# 한국어 동형이의어 태깅 시스템 구현*

김준수[0] 최호섭 이왕우 옥철영
울산대학교 컴퓨터정보통신공학과
{kimjunsu[0], hoseop, wwlee, okcy}@mail.ulsan.ac.kr

# Implementation Tagging System of Korean Homonym

Jun-Su Kim, Ho-Seop Choe, Wang-Woo Lee, Cheol-Young Ock
Dept. of Computer Engineering and Information Technology, University of Ulsan

**요 약**

본 논문은 한국어 정보처리에서 발생하는 단어 중의성 문제를 해결하기 위하여, 사전 뜻풀이 말뭉치에서 추출하여 구축한 의미정보 데이터베이스(Semantic Information Database)와 이를 활용한 단어 중의성 해결 모델을 이용한 실용적인 동형이의어 태깅 시스템 개발을 목적으로 한다. 중·소규모의 국어사전 150,000여 개의 표제어 전체의 뜻풀이에 품사 태그를 부착한 117만 어절 규모의 뜻풀이 말뭉치를 구축한 후 사전에 등재된 14,000여 개의 동형이의어 중 뜻풀이에 나타나는 8,164개의 동형이의어에 표제어 어깨번호를 이용한 의미 태그를 부착함으로써, 대량의 동형이의어 분별을 위한 공기(cooccurrence)하는 단어와 빈도(frequency)정보를 추출하여 데이터베이스화 할 수 있었다. 본 동형이의어 태깅 시스템의 정확률 측정과 문제점 발견을 위해 [21세기 세종 계획] 프로젝트에서 제공하는 150만 어절 의미 태그 부착 말뭉치를 대상으로 실험한 결과 세종 150만 의미 태그 부착 말뭉치에 고빈도로 출현하는 469개 어휘(총 출현 횟수 249,249개)에 대한 정확률이 91.58%로 나타났다.

## 1. 서론

자연언어처리의 목표는 "컴퓨터가 자연언어로 인간과 의사소통이 가능하게 하는 효율성 있는 기술"을 개발하는 것으로, 품사 태깅, 형태소 분석, 구문 분석, 의미 분석, 담화 분석 등의 자연언어처리 기술을 비롯하여 기계번역, 정보검색, 자연언어 인터페이스, 문서교정 및 되고, 문서분류, 자동요약, 음성인식, 음성대화, 언어학습 등과 같은 자연언어처리 응용 분야에 이르기까지 자연언어처리와 관련한 다양한 연구가 진행되고 있다. 그러나 자연언어처리 과정에서 발생되는 여러 형태의 중의성(ambiguity) 문제는 자연언어처리 기술 연구뿐만 아니라 응용 분야에까지 자연언어처리의 정확률을 저하시키거나 시스템의 효율성을 저해시키는 요인이 되고 있다. 따라서 자연언어처리의 중점적인 연구 목표 중의 하나가 중의성 해결이라 할 수 있다. 이러한 중의성 문제 중 품사적 중의성, 구문적 중의성에 대한 해결 방안 연구는 어느 정도 진행되었으나, 의미적 중의성 문제에 대한 연구는 아직 미흡한 상태이다.

자연언어처리에서의 의미적 중의성은 특히 동형이의어(homograph)나 다의어(polysemous)처럼 형태적 동일성으로 인해 발생되는 문제이다. 동형이의어란 단어의 형태만 같을 뿐 각 단어가 갖는 의미가 전혀 다른 언어를 말하며, 다의어는 하나의 의미에서 세분화된 의미를 갖는 단어를 말한다. 외국의 단어 중의성 해결(Word Sense Disambiguition: WSD) 연구들을 보면 주로 다의어 해결을 목적으로 하고 있다. 하지만 한국어의 특성상 많은 어휘들이 한자어에서 유래되어 동형이의어가 많이 발생하게 되었다. 그러므로 한국어 정보처리에서는 다의어 수준의 의미 중의성 해결에 앞서 동형이의어에 대한 정확한 분별이 선행되어야 한다.

본 논문의 목적은 한국어 동형이의어 중의성 해결에 있어 실험실 수준의 소량의 어휘를 대상으로 하던 기존의 국내 연구와는 달리, 중·소규모의 국어사전 150,000여 개의 표제어 전체의 뜻풀이에 품사 태그를 부착한 117만 어절 규모의 뜻풀이 말뭉치를 구축한 후 사전에 등재된 14,000여 개의 동형이의어 중 뜻풀이에 나타나는 8,164개의 동형이의어에 표제어 어깨번호를 이용한 의미 태그를 부착함으로써, 뜻풀이에서 동형이의어와 공기(cooccurrence)하는 단어와 그 빈도(frequency) 정보를 추출 할 수 있으며, 이들 빈도 정보를 이용하는 동형이의어 중의성 해결 모델을 통해 다량의 동형이의어 분별이 가능하도록 한다. 또한 동형이의어 중의성 해결 시스템 구축의 핵심적인 데이터인 품사 태그 부착 뜻풀이 말뭉치, 의미 태그 부착 뜻풀이 말뭉치, 의미정보 데이터베이스 등을 자연언어처리 각 분야에 적절히 이용할 수 있도록 하는 것도 부수적인 목적으로 한다.

## 2. 관련 연구

단어 중의성 해결 방법은 다음과 같은 기본적인 접근 방법이 있다. 첫째는 통합된 규칙 대 규칙(rule-to-rule) 접근 방식으로서, 잘못 형성된 의미 표현을 의미 분석 과정에서 함께 제거하는 방법이다. 둘째는 독립 접근 방식으로서, 단어 중의성 해결을 복합적인 의미 분석과는 독립적으로 수행하는 방법이다. 전자에 속하는 대표적인 WSD 방법이 선택 제약(selectional restriction)에 기반한 WSD이며, 후자에 속하는 것은 통계 기반 WSD이다. 통계 기반 WSD는 다시 학습시킬 데이터의 형태에 따라 세 가지의 방법으로 구분되는데, 정보 이론(information theory) 기반 중의성 해결, 베이지안 분류(Bayesian classification) 등과 같은 레이블(label) 처리된 학습 집합(training set)에 기반한 지도 학습 중의성 해결

* 이 연구는 정보통신부 대학 기초연구 지원 사업(2002년)의 지원으로 수행된 결과임.

(supervised disambiguition) 방법과 아무런 처리가 되어 있지 않은 문장 말뭉치만을 학습시키는 비지도 학습 중의성 해결 (unsupervised disambiguition) 방법, 그리고 사전(dictionary)이나 시소러스(thesaurus)와 같은 언어 자원(language resource)에 기반한 사전 기반 중의성 해결(dictionary-based disambiguition) 방법으로 나눌 수 있다. 최근 국내에서도 이러한 다양한 방법을 이용하여 단어 중의성을 해결하고자 하고 있으나, 소량의 동형이의어를 대상으로 평가하는 데 그치고 있는 실정이다. 기존의 동형이의어 중의성 해결과 관련된 국내의 몇몇 연구들을 정리하면 다음과 같다.

① 조정미(1998)는 말뭉치와 사전을 이용한 단어 중의성 해결 방법을 제안하였다. 간단한 사전 분석을 통해 의미 분별하고자 하는 단어의 의미 지시자와 단어의 분류 정보를 추출하였으며, 말뭉치로부터는 정규화 과정을 거친 다음, 목적어 관계의 선택 제약을 이용하여 단어간 유사성을 학습하였다. 또한, 말뭉치의 자료 부족 현상을 해소하기 위하여 선택 제한 지식을 명사 분포와 동사 분포로 표현하고 이를 이중적으로 의미 분별에 적용하였다. 10개의 한국어 타동사를 대상으로 실험하였는데, 명사 분포와 동사 분포를 함께 이용한 실험이 명사 분포만을 이용한 것보다 재현율이 약 22.1% 향상되었고, 사전 정보를 이용함으로써 정확률이 약 25.5% 정도 향상되었다. 그러나, 선택 제한 지식으로 목적격만 사용하였으며, 또 사전으로부터의 지식 획득이 수작업에 의존하기 때문에 정보의 구축이 어렵다는 문제점이 있다.

② 박영자(1998)는 사전의 의미 기술 문장에서 각 명사 의미에 대한 속성을 자동으로 추출하여 의미를 클러스터링 하는 새로운 방법을 제안하였다. 먼저 의미 기술 문장에서 명사들의 의미 연관 관계를 나타내는 의미 참조 네트워크를 구축한다. 의미 참조 네트워크로부터 의미 속성을 추출하고, 속성값은 Jaccard 측정식을 이용해 주어진 의미간의 유사도를 기반으로 한 퍼지릴레이션을 이용하여 계산된다. 의미 속성과 속성값의 쌍을 속성 공간의 한 벡터로 정의한 후, 유전자 알고리즘을 이용하여 최적의 클러스터링을 산출한다. 의미 참조 네트워크를 구성할 때, 의미 중의성을 해결하는데, 의미 분별하고자 하는 단어가 포함된 문장의 단어들과 의미 분별하고자 하는 단어의 의미 기술 문장들에 포함된 단어들이 많이 공유하는 의미를 의미 분별하고자 하는 문장의 의미로 선택한다. 그러나, 의미 분별하고자 하는 단어의 뜻풀이만을 가지고 했으므로 자료 부족 현상이 심각하고, 용언류를 제외한 명사만을 의미 정보로 이용하여 의미 분별의 한계가 있다.

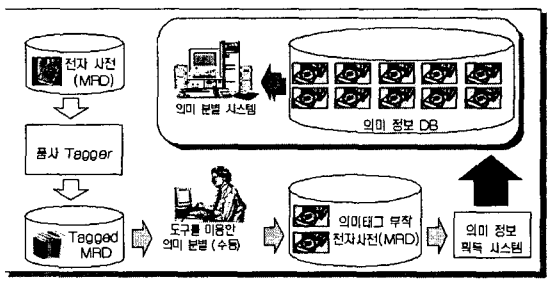
③ 서희철(1999)은 의미 계층 구조에 나타나는 유사어들의 공통적인 특징과 개별적인 특징을 모두 고려하여 의미 중의성을 해결하는 방법을 제안하였다. 품사 부착된 1,000만 어절의 말뭉치에서 의미 계층 구조를 이용하여 추출된 유사어의 용례를 추출하여 학습 말뭉치로 이용한다. 학습 말뭉치를 통해서 구축된 유사어 벡터의 자질값을 이용하여, 의미 중의성을 해결한다. 그러나, 시스템의 확장을 위해서는 보다 신뢰성 있는 의미 계층 구조의 정의와 더 많은 양의 학습 데이터가 필요하다는 문제점이 있다.

④ 허정(2000)은 사전의 뜻풀이에서 추출한 통계적 의미 정보에 기반한 동형이의어 중의성 해결 시스템을 제안하였다. 9개의 동형이의어를 포함하는 5,246 문장의 사전 뜻풀이를 학습 코퍼스로 하여 문장 내에서 동형이의어와 함께 쓰인 체언 및 용언을 의미 정보로 추출하였다. 이 의미정보를 확률 통계적인 방법을 적용하여 동형이의어 의미를 분별하는 모델을 제안하였다. 학습 코퍼스를 대상으로 체언와 용언의 가중치를 실험한 결과 0.9/0.1로 주었을 때 96.11%의 정확률을 나타내었으며, 학습되지 않은 외부 데이터(국어 정보 베이스 ver1.0 과 ETRI의 품사 부착 말뭉치에서 추출한 1,896문장)로 실험한 결과 평균 80.73%의 정확률을 보였다.

⑤ 강신재(2002)는 온톨러지(ontology)를 구축하는 방법과 온톨러지에 기반한 단어 중의성 해결 알고리즘을 제시하였다. 포항공대의 LIP(language independent and practical) 온톨로지를 이용하여 그래프에서 최소 비용 경로를 찾는 형태로 중의성이 있는 단어의 후보 개념간 선택 제약이 얼마나 잘 만족되는가를 평가한다. 실용 기계번역 시스템(COBALT-J/K, COBALT-K/J)에서의 실험 결과, 온톨로지를 사용하여 일본어 분석에서는 6.0%, 한국어 분석에서는 9.2%의 정확률 향상을 보였다.

## 3. 동형이의어 중의성 해결을 위한 의미정보 DB 구축

동형이의어의 중의성은 동형이의어가 사용된 문맥에서의 다른 단어와의 의미적 관계에 의해서 해결된다. 따라서, 동형이의어 중의성 해결 시스템의 정확률은 획득한 의미정보의 정확률 및 충실도에 의해 결정된다. 의미 정보 추출에 사용되는 온라인 자원은, 크게 사전과 말뭉치로 나눌 수 있는데, 현재 국내에는 한국어 동형이의어에 대한 충분한 규모의 의미 태그 부착 말뭉치가 없으므로, 본 연구에서는 국어사전을 말뭉치[1]로 변환하여 품사 태그와 의미 태그를 부착하여 동형이의어 중의성 해결에 필요한 의미정보를 추출하고자 하였다.




[그림 1] 사전 뜻풀이를 이용한 의미정보 DB 구축 개요

### 3.1 품사 태그 부착 뜻풀이 말뭉치 구축

가. 품사 태거를 이용한 자동 품사 태깅

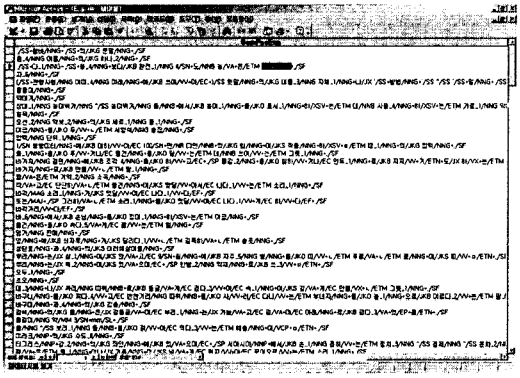
품사 태그 부착 뜻풀이 말뭉치를 구축하기 위해서는 품사 태거(part-of-speech tagger)가 있어야 하는데, 본 연구에서는 문화관광부의 [21세기 세종 계획] 프로젝트에서 제공하는 지능형 형태소 분석기 내의 품사 태거를 이용하여 국어사전의 뜻

1) 국어사전의 뜻풀이의 경우 어느 정도 정제된 언어 표현으로 기술되어 있으므로, 문학 작품이나 신문 기사 등을 말뭉치로 삼는 것보다 훨씬 더 정확한 언어 표현 양상을 살필 수 있다. 따라서 국어사전은 한국어정보처리 과정에서 의미 분석에 필요한 의미정보를 구축하기 위한 가장 기본적인 언어 자원(language resource)이며 자체적으로 훌륭한 말뭉치라 할 수 있다.

풀이와 용례에 품사 태그를 부착하였다.

나. 품사 태그 부착 뜻풀이 말뭉치 수작업 수정 작업

품사 태거를 이용한 자동 품사 태그 부착에서 발생되는 분석 오류를 확인·수정함으로써 고품질의 의미정보를 추출할 수 있다. 이를 위해 품사 태그 부착 뜻풀이 말뭉치를 사람이 직접 확인하는 작업은 반드시 필요한 작업 중의 하나이다. 그러나 품사 수정 작업은 어떻게 기준을 설정하느냐에 따라 품사 태그를 여러 가지로 부착시킬 수 있으나, 본 연구에서는 명사와 용언 중심의 의미정보 추출이므로 명사(일반명사, 고유명사)와 용언(동사, 형용사)을 중심으로 품사 태그를 확인·수정하였다.



[그림 2] 의미 태그 부착 뜻풀이 말뭉치

3.2 사전 기반 의미 태그 부착 뜻풀이 말뭉치 구축

의미 태그 부착 뜻풀이 말뭉치는 품사 태그 부착 뜻풀이 말뭉치에 의미 태그를 부착한 말뭉치를 말한다. 본 연구의 결과 중 하나인 동형이의어 태깅 시스템(또는 자동 의미 태그 부착 시스템)을 개발하기 위해서는 기초 자료를 확보를 위해서 반드시 사람이 직접 의미 태그를 부착하여 구축한 의미 태그 부착 말뭉치가 필요하다. 본 연구에서는 품사 태거에 의한 태그 부착 작업과 품사 태그 확인·수정 작업을 통해 구축된 품사 태그 부착 뜻풀이 말뭉치에 의미 태그를 수작업으로 부착하여 의미 태그 부착 말뭉치를 구축하고자 하였다.

[눈1] ① 빛의 강약과 파장을 느끼어 뇌에 시각을 전달하는 감각 기관
② 시력
③ 안목
④ 무엇을 보는 '모양이나 태도'를 뜻하는 말
[눈2] ① 풀이나 나무의 싹이 막 터져 돋아나는 자리. 또는 그 싹
[눈3] ① 눈금
[눈4] ① 그물 따위의 구멍
② 당혜, 운혜 등의 코와 뒤울의 꾸밈새
[눈5] ① 대기 중의 수분기가 찬 기운을 만나 엉어서 땅 위로 떨어지는 현상의 결정

[그림 3] "눈"의 뜻풀이에 나타나는 의미정보

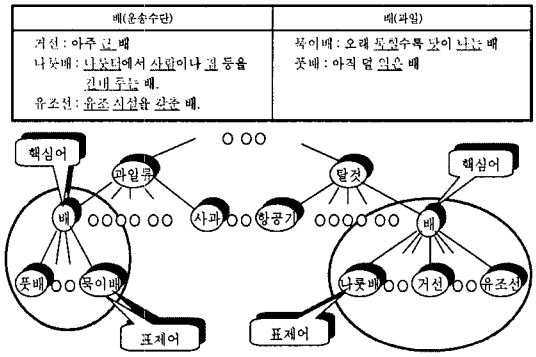
국어사전에서 동형이의어를 구분하기 위해 사용하는 표제어 오른쪽 상단에 부착된 어깨번호를 본 연구에서는 동형이의어 구분을 위한 의미 태그 표시로 사용하고 있다. 즉 "눈[1]", "배[1]" 등과 같은 표시를 "눈_1", "배_1" 등의 의미 태그 형식으로 바꾸어 품사 태그 부착 뜻풀이 말뭉치에 포함시켜 의미 태그 부착 뜻풀이 말뭉치를 구축하였다.

3.3 의미정보 추출 및 DB 구축

국어사전에서 추출 가능한 의미정보는 표제어의 뜻풀이/개념, 동의어, 유의어, 반의어, 전문용어, 특수어, 관련어, 방언, 어원, 관용구(idiom) 등 다양하다. 본 연구에서는 동형이의어 중의성 해결에 필요한 의미정보를 추출하기 위하여 뜻풀이에 쓰인 단어들을 이용하고자 하였다. 그 이유는 뜻풀이가 앞에서 서술한 의미정보를 상당수 담고 있기 때문이다. 그렇지만 뜻풀이에 있는 모든 단어를 의미정보로 설정할 수는 없다. 그래서 본 연구에서는 뜻풀이 내에서 동형이의어 중의성 해결에 필요한 의미정보의 자격을 독립적으로 일정한 의미를 가지고 있는 명사, 동사, 형용사, 부사 등과 같은 실질형태소이면서 어휘형태소로 설정하였다. 이것은 일정한 의미를 지니고 있지 않고 문법적인 성격을 부여하는 조사, 어미 등과 같은 형식형태소나 문법형태소는 한 단어의 의미를 분별하는 의미적 역할이 부족함으로 의미정보로서의 자격을 부여하기 어렵기 때문이다. 그러므로 본 연구에서는 동형이의어 중의성 해결에 핵심적인 역할을 하는 명사, 동사, 형용사 중심의 의미정보 수집 및 추출에 역점을 두었다. [그림 3]은 뜻풀이 속에 포함된 다양한 의미정보를 단적으로 보여주는 예이다.

이와 같이 뜻풀이에서 발견되는 의미정보를 이용하여 동형이의어 중의성 해결에 필요한 의미정보를 추출하고자 하였다.

본 연구에서의 의미정보 추출 방법은 2단계로 이루어진다. 먼저 [그림 4]와 같이, 단어의 상하관계를 중심으로 구축되는 시소러스(thesaurus)와 의미 계층 구조(sense hierarchical structure)의 원리를 이용한 의미정보를 추출하는 방법이다. 이 방법은 핵심어 A, B가 일반적으로 단어의 의미 관계(semantic relation) 중 상하 관계 정보를 가진 단어가 표현되거나 부분·전체 관계 정보를 가진 단어가 표현되기 때문에, 이러한 정보를 이용하여 의미정보를 추출할 수 있다.




[그림 4] "배"와 상하관계에 있는 표제어의 뜻풀이에서 추출 가능한 의미정보

[그림 4]에서 보면 "배_3(사람·물건을 싣고 물 위로 띠 다니는 물건)"과 "배_4(배나무의 열매)"의 의미정보를 추출하기 위하여 동형이의어의 사전 뜻풀이를 비롯하여, 뜻풀이의 패턴과 단어의 상하관계를 이용하여 의미정보를 추출하는 것을 알 수 있다. 이 방법으로 추출되는 의미정보를 보면 [표 1]과 같다.

[표 1] "배_3"과 "배_4"의 의미정보 추출

| 동형이의어 | 의미정보 |
|---|---|
| 배_3 | 사람, 물건, 싣다, 물, 위, 뜨다, 다니다, 물건, 크다, 나룻터, 사람, 짐, 건내다, 유조, 시설, 갖추다... |
| 배_4 | 배나무, 열매, 묵히다, 맛, 나다, 익다... |

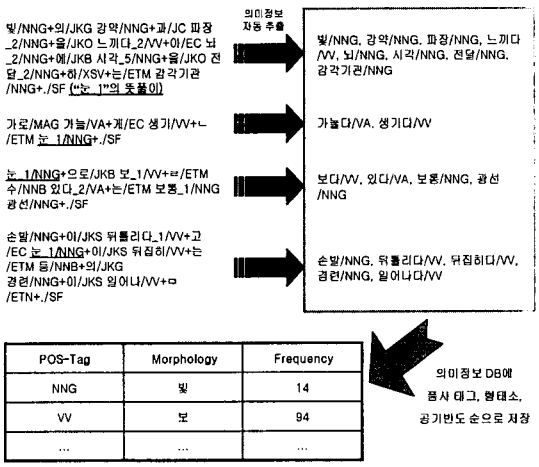
다음으로 동형이의어가 포함되어 있는 뜻풀이에서 공기하는 다른 단어들을 의미정보로 추출하는 방법이다. 이 방법은 위의 1단계 방법의 상하관계를 이용한 방법과는 다른 것으로 뜻풀이에 동형이의어가 포함되어 있을 경우 주위에 나타나는 단어를 의미정보로 추출하는 방법이다. [그림 5]은 "배"가 포함된 뜻풀이와 공기하는 단어들을 보여준다.

| |
|---|
| 난파 : 배가 항해 중에 폭풍우 등을 만나 깨어짐. |
| 운하 : 육지를 파서 강을 내고 배가 다니게 한 수로. |
| 배지느리미 : 물고기의 배에 달린 지느러미. |
| 배앓이 : 배를 앓는 병. |

[그림 5] "배"가 포함된 뜻풀이와 공기하는 단어들

이 방법으로 동형이의어에 대한 추가적인 의미정보를 확보할 수 있다. 결국 "배_3"의 경우 [표 1]에서 제시된 의미정보 외에 "항해, 폭풍우, 만나다, 깨어지다, 육지, 파다, 강, 내다, 다니다, 수로" 등과 같은 의미정보가 추가되는 것이다.

[그림 6]은 동형이의어의 뜻풀이를 비롯하여 의미 태그로 구별이 된 동형이의어가 들어간 의미 태그 부착 뜻풀이 말뭉치에서 자동으로 의미정보 데이터베이스를 구축하는 간략한 과정을 보인 것이다. 의미정보 데이터베이스에는 기본적으로 의미정보에 대한 품사 태그와 형태소가 들어감과 동시에 동형이의어 중의성 해결 모델에 필요한 동형이의어와 의미정보간의 공기빈도를 담고 있다.




[그림 6] "눈_1"의 의미정보 추출과 데이터베이스

## 4. 통계기반 동형이의어 중의성 해결 모델

사전 뜻풀이에서 추출한 동형이의어 의미정보 빈도를 이용하여 Bayes 정리를 적용한 의미분별 모델에서는, 임의의 문장 $C$ 에서 나타나는 동형이의어 $H$ 은 다음 수식에 의하여 의미 $H_{s_1}, H_{s_2}, \cdots, H_{s_n}$ 중 하나로 분별된다.

$$WSD(H, C) = \arg\max_{H_{S_k}} \sum_{j=1}^{m} P(H_{S_k} | w_j) \quad (1)$$

$$P(H_{S_k} | w_j) = \frac{P(w_j \cap H_{S_k})}{\sum_{i=1}^{n} P(w_j \cap H_{S_i})} \quad (2)$$

$$P(w_{N_i} \cap H_{S_k}) = \frac{\text{체언 } w_{N_i} \text{의 공기빈도}}{H_{S_k}\text{의 체언 공기빈도의 합}} \quad (3)$$

$$P(w_{P_i} \cap H_{S_k}) = \frac{\text{용언 } w_{P_i} \text{의 공기빈도}}{H_{S_k}\text{의 용언 공기빈도의 합}} \quad (4)$$

수식(1)에서 $H_{S_k}$ 는 동형이의어 $H$ 의 $k$-번째 의미이며 문장 $C$ 에서 출현하는 $w_j$ 는 $H_{S_k}$ 의 의미정보에 속하는 어휘로 빈도 정보를 가지고 있다. 또한 $w_j$ 는 다른 의미정보에서 다른 빈도로 출현하기도 한다. 수식(2)는 수식(3)과 수식(4)에서 구해진 체언, 용언 각각의 출현 어휘들에 대한 확률값에서 의미 $H_{S_k}$ 로 판단할 확률의 합을 나타낸다. 그리고 수식(1)은 수식(2)에서 구해진 의미별 확률의 합 중에서 가장 큰 값을 문장 $C$ 에 출현하는 동형이의어 $H$ 에 대한 의미로 결정하는 방법이다.

동형이의어 '다리'가 포함된 예문["다리 하나가 없는 사람"]을 본 통계적 모델을 이용하여 분석하는 과정 및 결과는 [표 2], [표 3]과 같다.

[표 2] 예문에서 추출한 의미정보

| 의미 | 품사별(빈도 합) | 추출 의미정보(빈도) |
|---|---|---|
| 다리_1 (신체) | 체언(1633) | 하나/NNG(1), 사람/NNG(14) |
| | 용언(830) | 없/VV(17) |
| 다리_2 (교각) | 체언(467) | 하나/NNG(1), 사람/NNG(5) |
| | 용언(185) | |

[표 3] 예문에 대한 동형이의어 분별 결과

| 의미 / 의미정보 | 다리_1(신체) | 다리_2(교각) |
|---|---|---|
| 하나/NNG | 0.2218 | 0.7782 |
| 사람/NNG | 0.4456 | 0.5544 |
| 없/VV | 1.0 | 0.0 |
| WSD 분별 | **1.6674** | 1.3326 |

## 5. 동형이의어 태깅 시스템

### 5.1 시스템 구성

본 연구의 동형이의어 태깅 시스템은 품사가 부착된 문장을 이용하여 동형이의어 분별을 시도한다. 그렇기 때문에 먼저 품사가 부착되지 않은 입력문장이나 텍스트 파일에서 형태소 분석을 시도한다. 형태소 분석된 결과는 어절 단위로 보여주게 되고 각 어절마다 여러 가지 결과가 나올 수 있기 때문에 사용자가 분석되어 나온 결과들 중 하나를 선택하거나 수정할 수 있도록 하였다. 품사 태그 부착 단계를 거치고 나면 동형이의어 분별을 시도하게 된다. 동형이의어 분별의 과정은 형태소 분석된 결과에서 동형이의어를 검색한다. 동형이의어가 검색되면 입력된 문장의 체언과 용언을 해당하는 동형이의어의 의미정보와 비교하여 일치되는 것들 중에서 확률값의 합이 제일 큰 값을 갖는 동형이의어를 분석 결과로 내보낸다. 분석 결과는 '동형이의어_동형이의어 어깨번호/품사' 형태로 출력된다.

[그림 7]의 동형이의어 태깅 시스템의 구성 및 기능은 다음과 같다.

가. 문장 입력

동형이의어 태깅을 위한 문장 입력은 [그림 7]의 왼쪽 상단의 에디트뷰(edit view)에 직접 문장을 입력하거나 [파일]메뉴의 [열기]를 이용하여 원시 말뭉치(raw corpus)나 품사 부착 말뭉치(tagged corpus)를 읽어 들이면 왼쪽 상단의 에디트뷰에 문장이 나타나게 된다.

나. 품사 태깅

사용자가 입력한 문장이나 원시 말뭉치 상태의 문장은 먼저 올바른 품사 부착이 필요하다. 품사를 부착하기 위해 형태소 분석을 하게 되는데 형태소 분석하는 방법은 [그림 7]상단의 툴바에 있는 [TAG] 버튼을 누르면 된다. [TAG] 버튼을 누르게 되면 입력된 문장들을 문장단위로 분리하여 각 문장에 대해서 형태소 분석을 시도한다. 분석된 결과는 어절단위로 왼쪽 가운데 리스트 뷰에 보여진다. 분석된 어절들은 여러 가지 품사 중의성을 가질 수 있는데 그 중 말뭉치에 많이 사용되는 결과를 보여주게 된다. 만약 사용자가 형태소 분석 결과를 수정하기 위해서는 2가지 방법을 이용할 수 있다. 첫 번째는 수정하려는 어절이 마우스 왼쪽 버튼을 눌러 분석된 다른 후보들을 선택하는 방법이다. 두 번째는 수정하려는 어절에 마우스 오른쪽 버튼을 눌러 사용자가 직접 품사를 고치는 방법이다. 사용자가 품사를 수정하고 나면 수정된 결과가 저장되고 이 결과를 바탕으로 동형이의어 분별을 하게 된다.

다. 동형이의어 의미 태깅

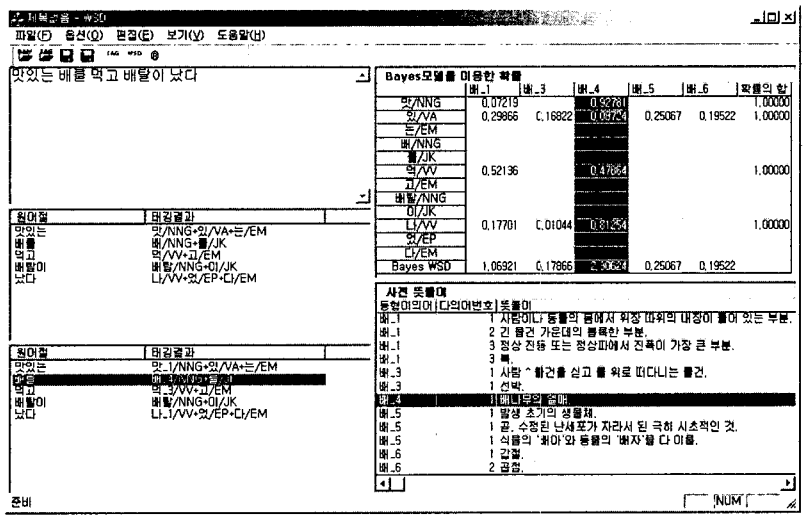
동형이의어 의미 태깅은 품사가 부착된 결과를 바탕으로 이루어진다. 동형이의어 분별을 위해서는 툴바의 [WSD]버튼을 이용한다. [WSD]버튼을 누르게 되면 왼쪽 하단의 리스트뷰에 품사 태깅 결과와 비슷하게 어절단위로 분석된 결과가 보여진다. 동형이의어는 사전의 동형이의어 어깨번호로 구분되는 태그로 부착하게 된다. 사용자가 분석된 결과를 수정하고자 하면 수정하려는 어절에 마우스 오른쪽 버튼을 클릭하여 어깨 번호를 수정할 수 있다.

라. 의미정보를 이용한 분석 결과

동형이의어 의미 분별은 품사가 부착된 문장에서 동형이의어와 공기하는 명사와 용언을 의미정보 데이터베이스에서 검색하여 통계적 확률값을 구하고 이를 이용하여 분석한다. [그림 7]의 오른쪽 상단 동형이의어 분석 뷰는 어떤 단어가 의미정보로 동형이의어 분별에 이용되었는지를 확률값과 함께 보여준다.

마. 사전 뜻풀이 검색

동형이의어 태깅 결과는 사전의 어깨번호로 구분되기 때문에 사용자는 태깅된 의미를 쉽게 알지 못한다. 왼쪽 하단의 동형이의어 태깅 결과 뷰에서 한 어절을 선택하면 오른쪽 하단




[그림 7] 동형이의어 태깅 시스템의 화면

의 사전 뜻풀이 검색 뷰에서 선택한 동형이의어의 뜻풀이를 보여줌으로써 해당 의미를 쉽게 파악할 수 있다.

5.2 시스템 실험 및 분석

본 동형이의어 태깅 시스템의 정확률 측정 및 분석을 위하여 다음과 같은 실험을 수행하였다.

① 실험 말뭉치 및 대상 동형이의어 선정 기준

- 실험 말뭉치: 정확률 실험을 위해서는 동형이의어가 분별된 대량의 비학습 말뭉치가 필요하다. 본 연구에서는 [21세기 세종 계획] 프로젝트에서 제공하는 150만 어절 의미 태그 부착 말뭉치(이하 '세종 의미 말뭉치'라 부름)를 실험 대상으로 선정하였다.

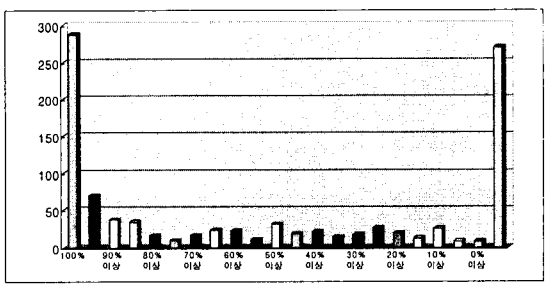
- 실험 대상 어휘 선정: 세종 의미 말뭉치에 출현하는 어휘 중 빈도수가 150회 이상인 어휘에 포함된 동형이의어 469개를 정확률 실험 대상 어휘로 선정한다. 대상 어휘는 평균 2.11개 즉 2개 정도의 의미를 가지고 있으며(총 993개의 의미 ex.배 → 배_1, 배_2, ...) 단 하나의 의미만 출현하는 210개 어휘를 제외한다면 하나의 동형이의어가 평균 3개의 의미로 사용되고 있다.

정확률 실험 대상 어휘의 제한은 본 태깅 시스템 구축에 이용된 국어사전과 세종 의미 말뭉치 구축에 이용된 사전이 서로 달라 정확률 분석에 앞서 사전간 동형이의어 의미 매핑 작업이 필수적으로 요구된다. 본 논문에서는 실험 대상 469개의 동형이의어에 대해서 의미 매핑 작업을 수행하였으며, 차후 사전간 동형이의어 매핑이 완료되면 태깅 시스템이 분석할 수 있는 전체 동형이의어에 대한 정확률을 측정할 수 있을 것이다.

② 실험 결과 및 분석

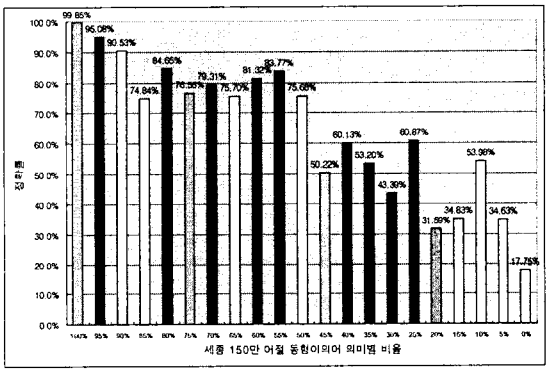
실험에 사용한 469개 어휘의 993개 동형이의어 의미별 정확률은 세종 의미 말뭉치의 의미 태그와의 일치 여부로 결정한다. 정확률은 수식(5)를 사용하였다. 동형이의어의 총 출현 횟수 249,249개이며 이중 228,250개를 정확하게 분석하여 전체 정확률은 91.58%로 나타났다. 세종 의미 말뭉치에 정확률에 대한 분포 현황은 다음 [그림 8]과 같다.

$$\%\ \text{correct} = 100 \times \frac{\text{exactly matched sense tags}}{\text{assigned sense tags}} \qquad (5)$$




[그림 8] 세종 의미 말뭉치 150만 어절 동형이의어 분별 정확률 분포 현황

[그림 9]는 세종 의미 말뭉치에서 동형이의어 의미별 비율을 기준으로 구한 정확률 현황이다. 90%이상의 비율을 가지는 동형이의어 의미의 경우 평균 97.50%라는 높은 정확률을 보이는 반면 10% 미만의 비율로 사용되는 경우 24.05%라는 낮은 정확률을 보이고 있다. 이는 동형이의어 분별용 의미정보에서 정보가 발견되는 경우와 그렇지 않은 경우를 생각해 볼 수 있다. 첫째, 의미정보가 발견은 되지만 사전 뜻풀이에 출현하는 횟수가 적어 낮은 공기빈도를 가지게 되고, 또한 다른 의미의 의미정보에도 포함되어 동형이의어 분별 모델에서 간섭현상이 발생하는 경우로 동형이의어 모델을 정교화 할 필요가 있다. 둘째, 의미정보가 발견되지 않는 경우 부족한 의미정보의 확장이 필요하다.




[그림 9] 세종 의미 말뭉치 동형이의어 의미별 사용 비율 및 비율별 평균 정확률

## 6. 결론 및 향후 과제

본 연구는 한국어정보처리의 여러 응용 분야(정보검색, 기계번역, 자연언어 인터페이스)에서 절실히 요구되는 동형이의어 중의성 문제의 해결에 사용될 다량의 의미 태그 부착 말뭉치를 구축하기 위한 실용적인 동형이의어 태깅 시스템 개발이 목적이다. 본 연구에서는 동형이의어 중의성이 동형이의어가 사용된 문맥에서의 다른 단어와의 의미적 공기 관계에 의해서 해결된다는 언어적인 원리를 이용하여, 동형이의어 중의성 해결 시스템을 개발하고자 하였다. 그리하여 많은 의미적 정보를 담고 있는 국어사전을 기반으로 하여 품사·의미 태그 부착 뜻풀이 말뭉치를 구축하여 동형이의어 중의성 해결을 위한 의미정보 데이터베이스를 구축하고, 이를 기반으로 한 동형이의어 태깅 시스템을 개발하였다. 본 태깅 시스템 성능 측정을 위하여 세종 의미 말뭉치에서 출현빈도 150회 이상인 동형이의어 467개에 대한 분별 실험에서 90%이상의 높은 정확률을 나타내었다.

동형이의어 분별 정확률 향상 및 태깅 시스템의 사용자 환경 개선을 위해서는 다음과 같은 문제점들이 해결되어야 할 것이다.

① 본 통계적 동형이의어 분별 모델의 문제점 중 하나는 문장 전체에서 동형이의 의미정보를 추출하게되어 한

문장 내에 동일한 동형이의어가 둘 이상 출현 할 때 이를 개별적으로 분석할 수 없다. 따라서 인접 어절에 대한 가중치나 어절 제한 등의 방법을 적용해야 한다.

② 본 동형이의어 태깅 시스템에서는 동형이의어 분석 결과를 보여 주고 있다. 분석에 이용된 의미정보를 사용자가 판단하여 불필요한 정보는 제거하며, 또한 부족한 정보는 추가하는 기능을 부가한다면 지속적으로 의미정보를 개량할 수 있을 것이다.

③ 본 동형이의어 태깅 시스템에서는 통계적 분석 모델만을 이용할 수 있다. 따라서 다양한 동형이의어 분별 모델을 개발하여 사용자가 이들 모델 중 하나를 선택하거나 복합적으로 이용하는 사용자 환경을 제공해야 한다.

## 7. 참고문헌


[1] P. O. Cho, and C. Y. Ock(1999), "A Korean Noun Semantic Hierarchy based on Semantic Features", In Proceeding of the 18th International Conference on Computer Processing of Oriental Languages(ICCPOL '99) vol.1

[2] 허정, 옥철영(2001), "사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 정보과학회 논문지: 소프트웨어 및 응용, 28권 9호, pp. 688-698

[3] Ide, Nancy and Veronis(1998), "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art", Computational Linguistics, Vol. 24, No. 1

[4] 이왕우, 이재홍, 이수동, 옥철영, 김현기(2001), "Bayes 정리에 기반한 개선된 동형이의어 분별 모델", 제13회 한글 및 한국어 정보처리 학술대회, pp.465-471

[5] 김준수, 김창환, 이왕우, 이수동, 옥철영(2001), "의미범주 및 거리 가중치를 고려한 통계기반 동형이의어 분별 시스템" 제13회 한글 및 한국어 정보처리 학술대회, pp.487-493

[6] Niwa, Y. and Nitta, Y.(1994), "Co-occurrence vectors from corpora vs distance vectors from dictionaries", In proceedings of the 15th International Conference on Computational Liguistics(COLING '94), Kyoto, Japan, pp.304-309

[7] Alpha K, Luk(1995), "statistical Sense Disambiguation with relatively Small Corpora Using Dictionary Definitions", In Proceeding of the 33th Annual Meeting of the Association for Computational Linguistics(ACL '95), Cambridge, MA

[8] 조정미(1998), "코퍼스와 사전을 이용한 동사 의미 분별", 박사학위 논문, 한국과학기술원 전산학과

[9] 박영자(1997), "사전을 이용한 단어 의미 자동 클러스터링: 유전자 알고리즘 접근법" 박사학위논문, 연세대학교

[10] 서희철, 이호, 백대호, 임해창(1999), "유사어를 이용한 단어 의미 중의성 해결", 제 11회 한글 및 한국어 정보처리 학술대회, pp.304-309

[11] 강신재(2002), "실용적인 온톨러지의 반자동 구축 및 어휘 의미 중의성 해소를 위한 응용", 박사학위논문, 포항공과대학교