여 본 논문에서 제시한 선호음절의 정보를 적용하는 방법을 이용한다.

본 논문의 구성은 다음과 같다. 2장에서는 복합명사 분해와 관련된 선행된 관련 연구들에 대해 살펴보고, 3장에서는 음절간 선호도를 이용한 복합명사 분해 알고리즘을 제시하며, 4장에서는 제시한 알고리즘을 이용하여 실험한 결과에 대해 기술하고, 마지막 장에서는 결론과 향후 진행할 부분을 제시한다.

2. 관련연구

복합명사 분해 방법은 대부분 휴리스틱을 이용한 방법으로, 이들을 좀 더 세분하면 사전에 기초한 방법[4]과 통계적 처리에 기초한 방법[5]으로 나눌 수 있다. 먼저 사전에 기초한 방법은 전자사전에 수록된 단위명사의 존재 유무를 이용하여 복합명사를 여러 개의 단위 명사로 분해한 후, 중의성 배제를 위해 경험적 처리방법을 이용한다. 이러한 휴리스틱 처리방법은 처리과정보다 결과의 우수성을 찾는 것으로서 이론적 증명이 어려우며, 미등록어가 포함되어 있는 복합명사의 경우에는 성능저하가 심하다는 단점이 있다. 그러나 역방향 분해방법은 사전에 기초한 방법이지만 기존의 방법보다 우수한 결과를 보이고 있으며 미등록어에 대한 처리도 비교적 우수한 결과를 보이고 있다. 두 번째로 통계적 처리에 기초한 방법은 코퍼스를 이용하여 각 단어의 출현 확률을 구하여 이의 특징으로부터 통계적 정보를 먼저 구한 뒤 이를 이용하여 복합명사를 분해하는 방법이다.

윤보현[12]은 통계 정보와 선호 규칙을 이용하여 복합명사를 단위명사로 분해하는 알고리즘을 제안하였다. 통계 정보에는 1음절 접사의 빈도, 2음절 또는 3음절 단위명사가 복합명사에서 사용된 위치 및 빈도정보를 이용하였다. 선호 규칙에는 중의적 분해를 발생하는 단위명사의 개수가 서로 다른 경우, 단위명사의 개수가 적은 복합명사 분해를 올바른 분해 패턴으로 선호하는 규칙을 제시하였다.

강승식[7]은 4가지 분해규칙 및 2가지 예외규칙을 사용하여 복합명사에 대한 분해 가능한 후보들을 먼저 생성한 후 이들에 대해 가중치를 부여하여 최적 후보를 선택하는 알고리즘을 제안하였다. 이 알고리즘의 특징은 단어의 길이와는 상관없이 미등록어를 포함한 복합명사의 분해가 가능하다. EH한 이 알고리즘에서는 분해 도중에 중의성을 가지는 문제를 만나면 가중치를 이용하여 높은 후보를 선택하는 방법을 사용 하였다.

심광섭[8]은 4가지 유형의 음절간 상호 정보를 합성한 것을 이용하여 복합명사내의 단위명사 분해 위치를 선택하는 알고리즘을 제시하였다. 분해 위치는 모두 단위명사가 될 때까지 반복하며, 분해된 명사가 단위 사전에 등록된 명사가 아닌 경우 한 음절씩 추

가하면서 사전을 탐색하는 방법으로 분해위치를 결정하도록 하였다. 이 과정에서 효율을 높이기 위해 2음절 분해방법을 사용하였다.

이현민[6]은 복합명사의 구조상 중심어가 후반부에 존재하는 점을 이용하여 분해방법을 끝 방향에서 시작하여 앞쪽으로 분해를 시도하는 역방향 분해 방법을 제시하였다. 이는 대부분의 복합명사들이 중심어의 위치가 후반부에 존재하는 특성으로 인해 높은 분해율을 보이고 있다. 또한 이 방법에 선호 음절정보를 적용하여 분해율을 더 높이는 방법[13]도 제시되었다.

상기의 복합명사 분해 알고리즘들은 각자의 방법대로 비교적 우수한 결과를 나타내고 있다. 하지만 복합명사의 중의적 분해 문제와 복합명사에 미등록어가 포함되어 있을 경우에는 처리결과가 상대적으로 낮게 나타나고 있다. 또한 역방향 분해를 제외한 알고리즘에서는 중의성을 증폭시키는 1음절 단위명사에 대한 처리문제를 내포하고 있다. 뿐만 아니라 기존의 논문들에서는 자음접변, 두음법칙, 연음법칙 등 한국어에서 나타나는 특징을 활용한 중의성 해결에 대한 고려가 미약하다고 할 수 있다.

3. 사전 미등록 단위명사가 포함된 복합명사의 분해 방법

복합명사를 단위명사로 분해하는 경우, 중의적 분해로 인한 선택문제와 사전에 존재하지 않아서 발생하는 미등록어를 포함한 문제들이 나타날 수 있다.

복합명사의 중의적 분해시 올바른 선택중의 한 가지는 하나의 복합명사에서 추출 가능한 여러 개 단위명사 집단 중 한 가지 조합을 선택해야 하는 문제이다. 예를 들면, '특기적성교육'이란 복합명사를 분해하면, 두 가지 조합인 '특기+적성+교육'과 '특기적+성교육'으로 분해할 수 있지만 올바른 분해 조합인 '특기+적성+교육'을 선택할 수 있어야 한다.

미등록어가 포함된 복합명사를 분해하는 문제는 복합명사를 구성하고 있는 단위명사가 사전에 등록되지 않아서 복합명사를 분해하는 데 실패하는 경우이다. 예를 들어, '정리해고자'에서 단위명사 '정리' 및 '해고'는 사전에 등록되어 있으나 '자'에 대한 처리가 되지 못하여 복합명사 분해 실패를 발생한다. 복합명사를 정확히 분해하기 위해서는 '정리+해고자'로 분해하여 '해고자'는 미등록어로서 분해되어야 한다. 그러나 조합이 가능한 모든 형태의 단위명사를 사전에 등록하는 경우 중의성을 증가시키는 문제를 안고 있다. 가령 잘못된 분해로 인해 '정리해'라고 분해되어 이 단어가 사전에 존재하지 않는다고 등록한다면, '정리해+고자'라고 분해되어 '해고자'와 '고자' 중 올바른 것을 선택하는 문제가 발생한다. 이와 같이 미등록어가 포함된 복합명사에 대한 올바른 분해 또한 제대로 처리될 수 있어야 한다.

이런 문제점을 해결하고자 본 논문에서는 역방향 분해방법으로 전처리 과정을 거친 후 선호음절을 이용한 복합명사 분해 방법을 제안한다.

3.1 복합명사 분해 과정

3.1.1 전처리 과정

1) 단음절 명사를 제외한 명사사전 이용

복합명사를 분해하기 위해서 단음절 명사를 제외한 명사로만 구성된 명사사전을 이용한다. 여기서 단음절 명사를 제외한 이유는, 한국어에서의 의미를 구성하는 단어의 대부분은 단음절로 구성되는 경우는 거의 없고, 2글자 이상의 한자어로 구성되어 있기 때문이다. 신조어 또한 단음절만으로 구성되는 경우가 거의 없기 때문이다. 또한 단음절의 한자음은 다양한 의미를 나타내므로 중의성을 급격히 증가시키기 때문에 제외한다.

2) 최장일치 수록명사 우선 분해

사전을 탐색 할 때 최장일치 명사를 우선 분해하도록 처리한다. 예를 들어, '민주주의'는 단위명사 사전에 존재하는 명사이면서 단위 명사인 '민주'와 '주의'도 단위명사 사전에 존재한다. 이런 경우는 최장일치 단위명사인 '민주주의'를 선택한다. 이 처럼, '민주주의'라는 단어 자체가 완전한 하나의 의미를 가진 단위명사인 경우에는 다시 분해하지 않는다. 만약 단위명사 사전에서 분해 가능한 후보를 찾지 못한 경우에는 끝 위치에서부터 1음절씩 제외시킨 후 사전에 존재하고 있는가의 여부를 반복적으로 탐색한다.

3) 접사 및 접사화 정보를 이용한 역방향 분해

최장일치 수록명사를 찾지 못한 경우에는 끝 방향에서 1음절씩 제외시켜 나가는 과정에서 다시 사전

에 수록된 단어가 나타나면 접사 사전을 이용하여 접사에 의한 파생어를 포함한 복합명사를 분해 가능하도록 한다. 접사는 접미사, 접두사 순으로 가능성을 검사하여 가장 높은 가능성에 대한 처리를 한다. 대부분의 미등록어는 접사에 의한 파생어들이 대다수이기 때문 이에 대한 처리를 먼저 행하면 중의성 문제를 줄일 수 있다. 경우에 따라서는 접미사의 경우 명사가 아닌 단어에 접미사를 추가하여 명사화하는 경우가 있으므로 이에 대한 처리 또한 필요하다.

3.1.2 선호 음절 정보를 이용한 분해 과정

1) 선호 음절 정보를 이용한 분해

접사사전에 없는 경우, 등록된 앞단어나 뒷단어로 연결해야 하는 미등록어가 존재한다면 이들 단어들은 미등록어로 처리할 것이 아니라 앞단어나 뒷단어로 연결이 가능한 미등록어인지 여부를 먼저 결정해야 한다. 한국어에서는 두음법칙, 자음접변, 순행동화, 역행동화 등의 발음과 관련한 나름대로의 규칙을 가지고 있다. 이러한 규칙은 발음과 관련한 오랜 관습과 관련된 발음법칙으로 인해 서로 간에 선호하는 음절들로 구성된 단어의 정보를 이용하여 자음과 자음간 및 받침간의 선호 통계치를 이용하는 경우 미등록어 분해시 비교적 우수한 분해율을 얻을 수 있다.

2) 음절 패턴에 의한 강제 분해

대부분의 복합명사 형태는 한자를 기반으로 하고 있는 복합명사가 주류를 이루고 있으며, 결합 구조는 선호하는 음절이 인접해 있다. 미등록어로 분리된 단어들에 대해서는 마지막 단계로 통계적 선호음절 정보를 이용하여 분해한다. 외래어를 제외한 미등록어의 구조 대부분은 주로 2음절이나 3음절 형태의 조합된 구조를 취하고 있으며, 간혹 4음절이상으로 구성된 경우는 다시 2음절이나 3음절 단위명사로 분해가 가능한 구조로 되어 있다. 그러나 2음절이나 3음절로의 분해방법은 분해된 2음절이나 3음절이 단위 사전에 반드시 존재하는 경우에만 분해하고, 단위 사전에 하나라도 존재하지 않는다면 미등록어로 처리한다.

3.2 사전 및 선호 음절 정보의 구성

3.2.1 단위사전

본 논문에서는 국어연구원의 단일명사 사전, 의존명사사전, 복합명사 사전, 접사사전을 토대로 단위명사들을 추출하여 단위명사 사전으로 이용하였다. 앞서 언급한 대로 단음절 명사를 단위명사 사전에서 제외한 것은 단음절이 대부분 한자어의 다양한 의미를 수반하기 때문에 이로 인해 중의성을 증폭하는 결과를 초래한다. 예를 들어 '대한민국'의 경우 '대', '한','민','국'이 모두 1음절 명사이기 때문에, '대+한+민+국'이라 분해가능하다. 또한 '천재볼프강'의 경우 '천재볼프강'이 미등록어이기 때문에 단위 명사로 분해되는 과정에서 2음절 단위명사인 '천재' 및 미등록어 '볼프강'이 의미를 가지는 단음절 명사인 '볼', '프', '강'으로 인식되는 오류가 발생한다.

3.2.2 접사 사전

복합명사를 단위명사로 분해하는 경우 '친구', '저녁' 등의 단순한 형태를 가지는 명사에서부터 이들 단순명사의 앞뒤에 접사를 첨가하여 많은 양의 파생명사를 생성할 수 있게 된다. 예로서 '세계'라는 단순명사에 접미사 '화', '성', '적' 등의 다양한 종류를 생성할 수 있다. 이처럼 단위명사인 경우 거의 접사를 수반할 수 있는 특성이 있기 때문에 생성 가능한 단위명사를 구축하는 것은 거의 불가능하다고 볼 수 있다. 최근에는 고유명사에도 접사를 붙여 새로운 신조어를 만드는 경우도 있다.

본 논문에서는 접사에 대해서는 복합명사 내에서 접사의 사용위치에 따라, 복합명사의 첫머리에 나타나면 접두사로, 마지막에 나타나면 접미사로, 복합명사 중간에 위치하면 접미사로 간주하였다. 이는 한국어의 구조상 접두사는 중간보다는 첫머리에 위치하며, 접사의 출현 빈도에서 접미사의 비중이 접두사보다 높기 때문이다.[6]

숫자의 경우 복합명사 내에서 사용되는 대부분의 숫자 뒤에는 숫자를 대신하는 단음절이거나 단위명사를 동반한다. 따라서 숫자로 시작하는 단어에 단위를 지칭하는 접미사가 결합된 형태로 처리한다. 또한 접두사나 접미사도 아니지만 통계적 자료에서 특정한 음절이 첫머리를 선호하는 음인지, 끝을 선호하는 음인지를 검사하여 첫머리와 끝에 나타나는 비율이 서로 5배 이상이면 강제로 접두사로 처리한다.

다음의 (표 1)은 본 논문에서 사용한 접사의 종류와 접사화 규칙에 대한 내용이다.

표 1. 접사 및 접사화 관련 규칙 대조표

구분	접사종류	접사화 규칙
접두사	가, 고, 과, 당, 대, 명, 무, 미, 반, 부, 불, 비, 생, 소, 신, 역, 재, 저, 전, 정, 주, 준, 초, 총, 최, 타, 탈, 피, 한, 항, 헛	머리음 >= 꼬리음 * 5
접미사	가, 각, 간, 계, 고, 곡, 관, 구, 국, 권, 금, 기, 군, 꾼, 네, 님, 답, 대, 댁, 도, 력, 령, 로, 록, 론, 료, 류, 률, 만, 망, 물, 미, 민, 방, 배, 법, 보, 복, 부, 비, 사, 산, 상, 서, 석, 선, 설, 성, 소, 수, 술, 식, 실, 액, 어, 용, 째, 적, 제, 차, 측, 풍, 학, 해, 행, 형, 호, 화	머리음 * 5<= 꼬리음

3.2.3 음절간 선호도 자료

선호 규칙을 위한 통계자료는 중의적 분해 문제에서의 올바른 선택과 관련한 처리를 하거나, 미등록어 처리시 분해 가능한 경우인지에 대한 정보를 제공한다. 이 자료를 이용하여 미등록어에 대한 분해율을 높이는데 유용하게 사용할 수 있다. 통계자료를 위해 2음절의 단위명사만을 대상으로 첫머리에 나타나는 빈도수와 끝에 나타나는 빈도수를 계산하여 이를 테이블로 저장한다. 이 정보를 이용하여 해당 음이 첫머리를 선호하는지 끝에 나타나기를 선호하는지에 대한 정보를 얻을 수 있다. 즉, 접두사나 접미사가 아니지만 접두사처럼 첫머리에 자주 나타나거나 접미사처럼 끝에 자주 나타나는 음절에 대한 정보를 제공한다.

또한 2음절의 단위명사에서 첫음절의 자음정보와 끝음절의 자음정보간의 관계를 계산하여 자음간의 선호관계를 설정한다. 한국어에서 두음법칙이나, 자음접변 등으로 인해 서로간 선호하는 자음이 존재하므로 이 정보를 이용하여 분리할 위치나 분리가 불가능한 위치를 결정하는데 사용한다. 뿐만 아니라 이 자료는 미등록어에 대한 처리시 유용한 정보를 제공한다. 21세기세종계획의 단일어사전과 파스칼세계대백과사전의 표제어를 대상으로 2음절 명사만을 추출하여 앞뒤음절간의 자음간 빈도수를 조사한 내용을 아래 (표 2)에 나타내었다.

3.4 선호음절정보를 이용한 분해 방법

본 논문에서 제시하는 음절간 선호도를 이용한 복

표 2. 추출된 인접 자음간의 발생빈도

앞 \ 뒤	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅅ	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
ㄱ	4148	93	484	1400	46	1731	1305	1594	17	4127	25	3391	4258	16	1400	36	390	522	1964
ㄲ	86	11	11	20	5	68	42	31	5	41	4	48	30	0	15	6	5	14	6
ㄴ	662	49	89	376	7	397	330	397	9	564	12	704	608	14	203	11	97	141	278
ㄷ	1445	52	164	423	16	830	625	700	15	1515	10	1261	1241	11	601	14	119	289	632
ㄸ	53	5	8	12	2	41	54	24	0	29	3	49	38	0	17	6	7	16	2
ㄹ	19	1	7	22	0	14	10	15	0	35	2	48	19	1	10	9	19	13	6
ㅁ	1312	47	174	499	26	768	631	503	13	1497	36	1346	1468	7	376	20	131	233	866
ㅂ	2086	66	200	758	41	964	688	908	16	2635	15	2054	2032	12	705	29	282	446	1439
ㅃ	23	1	1	6	0	13	10	5	2	14	0	12	9	0	2	0	4	11	3
ㅅ	3297	60	350	1131	33	1492	1661	1376	17	3455	30	3133	3031	17	1098	40	526	748	2187
ㅆ	77	8	2	29	2	50	42	42	1	55	0	61	56	0	16	7	23	11	9
ㅇ	3833	92	454	1530	29	2164	1428	1627	29	4447	51	3428	3618	46	1402	41	402	563	2371
ㅈ	4137	60	239	1498	20	1358	1097	1726	14	3773	18	3269	3312	5	1524	26	303	664	1785
ㅉ	14	0	2	3	1	10	7	6	1	7	2	7	6	0	2	0	0	0	2
ㅊ	1351	13	179	429	13	596	457	579	0	1165	9	1012	1146	6	439	15	129	228	551
ㅋ	48	3	26	32	1	83	40	19	2	60	2	55	44	0	20	12	58	18	12
ㅌ	530	18	35	113	7	208	146	271	6	758	4	502	608	1	146	16	47	128	281
ㅍ	678	19	51	255	10	326	309	262	4	622	4	494	564	0	171	10	115	132	332
ㅎ	1928	28	148	652	12	803	624	713	7	2376	14	1596	1574	3	416	14	152	275	789

합명사 분해의 전체흐름은 (그림 1)과 같다. 복합명사 분해 과정은 복합명사의 끝음절에서 역방향으로 단위명사 사전을 이용하여 최장일치 되는 단위명사를 추출한다. 만일 단위사전에 존재하지 않을 경우는 끝음절을 임시 음절열에 추가하고, 나머지 음절을 가지고 사전탐색을 다시 시작한다. 이러한 과정을 재귀적 호출로 반복한다. 이 과정에서 사전에 등재된 명사를 발견하지 못한 경우는 그 다음 단계인 접사형 명사인지를 검사한다. 첫음절과 끝음절을 각각 분리한 후 사전탐색을 진행한다. 그러나 접사처리 단계에서도 분리하지 못한 복합명사는 본 논문에서 제안하는 선호음절 정보를 이용하여 각 음절에 대한 분리점을 설정한다. 설정된 분리점을 기준으로 분리된 음절을 대상으로 사전탐색을 진행한다. 만약 탐색과정에서 실패한 음절이 존재하는 경우에는 더 이상의 분해 정확도를 보장할 수 없다. 여기까지 도달한 미등록 복합명사는 분해 효율을 떨어뜨리는 명사들로서 마지막으로 강제분해를 진행한다. 이 경우에는 분해된 음절들 중 부분적으로 사전에 탐색되는 명사들이 존재하게 된다.

복합명사 '정리해고자'를 예로해서 본 논문에서 제안한 방법으로 분해를 해보면, 먼저 최장길이의 명사가 단어사전에 존재하는지 검사한다. 존재하지 않는다면 접사사전에서 '자'가 존재하는지를 검사한다. 만약 실패하면 다시 끝 2음절의 단어인 '고자'를 단어사전에서 검색한다. 사전에 존재한다면 나머지 열의 '정리해'가 사전에 존재하는지를 검사하여 존재하지 않는 단어이면 잘못된 분해로 인식하여 다시 분해를 시도하여 '정리+해+고자' 와 '정리+해고+자'의 경우를 가지고 (표 2)의 선호도를 이용하여 최종 '정리+해고자'로 분리 한다.

미등록어가 포함되면 분해된 명사들 중의 하나를 선택해야 하는 중의성 문제가 분해율에 영향을 크게 미치게 된다. 그러나 음절간 선호도를 이용한 정보를 중의성의 해소에 이용하면 분해 정확률을 향상시킬 수 있다. 실제 '정리해고자'를 실험하여 그 결과를 얻은 화면은 그림 2와 같다.

그림 2에서 나타난 자료에서 우선 분해를 시도하면 '정리+해+고자' 와 '정리+해고+자'로 분리되어지며 이 중 올바른 선택을 위해 '해'와 '자'의 선호도를 검사하여 우수한 것을 선택한다. 먼저 '해'의 처리인 경우 '리'와 '해'의 선호도는 6이고, '해'와 '고'의 선호도는 1928이므로 '해'는 '정리해' 쪽 보다는 '해고자' 쪽을 더욱 선호함을 알 수 있다. 또한 '자'는 앞선 '고'와의 선호도는 4258로서 '고'와 '자'의 선호도가 높기 때문에 결국은 '정리+해고자'가 선택되고 이 중 '해고자'는사전 미등록어지만 하나의 명사로 분리할 수 있다.

4. 실험 및 분석

4.1 실험자료

본 논문에서 제안한 분해 방법의 성능 평가를 위해 국립국어연구원의 국어빈도조사정보를 이용하여 통계자료를 추출하였고, 복합명사는 국내 국어사전

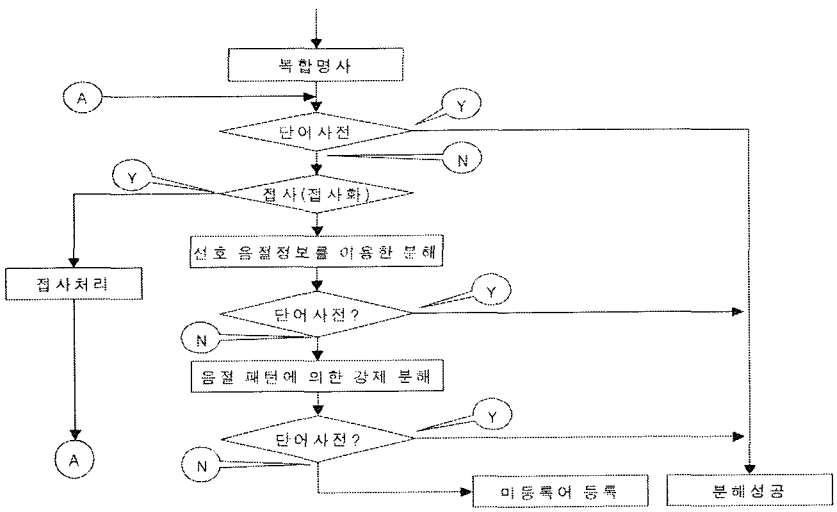


그림 1. 복합명사 분해 절차

그림 2. 선호 음절 정보를 이용한 복합명사의 분해 화면

에 등재되어있는 복합명사 정보와 인터넷에서의 정보검색 도구를 이용하여 36061개의 복합명사를 추출하여 이 중 사전에 포함되어 있지 않은 미등록어를 포함한 복합명사 1964개를 대상으로 실험하였다. 사전정보 중 추출된 복합명사 중에서 역사적 사건이나 고유명사 등에서 파생한 복합명사는 가급적 배제하였다.

4.2 실 험

추출한 복합명사를 가지고 전처리 과정인 역방향 분해기법 및 선호 음절 정보를 이용하여 실험한 결과 각각 81.52%, 88.49%의 분해 성공률을 얻었다. 이 실험에 사용한 복합명사의 음절수별 정확률은 (표 3)과 같다.

실험데이터의 결과에서 10음절에서는 역방향 분해 알고리즘으로 실험한 결과가 선호음절정보를 이용한 결과보다 약간 상위의 값을 가진 이유는 10음절 이상의 미등록 복합명사의 수가 상대적으로 적은 양이고, 실험에 사용된 미등록복합명사에 따라 차이가 나타날 수 있기 때문에 실패하는 비율에 매우 민감한 수치를 보이고 있다.

4.3 분 석

(표 3)에서 보는 바와 같이 본 논문에서 제시한 선호도를 이용한 분해 정확률이 기존의 분해 기법의 분해 정확률보다 높게 나타남을 알 수 있다. 사전에 존재하지 않는 미등록어라 할지라도 명사의 구성상 음절간의 선호도를 이용하여 분리할 수 있음을 보이고 있다. 이는 음절간에 발생하는 각종 문법상의 규칙에 기인한 것으로 이를 이용하면 미등록어라 할지

표 3. 복합명사 분리 정확률

음절수	복합명사 수	역방향 방법의 전처리과정		선호 음절 정보를 이용한 결과	
		분리 성공	비율	분리 성공	비율
3	9	9	100.00	9	100.00
4	1068	928	86.89	951	89.04
5	622	465	74.76	549	88.26
6	139	108	77.70	124	89.21
7	64	49	76.56	59	92.19
8	19	15	78.95	16	84.21
9	17	12	70.59	14	82.35
10	12	9	75.00	8	66.67
11이상	14	6	35.71	8	57.14
총 계	1964	1601	81.52	1738	88.49

표 4. 음절별 분해실패 원인 예

음절수	정확한 분해 예	실패한 분해 결과	분리 실패 원인
4	덩+샤오핑	덩샤+오핑	빈도 분리
5	다뉴+세문경	다뉴세+문경	등제어 분리
6	단측파대+전송	단측파+대전송	접사화 정보 분리
7	다이버시티+방식	다이버+시티+방식	빈도 분리
8	전진사지+삼층+석탑	전진사+지삼층+석탑	빈도 분리
9	층상+함동+황화철+광상	층상함+동황화+철광상	접미사 빈도 분리
10	오일팔+광주+민주화+운동	오일+팔광+주민+주화+운동	빈도 분리
11이상	법천사+지광국사+현묘+탑비	법천+사지광+국사현+묘탑비	접사화 정보 분리

표 5. 분리 실패한 10음절 복합명사

분리 실패한 10음절	정확한 분리 예	잘못된 처리 결과
법천사지광국사현묘탑	법천사+지광+국사+현묘탑	법천+사지광+국사현+묘탑
오일팔광주민주화운동	오일팔+광주+민주화+운동	오일+팔광+주민+주화+운동
창녕신라진흥왕척경비	창녕+신라+진흥왕+척경비	창녕+신라+진흥+왕척+경비
팔레스티나민족평의회	팔레스티나+민족+평의회	팔레+스티나+민족평+의회

라도 효율적인 분리가 가능하다.

그러나 본 논문에서 제시한 기법도 기존의 방법과 마찬가지로 음절 길이가 길어지면서 분해정확률이 감소하는 현상을 보였다. 음절길이가 10을 넘기는 경우 복합된 명사의 데이터를 구하기도 힘들었고 적은 양의 데이터에서 소량의 실패에도 결과 값에 영향을 크게 미치는 결과를 초래하였다. 이 경우의 대부분은 복합명사 내에 존재하는 고유 명사 등에 사전에 등록된 단위명사가 포함되어 있어, 사전에 등록된 단위명사의 앞뒤를 분리해 미등록어로 처리하기 때문이다.

본 논문에서 제시한 기법 중 접사의 처리에서도 오류가 발생하였다. 접두사 혹은 접미사로 쓰일 수 있는 접사에 대해서 복합명사의 처음에 위치할 때는 접두사로 인식하고, 끝음절은 항상 접미사로 인식하도록 제시하였지만 접사와 관련한 음을 포함한 일반 명사에 대한 오류도 발견되고 있으며 이에 대한 정확한 처리가 필요하다.

(표 4)는 분해에 실패한 복합명사에 대한 잘못된 분해결과 및 실패 원인에 대한 음절별 예를 보인 것이고, 특히 10음절에서는 기존의 분해율 보다 낮은 분해 실패 10음절 복합명사의 잘못된 분해 결과의 예를 (표 5)에 나타낸 것이다.

5. 결 론

인터넷 상에서의 검색이나 번역프로그램등과 같은 도구를 사용해서 번역하는 경우, 이에 필요한 자연어를 처리하기 위해서는 사전의 정보가 필수적이다. 이 과정중 사전에 존재하지 않는 복합명사를 처리하는 문제로 인해 시스템의 성능에 커다란 영향을 미칠 수 있다. 왜냐하면 한국어에서 복합명사는 한글 맞춤법상에서 명사들 간의 결합을 허용하고 있기 때문에 그로 인해 번역이나 검색시 단위 명사사전 검색에 실패할 확률이 높기 때문이다.

이를 개선하고자 본 논문에서는 복합명사의 선호도를 이용한 분해 방법을 제안하고 이를 실험하여 그 결과를 보였다. 분해 명사군은 사전을 이용한 탐색을 사용하고, 1음절 명사로 인한 중의성의 증폭을 방지하기 위해 2음절 이상의 명사로만 구성된 단위명사 사전을 이용하였다. 또한 접사 처리를 위해서는 접사 사전을 구축해서 사용하였다. 그리고 본 논문에서의 실험에 사용된 복합명사들은 국어사전 및 백과사전과 인터넷 검색으로부터 추출한 복합명사를 대상으로 실험하였으며 그 결과 약 88.49%의 정확도를 얻었다. 실험에 사용된 복합명사들 중 사전 미등록명사는 대부분 접사파생어로서, 제안한 복합명사 분해 기법을 이용하여 미등록어를 대상으로 분해한 결과 다른 방법들에 의한 것보다 비교적 높은 분해 정확도를 얻을 수 있었다.

그러나 본 논문에서도 선호도를 이용한 역방향 분해를 우선 적용하여 사용하기 때문에 최장일치 분해를 시도하는 과정에서 나타나는 오류를 막기 위해

접사와 관련한 정보를 활용하였지만, 이 처리과정에서 발생한 분해 결과가 정확한 것인가에 대한 자료는 이 방법에서는 확인할 수 없었다. 더불어 외래어로 표기된 미등록어가 2개 이상 포함되는 경우, 본 논문에서 제안한 방법을 적용한 결과 정확율이 다소 감소하는 현상을 초래하였기 때문에 이를 개선하고 보완하기 위해 이 부분의 연구가 추가적으로 요구된다.

참 고 문 헌

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

[2] Eugene Charniak, "Statistical Language Learning," *The MIT Press*, 1993.

[3] T. Hisamitsu And Y. Nitta, "Analysis of Japanese Compound Nouns by Direct Text Scanning," *Proceeding of the 16th International Conference on Computational Linguistics*, pp. 550-555, 1996.

[4] Bo-Hyun Yun, Ho Lee, and 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.

[5] K. Yosiyuki and T. Hozumi, "Analysis of Japanese Compound Nouns using Collocational Information," *Proceedings of the 15th International Conference on Computational Conference Linguistics*, pp. 865-869, 1994.

[6] 이현민, 박혁로, "복합명사의 역방향 분해 알고리즘", *한국 정보처리학회 논문지(B), 제8-B권 제4호*, 2001.

[7] 강승식, "한국어 복합명사 분해 알고리즘," *정보과학회 논문지(B), 25권 1호*, pp. 172-182, 1998.

[8] 심광섭, "합성된 상호정보를 이용한 복합명사의 분리," *정보과학회 논문지(B), 24권 117호*, pp. 1307-1317, 1997.

[9] 박혁로, 신중호, "비터비 학습알고리즘을 이용한 한국어 복합명사 분석," *한국 정보과학회 학술발표 논문집*, 1997.

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

[11] 최재혁, "음절수에 따른 한국어 복합명사의 분리 방안," *제8회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 262-267, 1996.

[12] 윤보현, 조정민, 임해창, "통계정보와 선호규칙을 이용한 한국어 복합명사의 분해," *정보과학회 논문지(B), 24권 8호*, pp. 925-928, 1995.

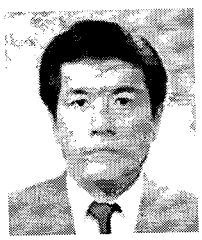
[13] 박찬이, 류방, 정원창, 백현철, 구명모, 김상복, "미등록 복합명사의 분해," *한국멀티미디어학회 학술발표 논문집, 제8권, 제1호*, pp. 368-371, 2005.

박 찬 이

1993 진주교육대학교 초등교육학과 학사
2001 진주교육대학교 컴퓨터교육학과 석사
현재 경상대학교 컴퓨터과학과 박사과정 수료
현재 진주교육대학교부설 초등학교 교사
관심분야 : 유무선통신, 한국어 정보처리, 멀티미디어통신

류 방

1984 경상대학교 전산통계학과 학사
1993 경상대학교 전자계산학과 석사
2004 경상대학교 컴퓨터과학과 박사
현재 진주보건대학 의약복지 정보계열 부교수
관심분야 : 유무선통신, 한국어 정보처리, 컴퓨터프로그래밍



김 상 복

1989 중앙대학교 전자공학과 박사
현재 경상대학교 컴퓨터과학과 교수, 경상대학교 컴퓨터정보통신연구소 연구원
관심분야 : 멀티미디어 통신, 한국어 정보처리, 컴퓨터구조