미등록어의 의미 범주 분석을 이용한 복합명사 분해

강 유 환* · 서 영 훈**

Segmentation of Korean Compound Nouns Using Semantic Category Analysis of Unregistered Nouns

Yu-Hwan Kang* · Young-Hoon Seo**

Abstract

This paper proposes a method of segmenting compound nouns which include unregistered nouns into a correct combination of unit nouns using characteristics of person's names, loanwords, and location names. Korean person's name is generally composed of 3 syllables, only relatively small number of syllables is used as last names, and the second and the third syllables combination is somewhat restrictive. Also many person's names appear with clue words in compound nouns. Most loanwords have one or more syllables which cannot appear in Korean words, or have sequences of syllables different from usual Korean words. Location names are generally used with clue words designating districts in compound nouns. Use of above characteristics to analyze compound nouns not only makes segmentation more accurate, helps natural language systems use semantic categories of those unregistered nouns. Experimental results show that the precision of our method is approximately 98% on average. The precision of human names and loanwords recognition is about 94% and about 92% respectively.

Keywords : Natural Language Processing, Compound Noun Analysis, Unregistered Word Recognition

논문접수일 : 2004년 2월 27일 논문게재확정일 : 2004년 11월 10일

※ 본 연구는 한국과학재단 목적기초연구(R05-2003-000-11978-0) 지원으로 수행되었음.

* 충북대학교 컴퓨터공학과 박사과정

** 충북대학교 전기전자컴퓨터공학부 교수

1. 서 론

한국어에서 명사와 명사는 띄어쓰는 것을 원칙으로 하나 붙여써도 무방하기 때문에 복합명사의 형태가 다양하다. 또한 의미적 분석 없이 복합명사를 단위 명사들로 분해할 경우 여러 형태로 분해될 수 있기 때문에 많은 어려움이 있다. 복합명사를 올바르게 분해하는 것은 기계번역, 정보검색, 맞춤법 검사 등과 같은 자연어 처리 시스템의 성능에 큰 영향을 줄 수 있기 때문에 매우 중요하다.

복합명사 분해 문제를 해결하기 위한 가장 간단한 방법은 모든 복합명사의 형태를 사전에 등록하는 것이다. 그러나 모든 복합명사를 사전에 등록하는 것은 불가능하기 때문에 복합명사 분해를 위한 많은 연구들이 수행되어 왔다.

최재혁[1996]은 음절수에 따라 미리 정의해 놓은 복합명사의 분해 패턴을 이용한 복합명사 분해 방법을 제안하였다. 윤보현[1997]은 통계 정보와 선호 규칙을 이용하여 한국어 복합명사를 단위 명사로 분해하는 방법을 제안하였다. 또한 미등록어가 포함된 복합명사를 분해하기 위해 네 가지의 휴리스틱을 사용하였다. 심광섭[1997]은 합성된 상호 정보를 이용하여 띄어쓰기가 되어 있지 않은 한국어 복합명사를 단위 명사로 분리하는 알고리즘을 제안하였다. 합성된 상호 정보는 네 가지 유형의 음절간 상호 정보를 합성한 것으로 주어진 복합명사에서 단위 명사로 분리 가능한 지점을 선택하는데 사용된다. 강승식[1998]은 형태소 분석 결과로 추정된 복합명사를 단위 명사들로 분해하는 방법으로 네 개의 분해 규칙과 두 가지 예외 규칙을 사용하여 가능한 분해 후보들을 생성하고, 분해 후보들에 대해 가중치를 부여함으로써 최적 후보를 선택하는 알고리즘을 제안하였다. 정래정[1996]은 고유 명사 출현 패턴 정보와 부가 정보를 이용한 미등록 고유 명사의 색인 방법을 제안하였다. 이 연구는 고유 명사의 출현 패턴을 조사하여 고유 명사가 존재하는 부근에 나타나는 실마리 단어로 고유 명사를 인식한다. 박봉래[1998]는 동일한 미등록어가 사용된 용례 어절 또는 용례 구의 비교 분석을 통한 미등록어 인식 방법을 제안하였다. 동일한 미등록어가 사용된 어절은 유사한 형태를 가진다는 점을 이용하여 용례 어절을 추출하고, 동일한 미등록어는 각 용례 어절에서 동일한 문법 기능을 가진다는 점을 이용하여 미등록어를 인식한다. 이재성[2001]은 번역문에서 외래어 표기 용례를 자동 구축하기 위한 외래어 인식에 관한 연구에서 음절 정보를 이용한 외래어 인식을 제안하였다.

많은 연구들이 복합명사 분해와 미등록어 인식을 위해 수행되었고, 대부분의 복합명사 시스템들은 98% 이상의 높은 분석 정확률을 보인다. 그러나 사람 이름이나 외래어, 지명과 같은 미등록어를 포함하고 있는 복합명사를 분해할 경우에는 분석 정확률이 떨어진다. 또한 이전 연구들은 미등록어와 단위 명사간의 분해 위치만 결정할 뿐 미등록어의 의미 범주를 결정짓지 못한다는 단점이 있다. 미등록어를 사람 이름, 외래어, 지명 등으로 구분하여 인식할 수 있다면 개체명 인식기나 질의-응답 시스템 등 다른 자연어처리 시스템에서 보다 유용하게 사용될 수 있다.

본 논문에서는 사람 이름, 외래어, 지명과 같은 미등록어를 포함하고 있는 복합명사를 올바르게 분해함으로써 복합명사 분해의 정확률을 높이고, 미등록어의 품사 정보뿐만 아니라 미등록어가 사람 이름, 외래어, 지명인지에 대한 의미 범주 정보를 함께 제공해 줄 수 있는 방법을 제안한다. 미등록어를 포함하고 있는 복합명사를 효율적으로 분해하기 위하여 사람 이름, 외래어, 지명에 나타나는 음절의 다양한 특성과

실마리 정보를 이용한다.

2. 복합명사에서 미등록어로 인한 오류

복합명사에 포함되어 있는 미등록어는 복합명사 분해 시에 분석 정확률을 떨어뜨리는 원인 중의 하나이다. 미등록어의 일부 음절이 사전에 등재되어 있는 한 단어와 일치할 경우 잘못 분해될 수 있다.

예를 들어, '김대중대통령'에서 '김대중'이 미등록어이고, '대중'과 '대통령'이 등록어일 경우 '김 + 대중 + 대통령'으로 분해될 수 있다. 그러나 '김대중'이 사람 이름으로 인식될 수 있다면, '김대중 + 대통령'으로 올바르게 분해될 수 있다.

'브루나이국왕'은 외래어가 미등록인 경우의 예이다. '브루나이국왕'은 '브루 + 나이 + 국왕'으로 분해될 수 있다. 이러한 분석 오류는 '브루나이'라는 미등록어에 '나이'라는 단위 명사가 포함되어 있기 때문이다.

'사창동사거리'는 지명이 미등록어로 사용된 경우의 예이다. '사창동사거리'에서 '사창동'은 미등록어이고, '사창', '동사', '거리'는 등록어이기 때문에 '사창 + 동사 + 거리'로 잘못 분해될 수 있다.

따라서, 미등록어를 올바르게 인식할 수 있는 방법에 대한 연구는 복합명사 분해의 분석 정확률을 높일 수 있는 좋은 연구이다.

3. 외래어 인식

본 논문에서는 외래어 음절의 출현 특성과 음소 결합 특성을 이용한 외래어 인식 방법을 제안한다.

외래어는 원래 외국어였던 것이 국어의 체계에 동화되어 사회적으로 그 사용이 허용된 단어로써 한국어의 음절 중에 비교적 사용 빈도가 낮은 음절들로 구성되어 있다. 외래어는 대부분 인명 또는 지명이기 때문에 색인어로써 중요한 역할을 한다. 하지만 사전 검색으로 외래어를 인식하기 위해 모든 외래어를 사전에 등록하는 것은 현실적으로 어려운 일이다. 왜냐하면, 외래어의 특성상 많은 단어들이 계속 생성되고 소멸되기 때문이다. 따라서 외래어 인식을 위한 별도의 외래어 인식 알고리즘이 필요하다.

3.1 음절 출현 특성

음절은 외래어 판단의 중요한 정보로 이용된다. 음절들 중에서 몇몇 음절은 한국어에서 많이 사용되고, 몇몇 음절은 외래어에서 많이 사용된다. 또한 몇몇 음절 중에는 한국어에서만 사용되는 음절이 있고, 외래어에서만 사용되는 음절이 있다. 이러한 음절의 특성은 한 단어가 외래어인지 아닌지를 구분 짓는 좋은 정보로 이용된다.

음절의 특성을 수집하기 위해 본 연구실의 전자사전에서 한국어만으로 구성된 단어들을 추출한 후, 한국어에서 많이 출현하는 음절들에 대한 통계 정보를 수집하였다. 또한 엔사이버[엔사이버, 2003]와 다른 웹 사이트들로부터 약 5,000개의 외래어 단어를 수집한 후, 외래어에서 많이 출현하는 음절들에 대한 통계 정보를 수집하였다. 또한 수집한 단어들을 이용하여 한국어에서만 나타나는 음절 집합과 외래어에서만 나타나는 음절 집합을 구축한 후 각각 OK (Only-Korean) 셋과 OF(Only-Foreign) 셋으로 이름지었다.

본 시스템에서는 한 단어에서 하나 이상의 음절이 OK 셋에 포함되어 있다면 그 단어를 한국어로 인식하고, 그 반대의 경우에는 외래어로 인식한다. 예를 들어, 음절 '랠'은 외래어에서만 나타나는 음절이므로 시스템은 단어 '랠리'를

외래어로 인식한다.

'월드컵예선'은 외래어 인식의 좋은 예이다. 만일 '월드컵'이 미등록어라면, 시스템은 분석을 실패한다. 이러한 경우에, 시스템은 미등록어를 외래어로 인식하기 위해 다시 분석을 시도한다. '월드컵'에서 음절 '월'은 한국어에서 비교적 많이 나타나는 음절이고, 음절 '드'와 '컵'은 외래어에서 자주 나타나는 음절이다. 다음은 '월드컵'의 각 음절에 대한 통계값을 보여준다.

월(worl) : -15.2
드(de) : 183.7
컵(cup) : 1.6

이러한 통계값은 아래 수식을 통해 구한다.

- 음절 통계값 = (외래어에서 음절의 출현 빈도수 / 총 외래어 음절 수) - (한국어에서 음절의 출현 빈도수 / 총 한국어 음절 수)

시스템은 각 음절의 통계값을 더한 후, 그 값이 임계치(0)보다 크기 때문에 '월드컵'을 외래어로 인식한다. 그 결과, '월드컵예선'은 '월드컵+예선'으로 분해된다.

3.2 음소 결합 특성

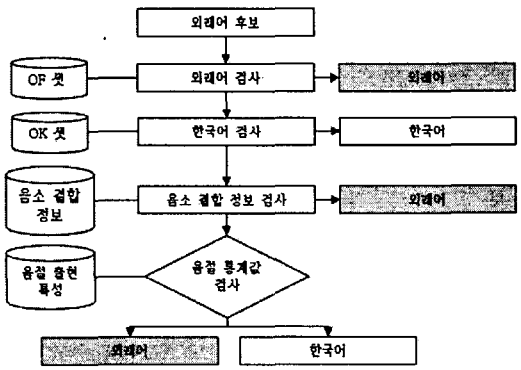
음소 결합 특성은 음절의 출현 특성과 비슷하게 외래어 인식에 이용된다. '커피숍'의 마지막 음절 '숍'은 중성 'ㅛ'와 종성 'ㅂ'으로 구성되어 있으며, 이러한 결합 특성은 외래어에서 자주 발생한다. 음소 결합 특성은 한국어에서 나타나는 음절 셋과 외래어에서 나타나는 음절 셋의 비교를 통하여 획득하며, 외래어를 보다 정확히 인식할 수 있도록 도와준다.

<표 1>은 음소 결합 규칙의 일부를 보여준다. 규칙 1번은 초성에 'ㅂ'이 오고, 중성에 'ㅠ'가 오면 종성에 상관없이 외래어로 인식함을 나타낸다. 규칙 20번은 중성에 'ㅛ'가 오고, 종성에 'ㅂ'이 오면 초성에 상관없이 외래어로 인식함을 나타낸다.

〈표 1〉 음소 결합 규칙

규칙번호	초성	중성	종성	타입
1	ㅂ	ㅠ	All	1
...				
20	All	ㅛ	ㅂ	2
...				

타입 1은 초성과 중성간의 음소 결합 특성을 나타내고, 타입 2는 중성과 종성간의 음소 결합 특성을 나타낸다. 본 연구에서는 타입 1에 해당하는 조합 규칙 19개와 타입 2에 해당하는 조합 규칙 14개를 구축하였다.



〈그림 1〉 외래어 인식 순서도

<그림 1>은 외래어 인식을 위한 순서도이다. 미등록어가 들어오면 먼저 외래어에서만 사용되는 음절이 있는지 조사한 후, 외래어에서만 사용되는 음절이 있다면 미등록어를 외래어로 인식하고, 한국어에서만 사용되는 음절이 있다면 한국어로 인식한다. 위의 두 경우를 모두 만족하지 못할 경우에는 음소 결합 정보를 이용하여 외래어인지를 판단한다. 마지막으로 각 음절의 출현 특성 정보를 이용하여 음절 통계값이

임계치(0) 이상일 경우 외래어로 인식한다.

4. 이름 명사 인식

본 연구에서는 사람 이름에 나타나는 음절의 특성과 실마리 정보를 이용하여 미등록어를 사람 이름으로 인식한다. 한국 사람 이름의 특성은 95% 이상이 3음절로 되어 있고, 성씨로 사용되는 음절의 수가 제한적이며, 이름에 한자 독음을 많이 사용한다는 것이다. 또한 실마리 단어가 복합명사에서 사람 이름과 함께 자주 나타난다.

한국 통신에서 제공하는 전화번호부로부터 외국인 이름을 제외한 사람 이름을 추출하였다. 그 결과 성씨로 사용될 수 있는 400여개의 음절을 추출하였고, 이중 상위 50개의 음절만을 성씨 사전으로 구축하였다. 50개의 성씨만으로도 대부분의 이름을 인식할 수 있으며, 모든 음절을 사용하는 것보다 일부 음절만을 사용하는 것이 효율적이다.

성씨 사전 이외에 이름의 두 번째 와 세 번째에 사용되는 음절에 대한 통계 정보를 각 음절의 출현 빈도수를 이용하여 구축하였으며, 두 음절간의 bigram 정보를 구축하였다. 마지막으로 복합명사에서 사람 이름과 함께 나타나는 실마리 단어를 수집하였다.

4.1 실마리 단어 구축

실마리 단어는 이름 명사 인식의 효율을 높이기 위해 사용된다. 본 연구에서는 복합명사에서 이름과 함께 나타나는 50개 이상의 2음절 실마리 단어를 수집하였다.

복합명사에서 실마리 단어가 나타나면 실마리 단어 앞의 3음절은 이름 명사일 가능성이 높다. 하지만 모든 경우에 있어서 이름 명사이지는 않다. '현대차사장'의 경우 '사장'이 실마리 단어이지만 '현대차'가 사람 이름은 아니다. 따라서 '현대차'가 사람 이름인지의 여부를 이름 인식을 통하여 확인한다.

2음절 실마리 단어 이외에도 1음절과 3음절 실마리 단어를 수집하였다. 1음절 실마리 단어에는 '씨', '군', '옹', '님', '양', '전' 등이 있다. '님'과 같은 경우 '성철스님'과 같이 마지막 음절에 '님'이라는 실마리 단어가 나타나지만 이러한 경우에는 '스님'이라는 단어를 이용하여 예외 처리를 수행함으로써 이름으로 잘못 인식되는 경우를 방지한다. 3음절 실마리 단어의 경우에는 대부분이 2음절 실마리 단어에 접미사가 결합된 형태이기 때문에 이들을 제외한 '대변인'과 같이 2음절 실마리 단어에 포함되지 않은 단어만을 실마리 단어로 수집하였다.

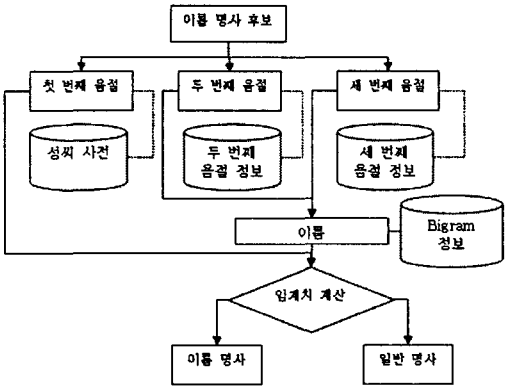
〈표 2〉 실마리 단어와 예외 처리의 예

실마리 단어	예외 처리
님	형님, 스님, …
판 사	출판사
사 장	이사장, 간사장, …

<표 2>는 1, 2음절 실마리 단어에 대한 예외 처리의 예를 보여준다. '님'이라는 실마리 단어가 나온 경우에는 실마리 단어 앞에 있는 음절을 조사하여 '형님', '스님'과 같이 예외 처리할 상황인지 판단한다. 예를 들어, '영진출판사'의 경우, 마지막에 오는 '판사'라는 실마리 단어만을 보면 '영진출'을 이름 또는 추정 명사로 인식할 수 있다. 따라서, '판사'라는 실마리 단어가 나타난 경우 앞 음절을 조사하여 예외 처리 상황인지를 판단함으로써 '영진+출판사'로 올바르게 분해한다.

<그림 2>는 이름 명사 인식을 위한 순서도이다. 이름 명사 후보가 입력되면 첫 번째 음절이 성씨 사전에 들어 있는지 조사하고, 두 번째

음절과 세 번째 음절 그리고 bigram 정보에 대한 통계값을 계산한다. 통계값은 아래와 같이 구한다.



〈그림 2〉 이름 명사 인식 순서도

- 통계값 = bigram값 * (두 번째 음절의 빈도수 + 세 번째 음절의 빈도수) / 1000

첫 음절이 성씨로 사용되는 음절이고, 통계값이 임계치(5) 이상이면 이름 명사로 인식한다.

5. 지명 인식

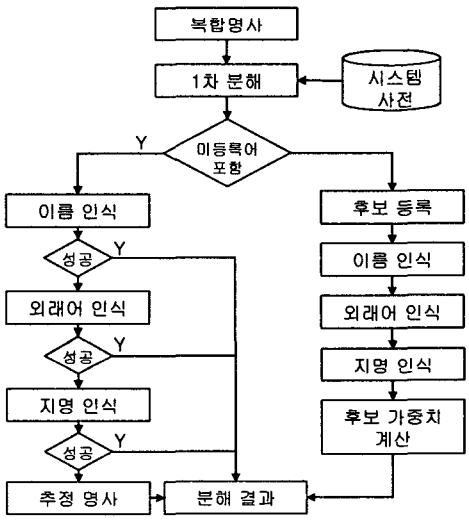
지명 인식은 지명 사전과 실마리 단어 사전을 이용한다.

지명 사전은 우리나라 행정구역명과 널리 알려져 있는 산, 강 이름을 포함하여 구축하였다. 행정구역명은 웹 사이트[우정사업본부, 2003]에서 우편 번호부를 다운 받아 마지막 음절이 '도', '시', '구', '군', '동', '면', '읍', '리'로 끝나는 어절만을 이용하여 15,860개의 행정구역명을 추출하였다.

'사창동사거리'에서 '사창', '동사', '거리'는 모두 사전에 등재되어 있는 단어이므로, '사창' + '동사' + '거리'로 분해될 수 있지만, 지명 사전과 실마리 단어 사전을 이용하면 '사창동' + '사거리'로도 분해될 수 있다. 이러한 분해 중의성은 6음절 복합명사의 경우 3/3 분해가 2/2/2 분해보다 우선한다는 통계 정보를 이용하여 분해 중의성을 해결한다.

6. 시스템 구성

<그림 3>은 복합명사 분해 과정을 보여준다. 복합명사가 입력되면 시스템 사전을 이용하여 단위 명사들로 분리한다. 만일 복합명사에 미등록어가 포함되어 있다면 이름 인식, 외래어 인식, 지명 인식의 순서로 미등록어 인식을 시도하고, 미등록어 인식에 실패했을 경우에는 추정 명사로 결과를 제시한다. 미등록어를 포함하지 않는 경우에는 분해 중의성 문제를 해결하기 위하여 가능한 분해 후보를 모두 구한다. 최종적으로, 분해 후보들에 대한 가중치를 계산함으로써 최적의 분해 후보를 선택한다. 후보에 대한 가중치는 복합명사의 음절 길이별 분리 패턴에 대한 통계 정보와 최장일치법을 이용한다.



〈그림 3〉 복합명사 분해 시스템

7. 실험 및 분석

사람 이름 인식과 외래어 인식의 정확도를 먼

저 테스트하였다. 외래어 인식을 테스트하기 위해 ETRI의 품사 부착 말뭉치[MATEC99, 1999]로부터 6,828개의 명사를 수집한 후 일반 명사와 외래어로 구분하였다. 6,828개의 명사 중 외래어의 수는 551개였고, 그 중 시스템에 의해 외래어로 인식된 명사의 수는 507개였다.

<표 3>은 외래어 인식의 정확률과 재현율을 보여준다. 정확률과 재현율은 다음과 같이 계산한다.

- 정확률 = 시스템이 올바르게 인식한 외래어 명사 수 / 시스템이 외래어로 인식한 명사 수
- 재현율 = 시스템이 올바르게 인식한 외래어 명사 수 / 총 외래어 명사 수

〈표 3〉 외래어 인식의 정확률과 재현율

구 분	정 확 률	재 현 율
성 능	94%	87%

사람 이름과 일반 명사 사이에는 종종 모호성이 존재한다. 예를 들어, '현미경'은 일반 명사이지만 이름으로도 인식될 수 있다. 그러나 이와 같은 경우에는 사람 이름으로 간주하지 않는다.

사람 이름 인식을 테스트하기 위해 웹 문서로부터 2,190개의 3음절 명사를 수집한 후 일반 명사와 사람 이름으로 구분하였다. 2,190개의 명사 중 사람 이름의 수는 709개였고, 그 중 시스템에 의해 사람 이름으로 인식된 명사의 수는 697개였다. 이름 명사와 일반 명사로 동시에 인식될 수 있는 23개의 명사는 이름에서 제외하였다.

<표 4>는 이름 명사 인식의 정확률과 재현율을 보여준다.

〈표 4〉 이름 명사 인식의 정확률과 재현율

구 분	정 확 률	재 현 율
성 능	92%	90%

정확률과 재현율은 다음과 같이 계산한다.

- 정확률 = 시스템이 올바르게 인식한 사람 이름 수 / 시스템이 사람 이름으로 인식한 명사 수
- 재현율 = 시스템이 올바르게 인식한 사람 이름 수 / 총 사람 이름 수

마지막으로, 위의 실험을 토대로 복합명사 분해의 성능을 테스트하였다. 인터넷 신문기사로부터 4음절에서 6음절 길이의 복합명사 1,561개를 수집한 후 음절 길이별로 복합명사 분해의 정확도를 측정하였다. 1,561개의 복합명사 중 사람 이름, 외래어, 지명을 포함하고 있는 복합명사의 개수는 437개이다.

〈표 5〉 복합명사 분해 정확률

구분 / 음절	어절 수	오분석 어절 수	정확률
4음절	755	1	98.9%
5음절	437	14	96.8%
6음절	369	5	98.6%
전체	1,561	29	98.1%

<표 5>는 복합명사의 음절 길이별 분석 정확도를 보여준다. 미등록어를 포함하고 있는 복합명사의 분해 정확률은 평균 98.1%이다.

분석 오류는 대부분 접미사를 잘못 분해하였거나 분해 중의성으로 인하여 발생했다. '진료비실사'는 접미사로 인한 분해 오류의 예이다. '진료'와 '비실', '실사'가 등록어이나 '비'보다는 '사'가 접미사로 자주 쓰이기 때문에 '진료/nc + 비실/nc + 사/xsn'로 잘못 분석되었다. 앞의 오류 유형 외에 '경협'과 같이 축약형태가 포함된 복합명사 분해 시에 오류가 발생했다.

8. 결 론

본 논문에서는 사람 이름, 외래어, 지명과 같

은 미등록어를 포함하고 있는 복합명사를 단위 명사들로 분해하고, 미등록어의 의미 범주를 결정지을 수 있는 방법을 제안하였다.

미등록어 인식을 위해 사람 이름, 외래어, 지명에서 나타나는 음절 특성과 실마리 단어를 이용하였다. 사람 이름, 외래어, 지명과 같은 미등록어를 인식함으로써 미등록어를 포함하는 복합명사의 분해 정확률을 높일 수 있었다.

복합명사 분해 실험 결과 평균 98%의 분석 정확률을 보였으며, 사람 이름 인식과 외래어 인식 실험에서는 각각 94%와 92%의 정확률을 보였다.

현재는 외래어 인식의 정확률 향상을 위해 bigram 혹은 trigram 정보를 활용하는 방법과 일본인 이름과 같은 외국인 이름 인식을 위한 방법을 연구 중이다.

참 고 문 헌

[1] 강승식, "한국어 복합명사 분해 알고리즘", *정보과학회논문지(B)*, 제25권 1호, 1998년 1월, pp. 172-182.

[2] 박봉래, 황영숙, 임해창, "용례 분석에 기반한 미등록어의 인식", *정보과학회논문지(B)*, 제25권 제2호, 1998년 2월, pp. 397-407.

[3] 심광섭, "합성된 상호 정보를 이용한 복합명사 분리", *정보과학회논문지(B)*, 제24권 11호, 1997년 11월, pp. 1307-1317.

[4] 윤보현, 조민정, 임해창, "통계정보와 선호규칙을 이용한 한국어 복합명사의 분해", *정보과학회논문지(B)*, 제24권 8호, 1997년 8월, pp. 900-909.

[5] 이재성, "번역문에서의 외래어 표기용례 자동 구축", *충북대학교 컴퓨터정보통신연구논문지*, 제9권 2호, 2001년 11월, pp. 25-33.

[6] 정래정, 김준태, "고유명사의 출현 패턴을 이용한 색인의 성능 향상에 관한 연구", *제8회 한글 및 한국어 정보처리 학술대회*, 1996년 10월, pp. 68-72.

[7] 최재혁, "음절수에 따른 한국어 복합 명사 분리 방안", *제8회 한글 및 한국어 정보처리 학술대회*, 1996년 10월, pp. 262-267.

[8] 엔사이버, "두산세계대백과", http://www.encyber.com, 2003년

[9] 우정사업본부, "우편번호안내", http://www.koreapost.go.kr, 2003년

[10] MATEC99, "ETRI 품사 태그 부착 말뭉치: MATEC99", *한국전자통신연구원 컴퓨터·소프트웨어 기술연구소*, 1999년.

저자소개

강 유 환

충북대학교 컴퓨터공학과 박사과정을 수료하였으며, 컴퓨터공학과 학사(1998), 컴퓨터공학 석사학위(2000)를 취득하였다. 주요 관심분야는 자연언어처리, 기계 번역, 질의-응답 시스템 등이다. E-mail은 eric@nlp.chungbuk.ac.kr이다.

서 영 훈

현재 충북대학교 전기전자컴퓨터공학부 교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 석·박사학위(1991)를 취득하였다. 주요 연구분야는 자연언어처리, 기계 번역, 구문 분석, 정보 검색 등이다. E-mail은 yhseo@chungbuk.ac.kr이다.

◈ 이 논문은 2004년 2월 27일 접수하여 1차 수정을 거쳐 2004년 11월 10일 게재확정되었습니다.