Journal of Korea Multimedia Society Vol. 7, No. 3, March 2004(pp. 418-426)

접사정보 및 선호패턴을 이용한 복합명사의 역방향 분해 알고리즘

류 방[†], 백현철[††], 김상복[†††]

요 약

본 논문에서는 음절간 상호 정보를 이용하여 한국어 복합명사의 역방향 분해 알고리즘을 제안한다. 한국어 복합명사는 그 구조가 한자어에 의해 파생한 것이 대부분이며 음절 상호간 선호 음절이 존재하므로, 이 정보와 접사정보를 복합명사의 분해규칙으로 이용한다. 성능을 평가하기 위해 36061개의 복합명사를 이용하여 본 논문에서 제안한 알고리즘의 분해한 결과 99.3%의 분해 정확율을 얻었다. 실험과 관련한 기존 알고리즘간의 비교에서도 우수한 결과를 얻었으며, 특히 4음절과 5음절 복합명사의 경우 대부분 정확한 분해 결과를 얻었다.

A Reverse Segmentation Algorithm of Compound Nouns Using Affix Information and Preference Pattern

Ryu Bang[†], Baek Hyun-Chul[††], Kim Sang-Bok[†††]

ABSTRACT

This paper suggests a reverse segmentation algorithm using affix information and some preference pattern information of Korean compound nouns. The structure of Korean compound nouns are mostly derived from the Chinese characters and it includes some preference patterns, which are going to be utilized as a segmentation rule in this paper. To evaluate the accuracy of the proposed algorithm, an experiment was performed with 36061 compound nouns. The experiment resulted in getting 99.3% of correct segmentation and showed excellent satisfactory result from the comparative experimentation with other algorithm, especially most of the four or five-syllable compound nouns were successfully segmented without fail.

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

1. 서 론

한국어에서의 복합명사란 둘 이상의 단위 명사가 띄어씀 없이 연결된 명사를 말한다. 한글 맞춤법에서는 연속된 명사 사이는 띄어쓰기가 원칙이나 붙여쓰기를 허용하므로 붙여씀이 자유롭고 글쓰는 사람에 따라 다양한 형태로 나타난다. 이러한 복합명사의 경우 실생활에서는 정보의 전달이나 사용상 문제가 되지 않지만 사전 의존적인 정보 검색 시스템이나 기계번역 시스템에는 심각한 문제를 발생시킨다. 그러므로 복합명사의 적절한 처리가 이루어져야 한다. 이때 고려할 수 있는 방법은 복합명사 자체를 일일이 사전에 등록하는 것과, 복합명사를 명사사전에 존재하는 단위명사 수준으로 분해하는 방법이 있다. 전자의 경우는 단위 명사들이 서로 결합하여 발생하는 복합명사는 수 없이 많기 때문에 사전 등록방법은

※ 교신저자(Corresponding Author) : 김상복, 주소 : 경남 진주시 가좌동 900(660-701), 전화 : 055)751-5994
E-mail : ksb5994@hanmail.net
접수일 : 2003년 6월 3일, 완료일 : 2003년 8월 21일
[†] 진주보건대학 컴퓨터정보기술계열 조교수
(E-mail : bianc@korea.com)
[††] 진주의료원 전산실장
(E-mail : dosi_gas@lycos.co.kr)
[†††] 정회원, 경상대학교 컴퓨터과학과 교수

현실적으로 불가능하다. 따라서 단위명사 사전을 이용하여 주어진 복합명사를 단위명사로 분리하는 방법에 대해 여러 가지 알고리즘이 제시되었다[6-12].

복합명사를 분해할 때 고려해야 할 상황은 중의성 및 미등록어 처리 문제이다. 이 중에서 가장 어려운 점은 중의성 문제이다. 예를 들어, 복합명사 '특기적성교육'을 분해하는 경우 '특', '기', '적', '성', '교', '육', '특기', '기적', '적성', '교육', '특기적', '성교육' 등의 명사가 추출될 수 있다. 즉 단위명사로 '특기적성교육'을 분해했을 때, 단음절의 명사를 제외한 '특기+적성+교육', '특기적+성교육' 등으로 분해가 가능하게 되어 올바른 분해결과를 추출하는데 어려움이 발생하게 된다. 이러한 문제는 한국어에만 존재하는 것이 아니라 한국어와 문법적으로 유사한 구조를 가진 일본어나 한자문화권 국가들에서도 나타나는 현상이다. 중국어의 형태소의 분석에서는 띄어 씀이 없으므로 단어와 단어사이의 경계를 구분하는 과정이 필요하다. 그러므로 입력 문장으로부터 단어의 정확한 분리가 중요하다[1,2]. 일본어도 띄어쓰기를 하지 않기 때문에 복합명사의 분리 과정이 필요하며 이를 위해 여러 가지 의미적, 통사적 자료를 이용하여 복합명사의 구조를 규명하는 연구가 진행되고 있다[3,4].

또 하나는 단어사전에 존재하지 않는 미등록어를 포함한 복합명사의 분해 문제이다. 대표적 미등록어는 수치값, 고유명사, 외국어의 한글표기, 새로 생성된 용어 등을 들 수 있다. 예로서 8음절의 복합명사 '피아노천재볼프강'의 경우, 올바른 의미인 '피아노+천재+볼프강'으로 분리되어야 한다. 이 경우 '볼프강'은 고유명사이므로 단위명사 사전에 검색되지 않기 때문에, 단위사전에 나타나는 단위명사인 '강'이 먼저 분해되고 '볼프'를 미등록어로 처리한다. 이처럼 복합명사 분해 과정에서 중의적 분해가 발생하는 경우 올바른 선택을 하는 문제와 더불어 미등록어를 효율적으로 처리하는 문제를 해결해야 한다.

본 논문에서는 한국어 복합명사의 기반이 **대부분 한자어에서 유래**한 것이 많고, 한자는 **음절마다 의미를 가지고 있으며**, 같은 의미를 지닌 한자어가 대부분 여러 개 존재하지만 **선호하는 음절의 조합으로 단어가 구성된 점**을 착안하여 선호패턴[1]을 활용한 복합명사 분리 방법을 제시한다. 이 방법을 위해 기존의 복합명사 분리 알고리즘 중 가장 분해율의 우수한 역방향 분해 알고리즘[6]을 기반으로 하여 본 논문에서 제시한 **선호패턴 및 접사정보**를 적용하는 방법을 이용한다.

본 논문의 구성은 다음과 같다. 2장에서는 복합명사 분해와 관련된 선행된 연구들에 대해 살펴보고, 3장에서는 선호음절을 이용한 복합명사 분해 알고리즘을 제시하며, 4장에서는 제시한 알고리즘을 이용하여 실험한 결과에 대해 기술하고, 마지막 장에는 결론 및 미진한 부분에 대한 연구부분을 제시한다.

2. 관련연구

복합명사 분해 방법은 대부분 휴리스틱을 이용한 방법으로, 이들을 좀 더 세분하면 사전에 기초한 방법[7,11]과 통계적 처리에 기초한 방법[5,8,12]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. 이 방법은 처리과정보다 결과의 우수성을 찾는 휴리스틱 방법으로서 이론적 증명이 어려우며, 미등록어가 포함되어 있는 복합명사의 경우에는 성능저하가 심하다는 단점이 있다. 그러나 역방향 분해방법은 한국어의 특성상 중심어가 뒤에 나타나기 때문에 사전에 기초한 방법이지만 기존의 방법보다 우수한 결과를 보이고 있으며, 미등록어에 대한 처리도 비교적 우수한 결과를 보이고 있다. 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

1) 국립국어연구원의 복합명사전자사전의 정보를 이용하여 길이별 음절패턴을 추출하여 빈도수가 높은 상위 3가지의 음절분리정보 패턴을 의미함.

후보를 선택하는 알고리즘을 제안하였다. 이 알고리즘의 특징은 길이와 관계없이 미등록어를 포함한 복합명사의 분해가 가능하다. 만약 중의성을 가지는 문제를 만나면 가중치를 이용하여 높은 후보를 선택하는 방법을 이용한다.

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 추가하면서 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.

위 연구들의 복합명사 분해 알고리즘들은 각자의 방법대로 비교적 우수한 결과를 나타내고 있다. 하지만 복합명사의 중의적 분해 문제와 복합명사에 미등록어가 포함되어 있을 경우에서의 처리결과는 상대적으로 낮게 나타나고 있다. 또한 역방향 분해를 제외한 알고리즘에서는 중의성을 증폭시키는 1음절 단위명사에 대한 처리문제를 내포하고 있다. 그리고 기존의 논문들에서는 자음접변, 두음법칙, 연음법칙 등 음절간 상호정보를 활용한 중의성 해결에 대한 고려가 미약하다고 할 수 있다.

3. 복합명사 분해 알고리즘

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

첫째, 복합명사의 중의적 분해시 올바른 선택 문제는 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.

다음으로 미등록어가 포함된 복합명사 분해 문제는 복합명사를 구성하고 있는 단위명사가 사전에 등록되지 않아서 복합명사를 분해하는 데 실패하는 문제이다. 예를 들어, '시운전속력'에서 단위명사 '운전' 및 '속력'은 사전에 등록되어 있으나 '시'에 대한 음절 때문에 복합명사 분해에 실패한다. 복합명사를 정확히 분해하기 위해서는 '시운전+속력'으로 분해하여야한다.

3.1 복합명사 분해 방법에 대한 제안

조합이 가능한 모든 형태의 단위명사를 사전에 등록하는 경우 동일한 단어에 대해 중의성을 증가시키는 문제를 안고 있다. 위 예처럼 '시운전속력'에 대한 분해시 '시운전'과 '전속력'의 2가지 형태로 분해될 경우 '시운전'과 '전속력' 중 올바른 것을 선택하는 문제가 발생한다. 이와 같이 미등록어가 포함된 복합명사에 대한 올바른 분해 또한 제대로 처리될 수 있어야 한다.

이 문제점을 해결하고자 본 논문에서는 다음의 처리 방법을 제안한다.

3.1.1 단음절 명사를 제외한 명사사전 이용

복합명사를 분해하기 위해서 단음절 명사를 제외한 명사로만 구성된 명사사전을 이용한다. 단음절 명사를 제외한 이유는, 한국어에서의 의미를 구성하는 단어의 대부분은 단음절로 구성되는 경우는 거의 없고, 한자어를 2자, 또는 3자를 결합하여 하나의 단어를 구성한다. 신조어 또한 단음절만으로 구성되는 경우가 거의 없기 때문이다. 또한 단음절의 한자음은 다양한 의미를 나타내므로 중의성을 급격히 증가시키기 때문에 제외한다.

3.1.2 최장일치 수록명사 우선 분해

사전 탐색시 최장 일치 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. 이런 경우의 대부분은 '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 존재하는지 탐색을 반복한다.

3.1.3 접사정보를 이용한 역방향 분해

최장일치 수록명사를 찾지 못한 경우에는 끝 방향에서 1음절씩 제외해 나가는 과정에서, 다시 사전에 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 파생어들로서, 이 경우 접사 처리를 선행 하면 중의성 문제를 줄일 수 있다. 특히 접미사의 경우, 명사가 아닌 단어에 접미사를 추가하여 명사화하는 경우가 있으므로 이의 처리도 필요하다. 예를 들면 '않음','같음'등이 이에 해당된다.

접사사전에 없는 경우, 등록된 앞 단어나 뒷단어로 연결해야 하는 미등록어가 존재한다면 이를 미등록어로 등록할 것이 아니라 앞 단어나 뒷단어로 연결 가능한 미등록어인지를 결정하여야 한다.

3.1.4 선호패턴을 이용한 분해

대부분의 복합명사의 형태는 한자 기반의 복합명사가 주류를 이루고 있으며, 결합 구조는 선호하는 길이의 음절을 가지고 있다. 한국어에서는 두음법칙, 자음접변, 순행동화, 역행동화 등의 발음과 관련한 오랜 관습과 관련한 규칙을 가지고 있다. 미등록어로 분리된 단어들에 대해서는 국문법에 기반한 통계적 선호음절 정보를 이용하여 분해한다. 미등록어의 구조 대부분은 주로 2음절이나 3음절 형태의 조합된 구조를 취하고 있으며, 간혹 4음절이상으로 구성된 경우는 다시 2~3음절의 단위명사로 분해 가능한 구조로 되어 있다. 예를 들면 '수신제가치국평천하'는 '수신'+'제가'+'치국'+'평천하'의 2와 3음절로만 구성된 패턴으로 나누어진다.

그러나 2~3음절 분해방법은 반드시 분해된 2~3음절이 단위 사전에 존재하는 경우에만 분해하고, 만약 분리된 음절이 단위 사전에 한 개라도 존재하지 않는다면 미등록어로 처리한다.

본 논문에서는 위에서 순서대로 역방향 분해를 먼저 시도한 후, 실패한 경우 접사 및 선호패턴을 이용한 분해를 시도한다. 이 경우에도 실패하면 통계정보에 의한 2~3 음절 분해 규칙을 이용하여 복합명사를 강제 분해하는 방법을 적용한다.

3.2 사전구성 및 통계정보

3.2.1 단위사전

본 논문에서는 국립국어연구원에서 구축한 단일명사 사전, 의존명사 사전, 복합명사 사전, 접사사전을 토대로 단위명사를 추출하여 단위명사 사전으로 이용하였다. 앞서 언급한 대로 단음절 명사를 단위명사 사전에서 제외한 것은 단음절이 대부분 한자어의 다양한 의미를 수반하기 때문에 중의성을 증폭하는 결과를 초래한다. 예를 들어 '대한민국'의 경우 '대','한','민','국'이 모두 1음절 명사이기 때문에, '대+한+민+국'이라 분해가능하기 때문이다. 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 형태의 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.

접사 사전은 국립국어연구원 세종21 프로젝트에서 구축된 접사사전을 토대로 접두사로 사용되는 것과 접미사로 사용되는 접사, 접두사 및 접미사로 동시에 사용되는 접사 및 접사는 아니지만 첫음절을 선호하거나 끝음절을 선호하는 음절을 접사화 하여 사전을 구성하였다.

본 논문에서는 복합명사 내에서 접사의 사용위치에 따라, 복합명사의 첫머리에 나타나면 접두사로, 마지막에 나타나면 접미사로, 복합명사 중간에 존재하면 접미사로 간주하였다. 예로서 접사 소에 대한 복합명사의 분해는 표 1과 같다.

이는 한국어의 구조상 접두사는 중간보다는 첫머리에 위치하며, 접사의 출현 빈도에서 접미사의 비중이 접두사보다 높기 때문이다.[6]

표 1. 위치에 따른 접사의 분해 예

위치	복합명사	분리 결과	구 분
접두	소시민애환	소+시민+애환	접두사
중간	분향소설치	분향소+설치	접미사
접미	직업소개소	직업+소개소	접미사

숫자를 포함한 복합명사는 해당 숫자를 나타내는 단음절이거나 단위를 지칭하는 명사가 포함되어 있다. 따라서 이의 처리는 숫자를 시작으로 하는 단어에 단위를 지칭하는 접미사의 결합형으로 처리한다. 또한 접두사나 접미사도 아니지만 통계적 자료에서 특정한 음절이 첫머리를 선호하는 음인지, 끝을 선호하는 음인지 검사하여 첫머리와 끝에 나타나는 비율이 5배 이상이면 강제로 접두 처리한다.

다음의 표 2는 본 논문에서 사용한 접사의 종류와 접사화 규칙에 대한 내용이다.

표 2. 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 분리 선호 음절 수

선호 규칙을 위한 통계자료는 중의적 분해 문제에서의 올바른 선택과 관련한 처리를 하거나, 미등록어 처리 시 분해 가능한 경우인지에 대한 정보를 제공한다. 즉 이 자료를 이용하여 미등록어에 대한 분해율을 높이는데 유용하게 사용할 수 있다.

음절의 길이에 대한 통계자료로는 각 복합명사의 길이에 대한 패턴을 국립국어연구원의 복합어 사전과 시중 6개 백과사전 표제어에서 고유명사, 역사자료 및 전문용어를 제외한 자료를 이용하여 통계적으로 조사하여 가장 빈도수가 높은 3개의 패턴을 추출하였다. 추출된 분리 선호패턴은 표 3과 같다.

표 3. 복합명사 분리 선호 패턴

음절수	빈도수 높은 선호 패턴
3음절	1+2/ 2+1, Null
4음절	2+2/ 1+3/ 3+1
5음절	2+3/ 3+2/ 1+4
6음절	3+3/ 2+4/ 2+2+2
7음절	2+2+3/ 3+2+2/ 2+3+2
8음절	2+3+3/ 3+3+2/ 3+2+3
9음절	2+2+2+3/ 3+2+2+2/ 2+3+2+2
10음절	2+2+2+2+2/ 3+2+2+3/ 2+2+3+3
11음절 이 상	홀수개: 2+2+…+3/ 3+2+…+2+2/ n-2개의 첫 번째 패턴 짝수개: 2+2+…+2/ 3+2+…+2+3/ n-2개의 첫 번째 패턴

중요한 사항은 복합명사를 단위명사로 분해한 현재의 패턴이 올바른 것인지에 대한 정답을 알고리즘 자신은 확인할 수 없다는 것이다. 그러므로 검증을 위한 데이터를 이용하여 분해율을 구하여 이 값으로 성능을 결정하고 있다. 그러나 본 논문에서는 선호규칙 설정용 통계자료를 중의성 해결이나, 미등록어에 대한 분해 결정에 앞 뒤 자음간의 빈도정보와, 분리 선호 음절 패턴을 활용하고 있다. 그림 1은 실제 선호규칙용 통계정보를 이용하여 분해를 하기 위한 프로그램의 동작화면 예이다.

이 예에서 '면'의 음절은 첫음절로 326번, 끝음절로 1844번 출현한 것으로서 이 음절은 접미사화 할 수 있는 음절로서 강제 분해가 되는 지점이 된다.

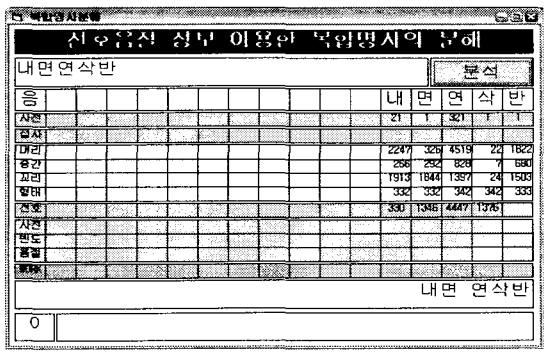


그림 1. 복합명사 분해 프로그램의 동작 화면

3.4 선호규칙을 활용한 분해 알고리즘

본 논문에서 제시하는 선호규칙을 활용한 복합명사 분해의 전체흐름은 그림 2와 같다. 복합명사를 분해하는 과정은, 길이가 N인 복합명사인 경우 끝음

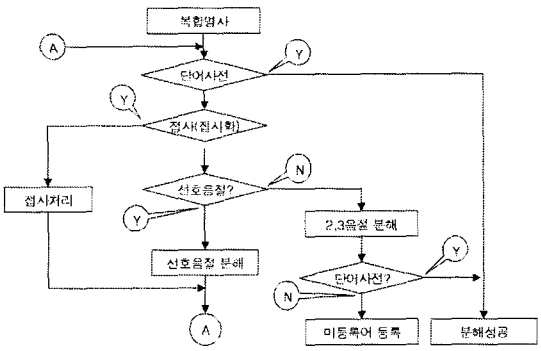


그림 2. 제안한 복합명사 분해 알고리즘

절에서 첫음절 방향으로 단위명사 사전을 이용하여 최장 일치되는 명사를 추출한다. 만약 추출에 실패하면, 끝음절을 임시 음절열에 추가하고, 나머지 [N-1]개의 복합명사열을 가지고 다시 사전탐색을 시작한다. 사전탐색에서 최장일치 단위 명사를 발견시 기존의 임시 음절열에 대해 접사 및 접사화 여부를 판별한다. 만약 임시 음절열이 접미사이거나 접미사화 음절이면 현재 추출된 명사에 임시 음절열을 추가하고, 접두사 또는 접두사화 음절이면 이전에 분해한 최장일치 명사의 첫머리에 임시 음절열을 추가한다. 만약 접미사도 아니고 접두사도 아닌 경우에는 미등록어 처리루틴에서 선호음절과 관련한 처리를 한다. 선호음절 처리루틴이 동작된다는 것은 사전에 존재하지 않는 미등록어를 포함하고 있다는 것을 의미한다. 이런 경우 대부분의 복합명사 분해 알고리즘은 미등록어로 처리한다. 반면 본 알고리즘은 분리된 미등록어를 표 2의 일정한 음절정보의 패턴을 이용하여 재차 분리작업을 시도한다. 이렇게 분리된 미등록어들이 단위명사 사전에 모두 존재하면 미등록어 분리작업을 종료한다. 만약 단위명사사전에 한 개라도 존재하지 않는 단어가 발생한다면 마지막 처리루틴인 강제분리루틴으로 작업을 진행한다. 이 때는 음절의 길이에 따라 2~3음절로 강제 분할을 시도한다. 위의 흐름을 그림으로 도식하면 그림 2와 같다.

복합명사 '바그다드점령임박'에 대해 본 논문에서 제안한 방법으로 분해를 해보면, 먼저 최장길이의 명사가 단어사전에 존재하는지 검사한다. 없다면 접사사전에서 '박'이 존재하는지를 검사한다. 없다면 끝에서 2음절의 단어인 '임박'을 단어사전에서 검색한다. 단위명사 '임박'이 탐색되어 분할 명단에 등록하고 복합명사로부터 '임박'을 제거한다. 다시 복합명사 '바그다드점령'이 단어사전에 존재하는지 검사한다. 실패하면 '령'이 접사사전에 있는지 검사한 후 다시 2음절의 '점령'을 검색한다. 단위명사 '점령'이 탐색되어 분할 명단에 기록하고 복합명사로부터 '점령'을 제거한다, 위의 과정을 반복하여 복합명사 '바그다드'의 끝음절 '드'으로 끝나는 단위명사를 단위명사 사전에서 탐색하지만 탐색에 실패하게 되어, '드'이 건너뛴 음절열에 추가되고 복합명사로부터 제거된다. 나머지 '바그다'의 '다'와 '그' 역시 사전 탐색에 실패하여 건너뛴 음절열에 추가되고 복합명사에서 제거된다. 더 이상 분해할 복합명사가 없으므로 마지막으로 건너뛴 음절열에 있는 '바그다드'를 미등록어로 간주하여 하나의 단위명사로서 분리를 해낸다. 이렇게 해서 최종적으로 '바그다드점령임박'은 '바그다드+점령+임박'으로 분해된다.

4. 실험 및 분석

4.1 실험자료

본 논문에서 제안한 분해 알고리즘의 성능 평가를 위해 국립국어연구원의 국어빈도조사정보와 6대 백과사전에 등재된 복합명사 정보와 인터넷에서의 정보검색 도구를 이용하여 36061개의 복합명사를 추출하였다. 사전에 포함되어 있지 않은 미등록어를 포함하는 실험 데이터는 2715개로 전체의 7.6%를 차지하고 있으며, 미등록어 중에서 특히 접사 파생어는 117개로 전체 데이터의 4.3%에 해당한다. 사전정보 중 추출된 복합명사의 구성 중 역사적 사건이나 고유명사 등에서 파생한 복합명사는 가급적 배제하였다. 이렇게 추출한 복합명사의 음절수별 전체 비율을 표 4와 같다.

표 4. 실험에 사용된 복합명사의 구성 비율

음절수	복합명사 수	구성비율(%)
3	114	0.32
4	23564	65.29
5	7817	21.66
6	2833	7.85
7	1137	3.15
8	371	1.03
9	161	0.45
10	47	0.13
11이상	48	0.13
총 계	36092	100.00

4.2 실 험

추출한 복합명사를 가지고 역방향 분해기법만으로 처리한 결과와 제안한 기법을 이용한 결과를 이용하여 실험한 결과 각각 98.2%, 99.3%의 분해 성공률을 얻었다. 실험한 복합명사의 음절수별 정확율은 표 5와 같다.

표 5. 제안한 알고리즘으로 얻은 결과

음절수	복합명사 수	분해성공 명사 수	비율
3	114	109	0.96
4	23564	23562	0.999
5	7817	7780	0.995
6	2833	2724	0.96
7	1137	1092	0.96
8	371	338	0.91
9	161	151	0.94
10	47	44	0.94
11이상	48	42	0.88
총 계	36092	35842	99.3

성능비교를 위해 같은 실험데이터에 대해 역방향 분해 알고리즘으로 실험한 결과 98.2%의 분해 정확율을 얻었다. 이 값이 역방향 분해 알고리즘에서 얻은 논문의 결과와 다르게 약간 상위의 값을 가진 이유는 실험용 복합명사 데이터가 서로 다르기 때문이다. 따라서 역방향 분해 알고리즘에서 제시한 결과와 직접적인 비교는 할 수 없었고, 역방향 분해 알고리즘에서 제시한 방법으로 구현하여 얻은 결과치와 본 논문에서 제시한 결과를 비교하여 얻은 결과 내용은 그림 3과 같다.

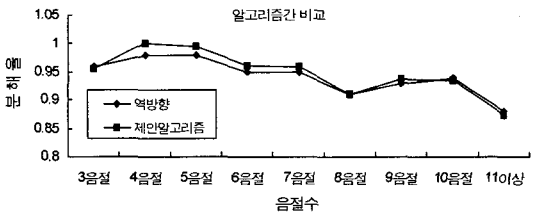


그림 3. 알고리즘간 결과치 비교

4.3 분 석

그림 3에서 보는 바와 같이 본 논문에서 제시한 선호규칙을 이용한 분해 정확율이 역방향 분해 기법의 분해 정확율보다 높게 나타남을 알 수 있다. 특히 4음절과 5음절에서는 복합명사의 분해 정확도는 거의 100%에 가까운 분해율을 보이고 있다. 예로서 '폐가처분'이라는 복합명사를 역방향 분해 알고리즘으로 분해할 경우 '가처분'이라는 최장 길이의 단어가 존재하므로 '처분'이라는 단어로 분리하지 못하기 때문이다. 반면 제시한 알고리즘에서는 먼저 역방향으로 분해한 후 분리된 음절 중 '폐'가 접두사이거나 접두사화 할 수 있는 정보인지 검사하여 이 경우를 충족시키지 못하면 다음 과정인 음절길이에 따른 분할을 시도한다. 그림 4에서 '폐'에 대한 접두사 체크와 접두사화 가능성에서도 모두 실패하여 음절분리 절차를 진행한다. '폐가처분'이 4음절 복합명사이므로 4음절 분리 패턴 중 2+2패턴으로 분리하여 단위사전에 존재여부를 체크한 결과 모두 존재하여 '폐가'와 '처분'으로 분리하기 때문이다. 이러한 구조로 인해 특히 4음절에서는 99.5%, 5음절에서는 99.9% 수준으로 복합명사의 분해가 이루어진다.

그러나 본 논문에서 제시한 기법도 기존의 방법과 마찬가지로 음절 길이가 길어지면서 분해정확율이 감소하는 현상을 보였다. 음절길이가 10을 넘기는 경우 복합된 명사의 데이터를 구하기도 힘들었고 적은 량의 데이터에서 소량의 실패에도 결과 값에 영향을 크게 미치는 결과를 초래하였다. 이 대부분은 복합명사에 고유 명사가 포함되어 있어, 등록된 단위명사의 앞뒤를 분리해 미등록어로 처리하기 때문이다. 이런 오류는 비단 10음절 이상의 복합명사뿐만 아니라 미등록을 포함한 다른 길이의 복합명사에서도 발생하였다.

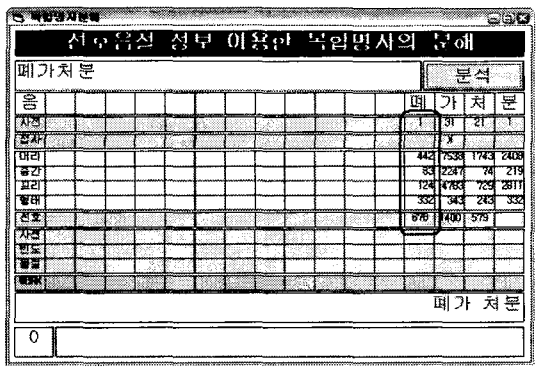


그림 4. 복합명사 분리 예시

본 논문에서 제시한 기법 중 접사의 처리에서도 오류가 발생하였다. 접두사 혹은 접미사로 사용할 수 있는 접사인 경우 복합명사의 처음에 위치할 때는 접두사로 인식하고, 끝음절은 항상 접미사로 인식하도록 하였기 때문이다. 예를 들면 '영업직수입제품'에서는 '영업직'+'수입'+'제품'으로 분리하는 오류가 발생한다. 이의 경우와 같이 접사로 등록된 음을 포함한 일반 명사에 대한 오류가 있으며 이에 대한 정확한 처리가 필요하다.

5. 결 론

기계번역분야에서 자연어 처리를 위해서는 사전의 정보를 필수로 하고 있다. 이 경우 사전에 존재하지 않는 복합명사의 처리 문제로 인해 시스템의 성능에 커다란 영향을 미친다. 한국어에서 복합명사는 한글 맞춤법의 허용으로 인해 명사간 결합을 허용하고 있으므로 번역이나 검색시 단위 명사사전 검색에 실패하게 된다.

본 논문에서는 복합명사의 선호규칙을 이용한 분해 알고리즘을 제안하고 실험하였다. 분해 명사군은 사전을 이용한 탐색을 사용하고, 1음절 명사로 인해서 중의성의 증폭을 막기 위해 2음절 이상의 명사로만 구성된 단위명사 사전을 이용하였다. 또한 접사의 처리를 위해 접사 사전을 구축 사용하였다. 실험에 사용된 복합명사는 국어사전 및 백과사전과 인터넷 검색으로부터 추출된 복합명사를 대상으로 실험한 결과 약 99.3%의 정확도를 얻었다. 실험에 사용된 복합명사들 중 사전 미등록명사는 대부분 접사파생어로서, 제안한 복합명사 분해 기법을 이용하여 미등록어에 대한 분해 결과 다른 방법에 의한 것보다 비교적 높은 분해 정확도를 얻을 수 있었다.

그러나 본 논문에서도 선호 규칙으로 역방향 분해를 우선 적용하기 때문에 최장일치 분해를 시도하는 과정에서 나타나는 오류를 막기 위해 접사와 관련한 정보를 활용하였으나, 이 과정에서 발생하는 분해 결과가 정확한 것인가에 대한 자료는 이 알고리즘에서는 확인할 방법이 없다. 또한 외래어 표기시 미등록어 중간에 등록된 명사가 존재하는 경우 이를 무시하는 처리가 필요하나 현재의 알고리즘으로 이를 적용한 결과 정확율을 더 감소하는 현상을 초래하였으며 성능향상을 위해 이러한 부분의 연구가 추가적으로 요구된다.

참 고 문 헌

[1] K. J. Chen and S. H. Liu, "Word Identification for Mandarin Chinese Sentences," Proceedings of the 14th International Conference on Computational Linguistics, pp.101-107, 1992.

[2] R. Sproat, C. Shih, W. Gale and N. Chang, "A Stochastic Finite-state Word Segmentation Algorithm for Chinese," Proceeding of ACL, 1994.

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

[4] 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.

[5] 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.

[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.

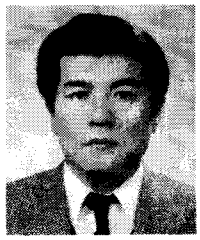
[12] 윤보현, 조정민, 임해창, "총계정보와 선호규칙을 이용한 한국어 복합명사의 분해", 정보과학회 논문지(B), 24권 8호, pp.925-928, 1995.

류 방

1984년 경상대학교 전산통계학과 학사
1993년 경상대학교 전자계산학과 석사
1999년 경상대학교 전자계산학과 박사과정 수료
1993년~현재 진주보건대학 컴퓨터정보기술계열 조교수
관심분야 : 유무선통신, 한국어정보처리, 컴퓨터프로그래밍

백 현 철

1988년 경상대학교 전산통계학과 학사
1998년 경상대학교 교육대학원 전산교육전공 석사
2003년 경상대학교 컴퓨터과학과 박사
1988년~현재 진주의료원 전산실장
관심분야 : 컴퓨터보안(방화벽, 침입탐지), 암호화, 휴먼인터페이스, 멀티미디어 통신



김 상 복

1989년 중앙대학교 전자공학 박사
현재 경상대학교 컴퓨터과학과 교수
관심분야 : 유무선통신, 한국어정보처리, 컴퓨터프로그래밍