정제된 의미정보와 시소러스를 이용한 동형이의어 분별 시스템

김 준 수[†] · 옥 철 영[††]

요 약

단어 의미 중의성 해소는 자연언어처리 분야에 매우 중요한 부분이다. 본 논문에서는 사전 뜻풀이 특성을 이용해 기존의 의미정보를 정제하고 유용한 정보인 확률정보, 거리정보 및 격정보 등을 추가한WSD 모델을 제안하였으며, 사전을 기반으로 구축된 "울산대학교 어휘 지능망(UOU-Word Intelligent Network : U-WIN)"상의 단어 계층적 구조(시소러스)를 이용하여 의미정보의 자료 부족 문제를 해소하는 모델을 제시하였다. "21세기 세종 계획"에서 제공하는 150만 어절 규모의 의미 태그 말뭉치를 대상으로 한 실험에서 최다 빈도 의미 결정(Maximum Frequence Class, MFC, 정확률 베이스라인)에 비해 18.87%(명사 21.73%, 동사 17.11%) 정확률 향상을 보였으며, 기존의 확률 가중치와 어절 거리 가중치를 이용한 모델에 비해서는 10.49%(명사 8.84%, 동사 11.51%)의 정확률 향상되었다. 또한 시소러스를 사용하지 않고 확률정보, 거리정보, 격정보 만을 이용한 모델에 비해 6.12%(명사 5.29%, 동사 6.64%) 높은 정확률을 보였다.

키워드 : 단어 의미 중의성, 동형이의어, 의미정보, 시소러스

A Korean Homonym Disambiguation System Using Refined Semantic Information and Thesaurus

Jun-Su Kim[†] · Cheol-Young Ock[††]

ABSTRACT

Word Sense Disambiguation(WSD) is one of the most difficult problem in Korean information processing. We propose a WSD model with the capability to filter semantic information using the specific characteristics in dictionary definitions, and with added information, useful to sense determination, such as statistical, distance and case information. we propose a model, which can resolve the issues resulting from the scarcity of semantic information data, based on the word hierarchy system (thesaurus) developed by Ulsan Universty's UOU Word Intelligent Network, a dictionary-based lexicological database. Among the WSD models elaborated by this study, the one using statistical information, distance and case information along with the thesaurus (hereinafter referred to as "SDJ-X model") performed the best. In an experiment conducted on the sense-tagged corpus consisting of 1,500,000 eojeols, provided by the Sejong Project, the SDJ-X model recorded improvements over the maximum frequency word sense determination (maximum frequency determination, MFC, accuracy baseline) of 18.87% (21.73% for nouns, 17.11% for verbs). The results were superior in accuracy to the model using statistical and inter-eojeol distance weights by 10.49% (8.84% for nouns, 11.51% for verbs). Finally, the accuracy level of the SDJ-X model was higher than that recorded by the model using only statistical information, distance and case information, without the thesaurus by a margin of 6.12% (5.29%for nouns, 6.64% for verbs).

Key Words : Word Sense Disambiguation, Homonym, Semantic Information, Thesaurus

1. 서 론

자연언어처리(Natural language processing : NLP)는 인간의 언어 능력 및 지식을 공학적이고 실험적인 접근을 통해 분석하고 처리함으로써, 인간과의 자연스러운 의사소통이 가능한 컴퓨터 시스템을 구현하는 것을 기본 목적으로, 형태소분석, 구문분석, 의미분석 그리고 담화분석의 단계로 크게 나누어 연구를 진행하고 있다. 한국어를 대상으로 한 자연언어처리 즉, 한국어정보처리는 대략 1980년 초반부터 시작하여 약 20년 동한 꾸준히 발전하여 형태소분석과 구문분석 등의 분야에서는 괄목할 만한 성과를 보이고 있다. 반면 의미분석과 담화분석에 대한 연구는 필요한 언어 자원의 부족 등으로 아직은 초보적인 연구 수준에 머물고 있다.

자연언어처리의 의미분석 단계에서 발생하는 단어 의미

※ 이 논문은 울산대학교에서 지원하는 2003년 교내연구비지원사업의 지원에 의하여 연구되었음.
† 정 회 원 : 울산대 컴퓨터정보통신공학부 객원교수
†† 종신회원 : 울산대 컴퓨터정보통신공학부 교수
논문접수 : 2005년 1월 10일, 심사완료 : 2005년 10월 24일

중의성의 문제를 해결하기 위해 NLP 분야에서는 다양한 언어 자원을 학습하여 문맥이나 상황에 적절한 의미를 자동으로 결정하려는 연구를 지속하고 있으며, 이를 NLP 분야에서는 단어 의미 중의성 해결(Word Sense Disambiguation : WSD)이라고 한다. WSD는 자연언어처리의 기반 요소로 기계번역 과정에서 의미 중의성 단어의 올바른 대역어 선정은 번역시스템의 성능을 좌우하며, 정보검색의 질의나 검색 결과에서 단어 중의성을 해소함으로써 사용자가 요구하는 양질의 정보를 제공할 수 있다는 이점이 있다. 또한 질의를 한국어로 하여도 질의의 요구에 맞는 다양한 외국어 문서와 검색된 외국어 정보까지도 한국어로 자동 번역하여 보여주는 다국어 정보검색(Multi-lingual IR)의 핵심 기술이 된다. 그 외에도 음성 인식 및 합성, 철자 교정, 품사 태깅 등의 여러 응용 분야의 기반 기술로 중요한 역할을 담당한다.

한국어에 나타나는 단어 의미 중의성의 심각성은 문화관광부의 국어정보화 사업인 "21세기 세종계획[1]"의 의미 부착 말뭉치(sense tagged corpus)를 분석함으로써 확인할 수 있다. 116,706문장(문장 당 평균 13어절)으로 구성된 150만 어절 규모의 의미 부착 말뭉치에는 총 12,287개(9,424종, 누적 빈도 559,559개)의 동형이의어(한 문장 당 평균 4.8개)가 포함되어 있으며, 이 중 하나의 의미로 90%이상 사용된 8,075개(누적 빈도 467,900개(83.62%)를 제외한 나머지 동형이의어 약 2,500 여개(한 문장 당 평균 0.78개 발생)가 단어 의미 중의성 해결의 대상이 된다. 특히 한국어의 경우 많은 단어들이 한자어에서 유래되어 영어에 비해 동형이의어가 많으며, 이로 인해 한국어 의미분석은 다의어 수준의 단어 의미 중의성 해결에 앞서 동형이의어에 대한 정확한 분별이 선행되어야 한다.

단어 중의성은 문맥에서 주변 단어들과의 의미적 관계에 의해서 그 중의성이 해소된다[1, 2]. 사람의 경우 비교적 좁은 영역의 주변 단어들이 주어질 때에도 단어 의미 중의성을 해소할 수 있다는 사실이 Yaacov&Lusignan (1985)[3]의 언어 심리학 실험을 통해 밝혀졌다. "21세기 세종계획"에서 제공하는 150만 어절 규모의 의미 부착 말뭉치를 대상으로 한 실험에서 대상 중의성 단어를 중심으로 전후 3어절 범위 내에 96.4%가 분포함을 알 수 있다[4]. 그러나 국내에는 아직 신뢰할 만한 대규모의 의미 부착 말뭉치가 부족하여, 중의성 단어와 의미적으로 연관성이 있는 정보를 얻기 어려운 실정이다. 본 논문에서는 허정(2001)[5]의 WSD 연구에서 도입한 의미 분별된 사전 뜻풀이·용례 말뭉치를 기반으로 구축된 의미정보를 정제하고 확장하여 이용한다. [5]의 의미정보는 조평옥(1999)[6]의 분석을 바탕으로 동형이의어와 함께 사용된 일반명사, 용언(동사, 형용사)를 동형이의어 의미 결정에 중요한 의미정보로 파악하였다. 이는 손쉽게 의미정보를 구축할 수 있다는 장점은 있지만 단어간의 순서, 구문관계, 특정 연어 등의 다양한 정보를 고려하기 어렵다는 단점이 있다. 따라서 사전 뜻풀이의 특성을 이해하고 구문적 정보를 바탕으로 한 선택 기준을 제시하여 일반명사, 동사, 형용사 그리고 부사를 격조사(주격, 목적격, 관형격, 부사격), 파생접미사(동사파생, 형용사파생)와 전성어미(명사형, 관형사형) 등과 함께 공기빈도(co-occurrence frequency), 어절간 인접 거리 정보 등을 WSD를 위한 의미정보에 확장하고자 한다. 또한 사전 말뭉치 기반 의미정보의 문제점인 자료부족 문제를 '표준국어사전'을 기반으로 구축된 '울산대학교 어휘 지능망(UOU-Word Intelligent Network : U-WIN)[2]' 상의 시소러스를 이용하여 해결하는 방안을 제시하고자 한다.

2. 사전 뜻풀이 특성을 이용한 의미정보 정제

2.1 의미정보 정제의 필요성

사전 뜻풀이 말뭉치의 특성상 일반 언어 말뭉치에 비해 다음과 같은 명사(총칭, 하나, 사람, 일, 말), 동사(가리키다, 하다, 되다, 이르다), 형용사(없다, 있다, 크다, 다르다) 등이 고빈도로 나타난다. 이들은 주로 문법 용어와 어떤 것을 정의하는데 쓰이는 사전 특성적인 단어들로 일반 문장을 분석 대상으로 하는 WSD의 의미정보로는 부적절한 것으로 보고 불용어(stopword)로 간주하여 제외시켜야 한다. <표 1>은 사전 뜻풀이 말뭉치 특성상 발생하는 불용어 목록 중 일부이다.

불용어와 함께 WSD에 부적절한 단어들을 선별하기 위한 방법으로 단어 유사도(similarity)에 기반한 통계적 접근 방법과 사전 뜻풀이 문장의 특성을 바탕으로 한 언어학적 접근 방법이 있다.

통계적 접근 방법은 공기빈도를 이용하여 단어 유사도를 측정하면서 특정한 공식 또는 계수를 사용하는데, 이를 유사계수라고 한다. 유사계수는 단순 공기빈도를 이용할 수도 있고 상대 공기빈도를 이용할 수도 있다. 일반적으로 단순 공기빈도를 이용하는 방식 대신 공기빈도와 개별 단어의 빈도, 전체 단어 빈도를 이용하여 단어 사이의 통계적인 연관성을 객관적으로 평가하는 상대 공기빈도 방식의 유사계수가 주로 이용된다.

과거에는 유사계수 공식으로 벡터공간 검색모형에서 출발한 코사인 계수(cosine coefficient : COS)나 클러스터링에 주로 쓰이는 자카드 계수(jaccard coefficient : JAC) 등의 공식을 많이 이용하였다[7]. 그러나 1990년 Church & Hanks

〈표 1〉 사전 뜻풀이 말뭉치 특성상사 발생하는 불용어 목록

불용어	빈도수	불용어	빈도수
총칭	1,248	일	8,842
하나	108	사람	8,261
부분	1,343	종류	392
전체	387	뜻	3,114
이름	937	나타내다	2,379
준말	4,428	뜻하다	121
말	6,583	가리키다	332
별칭	107	이르다	100

1) 문화관광부 국어정보화 사업(http://www.sejong.or.kr)

2) 울산대학교 한국어처리연구실(http://nlplab.ulsan.ac.kr)

(1990)[8]가 정보이론으로부터 도출된 상호정보량(mutual information : MI)을 연어(collocation) 분석에 적용할 것을 제시한 이후 많은 연구에서 상호정보량 공식을 이용하고 있다.

김준수(2002)[9]에서는 동형이의어 48개(의미정보 총 48,172개)에 대한 상호정보량을 분석하여 의미정보 정제 및 임계값 설정을 실험하였다. 그 결과 상호정보량 1.1 이상의 의미정보 25,076개(52.06%)로 '21세기 세종계획'에서 제공하는 350만 어절 규모의 품사 부착 말뭉치에서 추출한 대상문장 중 86.20%인 18,243문장을 분석할 수 있었다. 다음의 <표 2>는 "배_3(운송수단)"의 의미정보에 대한 상호정보량 측정 결과 이다. 앞 절에서 불용어 대상으로 지목한 "일", "사람", "말"과 "하다" 등은 뜻풀이 내에 폭넓게 사용되어 상호정보량이 0에 근사하거나 음수(단어 유사도가 낮거나 단어간 유사성이 없음을 나타냄) 값을 가지며, 상대적으로 낮은 공기빈도에도 불구하고 특정 단어와 공기하는 "운항", "입항", "띄우다"등은 높은 상호정보량을 보였다.

[9]의 WSD 실험 결과 상호정보량을 임계값을 통해 의미정보를 제한할 경우 분석이 가능한 문장(재현률)은 감소하지만 정확률은 향상됨을 알 수 있다. 특히 세종 말뭉치를 대상으로 한 실험에서 상호정보량 0.7 이상에서 재현률이 급격히 감소하였으며, 상호정보량 1.1 이상의 의미정보에서 재현률과 정확률이 역전되는 현상을 보였다. 일반적으로 1.1~1.2 이상의 상호정보량을 가지는 의미정보는 동형이의어 의미 분별에 유용하다. 그렇다고 상호정보량 만으로 유용성을 판단하기에는 어렵다. 그 예로 <표 3>을 보면 "조정", "해안", "뜨다", "향하다" 등은 "배_3(운송수단)"에 중요한 의미정보이다. 특히, "향하다", "세우다", "지나가다", "움직이다", "내리다" 등은 시소러스 상에서 "배"의 상위어에 해당하는 "운송수단"의 의미자질에 해당되며, "뜨다", "가라앉다" 등은 "배"의 의미자질에 해당된다. 그리고 "해안", "파도", "연안" 등은 "배"가 존재하는 공간인 "바다"의 하위어에 해당하는 단어들이다.

단어 상호간 의미적 관계는 상호정보량 등의 유사계수에 의한 단어 유사도 측정만으로는 어렵다고 본다. 명사간의 상호정보량은 명사와 용언 사이의 상호정보량 보다 상대적으로 높은 편이다. 명사와 용언 사이의 단어 유사도 측정은

〈표 2〉 "배_3(운송수단)"의 의미정보에 대한 상호정보량 측정

의미정보	개별 단어 빈도	공기빈도	상호정보량
하다	9,557	47	0.195
일	8,842	27	-0.032
사람	8,261	22	-0.042
말	7,169	15	-0.223
정박	11	10	2.414
젓다	33	8	1.831
띄우다	62	6	1.534
운항	20	5	1.844
원양	7	4	2.203
입항	3	1	1.967

〈표 3〉 상호정보량 임계값을 1.2로 지정했을 때 '배_3(ship)'에서 제거되는 단어 중 의미 분별에 유용한 의미정보

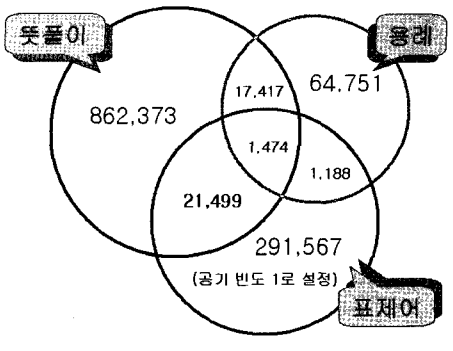
상호정보량	의미정보
0.0≤X<0.4	향하다(0.18), 돌다(0.21), 세우다(0.38), 지나가다(0.39), …
0.4≤X<0.8	조정(0.42), 해안(0.58), 움직이다(0.58), 뜨다(0.65), 물고기(0.67), 바람(0.71), 그물(0.77), …
0.8≤X<1.2	파도(0.87), 내리다(0.85), 선박(0.97), 군함(0.90),가라앉다(0.93), 보트(1.03), 연안(1.15), 바다(1.18), …
1.2≤X	다니다(1.25), 태우다(1.21), 운하(1.23), 운행(1.32), 흘러가다(1.37), 해상(1.43), 함선(1.44), 실리다(1.45), 승무원(1.46), 잠기다(1.45), 띄우다(1.53), 젓다(1.83) 입항(1.97), 원양(2.20), 정박(2.41), …

시소러스(단어간 계층적 연결 정보)나 개념망(개념 사이의 의미적 연관도 그래프)과 같은 언어 자원을 복합적으로 이용해야 한다.

2.2 사전 뜻풀이 특성을 이용한 의미정보 정제

사전 뜻풀이 말뭉치에서 동형이의어와 공기하는 명사, 동사, 형용사 그리고 부사를 의미 결정에 필요한 요소로 판단하고 추출한 의미정보는 총 902,763종(빈도합 : 1,786,521), 용례에서 추출한 의미정보는 총 84,830종(빈도합 : 100,337), 뜻풀이에 동형이의어가 포함된 표제어(여러 동형이의어에 표제어가 중복적으로 적용되며, 뜻풀이·용례에 중복되지 않는 경우는 빈도를 1로 지정하였다.) 총 315,728개 이다. 의미정보의 분포는 다음 (그림 1)과 같다.

(그림 1)과 같이 구축된 의미정보를 분석해 보면 의미결정에 중요한 정보와 불필요한 정보가 혼재되어 있음을 알 수 있다. 또한, 대상 동형이의어는 의미가 구분되어 있으나 공기하는 동형이의어에 대한 구분은 이루어지지 않았다. 사전(dictionary) 공기정보에 바탕을 두고 수집된 모든 의미정보가 유용한 것은 아니다. 구문적인 상황을 고려하지 않고 추출된 의미정보에는 의미적 연관성이 낮은 경우도 많이 발생한다. 불필요한 의미정보의 효율적인 정제를 위해 상호정보량(mutual information) 측정을 통한 단어 유사도 계산과 임계치 설정을 통해 의미정보를 정제한 결과 저빈도 어휘간



〈그림 1〉 뜻풀이·용례·표제어 의미정보 중복

의 연관성이 상대적으로 높게 평가되어 빈도에 의존적인 결과를 보였다[9]. 또한 다양한 명사와 결합하는 용언의 특징으로 명사와 용언간의 관계는 단어 유사도 측정만으로는 해결하기 어렵다. 본 논문에서는 사전 뜻풀이의 특성을 이해하고 구문적 정보를 바탕으로 한 선택 기준을 제시하여 양질의 의미정보를 선별할 수 있는 방법을 제안하고자 한다.

한국어는 교착어(agglutinative language)의 일종으로 조사나 어미 같은 기능어가 독립된 형태소로 잘 발달되어 있으며 문장의 구성단위인 어절은 내용어와 기능어의 복합체 구조를 갖는다. 한국어의 큰 특징으로는 중심어 후행성과 어순의 자유성을 들 수 있다[10]. 즉, 동사나 형용사와 같은 용언구가 문장의 제일 끝에 위치한다. 그러나, 서술어에 대한 보어(complement)와 부가어(adjunct)는 순서에 상관이 없으며 수식구나 수식절을 동반하는 문장 성분은 수식 성분 뒤에 위치하고 수식 성분이 둘 이상일 때 이들 간의 어순은 자유롭다.

한국어는 부분 자유 어순 언어(partially free word order language)이므로 독립된 문장에서 단어들의 많은 어순이 문법적으로 적법하다. 이렇게 다양하게 나타날 수 있는 어순을 제한하는 것은 문맥(context)이다. 즉, 한국어에서는 특정 문맥이 그 문맥을 반영하도록 단어의 어순을 제한한다. 서술어와 보어 사이에는 보어 격(case)을 결정하는 문법 관계가 존재한다. 주어 관계나 목적어 관계와 같은 문법 관계는 서술어와 보어 사이의 격을 결정함으로써 정해지며 이는 구문 분석에 매우 중요한 작업이다. 한국어에서는 이러한 격 관계를 결정하는 것이 조사와 같은 기능어인데 이들은 내용어를 보조하는 부수적 역할을 수행한다[11].

사전 기술(記述)의 특징상 하나의 뜻풀이에 사용되는 정보는 일반 문장에 비해 의미적 연관성이 매우 높다[5, 12]. 특히 공기하는 명사 사이에는 특별한 경우를 제외하고는 의미적 연관성이 매우 높아 특별한 선택 기준을 마련하지 않아도 될 것이다. 문제가 되는 것은 명사와 공기하는 용언(동사, 형용사)과의 의미적 관계를 설정하기 어렵다는 것이다. 한국어의 자유로운 어순과 격정보 생략 현상에 주요 원인이 있다. 기존 의미정보에서 고려하지 못했던 기능어(격조사, 파생접미사, 전성어미 등)를 활용하여 구문 분석의 초보적인 분석 방법을 도입함으로써, 완벽하지는 못하지만 동형이의어 의미 분별에 이용할 만큼의 효과를 얻을 수 있을 것이다. 뜻풀이 문장의 고찰을 통해 간략한 선택 제약 기준을 제시한다.

동형이의어 명사(N1)의 경우

- (N1, 명사 N2)
  - ① 동형이의어 명사 N1의 앞/뒤 5어절 범위 내에 공기하는 명사 N2
- (N1, 용언 V)
  - ① N1이 격조사 정보를 가지는 경우 뒤 어절에 처음으로 나타나는 용언
  - ② N1이 격조사 정보를 가지지 않는 경우 앞뒤 3어절 이내에 나타나는 용언
- (N1, 부사 M)

  부사는 용언 또는 다른 말 앞에 놓여 그 뜻을 분명하게 하는 품사
  - ① N1이 동사파생접미사(XSV)를 가지는 경우 앞 어절에 나타나는 부사
  - ② N1이 파생접미사 정보는 없고 격조사를 가지는 경우 앞 어절에 나타나는 부사

동형이의어 용언(V1)의 경우

- (V1, 용언 V2)
  - ① 앞뒤 1어절 범위 내에 발생하는 경우는 복합용언일 가능성이 높다.
    - 단어가 연어 또는 복합어로 나타날 경우 단어 중의성은 해결 된다.
- (V1, 명사 N)
  - ① V1이 관형사형 전성어미(ETM) 정보를 가지는 경우 다음에 나오는 명사의 의미를 제약한다.
  - ② 그 외의 경우 앞뒤 2어절 이내에 나타나는 명사와 의미적 연관성을 가진다.
- (V1, 부사 M)
  - ① 앞 2어절 이내에 나타나는 부사는 V1의 뜻을 분명히 한다.

공기 정보에 기반하여 구축된 의미정보에 앞에서 제시한 선택 제약 기준을 적용하게 되면, 용언의 의미정보 중 위의 조건을 만족하는 경우는 '갖추/VV(동사)'와 '다니/VV(동사)' 만이 동형이의어 '배'의 의미정보로서 의미적 연관성이 있게 된다. 전체 동형이의어에 이를 적용하게 되면 의미정보의 개수는 총 791,058종으로 제약 전 1,242,852종의 63.65% 수준으로 줄어들게 되다. 의미정보 정제가 WSD에 미치는 영향은 본 논문의 실험에서 확인하도록 하겠다.

3. 시소러스를 이용한 의미정보 확장

울산대학교 어휘 지능망(UOU-Word Intelligent Network : U-WIN)은 한국어정보처리를 비롯한 정보검색, 기계번역, 시맨틱 웹 등 다양한 분야에 이용될 수 있는 어휘 데이터베이스로서, '인간이 가지는 여러 관념 속에 공통적인 속성을 기반으로, 인간의 보편적인 인지 체계와 개념 관계를 파악하여 이것을 표현한 언어를 대상으로 한 형식적이고 명세적인 어휘 네트워크'를 말한다. 시소러스, 의미망, 온톨로지 등을 통합 관리하는 지식 베이스로 현재 12만 어휘로 구성된 어휘 의미망으로 구성되어 있다. 본 논문에서는 U-WIN의 시소러스 부분만을 우선 이용하여 단어 의미 중의성 해결 시 발생하는 의미정보의 자료 부족 문제를 해결하고자 한다.

3.1 U-WIN의 구성

U-WIN의 구축은 기초 어휘, 단어 표기 방법, 다의어 처리 방법, 상하 구조, 그리고 의미관계의 원리를 바탕으로 한다. 우선 기초 어휘는 언어생활에서 빈도수가 높고 분포가

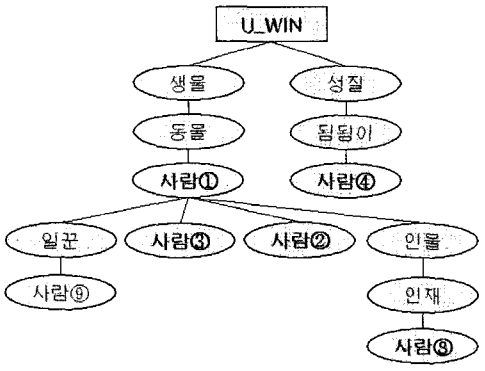
넓으며, 합성어나 파생어와 같은 이차 조어의 근간이 되는 최소한의 필수 어휘로서 임지룡(1991)[13]과 언어정보개발연구원[3] 등의 연구를 기반으로 국어사전 편찬시의 기초 어휘 분류 체계와 동의어 집합의 대표 단어 그리고 말뭉치를 이용한 빈도 정보에 기초하여 기초 어휘를 선정함으로써 U-WIN 최상위어 설정과 기본 구축 단어 선정에 중요한 정보로 활용하였다.

단어의 표기와 관련하여 국어사전은 대체로 한글맞춤법과 외래어표기법에 의거하여 단어를 표기하고 있다. 그러나 전문용어사전이나 백과사전 등은 표기법을 제대로 지키지 못하는 경우가 많으며, 예전의 국어사전은 표기법을 등한시한 경향이 있었다. 또한 일반 사용자의 경우 오용어를 사용하는 경우가 많이 발생하여 사용자의 표기 오류 문제[4] 등을 '표준국어사전'과 원어 정보 표기를 통해 수정하였다.

사전 표제어(단어)와 뜻풀이의 관계는 다음 (그림 2)와 같다. 한 단어는 다의적 특성에 의해 의미적 부담이 매우 높다. U-WIN에서는 단어와 뜻풀이의 관계를 1:n이 아닌 1:1로 설정 관리함으로써 다의적 특성에 의한 의미적 부담을 해소하고, 의미적 계층 구조의 세밀성 및 동의/유의/반의 관계 설정에 용이한 단어 계층적 구조로 구축하였다.

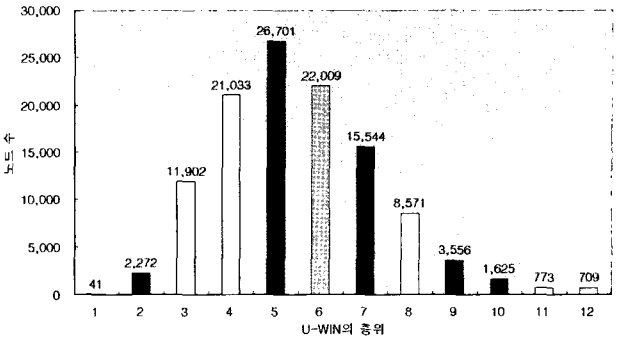
U-WIN의 기본 구조인 단어의 계층적 구조(hierarchical structure)는 의미론적 상하관계를 중심으로 층위(level)를 이룬 집합체로 상하위 층위(또는 노드)가 의미적으로 밀접한 연관성을 가짐과 동시에, 엄밀한 의미에서의 상하관계가 형성되고, 보편적인 개념화(추상화) 과정으로 인식되는 구조로 판단된다. 아래의 (그림 2)는 "사람"에 대한 다의어 수준의 계층 구조를 보여준다. (그림 2)의 "생물", "성질"과 같은 최상위 층위(노드)는 41개로 구성 되어 있으며, 깊이(depth)

사람 (표준국어대사전)
① 생각을 하고 언어를 사용하며, 도구를 만들어 쓰고 사회를 이루어 사는 동물.
② 어떤 지역이나 시기에 태어나거나 살고 있거나 살았던 자.
③ 일정한 자격이나 품격 등을 갖춘 이.
④ 사람의 됨됨이나 성질.
⑤ 상대편에게 자기 자신을 엄연한 인격체로서 가리키는 말.
⑥ 친근한 상대편을 가리키거나 부를 때 사용하는 말.
⑦ 자기 외의 남을 막연하게 이르는 말.
⑧ 뛰어난 인재나 인물.
⑨ 어떤 일을 시키거나 심부름을 할 일꾼이나 인원.
⑩ (수량을 나타내는 말 뒤에 쓰여) '①'을 세는 단위.
⑪ {법률} 권리와 의무의 주체인 인격자. 자연인과 법인을 포함한다.



〈그림 2〉 U-WIN의 단어 계층 구조 예(다의어 '사람')

3) 연세대학교 언어정보개발연구원(http://lexeme.yonsei.ac.kr)
4) {데이터베이스} → 데이터베이스, {길다랗다} → 기다랗다



(그림 3) U-WIN의 층위별 노드 수

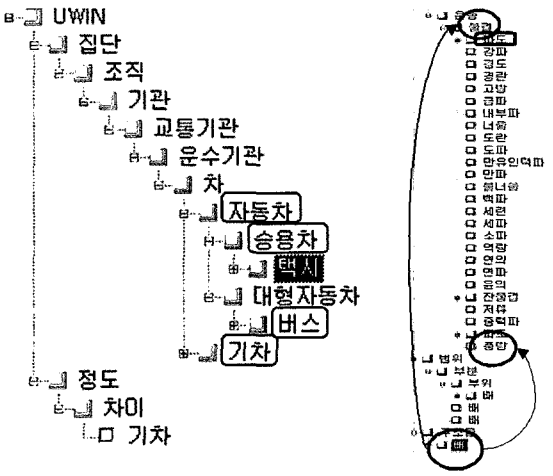
에 가장 많은 노드가 분포하는 구성을 보이고 있다. (그림 3)은 U-WIN의 계층 구조별 층위별 노드의 수를 보여 주고 있다.

3.2 U-WIN의 시소러스를 이용한 의미정보 확장

통계적 WSD 모델이 가지는 대표적인 문제점 하나는 문장에서 의미정보를 얻지 못하여 단어 의미 중의성을 해결하지 못하는 경우이다. 다음 예문 1은 문장 내에서 나타나는 동형이의어 "타다"와 "가다" 모두 의미정보를 얻지 못하는 경우이며, 예문 2는 동형이의어 "배"의 의미정보를 얻지 못하는 경우이다.

(예문 1) 철수는 택시를 **타**고 학교에 **간다.**
(예문 2) 파도는 기분 내키는 대로 **배**를 뒤흔들었다.

예문 1, 2와 같이 문장에서 동형이의어 분별에 필요한 의미정보를 얻지 못한 경우 U- WIN의 시소러스를 이용한 의미정보 확장을 시도할 수 있다. 예문 1의 문장 내에 공기하는 명사 '택시'는 의미정보에 나타나지 않고 있으나, U-WIN의 단어 계층 구조(시소러스)에서 [집단]-[조직]-[기관]-[교통기관]-[운송기관]-[차]-[자동차]-[승용차]의 계층 구조 아래 '택시'가 포함되어 있다. 상하 관계 노드인 '승용차', '자동차'와 이들의 자식 노드에 포함된 '버스', '기차' 등이 동형이의어 '타다_2(탈것에 오르다)'의 의미정보에 포함되어 '택시'



(그림 4) U-WIN의 시소러스를 이용한 의미정보 확장

를 대신하는 공기정보로 이용하게 되면 의미정보 부족 문제를 해결할 수 있다. 다음 (그림 4)에 U-WIN상의 '택시'의 계층 구조 및 '타다_2'의 의미정보 분포를 보여 준다. 또한 예문 2의 경우 U-WIN 상의 '파도'의 형제 노드 및 부모 노드에 해당하는 '물결', '너울', '풍랑' 등이 '파도'와 상하위 관계 및 동의어/유의어 관계를 가지며 동형이의어 '배_3(운송수단)'의 의미정보에 포함되어 있어 '파도'를 이들 단어로 대체하여 동형이의어 의미 분별을 시도 할 수 있다. 따라서 단어간의 상하관계, 동의관계, 유의관계, 그리고 부분-전체 관계 등으로 구성된 U-WIN의 시소러스를 활용함으로써, 사전 뜻풀이 및 용례를 기반으로 구축된 의미정보에서 발생하는 비공기 정보에 의한 자료 부족 문제를 해결할 수 있다.

4. 확장된 의미정보 시소러스를 이용한 WSD 모델

4.1 중의성 단어를 포함한 문장 분석

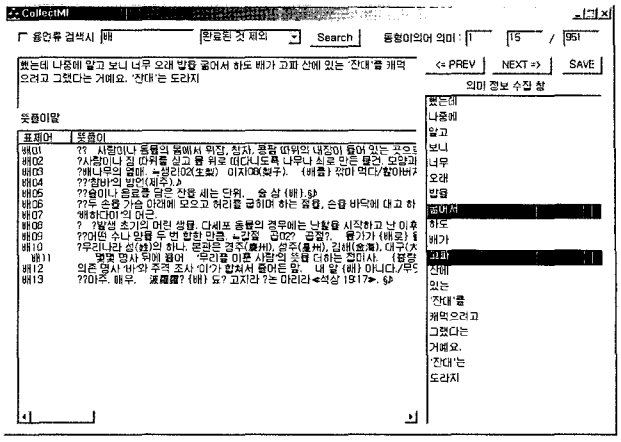
단어 중의성을 해소하기 위한 한 방법으로 중의적 단어의 한 가지 의미와 그 주변에 있는 단어들 사이의 "의미 연합"(semantical association) 즉 의미적 호응이 필요하다[14]. 특히, 한국어는 영어와 달리 어순이 자유로우며, 여러 다양한 통사적, 의미적 표지(marker)가 사용되며, 문장을 이해하는데 중요한 역할을 하는 동사나 형용사가 문장의 끝에 나온다. 이에 따라 한국어를 이해할 때 상대적으로 어휘 정보가 차지하는 비중이 크게 된다. 즉 한국인의 사고체계는 한국어 어휘의 정보를 기반으로 해서 진행되는 측면이 강하다[14, 15].

"한국어 사용자는 문장의 어떠한 정보를 바탕으로 중의적 단어의 의미를 결정하는가?"에 대한 의문을 실험적으로 확인하고 이용하기 위해 본 논문에서는 '21세기 세종 계획'에서 제공하는 150만 어절 규모의 의미 태그 부착 말뭉치[5](이하 '세종 의미 말뭉치'라 한다)를 대상으로 언어학적 지식을 가진 석·박사들이 피실험자로 참여한 연구를 수행 하였다. 작업의 효율성을 높이기 위해 다음 (그림 5)와 같은 도구를 이용하여 수작업으로 진행되었다.

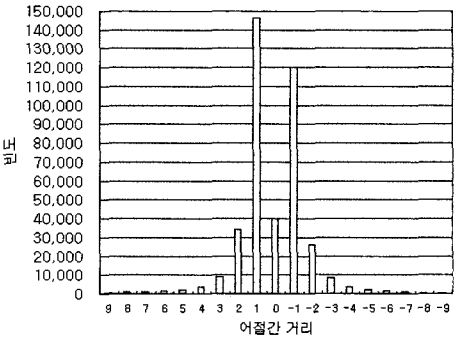
의미분별정보 수집도구는 말뭉치로부터 특정 중의성 단어가 포함된 문장을 추출하여 수작업의 효율성과 일관성을 유지하도록 하고, 피실험자에게는 중의성 단어의 의미를 결정하기 위해 가장 중요하다고 판단되는 어절을 복수로 선택할 수 있도록 하였다.

(예문 3) 나중에 알고 보니 너무 오래 밥을 굶어서 하도 배가 고파 산에 있는 '잔대'를 캐먹으려고 그랬다는 거예요.

예문 3의 문장에 나타나는 동형이의어 '배'의 의미[6]를 결정하기위한 주변의 요소를 판단하는 과정으로 피실험자는 주변에 나타나는 '굶어서'와 '고파'의 어절을 중요하게 선택하였다. 만약 중의성 해결에 연관이 있는 모든 어절을 선택하도록 하였다면 '밥을', '잔대를'과 '캐먹으려고'와 같이 의미적 연관이 있는 경우 모두를 선택하게 되었을 것이다. 본 실험의 목적은 의미결정에 핵심적인 요소만을 골라내는 것이므로 피실험자의 판단은 올바르다고 볼 수 있다.



(그림 5) 의미분별정보 수집도구



(그림 6) '세종 의미 말뭉치'에서 수작업으로 추출한 의미정보 분포도

예상과 같이 중의성 단어의 의미 결정에는 주변의 명사(일반명사, 고유명사)와 용언(동사, 형용사) 그리고 일부의 부사가 핵심적인 역할을 담당한다는 것을 확인할 수 있었다. 아래의 (그림 6)은 '세종 의미 말뭉치'의 실험 결과에 대한 어절 분포도이다. 실험 결과 중의성 단어를 중심으로 한 앞/뒤 2어절(window size: ±2) 범위에 91.6%가 분포하며, 앞/뒤 3어절(window size: ±3) 범위에 대부분인 96.4%가 분포하였다.

아래의 예문 4에는 '배', '맛' 그리고 '식사' 등이 중의성을 가진다. '배'를 포함한 어절과 다음 어절의 품사 부착은 다음 '배/NNG(일반명사)+가/JKS(주격조사) 고프/VA(형용사)+았/EP(선어말어미)+던지/EF(연결어미)'과 같다. 주격조사를 가지는 '배'는 다음에 나타나는 형용사 '고프다'에 의해 '배'의 첫 번째 의미인 배_1(신체 일부)로 결정 할 수 있다. '맛'은 ①음식물 따위가 혀에 닿았을때 일어나는 느낌 ②가리맛과

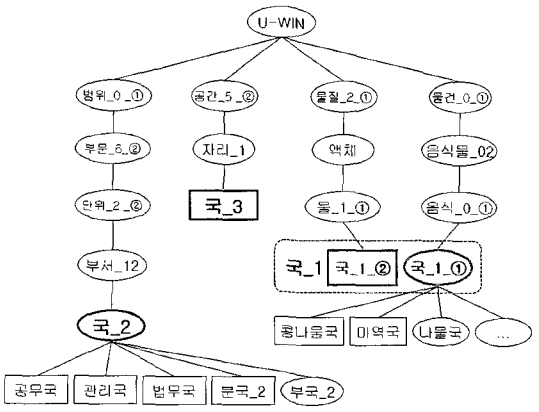
5) 의미 태그를 부착하는 방법에는 크게 두 가지로 분류할 수 있다. 첫째는 사전의 의미(sense) 구분 기호 즉, 표제어의 어깨번호로 구분되는 단어의 의미(동형이의어 단계의 구분으로 보는 것이 타당할 것이다.)를 이용하는 방식이다. 둘째 WordNet과 같은 어휘의 범주(category)를 이용한 의미(semantic) 구분 기호로 이용하는 방식이다. '세종 의미 말뭉치'는 '표준국어사전'의 표제어 어깨번호를 의미(sense)표지로 사용하고 있다.

6) 명사 '배'의 대표적인 사전적 의미 : ①배_1: (척추동물의) 위장 따위가 들어 있는 가슴과 골반 사이의 부분 ②배_2: 물 위에 따다니며 사람이나 짐 따위를 실어 나르게 만든 탈것 ③배_3: 배나무의 열매 ④배_4: 식물의 씨 속에서 자란 싹눈이 되는 부분 ⑤배_5: 갑절, 또는 곱절

와 죽합과에 딸리는 조개를 통틀어 이르는 말 의 두 가지 의미를 가지고, '식사'는 ①의식의 행사 ②식장(式場)에서 주최자가 그 식에 대하여 인사로 말함. 또는 그 말 ③끼니로 음식을 먹음. 또는 그 음식 ④식욕(食慾) ⑤음식물에 체하여 설사를 하는 증상. ≒식설01(食泄) ⑥남을 속이기 위하여 거짓으로 꾸밈 ⑦듣기 좋게 꾸며서 하는 말. ㊟식설02(飾說) 등의 일곱 가지의 의미를 가지게 된다. 문장의 구성하는 주요품사인 명사, 동사, 형용사와 부사만을 대상으로 할 경우 <맛_1, 식사_3>와 <맛_2, 식사_3>의 의미적 호응 관계를 생각할 수 있으며, 의미표지(격조사, 파생접미사, 전성어미 등)까지 고려하면 <맛_1, 식사_3>만이 가능하게 된다.

(예문 4) 몹시 배가 고팠던지 후루룩거리며 남자는 맛있게 식사를 하였다.

다음의 예문 5의 중의성 단어 '국[7]'이 두 번 쓰인 것을 볼 수 있는데, 첫 번째 '국그릇'은 '표준국어사전'에 표제어로 등록이 되어 그 쓰임이 이미 하나의 단어로 인식 된다. 한국어는 복합명사에 띄어쓰기가 자유로워 단순 명사 연결구와의 구별이 어렵다. 이는 NLP의 또 하나의 중요한 연구과제이다. '국그릇/NNG'과 같이 품사 부착이 가능하며, 이 경우 단의어(monosemy)가 된다. '국/NNG+그릇/NNG' 보고 분석 한다면, '국'의 첫 번째 의미(①고기, 생선, 채소 따위에 물을 많이 붓고 간을 맞추어 끓인 음식)과 '그릇'의 뜻풀이(음식이나 물건 따위를 담는 기구를 통틀어 이르는 말)을 이용하면 <국_1, 그릇> 판단할 수 있을 것이다. 두 번째 '국'의 경우는 앞의 '국'의 의미 결정처럼 사전 뜻풀이를 이용한 방법으로는 해결하기 어렵고, 목적격조사를 품고 있으므로 후위에 나타나는 용언(동사, 형용사)에 의해 의미가 결정될 가능성이 높아진다. 가능성을 실제로 확인하기 위해서는 말뭉치에 그러한 결합 관계가 나타나는지의 유무를 판단할 필요가 있다. 또한 단어들 사이의 계층적 구조를 이용하여 확장 분석을 시도할 필요가 발생한다. 다음의 (그림 7)은



(그림 7) '울산대학교 어휘 지능망'(U-WIN) 상의 '국'

7) 국_1 : ① 고기, 생선, 채소 따위에 물을 많이 붓고 간을 맞추어 끓인 음식 ②=국물①(①국, 찌개 따위의 음식에서 건더기를 제외한 물. ≒국01②)
국_2 : 관청이나 회사에서 일을 나누어 처리하는 단위의 하나. 보통 과(課)나 부(部)의 위에 둔다.
국_3 : {민속} 풍수지리에서, 명당에 흐르는 물과 그 주위의 형세가 합하여 이룬 자리. 풍수의 가장 중심 부분이 된다.

U-WIN상의 시소러스에 나타난 중의성 단어 '국'에 대한 다의어 수준의 계층적 구조를 보여 주고 있다.

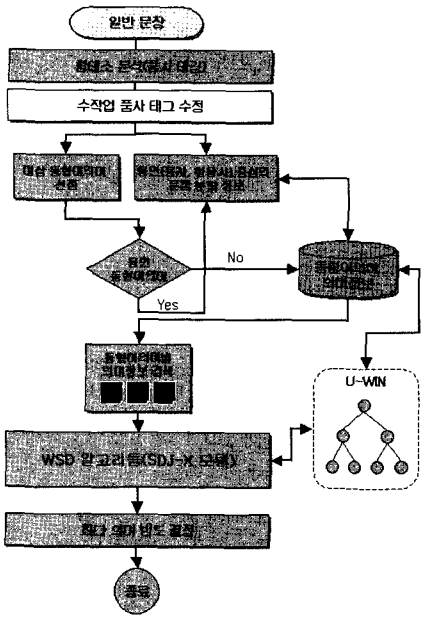
(예문 5) 식탁에 **놓**은 **국**그릇이 **배**의 롤링에 저절로 미끄러져 **국**을 쏟았다.
- '세종 의미 말뭉치' 결과
식탁/NNF+에/JKB 놓_1/VV+은/ETM **국_1/NNG+그릇/NNG**+이/JKS 배_2/NNG+의/JKG 롤링/NNG+에/JKB 저절로/MAG 미끄러지/VV+어/EC **국_1/NNG+을/JKO 쏟/VV+았/EP+다/EF+./SF**

예문 5의 또 다른 중의성 단어인 '배'는 다음에 나타나는 단어인 '롤링(rolling)'의 의미 태그가 부착된 사전 뜻풀이 '①배나 비행기가 좌우로 흔들리는 일(배_2/NNG+나/JX 비행기/NNG+가/JKS 좌우_1/NNG+로/JKB 흔들리/VV+는/ETM 일_1/NNG+./SF)'을 통해 쉽게 의미를 결정할 수 있다.

4.2 확장된 의미정보와 시소러스를 이용한 WSD

4.1절 "중의성 단어를 포함한 문장 분석"의 결과 확률정보, 거리정보, 격정보 그리고 용언(동사, 형용사)을 중심으로 문장을 분리하여 동형이의어의 의미를 결정하는 효율적이라는 결론을 얻었다. 본 논문의 2.2절 "사전 뜻풀이 특성을 이용한 의미정보 정제"를 통해 공기 어휘에 대한 확률정보, 거리정보 및 격정보 등을 구축하고 본 절에서 제안하는 WSD 모델에 이용한다. 특히 의미정보의 자료 부족 문제를 해결하기 위해 '울산대학교 어휘 지능망'(U-WIN)의 상하위 계층 구조의 이용을 제안한다.

확률정보, 거리정보, 격정보, 그리고 시소러스를 이용한 WSD 줄여서 SDJ-X 모델 이라고 한다. 모델의 전체적인 개념은 위의 (그림 8)과 같으며 모델에서 제안하는 알고리듬을 설명하면 다음과 같다.



(그림 8) 확률, 거리, 격정보 및 개념망의 계층 구조를 이용한 WSD 모델의 개념도

ㄱ. 전처리 과정

① 형태소 분석(품사 태깅) 및 수작업을 통한 품사 태그 수정

② 대상 동형이의어 선정

③ 선택 제약 정보에 의한 문장 분할 정보 획득

a. /SP(쉼표) 정보를 이용한 문장 분할

▶ 예외처리 : 단어 나열형 문장 분할에서 제외 (예, '배, 사과, 밤, 귤')

b. 용언(동사, 형용사)을 중심으로 한 문장 분리

④ 동형이의어별 의미정보 검색

a. 명사 동형이의어

▶ 문장 분할 정보(ㄱ.③.a)의 경우

- 문맥 범위내의 정보만 추출

▶ 문장 분할 정보(ㄱ.③.b)의 경우

- 의미정보가 명사인 경우 : 문장에 공기하는 모든 의미정보(명사) 추출
- 의미정보가 용언인 경우 : 문장 분할 정보를 이용하여 의미정보(용언) 제한 선택
- 의미정보가 부사인 경우 : 2.2절 선택 제약 기준 적용

b. 용언 동형이의어

▶ 문장 분할 정보(ㄱ.③.a)의 경우

- 문맥 범위내의 정보만 추출

▶ 문장 분할 정보(ㄱ.③.b)의 경우

- 의미정보가 명사인 경우 : 문장 분할 정보를 이용하여 의미정보(명사) 제한 선택
- 의미정보가 용언인 경우 : 2.2절 선택 제약 기준 적용
- 의미정보가 부사인 경우 : 2.2절 선택 제약 기준 적용

ㄴ. WSD 과정

① 첫 번째 동형이의어

② 의미정보가 있는 경우

a. 전처리를 통해 선별된 의미정보의 확률 및 거리정보를 다음 수식에 의해 계산

▶ $\max_{m} \sum_{j=1}^{m} P(H_{sk} \mid w_j) \times$ 평균거리정보

$$\text{평균거리정보} = \frac{1}{\sqrt{\frac{(\text{문장내 거리}) + (\text{의미정보 거리})}{2}}}$$

문장내 거리 : 동형이의어와 의미정보 사이의 어절거리

의미정보 거리 : 의미정보 구축 과정에서 얻은 어절거리의 평균값

▶ 예외처리 → 의미정보가 한개 밖에 없는 경우

- 의미정보에 나타나지 않는 문장 내 명사를 시소러스 상의 계층구조에 적용하여 얻은 어휘집합과 의미정보를 비교하여 발견되면 의미정보로 활용하여 (ㄴ.②)단계로 이동

③ 의미정보가 없는 경우

a. 명사 동형이의어

▶ 문장 내 명사를 시소러스 상의 계층구조에 적용하여 얻은 어휘집합과 의미정보를 비교하여 발견되면 의미정보로 활용하여 (ㄴ.②) 단계로 이동

▶ 전처리 과정의 문장 분할에 이용된 용언을 활용하기 위해 분석대상(명사 동형이의어)을 시소러스에 적용하여 얻은 어휘집합들이 가지는 용언 정보를 비교하여 분석

b. 용언 동형이의어

▶ 용언 인접한 명사(격정보 활용)를 시소러스에 적용하여 얻은 어휘집합들이 가지는 용언 정보를 비교하여 분석

④ 동형이의어 의미가 결정되면 다음의 (ㄴ.④.a)를 수행한 후 (ㄴ.①)로 이동(반복 수행)

a. 동형이의어가 하나의 의미로 결정된 경우

▶ 다음 동형이의어를 위해 추출된 의미정보에서 결정된 의미 이외의 경우를 삭제

ㄷ. 후처리 과정

① 의미정보와 시소러스를 이용하여 의미를 결정하지 못한 동형이의어는 최대빈도를 가지는 의미로 결정한다.

SDJ-X 모델을 이용한 단어 의미 중의성 해결의 과정을 살펴보도록 한다. 예문 6은 다음과 같으며 하나의 문장에 여러 개의 동형이의가 나타나고 있다. 예문은 '21세기 세종계획'의 의미 말뭉치(이하 '세종 말뭉치'라 한다.)에서 뽑은 것으로 세종 말뭉치 상에는 '날_1/NNG(일반명사), 척_8/NNB(의존명사), 배_2/NNG, 섬_3/NNG, 눈_1/NNG'이 중의성이 있는 것으로 판단하였다. 본 논문에서 구축한 동형이의어 리스트에 의하면 '날/NNG, 배/NNG, 섬/NNG, 오/VV(동사), 눈/NNG, 띄/VV' 등이 중의성이 있는 것으로 판단되어 이들을 분별 대상으로 삼는다. 세종 말뭉치의 경우 명사(일반명사, 의존명사)와 용언(동사, 형용사)를 대상으로 하고 특히, 명사에 초점을 맞추고 있다. 본 논문에서는 일반명사, 동사, 형용사의 출현 빈도 및 사전 말뭉치에 나타나는 어휘를 기반으로 하여 세종 말뭉치에 비해 용언 동형이의어를 많이 포함하고 있다.

(예문 6) 그러던 어느 날, 바다 저 먼 곳에서 한 척의 배가 섬 쪽으로 오는 것이 눈에 띄었다.

- 분석대상 -

세종 말뭉치 동형이의어 : 날_1/NNG, 척_8/NNB, 배_2/NNG, 섬_3/NNG, 눈_1/NNG

본 논문의 대상 동형이의어 : 날/NNG, 배/NNG, 섬/NNG, 오/VV, 눈/NNG, 띄/VV

동형이의어 분석 과정은 문장에 나타나는 동형이의어 순서로 진행된다. 첫 번째 동형이의어 '날/명사'에 대한 분석, 명사 동형이의어에 명사 공기 정보만 나타나는 경우로 두개의 공기쌍을 의미정보에서 얻을 수 있지만 문장이 /SP(쉼표)에 의

해 [그러던 어느 날][바다 저 ~]로 분리가 된다. 모든 /SP를 통해 문장이 분리 되는 것은 아니며, '배, 사과, 밤, 귤,...'등과 같은 단어 나열형이 있으며 이는 일반적으로 비슷한 부류의 단어(즉, 시소러스 상의 형제 노드에 해당할 수 있는 경우이다.)로 볼 수 있다. 따라서 단어의 일반적인 나열형을 제외하고는 일단 문장이 분리 되는 것으로 가정한다. 사람에 의해 의미가 결정된다면 어느/MM(관형사) 가 '날/NNG'을 꾸미는 것으로 판단하여 의미를 결정할 수 있겠지만 사전 말뭉치의 특징상 관형사(품사 태그:MM, 체언 앞에 놓여서, 그 체언의 내용을 자세히 꾸며주는 품사)의 쓰임이 매우 빈약하여 의미정보 구축과정에서 제외 되어, 결과적으로 의미 있는 공기정보를 얻지 못해 의미 결정을 유보하게 된다.

■1번 동형이의어 : 날__1/NNG => 결정 유보

동형이의어	품사	의미	공기어휘	품사	의미	거리	의미정보 평균거리	동형이의어 빈도	공기어휘 빈도	공기빈도	확률값
날	NNG	1	곳	NNG	0	4	4	546	2374	1	1.0000
날	NNG	2	배	NNG	1	7	1	107	387	1	1.0000

두 번째 동형이의어 '배/NNG'에 대한 분석, 의미 결정이 유보된 '날/NNG'은 의미정보에서 제외시킨다. '[]/NNG+[]/JKS' 형태의 격조사를 포함하는 경우 다음에 처음 나타나는 용언이 중요한 결정 요소가 되며, 예문 6에서는 (배/NNG, 오/VV)가 중요한 의미 결정 요소가 된다. (배_1/NNG, 오_1/VV)와 (배_2/NNG, 오_1/VV)로 배_1(신체), 배_2(선박)의 두 의미에 '오다/동사' 가 결합한다. 이 경우 '의미정보의 평균 거리 정보, 확률 정보'를 이용하여 하나의 의미정보로 축소시킨다. '회색'으로 표시된 행은 의미정보에서 제외된다. (배/NNG, 바다/NNG) 역시 평균 거리 정보와 확률 정보를 이용하여 계산하면 (배_1/NNG, 바다/NNG)가 의미정보에서 탈락됨을 알 수 있다. (배/NNG, 곳/NNG) 역시 동일한 방법으로 (배_1/NNG, 곳/NNG)를 탈락 시킨다. 최종적으로 남은 의미정보의 확률값을 의미별로 합하여 높은 쪽으로 의미를 결정한다.

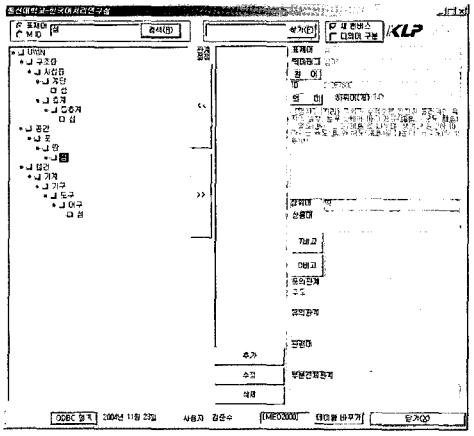
■2번 동형이의어 : 배__2/NNG => 결정 성공

동형이의어	품사	의미	공기어휘	품사	의미	거리	의미정보 평균거리	동형이의어 빈도	공기어휘 빈도	공기빈도	확률값
배	NNG	1	곳	NNG	0	-3	5	387	2374	1	.0625
배	NNG	1	날	NNG	2	-7	1	387	107	1	1.0000
배	NNG	1	눈	NNG	1	5	5	387	652	2	1.0000
배	NNG	1	바다	NNG	0	-6	8	387	468	1	.0741
배	NNG	1	오	VV	1	3	2	387	678	2	.2626
배	NNG	2	곳	NNG	0	-3	5	616	2374	24	.9375
배	NNG	2	바다	NNG	0	-6	3	616	468	20	.9259
배	NNG	2	오	VV	1	3	1	616	678	9	.7374

세 번째 동형이의어 '섬/NNG'에 대한 분석, '곳/NNG, 바다/NNG'와 '섬_3[8]'과 결합한다. '곳/NNG'은 공간을 의미하며 '배_3'의 뜻풀이와 같이 땅(육지)의 일부분으로 계층망의 정보와 의미정보가 결합 가능하다.

■3번 동형이의어 : 섬__3/NNG => 결정 성공

동형이의어	품사	의미	공기어휘	품사	의미	거리	의미정보 평균 거리	동형이의어 빈도	공기어휘 빈도	공기빈도	확률값
섬	NNG	3	곳	NNG	0	-4	5	123	2374	2	1.0000
섬	NNG	3	바다	NNG	0	-7	3	123	468	11	1.0000



(그림 9) '울산대학교 어휘 지능망'(U-WIN) 상의 '섬'

네 번째 '오/VV'에 대한 분석, 동사나 형용사와 같은 용언류의 분석은 명사의 의미결정 과정과 달리 수식-피수식 관계 등의 구문적 특성을 잘 파악해야 한다. 문장을 핵심 용언들을 중심으로 단문분할[9]적 기법을 적용하여 용언과 인접한 어절의 격정보 만을 고려하여 문장 분할을 시도한다[11]. 의미적 연관 관계에 초점이 맞추어진 단어 의미 중의성 해결에서는 정밀하고 복잡한 수준의 단문분할은 요구하지 않고 의미적 연결 관계의 제약만으로도 효과적인 결과를 얻을 수 있다. [...배가 섬 쪽으로 오는], [것이 눈에 띄었다]로 분할해 볼 수 있으며 이를 이용하면 '오/VV'의 의미정보에서 '눈/NNG' 모두를 탈락 시킬 수 있다. 본 논문에서는 이러한 문장 분할 방법을 용언과 명사의 의미적 관계 즉, (용언, 명사) or (명사, 용언)의 의미정보로 한정한다.

■4번 동형이의어 : 오/VV => 오__1/VV

동형이의어	품사	의미	공기어휘	품사	의미	거리	의미정보 평균 거리	동형이의어 빈도	공기어휘 빈도	공기빈도	확률값
오	VV	1	곳	NNG	0	-6	2	678	2374	25	1.0000
오	VV	1	눈	NNG	1	2	1	678	652	3	.0420
오	VV	1	눈	NNG	4	2	1	678	220	16	.6633
오	VV	1	눈	NNG	5	2	1	678	31	1	.2947
오	VV	1	배	NNG	1	-3	2	678	387	2	.2626
오	VV	1	배	NNG	2	-3	1	678	616	9	.7374

8) 배_3의 사전 뜻풀이: 주위가 수역으로 완전히 둘러싸인 육지의 일부. 분포 상태에 따라 제도(諸島)·군도(群島)·열도(列島)·고도(孤島)로 나누며, 생겨난 원인에 따라서는 육도(陸島)와 해도(海島)로 나눈다.

9) 단문분할이란, 한 문장에 용언이 여러 개 나타날 때 용언을 중심으로 문장을 나누는 방법이다. 단문분할을 위해 사용되어온 방법은 크게 구문 패턴을 이용하는 방법, 구문분석과 공기정보를 이용하는 방법, 하위범주화 사전을 이용하는 방법 등으로 구분할 수 있다. 구문 패턴을 이용하는 방법은 문장의 접속 구조를 찾아내고 문맥 정보를 포함하는 패턴으로 정의한 후 정의된 패턴에 따라 분할한다. 구문분석과 공기정보를 이용하는 방법은 통계적인 방법으로 조사의 격과 공기정보를 이용한다. 하위범주화 사전을 이용하는 방법은 수작업에 의한 하위범주화 사전을 이용하는 방법과 자동으로 구축하여 사용하는 방법으로 크게 나눌 수 있다[9].

다섯 번째 '눈/NNG'에 대한 분석, 앞에서 언급한 방법들을 참조하여 선행 동형이의어 의미 결정 과정에서 탈락한 의미정보는 탈락하게 되면 '곳/NNG, 띄/VV'의 정보만이 남게 되며 '[]/NNG+[]/JKS' 형태의 격조사를 포함하는 경우 다음에 최초로 나타나는 용언이 중요한 결정 요소가 된다는 정보에 가중치를 부여한다.

■5번 동형이의어 : 눈_1/NNG => 결정 성공

동형이의어	품사	의미	공기어휘	품사	의미	거리	의미정보 평균 거리	동형이의어 빈도	공기어휘 빈도	공기 빈도	확률값
눈	NNG	1	곳	NNG	0	-8	4	652	2374	18	.2611
눈	NNG	1	띄	VV	1	1	1	652	36	20	1.0000
눈	NNG	1	배	NNG	1	-5	5	652	387	2	1.0000
눈	NNG	1	오	VV	1	-2	1	652	678	3	.0420
눈	NNG	4	곳	NNG	0	-8	5	220	2374	3	.1287
눈	NNG	4	오	VV	1	-2	1	220	678	16	.6633
눈	NNG	5	곳	NNG	0	-8	5	31	2374	2	.6102
눈	NNG	5	오	VV	1	-2	1	31	678	1	.2947

마지막 동형이의어 '띄/NNG'에 대한 분석, 앞에서 의미가 결정된 의미정보 '눈_1/NNG'과 결합하게 되어 '띄_1/NNG'로 의미를 결정하게 된다.

■6번 동형이의어 : 띄/NNG => 의미 결정 띄_1/VV

동형이의어	품사	의미	공기어휘	품사	의미	거리	의미정보 평균 거리	동형이의어 빈도	공기어휘 빈도	공기 빈도	확률값
띄	VV	1	눈	NNG	1	-1	1	36	652	20	1.0000

5. 실험 및 결과

본 논문에서 제안하는 WSD 모델의 성능을 측정하기 위해서는 의미 분별된 대량의 비학습 말뭉치가 필요하다. 본 논문에서는 150만 어절 규모의 세종 의미 말뭉치를 그 대상으로 하였으며, 기존 연구에 이용한 명사 및 동사 동형이의어를 중심으로 세종 의미 말뭉치에 등장하지 않거나 하나의 의미로만 사용되는 것을 제외한 명사 15개(기원, 독, 배, 부정, 비, 상, 의사, 의지, 장수, 절, 지도, 차, 철, 판, 표)와 동사10개(갈다, 끼다, 달다, 들다, 쓰다, 붓다, 쉬다, 차다, 켜다, 타다)를 대상으로 정확률을 측정하고 기존 연구와 비교하였다.

의미정보 구축에 이용된 국어사전(금성사전)과 세종 의미 말뭉치 구축에 이용된 국어사전(표준사전)이 상이하여 정확률 측정 및 기타 활용을 위해서는 두 국어사전 사이의 의미를 연결시키는 작업이 선행되어야 한다. 의미 연결에 앞서 말뭉치에 나타나는 동형이의어를 살펴보면 다음 <표 4>와 같다. 말뭉치의 규모는 사전 뜻풀이 말뭉치(약 100만 어절), 세종 의미 말뭉치(약 150만 어절)로 포함된 동형이의어에 대한 종수와 의미 개수는 다음과 같다.

표제어로 등록된 동형이의어(금성사전 기준)는 14,447종(의미개수: 41,036)이며 사전 말뭉치에는 74.12%, 세종 의미 말뭉치에는 66.20%가 사용되어 사전 말뭉치에 동형이의어가 더욱 많이 사용됨을 알 수 있었다. 금성사전과 표준사전 사이의 의미 연결은 금성사전을 기준으로 총14,171종(명사 13,341, 동사 664, 형용사 186)에 38,215개 의미를 수작업을 통해 연결을 수행하였다. 대부분의 동형이의어에 대해 일대일(1:1)의 연결이 이루어 졌으며, 일부 동형이의어에 대해 일대다(1:N) 또는 다대일(N:1)의 연결이 이루어졌으나 전체에서 차지하는 비율이 낮고, 자주 사용하지 않는 의미에 국한되어 단의 의미 중의성 해결에는 큰 문제가 없었다.

<표 4> 사전 뜻풀이 말뭉치와 세종 의미 말뭉치에 포함된 동형이의어

	사전 뜻풀이 말뭉치(100만 어절)		세종 의미 말뭉치(150만 어절)	
품사	종수	의미개수	종수	의미개수
일반명사(NNG)	9,820	16,404	8,839	11,387
동사(VV)	535	995	271	464
형용사(VA)	99	111	31	35
부사(MAG)	154	189	273	325
합계	10,708	17,699	9,564	12,211

본 논문에서는 정확률을 이용해서 WSD 모델의 성능을 평가하며, 정확률은 다음과 같이 구한다.

$$\text{정확률}(\%) = \frac{\text{정확하게맞춘동형이의어수}}{\text{동형이의어수}} \times 100$$

기존 연구의 실험에 이용된 말뭉치는 '21세기 세종 계획'에서 제공하는 350만 어절 규모의 품사 태그 말뭉치이며, 사람의 수작업을 통해 일부 실험 대상 동형이의어에 대한 의미 태깅만 수행하여 실험을 수행하였으며, 일부 의미 태깅의 오류를 포함하고 있다. 따라서 본 논문에서는 작업의 효율성과 다양한 동형이의어에 대한 의미 구분을 위해 350만 어절 품사 태그 말뭉치의 일부인 150만 어절 세종 의미 태그 말뭉치를 대상으로 2장 "사전 뜻풀이 특성을 이용한 의미정보 정제"의 결과로 재구축된 의미정보를 이용하는 실험을 수행하였다.

본 논문에서 제안하는 모델은 확률정보, 거리정보 그리고 격정보를 이용하는 모델(SDJ)과 개념망을 이용해 확장하는 모델(SDJ-X)로 나누고 [4]의 WSD 모델 중 가장 정확률이 높은 NPH 및 어절 거리 가중치를 적용한 모델(이하, NEo 모델이라 한다.)과 성능 비교하였다. 기존 의미정보와 재구축한 의미정보의 비교는 NEo 모델에 기존 의미정보를 적용한 실험을 NEo-old라 하고, 재구축 의미정보를 적용한 실험을 NEo-new라 한다. 정확률 비교 결과는 다음 <표 5>, <표 6>과 같다. 실험 결과 정제된 의미정보를 실험한 NEo-new의 정확률이 기존 의미정보를 이용한 N때-old에 비해 2.82%(명사 : 3.15%, 동사 : 2.62%) 정확률이 향상되었다. 이는 사전 뜻풀이 특성을 반영한 선택 제약 정보에 의해 의미정보가 정제되었음을 보여준다.

SDJ-X 모델이 최다 빈도 의미 결정(Maximum Frequence Class, MFC, 정확률 베이스라인)에 비해 18.87%(명사 21.73%,

동사 17.11%) 정확률 향상을 보였으며, 확률 가중치와 어절 거리 가중치를 이용한 모델(NEo-new)에 비해서는 10.49% (명사 8.84%, 동사 11.51%)의 정확률 향상되었다. 또한 시소러스를 사용하지 않고 확률정보, 거리정보, 격정보 만을 이용한 SDJ 모델에 비해 6.12%(명사 5.29%, 동사 6.64%) 높은 정확률을 보였다.

116,706문장(문장 당 평균 13어절)으로 구성된 150만 어절 규모의 세종 의미 말뭉치에 하나의 의미가 80% 이상을 차지하는 동형이의어는 총 8,254종으로 그중 하나의 의미로만 사용된 동형이의어는 7,219종(누적 빈도 297,570(55.43%)으로 이때의 MFC는 100%가 된다. SDJ와 SDJ-X 모델의 정확률은 각각 90.92%, 91.59%이다. 하나의 의미가 80% 이상을 차지하는 경우의 MFC는 98.00%이고 SDJ와 SDJ-X 모델은 각각 88.92%, 89.78%를 나타내므로, 이 경우는 MFC에 의한 의미 결정이 효과적임을 알 수 있다.

〈표 5〉 세종 의미 말뭉치(150만)에 대한 실험 결과 비교 (동사 10개)

단어	의미수	출현 횟수	MFC	NEo-old	NEo-new	SDJ	SDJ-X
갈다	3	94	41.49	60.64	64.89	71.28	78.72
끼다	3	182	57.69	70.33	73.08	80.77	86.26
달다	3	175	70.86	68.00	68.57	73.71	83.43
들다	2	2,527	59.64	65.22	69.17	73.84	78.16
쓰다	3	2,262	49.12	49.82	51.68	57.74	67.55
붓다	2	72	70.83	72.22	77.78	77.78	88.89
쉬다	3	288	65.63	71.53	70.83	77.78	85.42
차다	3	333	71.47	60.96	63.66	65.77	77.48
켜다	2	70	81.43	82.86	74.29	90.00	90.00
타다	6	789	76.30	79.47	82.13	83.27	85.42
평균	3	6,792	59.23	62.21	64.83	69.70	76.34
표준 편차			12.44	9.60	8.43	9.08	6.75

〈표 6〉 세종 의미 말뭉치(150만)에 대한 실험 결과 비교 (명사 15개)

단어	의미수	출현 횟수	MFC	NEo-old	NEo-new	SDJ	SDJ-X
기원	3	151	47.02	60.26	62.25	64.90	74.17
독	2	55	80.00	87.27	90.91	92.73	96.36
배	4	949	44.15	66.17	67.23	69.76	72.92
부정	3	605	48.26	43.64	50.91	57.19	63.97
비	5	402	83.58	90.05	91.29	92.29	94.28
상	4	211	39.81	72.51	73.46	75.83	83.41
의사	3	354	49.72	62.99	66.10	68.64	76.55
의지	2	325	78.15	74.46	74.46	82.15	87.38
장수	3	61	59.02	42.62	55.74	75.41	81.97
절	3	160	53.13	68.75	71.25	76.25	81.88
지도	2	222	73.42	80.18	86.04	85.59	87.84
차	3	380	72.37	72.37	75.53	79.21	83.68
철	3	67	41.79	65.67	70.15	73.13	80.60
판	3	72	55.56	62.50	68.06	69.44	80.56
표	3	183	48.63	61.20	67.21	68.85	78.69
평균	3	4,197	56.99	66.74	69.88	73.43	78.72
표준 편차			15.06	13.34	11.48	9.83	8.18

〈표 7〉 세종 의미 말뭉치에 하나의 의미로 80% 이상 사용된 경우의 실험 결과

비율 (빈도)	100% (297,570)	100%>x≥90% (159,402)	90%>x≥80% (29,138)	합계/평균 (486,110)
MFC	100.00%	96.12%	87.94%	98.00%
SDJ	90.92%	85.93%	84.86%	88.92%
SDJ-X	91.59%	86.93%	88.64%	89.78%

6. 결론 및 향후 연구

본 논문에서는 동형이의어 수준의 단어 의미 중의성 해결을 목적으로 사전 뜻풀이 특성을 이용해 정제한 의미정보를 확률정보, 거리정보 및 격정보, 문장 분할정보 등을 기반으로 하는 선택 제약 정보 기반 WSD 모델로 발전 시켰다. 또한 사전 말뭉치 기반 의미정보의 문제점인 자료 부족 문제를 '표준국어사전'을 기반으로 구축된 '울산대학교 어휘 지능망'(U-WIN)의 단어 계층적 구조(시소러스)를 이용하는 WSD 모델을 제안하였다.

본 논문에서 제안하는 확률정보, 거리정보, 격정보, 그리고 시소러스를 함께 이용하는 모델(SDJ-X)이 "21세기 세종 계획"에서 제공하는 150만 어절 규모의 의미 태그 말뭉치를 대상으로 한 실험에서 최다 빈도 의미 결정(Maximum Frequence Class, MFC, 정확률 베이스라인)에 비해 18.87%(명사 21.73%, 동사 17.11%) 정확률 향상을 보였으며, 확률 가중치와 어절 거리 가중치를 이용한 모델에 비해서는 10.49%(명사 8.84%, 동사 11.51%) 정확률 향상되었다. 또한 시소러스를 사용하지 않고 확률정보, 거리정보, 격정보 만을 이용한 모델에 비해 시소러스를 이용한 SDJ-X 모델이 6.12%(명사 5.29%, 동사 6.64%) 향상되었다. 그리고 본 논문에서 제안하는 WSD 모델의 견고성 측정을 위해 의미가 편중된 경우(하나의 의미로 80% 이상 사용)의 WSD 정확률 측정결과 89.78% (MFC : 98%)로 최다 빈도 의미 결정보다 낮은 성능을 보였으나, 하나의 의미로 90-80% 정도 사용된 경우 88.64% (MFC : 87.94%)로 최다 빈도 의미 결정보다 좋은 결과를 보였다.

단어 의미 중의성 해결을 위해서는 다양한 언어 지식이 필요하며, WSD 모델의 실용화를 위해서는 신뢰할 수준의 동형이의어 분별과 함께 다의어 수준의 WSD가 요구된다. 향후 연구 과제는 첫째, 다의어 수준의 WSD를 위한 의미정보 세분화 작업과 이 과정에서 나타나게 될 자료 부족 문제의 해결 방안 연구이다. 그 방안 중 하나로 현재 단어 계층 구조(시소러스) 활용 수준인 U-WIN의 이용을 개념망 수준으로 확대하는 연구를 진행하고자 한다. 둘째, 실제 자연언어처리 응용 분야에 적용하여 문제점을 파악하여 응용 시스템에 적합한 WSD 모델의 개량 작업을 진행하고자 한다.

참 고 문 헌

[1] 이호, 백대호, 임해창, "분류 정보를 이용한 단어 의미 중의성 해결", 한국정보과학회 논문지(B), Vol.24, No.7, pp.779-789, 1997.

[2] 이승우, 이근배, "국소 문맥과 공기 정보를 이용한 비교사 학습 방식의 명사 의미 중의성 해소", 한국정보과학회 논문지(B), Vol.27, No.7, pp.769-782, 2000.

[3] Choueka, Yaacov and Serge Lusignan, "Disambiguation by short contexts, Computers and the Humanities," 19, pp.147-158, 1985.

[4] 김준수, 최호섭, 옥철영, "가중치를 이용한 통계 기반 동형이의어 분별 모델", 한국정보과학회 논문지(소프트웨어 및 응용), Vol.30, No.11, pp.1112-1123, 2003.

[5] 허정, 옥철영, "사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 한국정보과학회 논문지(소프트웨어 및 응용), Vol.28, No.9, pp.688-698, 2001.

[6] 조평옥, 옥철영, "사전 뜻풀이말에서 구축한 한국어 명사 의미 계층구조", 한국인지과학회 논문지, 제10권, 제4호, 1999.

[7] 이재윤, 단어 동시출현 기반 질의확장의 성능 최적화에 관한 연구, 박사학위 논문, 연세대학교 문헌정보학과, 2003.

[8] Kenneth ward Church, Patrick Hanks, "Word association norms, mutual information, and lexicography," Computational Linguistics, Vol.16, issue 1, pp.22-29, 1990.

[9] 김준수, 이왕우, 김창환, 옥철영, "상호정보량을 이용한 동형이의어 분별용 의미정보의 정제", 한국정보과학회 2002 봄 학술발표논문집(B), 제29권, 제1호, pp.460~462, 2002.

[10] 나동렬, "한국어 파싱에 대한 고찰", 정보과학회 논문지, 제12권, 제8호, pp.33~46, 1992.

[11] 박성배, 문장 분할을 이용한 한국어 분석, 석사학위논문, 서울대학교, 1996.

[12] 조정미, 코퍼스와 사전을 이용한 동사 의미 분별, 박사학위 논문, 한국과학기술원 전산학과, 1998.

[13] 임지룡, "다의어 인지적 의미 특성", 한국언어학회, 언어학, 제18권, 단일호, pp.229~261, 1997.

[14] 최경봉, "단어 의미의 구성과 의미 확장 원리-다의어 문제를 중심으로", 한국어학회, 한국어학 제9집, pp.307~331, 1999.

[15] 김태자, "다의어고(多義語攷)", 한국언어문학회, 한국언어문학, 제23권, 단일호, pp.195~212, 1984.

김 준 수

e-mail : kimjunsu@ulsan.ac.kr

1998년 울산대학교 수학과(학사)

2000년 울산대학교 수학과(이학석사)

2005년 울산대학교 컴퓨터정보통신공학과 (공학박사)

2001년~2003년 (주)시소러스 선임연구원

2004년~현재 울산대학교 컴퓨터정보통신공학부 객원교수

관심분야 : 한국어정보처리, 의미분석, 정보검색

옥 철 영

e-mail : okcy@ulsan.ac.kr

1982년 서울대학교 컴퓨터공학과(학사)

1984년 서울대학교 컴퓨터공학과(공학석사)

1993년 서울대학교 컴퓨터공학과(공학박사)

1994년 러시아 TOMSK 공과대학 교환교수

1996년 영국 GLASGOW대학교 객원교수

1984년~현재 울산대학교 컴퓨터정보통신공학부 정교수

관심분야 : 한국어정보처리, 지식베이스, 기계학습, 온톨로지