사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템

(A Homonym Disambiguation System based on Semantic Information Extracted from Dictionary Definitions)

허 정 † 옥 철 영 ††

(Jeong Hur) (Cheol-Young Ock)

요 약 동형이의어는 문장에서 그와 함께 사용된 체언, 용언에 의해서 그의 의미를 확정지을 수 있다.

본 논문에서는 사전의 뜻풀이말에서 추출한 통계적 의미정보에 기반한 동형이의어 중의성 해결 시스템을 제안한다. 의미정보는 동형이의어를 포함하고 있는 사전의 뜻풀이말에서 체언(보통 명사)와 용언(형용사, 동사)을 추출하여 구성된다. 정확한 의미정보를 추출하기 위해서 사전 뜻풀이말의 유형을 두 가지로 분류하였다. 첫 번째 유형은 의미분별할 동형이의어와 표제어가 의미적으로 상-하의어 관계를 이루고 있는 경우로, 표제어의 뜻풀이말에서 동형이의어가 의미적으로 중심어이다. 이러한 상-하의어 관계는 의미계층 구조가 없는 경우에 활용할 수 있으며, 자료 부족 문제를 해결하기 위한 의미정보의 확장에 유용하다. 두 번째 유형은 동형이의어가 뜻풀이말의 중간에 사용된 경우이다.

본 논문에서 제안하는 동형이의어 중의성 해결 시스템은 체언과 용언 의미정보를 모두 고려한 모델로, 체언과 용언이 동형이의어 중의성 해결에 영향을 주는 정도(가중치)를 결정하기 위하여 9개의 동형이의어 명사를 대상으로 실험하였다. 학습에 이용된 코퍼스(사전 뜻풀이말)로 실험한 결과, 체언과 용언의 가중치가 0.9/0.1일 때 평균 96.11%의 중의성 해결 정확률이 가장 높았다.

또한 제안하는 동형이의어 중의성 해결 시스템의 일반성을 측정하기 위해, 학습되지 않은 외부 데이타(국어 정보베이스 I과 ETRI 코퍼스 1,796 문장)로 실험한 결과 평균 80.73%의 정확률을 보였다.

Abstract A homonym could be disambiguated by another words in the context such as nouns, predicates used with the homonym.

This paper proposes a homonym disambiguation system based on statistical semantic information which is extracted from definitions in dictionary. The semantic information consists of nouns and predicates that are used with the homonym in definitions. In order to extract accurate semantic information, definitions are classified into two types. One has hyponym-hypernym relation between title word and head word (homonym) in definition. The hyponym-hypernym relation is one level semantic hierarchy and can be extended to deeper levels in order to overcome the problem of data sparseness. The other is the case that the homonym is used in the middle of definition.

The system considers nouns and predicates simultaneously to disambiguate the homonym. Nine homonyms are examined in order to determine the weight of nouns and predicates which affect accruracy of homonym disambiguation. From experiments using training corpus(definitions in dictionary), the average accruracy of homonym disamguation is 96.11% when the weight is 0.9 and 0.1 for noun and verb respectively.

And another experiment to measure the generality of the homonym disambiguation system results in the 80.73% average accuracy to 1,796 untraining sentences from Korean Information Base I and ETRI corpus.

· 이 논문은 1999년 한국학술진흥재단(1998-0001-E01184) 및 2000년 정보통신 우수시범학교 지원사업의 지원에 의하여 연구되었음.

† 비 회 원 : 한국전자통신연구원 언어공학연구부 연구원
jeonghur@etri.re.kr

†† 종신회원 : 울산대학교 컴퓨터정보통신공학부 교수
okcy@mail.ulsan.ac.kr

논문접수 : 2000년 11월 6일
심사완료 : 2001년 7월 5일

1. 서 론

자연언어처리의 과정은 형태소 분석(morphological analysis), 구문 분석(syntactic analysis), 의미 분석(semantic analysis), 담화 분석(discourse analysis)의 네 단계로 이루어진다[1]. 그러나 각 단계에서 발생되는 중의성은 자연언어처리의 정확률을 저해시키는 원인이다. 이와 같은 이유로 자연언어처리의 중점적인 연구 목표는 중의성 해결이다. 한국어에 대한 형태소 분석이나 구문 분석 시의 중의성 해결 방법은 그동안 활발히 연구되어 좋은 성과가 나타나고 있으며, 의미 분석 시의 중의성 문제는 최근에 연구가 활성화되고 있다.

문장 중의 단어는 다른 단어와의 관계에 의해 하나의 의미가 결정된다. 특히 해당 단어가 동형이의어나 다의어인 경우 문장에서 함께 사용된 다른 단어와의 의미 의존 관계에 의해 그 단어의 의미가 결정된다. 이와 같이 동형이의어와 다의어의 의미를 결정하는 과정을 의미 중의성 해결(WSD, Word-Sense Disambiguation)이라 한다. 예를 들면,

(1) 철수가 배(과일)를 먹었다.

(2) 철수가 배(운송수단)를 탔다.

(3) 철수는 배(신체부위)가 아프다고 했다.

위의 예에서 "배"는 동형이의어로 "과일(梨)", "운송수단(船舶)", "신체부위(腹部)" 등의 의미 중의성을 가진다. 문장 (1)에서의 "배"는 "먹을 수 있는 대상", (2)에서는 "탈 수 있는 대상", 문장 (3)에서의 "배"는 "아픈 대상"의 의미로 문맥의 다른 단어와의 의미적 관계에 의해 의미분별된다.

이러한 의미 중의성 해결을 위해 많은 방법론들이 제시되고 있다. 학습 데이타의 형태에 따라서 사전을 이용하는 방법과 코퍼스를 이용하는 방법으로 분류할 수 있고, 방법론에 따라서 규칙을 이용한 방법, 확률 통계를 이용하는 방법과 의미 계층 구조를 이용하는 방법으로 분류할 수 있다[17].

코퍼스를 이용한 의미 중의성 해결을 위해서는 대량의 의미 부착 코퍼스(semantic tagged corpus)가 필요한데, 신뢰성이 보장된 이용 가능한 코퍼스를 구하기가 힘들고 구축하기에는 많은 비용이 드는 단점이 있으나 언어의 동적인 특성을 잘 반영하는 장점이 있다. 사전을 이용한 방법은 언어의 동적인 특성을 반영하지 못하는 단점이 있으나, 개별 단어의 의미와 용례를 기술하고 있기 때문에 의미정보 추출이 쉽다는 장점이 있다.

일반적으로 통계 기반의 의미 중의성 해결 방법은 통계 데이타의 자료 부족(data sparseness) 문제가 야기되는데 이는 의미 중의성 해결의 정확률에 영향을 끼친다. 자료 부족 문제를 최소화하기 위해서 의미 계층 구조를 이용한 유사어들의 공기 정보를 활용하는 방법들이 연구되고 있다[4,13].

의미 계층 구조를 이용한 연구들은 WordNet, Roget Thesaurus 등의 시소러스를 활용하여 유사어들의 공기 정보 및 상-하의어 관계 등의 의미정보로 의미 중의성을 해결한다[6,8]. 그러나, 한국어로 작성된 이용가능하고 신뢰성이 보장되는 의미 계층 구조를 구하기도 힘들고, 구축하는 데에는 많은 비용이 든다. 이로 인해, WordNet과 Roget Thesaurus의 영어 단어들을 해당 한국어들로 번역하여 이용하는 연구가 진행되고 있다. 그러나, 서로 다른 두 자연언어 사이에는 어휘의 개념적 차이뿐만 아니라 표현의 차이가 심한 경우가 많기 때문에 영어권의 의미 계층망을 한국어에 적용하는 것은 한계가 있다.

David Yarowsky(1992)는 자료 획득 병목 현상[1]을 최소화하기 위해서, Roget Thesaurus의 카테고리를 기반으로 한 통계 모델을 제안하였다[13]. 1,042개의 카테고리 각각에 대해 다음의 과정을 통해 의미정보를 획득한다. 먼저, 각각의 Roget 카테고리를 표현하는 문맥(context)을 추출한다[2]. 추출된 문맥에서 대표어(salient word)를 추출하고 각 단어에 대한 가중치(weigth)를 결정한다. 이와 같이 의미정보를 추출한 다음, 의미분별을 하고자 하는 단어가 어느 카테고리에 속하는가를 문맥의 가중치 값을 통해 결정한다. 평균 3진 의미분별(3-way sense distinction)에서 92%의 정확률을 보였다.

Alpha K. Luk(1995)은 자료 부족 문제를 최소화하기 위해서, LDOCE(Longman Dictionary of Contemporary English; Procter, 1978)의 통제 어휘(controlled vocabulary)와 Brown 코퍼스를 이용한 의미 중의성 해결 모델을 제안하였다[12]. 먼저 LDOCE의 통제 어휘[3] 2,000개 중 1,792개를 정의 개념(defining concept)으로 추출하고, 이 정의 개념들에 대한 통계 의미정보를 Brown 코퍼스[4]로부터 추출한다. 평균 3진 의미분별에

---

1) 이용하고자 하는 지식의 획득에 많은 시간과 비용이 소요되는 것을 지식 획득의 병목(knowledge acquisition bottlenect)이라고 한다[10].

2) 1,000만 단어를 포함하는 1991년도 Grolier 전자 백과 사전에서 30,924 라인을 추출하였다.

3) 표제어에 대한 의미를 기술할 때 사용하는 의미 기술 단어를 이른다. LDOCE는 2,000개 가량의 통제 어휘를 이용하여 모든 표제어의 의미를 기술하고 있다.

4) 100만 단어로 구성이 되어 있고, 한 문장은 평균 19.4개의 단어로 구성되어 있다.

서 77%의 정확률을 보였다. 그리고 의미분별하고자 하는 단어가 포함된 문장을 구성하는 단어들의 목록을 무작위로 주고 사람이 의미분별을 하도록 실험한 결과 71%의 정확률을 보였다. 상기의 방법론은 통제 어휘 자체가 의미 중의성을 지닌 동형이의어가 많음으로 인해서 한계점이 있다.

조정미(1998)는 지식 획득의 병목 문제와 자료 부족 문제를 최소화하면서 효과적으로 의미분별을 하는 방법을 제안하였다[10]. 품사 부착 코퍼스로부터 구문 지식에 해당하는 선택 제약 지식을 추출하여 지식 획득의 병목 현상을 해소하며, 추출된 명사와 동사의 선택 제약 지식을 순환적으로 학습하여 자료 부족 문제를 해소한다. 또한, 사전 분석을 통해 의미분별을 하고자 하는 단어의 의미 지시자와 단어의 분류 정보를 추출한다. 이와 같이 추출된 정보에서 명사의 정보만을 가지고 동사의 의미 분별을 한 경우 48.3%의 정확률을 보인 반면, 명사와 동사의 분포를 순환적으로 학습하여 실험한 결과 61%의 정확률을 보여 자료부족 문제가 많이 완화됨을 알 수 있다. 또한, 사전의 의미 지시자와 단어 분류 정보를 포함한 실험에서 86.3%의 정확률로 코퍼스에서 추출한 명사와 동사의 분포 정보만을 이용한 경우보다 정확률이 크게 향상됨으로써 사전에 포함된 의미 정보가 중요함을 확인할 수 있다.

서희철(1999)은 의미 계층 구조에 나타나는 유사어들의 공통적인 특징과 개별적인 특징을 모두 고려하여 의미 중의성을 해결하는 방법을 제안하였다[4]. 품사 부착된 1,000만 어절의 코퍼스에서 의미 계층 구조를 이용하여 추출된 유사어의 용례를 추출하여[5] 학습 코퍼스로 이용한다. 학습 코퍼스를 통해서 구축된 유사어 벡터의 자질값을 이용하여, 의미 중의성을 해결한다. 세 단어(배, 밤, 고개)를 대상으로 실험한 결과 의미 벡터만을 이용한 의미 분별에서보다, 의미 계층 구조의 유사어를 이용한, 유사어 벡터의 자질값을 이용하였을 때 16%의 정확률 향상을 보임으로써 의미 계층 구조가 의미 분별의 중요한 자원임을 확인 할 수 있다. 그러나, 대용량의 우리말 의미 계층 구조를 구축하기 어려운 문제점이 있다.

박성배(2000)는 답이 알려져 있지 않은 데이타(unlabeled data)를 사용한 의미 중의성 해결을 위해 위원회(committee)에 의한 선택 샘플링(selective sampling) 알고리듬을 이용하는 방법을 제안하였다[2]. 답이 알려진 데이타(labeled data)를 이용한 의미 중의성 해결은 데이타 구축에 사람의 많은 간섭이 필요함으로 많은 비용이 든다. 이를 최소화하기 위해 답이 알려져 있지 않은 데이타를 이용한 의미 중의성 해결 방법을 제시했다. 답이 알려진 데이타를 이용하여 의미 분별한 결과가 87%이고, 답이 알려져 있지 않은 데이타를 이용하여 의미 분별한 결과가 85.2%로 1.8%의 미세한 정확률 차를 보임으로써, 많은 비용이 필요한 데이타를 이용하지 않고도 의미 분별에 정확률을 향상시킬 수 있음을 알 수 있다.

박영자(1998)는 사전의 의미 기술 문장에서 각 명사 의미에 대한 속성을 자동으로 추출하여 의미를 클러스터링 하는 새로운 방법을 제안하였다[3]. 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 피지 릴레이션을 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, 의미분별을 하고자 하는 단어가 포함된 문장의 단어들과 의미분별을 하고자 하는 단어의 뜻풀이말들에 포함된 단어들이 많이 공유하는 의미를 의미분별을 하고자 하는 문장의 의미로 선택한다. 의미 분별에 실패했을 때 한 단어의 여러 의미들 중에서 뜻풀이말이 긴 의미를 그 단어의 대표 의미로 가정한다. 세 단어의 의미 분별 결과 눈, 배, 차가 각각 81%, 74%, 83%의 정확률을 보였다. 실험 대상 문장이 사전 뜻풀이말로 문장의 길이가 짧고, 제한된 단어(controlled vocabulary)를 사용했으므로 정확률에 크게 의미를 부여할 수는 없으나, 사전 뜻풀이말의 공기 관계를 이용하여 의미 분별을 향상시킬 수 있음을 확인할 수 있다.

이상의 연구들에서 살펴본 바와 같이 사전의 뜻풀이말에서의 단어의 공기 관계를 이용하여 단어의 의미를 분별할 수 있다. 또한 명사와 용언의 정보를 함께 이용함으로써 자료 부족 문제를 완화시킬 수 있으며, 의미 계층 구조가 없는 상황에서 사전의 뜻풀이말을 이용한 의미 계층 구조 정보를 추출하는 것은 의미 분별의 향상에 크게 기여할 수 있다. 이와 같이 기존의 연구들은 의미계층망이나 코퍼스를 이용한 의미 중의성 해결 방법을 제시하였다. 그러나, 한국어에 대한 객관적이고 실용적인 의미계층망이 없고 의미 주석된 코퍼스가 없는 상태에서 이러한 연구방법론을 한국어에 직접 적용하기에는 한계가 있다.

5) 동일한 상위 개념을 가지면서 같은 레벨에 있는 단어들을 유사어로 간주하고, 유사어 중에서 의미 중의성이 없는 단어이면서 빈도가 100회 이상인 단어만 이용한다.

특히, 박영자(1998)는 의미 중의성 해결을 위해 의미 분별하고자 하는 동형이의어의 뜻풀이말 자체만을 이용함으로써 자료 부족 현상이 심각하고, 용언을 제외한 명사만을 의미 정보로 이용함으로써 의미 분별의 한계가 있다. 본 논문에서는 동형이의어의 중의성을 해결하기 위하여 의미분별하고자 하는 동형이의어를 포함하고 있는 사전 뜻풀이말 전체에서 표제어와 동형이의어간의 상-하의어 관계의 의미 계층 구조를 유추하고, 제한된 의미 계층 구조를 활용하여 체언과 용언의 공기 정보를 추출하여 의미정보로 구성하는 동형이의어 중의성 모델을 제안한다.

2. 동형이의어 중의성 해결을 위한 의미정보

의미분별을 위해서는 의미정보 획득 작업이 선행되어야 한다. 획득된 의미정보의 정확성과 정교성은 의미분별의 정확률에 크게 영향을 미친다. 문맥에서 중의성을 가지는 단어에 대한 의미분별은 문맥의 공기 관계뿐만 아니라 담화가 이루어지는 상황에서 사람의 추론이 개입하여 이루어진다. 따라서, 의미분별에 절대적인 의미정보를 획득하기란 어렵다.

기존의 연구에서는 선택 제약을 이용한 규칙 기반의 의미정보를 획득하는 방법과 공기 정보를 이용한 통계 기반의 의미정보를 획득하는 방법을 이용했다[3,4,6,8,10,12,13,18]. 선택 제약은 구문구조에 해당하는 정보이므로, 의미정보 추출을 위해서는 구문구조가 부착된 코퍼스가 필요하다. 공기 정보를 이용한 방법에서는 단어들의 공기 정보에 따른 빈도 및 확률값을 의미정보로 가지고, 이 정보를 이용한 의미 벡터를 통해서 의미분별을 한다. 최근 연구에서는 공기 정보를 이용한 의미분별에 대한 연구가 활발히 진행되고 있다[12,13].

다음 <표 1>은 사전의 뜻풀이말에서 동형이의어 "배"가 "신체부위"와 "운송수단"의 의미로 사용된 경우의 표제어와 해당 뜻풀이말을 보이고 있다.

<표 1>에서 알 수 있듯이 동형이의어 "배"는 의미별로 함께 사용되는 단어들이 다르며 이들 단어들과 의미적으로 공기 관계를 구성한다. 본 연구에서는 사전 뜻풀이말에서 동형이의어와 함께 사용된 단어들간의 이와 같은 특성을 이용한 공기 정보를 의미정보로 이용한다.

다양한 의미정보를 추출하기 위하여 본 논문에서는 뜻풀이말을 <표 2>와 같이 두 가지 유형으로 분류하였다.

표 2 의미정보 추출 대상 뜻풀이말 유형(배:운송수단)

	표제어	뜻풀이말
1차 유형	**거선**	아주 큰 **배**
	나룻배	나룻터에서 사람이나 짐 등을 건내주는 **배**.
	유조선	유조 시설을 갖춘 **배**
	:	:
2차 유형	**난파**	**배**가 항행 중에 폭풍우 등을 만나 깨어짐.
	운하	육지를 파서 강을 내고 **배**가 다니게 한 수로.
	:	:

<표 2>에서 1차 유형은 표제어의 뜻풀이말에서 동형이의어가 표제어의 핵심어인 경우이고, 2차 유형은 동형이의어가 뜻풀이말 중간에 나타나는 유형이다. 1차 유형은 <그림 1>과 같이 동형이의어와 표제어간에 상-하의어 관계를 가진 것으로, 한국어에 대한 WordNet이나 Roget Thesaurus와 같은 의미계층 구조가 없는 경우에 사전에서 직접 추출할 수 있는 1 단계(one-level) 의미계층 구조이다.

표 1 뜻풀이말에서 의미공기 정보(배 : 신체부위, 운송수단)

배(신체부위)		배(운송수단)	
표제어	뜻풀이말	표제어	뜻풀이말
똥배	: 똥똥하게 불러 나온 **배**.	**강선**	: 강철로 만든 **배**.
물배	: 물만 먹고 부른 **배**.	**객선**	: 손님을 태우는 **배**.
젖배	: 젖을 먹는 어린아이의 **배**.	**계류선**	: 부두나 바닷가에 매어놓은 **배**.
헛배	: (소화 불량 등으로) 음식을 먹지 아니하고도 부른 **배**.	**고주**	: 외로이 떠 있는 **배**.
가로막	: **배**와 가슴 사이에 있는 근육성의 막.	**난파선**	: 항해 중 폭풍우나 그 밖의 장애로 파괴된 **배**.
개복수술	: **배** 안에 있는 기관의 수술 또는 이물을 없애기 위하여 복벽을 갈라내는 수술.	**강나루**	: 강가의 **배**가 건너다니는 일정한 곳.
		귀항	: **배**가 출발하였던 항구로 돌아오거나 돌아가는 항해.
내장	: 가슴과 **배** 속에 있는 여러 기관의 총칭.	**닻**	: **배**를 한 곳에 머물게 하기 위하여, 줄을매어 물 밑바닥으로 가라앉히는, 갈고리가 달린 제구.
배불뚝이	: **배**가 불뚝하게 나온 사람.		
뱃살	: **배**를 싸고 있는 살이나 가죽.		

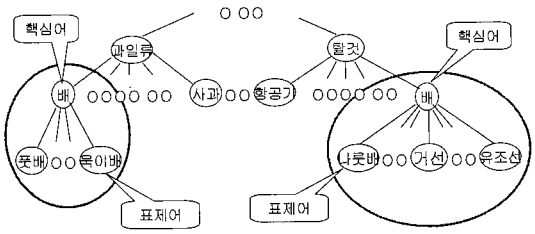

그림 1 동형이의어와 표제어간의 상-하의어 관계

조평옥(1997)은 표제어와 뜻풀이말에서의 핵심어간의 상-하의어 관계를 이용한 명사 의미계층구조를 구축하였다[11]. 이러한 계층구조의 상-하의어의 관계를 다음 하위 단계까지 확장한다면, 출현빈도가 적은 동형이의어의 의미정보 추출 시의 자료 부족 문제를 완화시킬 수 있다. 즉 <표 3>과 같이 표제어("나룻배")를 중심으로 다시 하위 단계로 확장함으로써 의미정보 구축 시의 자료 부족 문제를 해결할 수 있다. 그러나, 본 논문에서는 하위 단계로의 확장은 고려하지 않았다.

표 3 표제어를 중심으로 한 의미정보 확장

표제어	뜻풀이말("나룻배"를 포함.)
나루질	**나룻배**를 부리는 일.
나루터	**나룻배**를 부리는 곳.
나루턱	**나룻배**가 들어 닿는 곳.
:	:

본 논문에서 구축하는 의미정보는 동형이의어가 포함된 뜻풀이말의 표제어와, 해당 뜻풀이말에서 사용된 단어에서 체언(보통명사)과 용언(동사, 형용사)으로 구분된 단어와 각 단어의 출현 빈도로 구성한다. <표 4>는 의미정보의 구성 형태를 보이고 있는데 본 논문에서는 의미분별에 도움이 되지 않는 조사와 고유명사 등은 의미정보에 포함시키지 않았다. <표 5>는 실제 사전의 뜻풀이말에서 추출한 동형이의어 "배"의 의미정보 예를 보이고 있다.

표 4 의미정보의 구성 형태

[체언류](총빈도)
<< 품사 태그^{1} >>:단어^{11}(빈도^{11}):단어^{12}(빈도^{12}):단어^{13}(빈도^{13}):...

[용언류](총빈도)
<< 품사 태그^{2} >>:단어^{21}(빈도^{21}):단어^{22}(빈도^{22}):단어^{23}(빈도^{23}):...
<< 품사 태그^{3} >>:단어^{31}(빈도^{31}):단어^{32}(빈도^{32}):단어^{33}(빈도^{33}):...

표 5 의미정보의 예 (배:운송수단, 신체부위)

배 (운송수단)

[체언류](총 빈도수:2499)
<<NNG>>:강관(7):강(9):고기잡이(5):기관(5):**모양(4)**:**물(25)**:바람(9):승객(7):운항(6):정박(8):짐(24):키(6):항구(20):항로(7):...

[용언류](총 빈도수:858)
<<VV>>:가(16):건너(5):나르(20):나아가(7):내리(10):다니(15):만들(32):실(44):타(29):태우(6):...
<<VA>>:가볍(5):**넓(3)**:빠르(4):작(28):크(17):평평하(3):...

배 (신체부위)

[체언류](총 빈도수:1783)
<<NNG>>:가슴(34):꼬리(4):꽁지(8):날개(9):동물(6):등(68):머리(20):**모양(14)**:몸(32):**물(5)**:물고기(4):병(10):수술(6):음식(6):...

[용언류](총 빈도수:621)
<<VV>>:가르(7):곯(3): 띠(8):앓(2):차(4):...
<<VA>>:고프(8):**넓(4)**:뚱뚱하(3):많(7):부르(32):볼록하(5):아프(7):희(44):...

3. 동형이의어 중의성 해결 시스템

3.1 중복되는 의미정보에 대한 고려

동형이의어의 의미별 의미정보를 하나의 집합으로 본다면, <그림 2>와 같이 의미정보를 구성하는 공기 관계의 단어들은 의미정보 집합의 요소들이다.

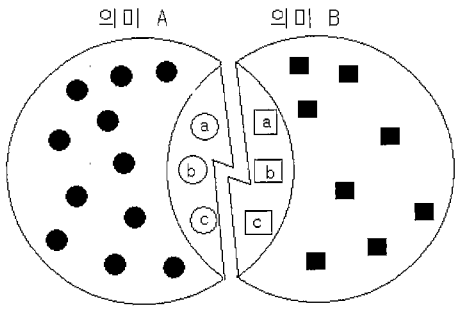

그림 2 의미 A와 의미 B를 가지는 동형이의어 W의 의미정보 집합 관계

<그림 2>에서 의미정보 집합들 사이에는 공기 관계에 있는 동일한 단어(A∩B)들이 존재할 수 있다. 여기서 교집합에 속하는 단어들의 개수가 전체 의미정보에서 차지하는 비중에 따라 의미분별의 정확성에 많은 영향을 미친다. 예를 들어, 의미 A와 의미 B의 교집합 A

∩B가 합집합 A∪B에 대해 많은 비중을 차지한다면, A의 의미가 B의 의미(혹은 B의 의미가 A의 의미)로 오분석될 가능성이 커진다. 그러나, 본 논문에서 구축하는 의미정보는 동형이의어의 의미에 따라 개별적으로 추출되므로, 의미별로 공기 관계에 있는 단어들의 출현 빈도가 다르다. 실제로 <표 5>에서 "모양", "물", "넓다"는 의미별로 동형이의어 "배"와 공기 관계에 있으나, 각 의미별 출현 빈도는 다름을 알 수 있다.

본 논문의 동형이의어 중의성 해결 모델에서는 교집합에 속하는 단어들의 의미분별력을 개별 의미정보 집합 내에서의 상대 빈도로 계산한다(3.2절 수식 (3), (4) 참조)

3.2 동형이의어 중의성 해결 모델

<그림 3>은 "돛 없는 작은 배"라는 문장에서 동형이의어 "배"의 의미를 분별하는 과정을 보이고 있다. 입력 문장이 형태소 분석 및 태깅 과정을 거쳐 어절별 품사가 결정되면, 입력 문장에서 공기 관계에 단어의 의미정보 값을 "배"의 의미별 의미정보에서 획득하고, 의미 중의성 알고리듬에 따라 의미공기 유사도를 구한 후 동형이의어의 의미를 결정한다.

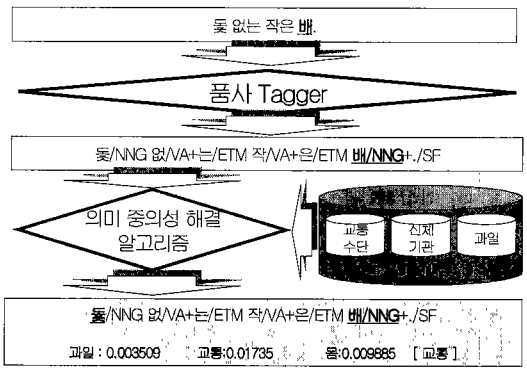

그림 3 동형이의어 중의성 해결 과정

동형이의어에 대한 의미분별은 입력 문장과 의미정보를 이용한 수식(1)에 의해서 이루어진다.

$$WSD(C, S_i) = \arg MAX_{S_i} Sim(C, S_i) \quad (1)$$

수식 (1)에서 $Sim(C, S_i)$는 문장 C 에서의 동형이의어와 공기 관계에 있는 단어(명사, 용언)들과 동형이의어의 의미별 의미정보 S_i와의 유사도(similarity)를 나타내며, 최대 유사도를 가지는 의미 S_i 를 선택하여 동형이의어의 중의성을 해결한다.

유사도 $Sim(C, S_i)$는 다음 수식 (2)에 의해서 구해진다.

$$Sim(C, S_i) = \alpha \times Noun(C, S_i) + \beta \times Pred(C, S_i) \quad (2)$$

$Noun(C, S_i)$는 문장 C에서 출현하는 명사와 의미정보 S_i 내에서의 명사와의 의미공기 유사도이고, $Pred(C, S_i)$는 문장 C에서 출현하는 용언과 의미정보 S_i 내에서의 용언과의 의미공기 유사도이다. α와 β는 명사와 용언이 동형이의어 의미분별 시에 영향을 주는 가중치(weight)로, 학습 코퍼스(의미분별된 뜻풀이말)를 이용한 의미분별 정확률에 의해 결정된다(3.3절 참조). 본 논문에서 제안하는 동형이의어 중의성 해결 모델에서 명사와 용언을 동시에 고려하는 이유는 동형이의어는 명사뿐만 아니라 용언(예, "감다", "고르다", "배다", "찌다" 등)에서도 발견되며, <표 1>과 <표 5>에서 알 수 있듯이 특정 동형이의어의 중의성을 해결하는 결정적인 의미공기 관계의 단어는 명사와 용언에서 동시에 발견되기 때문이다.

문장 C에서 출현하는 명사 및 용언과 의미정보 S_i내에서의 명사 및 용언과의 의미공기 유사도 $Noun(C, S_i)$과 $Pred(C, S_i)$는 수식 (3)과 수식 (4)에 의해서 구해진다.

$$Noun(C, S_i) = Match(C_n, S_i) \times \sum_j P(W_{nj} \mid S_i) \quad (3)$$

$$Pred(C, S_i) = Match(C_v, S_i) \times \sum_j P(W_{vj} \mid S_i) \quad (4)$$

여기서 $P(W_{nj} \mid S_i)$는 문장 C 에서 출현한 명사 W_n이 의미별 의미정보 S_i내에서의 의미공기 관계를 가지는 확률이다. 즉, $\frac{S_i\text{내에서 } W_n\text{의 출현빈도}}{S_i\text{내에서 체언류의 총출현빈도}}$ 이다.

또한, $Match(C_n, S_i)$는 문장 C에서 의미별 의미정보 S_i와 의미공기 관계를 가지는 명사의 개수이고, $Match(C_v, S_i)$는 용언의 개수이다. 이처럼 의미공기 관계를 가지는 단어의 수를 곱하는 이유는, 의미분별의 중요 인수로 $P(W_j \mid S_i)$ 값이 상대적으로 높은 단어(즉 의미분별에 결정적인 의미공기 관계에 있는 단어)뿐만 아니라, 비록 $P(W_j \mid S_i)$ 값이 낮더라도 문장 C에서 의미공기 관계에 있는 단어가 많이 출현한 경우 의미 S_i와 유사하기 때문에 이를 고려하기 위해서이다.

<표 6>은 "배"의 의미가 서로 다른 두 문장의 예에서 수식이 적용되는 결과를 보이고 있다. 문장[1]에서 "물/NNG", "만들/VV", "띄우/VV"는 분별의 대상이 되는 세 개의 의미정보 집합 모두에 포함된 교집합 요소들이다. 그러나, 각 의미정보 집합 내에서 출현빈도는 의미에 따라 다르다. 명사 "물"에 대해 적용된 수식 (3)의 결과와 두 개의 용언 "만들다"와 "띄우다"에 적용된 수식 (4)

의 결과에서 출현 단어에 대한 의미공기 유사도는 "운송수단"의 의미가 가장 높다. 결과적으로 "운송수단"의 의미로 분별이 된다. 반면 문장[2]에서는 명사 "물/NNG"에 대한 수식 (3)이 적용된 결과는 "운송수단"의 의미공기 유사도가 가장 높으나, 용언이 적용된 수식 (4)에서는 "신체부위"의 의미공기 유사도가 가장 높았다. 명사와 용언에 대한 의미공기 유사도 가중치가 적용된 수식 (2)의 결과 "신체부위"의 의미로 의미분별된다. 문장[2]의 예에서도 보는 바와 같이 동형이의어의 의미분별을 위해서는 명사뿐만 아니라 용언도 고려해야 함을 알 수 있다. 수식 (2)에 적용된 명사와 용언의 가중치 α와 β는 실험에 의해 결정된 0.9와 0.1를 적용하였다(3.3절 참조).

표 6 의미분별에 적용되는 수식 결과의 예
(배:운송수단, 신체부위)

의미		[1] 새로 만든 배를 처음 물에 띄움.	[2] 물만 먹고 부른 배.
과일	체언류	물/NNG(1)	물/NNG(1)
	용언류	만들/VV(1), 띄우/VV(1)	먹/VV(2)
운송수단	체언류	처음/NNG(2), 물/NNG(25)	물/NNG(25)
	용언류	만들/VV(32), 띄우/VV(5)	
신체부위	체언류	물/NNG(5)	물/NNG(5)
	용언류	만들/VV(2), 띄우/VV(1)	먹/VV(22),부르/VA(32)

예문	의미	수식(3)	수식(4)	수식(2)	수식(1)
문장[1]	과일	0.007	0.067	0.013	운송수단
	운송수단	**0.022**	**0.086**	**0.0284**	
	신체부위	0.003	0.01	0.0037	
문장[2]	과일	0.007	0.033	0.0096	신체부위
	운송수단	**0.01**	0	0.009	
	신체부위	0.003	**0.174**	**0.0201**	

3.3 명사와 용언의 가중치 α, β 결정

동형이의어의 의미분별에 적용될 명사와 용언의 가중치를 결정하기 위해 사용된 학습 코퍼스는 <표 7>의 9개의 동형이의어를 포함하고 있는 사전 뜻풀이말 5,246 문장(1차 유형의 뜻풀이말은 2,065 문장)이다. 총 어절수는 38,296개로 한 문장당 평균 7.3개의 어절로 구성되어 있다.

가중치 결정을 위한 실험은, 체언류와 용언류의 가중치를 0.1씩 변화시키면서 그 정확률의 변화를 관찰하였다. <표 8>에 의하면 의미정보에서 용언만을 이용한 의미분별은 정확률이 평균 75.54%인 반면, 명사만을 이용한 의미분별은 정확률이 89.44%였다. 그리고, 명사와 용언의 가중치를 0.9/0.1로 하였을 때 의미분별의 정확률이 96.11%로 가장 높았다. 이러한 결과는 다음의 이유에 의한 것으로 분석된다.

첫째, 용언의 의미정보 집합보다 체언의 의미정보 집합이 상대적으로 크다. 이는 용언의 개수가 명사에 비해 상대적으로 적기 때문에 자료 부족 현상이 체언보다는 용언의 의미정보에서 더욱 심하다고 할 수 있다.

둘째, 용언의 의미정보 집합간의 교집합이 체언 의미정보의 교집합에 비해 상대적으로 크다. 이는 대부분의 용언이 다의적으로 사용되어 많은 단어들과 의미공기 관계를 이루고 있기 때문에, 용언만을 이용한 의미분별에서 교집합을 이루는 다른 의미의 집합에 간섭을 많이 받고, 결과적으로 의미분별의 정확률이 낮다고 분석할 수 있다.

표 7 명사와 용언의 가중치를 결정하기 위한 학습 코퍼스

단어	의미	1차 유형 학습 데이타 수	전체 학습 데이타 수
기관[*1]	몸(器官)	114	212
	조직(機關)	224	453
	장치(機關)	26	112
기구	장치(機具)	336	472
	조직(機構)	48	78
눈	신체부위(目)	107	553
	식물(木)	10	31
	기상현상(雪)	44	181
다리	교각(橋脚)	49	81
	발(下肢)	33	326
병	그릇(甁)	26	48
	병사(兵)	2	3
	질병(病)	463	880
배	과일(梨)	3	24
	운송수단(船)	178	513
	신체부위(腹)	12	322
비	청소도구	18	29
	기상현상(雨)	82	288
	비석(碑)	12	21
	비율(比)	40	70
신	신발	46	89
	종교(神)	99	236
차	운송수단(車)	47	107
	음료(茶)	20	56
	차이(差)	26	61
합계		2,065	5,246

<표 8>에서 보는 바와 같이 동형이의어의 의미분별에서 정확률의 향상을 위해서는 체언과 용언의 의미정보를 모두 고려해야 함을 알 수 있다. 그러나, 적용되는 가중치에 따라서 동형이의어별 정확률에는 다소 차이가 있다.

표 8 명사와 용언의 가중치 결정을 위한 실험 결과

(단위:%)

α/β	기관	기구	눈	다리	병	배	비	신	차	평균
0/1	55.47	72.73	77.52	85.51	85.61	74.39	82.60	73.23	79.91	75.54
0.1/0.9	88.29	92.73	88.89	93.36	96.34	92.20	93.63	96.61	97.32	92.66
0.2/0.8	90.99	93.27	89.90	93.86	96.77	92.20	93.87	96.61	97.32	93.35
0.3/0.7	93.05	93.64	90.33	93.86	96.77	92.66	94.36	96.93	97.32	93.90
0.4/0.6	93.95	94.00	90.72	93.86	96.89	93.13	94.85	96.93	97.77	94.28
0.5/0.5	94.60	94.18	91.50	95.09	96.89	93.83	95.35	96.93	97.77	94.76
0.6/0.4	95.75	94.73	92.29	95.09	**96.99**	94.06	95.59	96.93	**98.21**	95.19
0.7/0.3	96.14	95.09	93.07	95.09	96.88	94.06	97.30	96.61	**98.21**	95.50
0.8/0.2	96.39	96.36	93.46	**95.58**	96.88	94.06	97.30	96.61	**98.21**	95.77
0.9/0.1	97.04	**97.64**	**93.73**	**95.58**	**96.99**	**94.29**	**97.31**	**97.23**	97.77	**96.11**
1/0	**97.69**	95.46	78.82	87.71	89.91	87.78	86.52	92.30	91.97	89.44

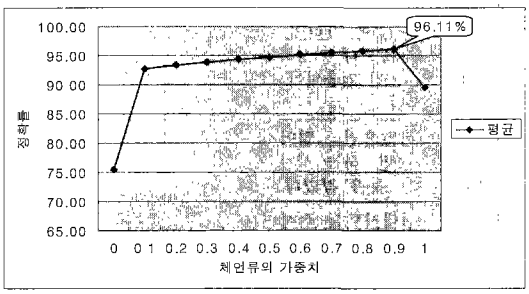

그림 4 가중치 결정 전체 실험의 평균 결과

3.4 가중치 결정 실험에서 의미분별 실패한 예와 원인

가중치 결정을 위해 사용된 학습 코퍼스는 사전의 뜻풀이말로서, 표제어의 의미를 기술하기 위한 가장 핵심적인 단어들만을 사용하고 있으며 정확한 의미 전달을 위해 대부분 단문 형태이다(평균 7.3 어절로 구성). 따라서, 뜻풀이말의 구문구조 분석 오류로 인한 의미분별 실패보다는 의미정보 집합을 구성하는 단어들 자체가 동형이의어임으로 인한 오분석과 빈도수 높은 소수의 단어에 의한 오분석이 의미분별 실패의 가장 큰 원인이었다.

<표 9>는 의미정보 집합을 구성하는 단어들 자체가 동형이의어임으로 인한 오분석 예를 보이고 있다.

표 9 의미정보 집합을 구성하는 단어들 자체가 동형이의어임으로 인한 오분석 예

오분석 문장	의미분별대상단어	오분석 결과	올바른 결과
배가 통과할 수 있도록 **다리**의 한 끝 또는 양쪽이 들리게 된 **다리**.	배	신체부위	운송수단
배가 자유로이 지나다닐 수 있도록 **다리**의 일부 또는 전체를 움직일 수 있게 만든 **다리**.	배	신체부위	운송수단
기관과 돛이 갖추어져 있는 작은 **배**.	기관	신체부위	장치

<표 9>의 "배가 통과할 수 있도록 **다리**의 한 끝 또는 양쪽이 들리게 된 **다리**"라는 예에서 "다리"는 "교각"과 "발"의 의미를 지닌 동형이의어이다. "배"의 의미정보 집합 중 "신체부위"의 의미 집합에서 "다리(발)"라는 단어가 의미공기 관계의 요소로 출현하고, "운송수단"의 의미 집합에서도 의미공기 관계의 요소로 출현한다. 그러나, 본 논문에서의 의미정보는 의미공기 관계에 있는 단어들과 그의 빈도정보만으로 구성되어 있을 뿐, 의미정보 집합의 요소들을 의미분별하지 않았다. 이는 의미분별하고자 하는 동형이의어의 의미정보 집합내에서의 의미공기 관계를 가지는 단어들이 또 다른 동형이의어일 경우 이를 분별하기 위해서는 두 개의 동형이의어간의 복합적인 의미공기 유사도를 계산해야 하는 어려움이 따른다. 이에 대해서는 추후 연구가 더 진행되어야 할 것이다. 위 <표 9>에서 오분석의 원인이 되는 중의성 단어들은 **밑줄**로 표시하였다. <표 10>은 빈도수가 높은 소수의 단어에 의한 오분석 예이다.

<표 10>의 "네모 반듯한 모양의 배"에서 "모양"이라는 단어가 "신체부위"의 의미정보 집합에서 높은 빈도

표 10 빈도수가 높은 소수의 단어에 의한 오분석 예

오류 발생 문장	의미분별 대상단어	오분석 결과	올바른 결과
네모 반듯한 **모양**의 배.	배	신체부위	운송수단
빛은 **갈색**으로 다른 큰 **물고기**나 배의 바닥에 붙어 삶.	배	신체부위	운송수단
바닥에 배를 붙이고 **움직여** **나아가다**.	배	운송수단	신체부위
말이나 소의 등에 **실은** **짐**을 배와 얼러 매는 줄.	배	운송수단	신체부위
인위적으로 **만들**어 놓은 신의 형상.	신	신발	종교
목이 말라서 **물**이 자꾸 먹히는 병.	병	그릇	질병

로 나타난다(<표 5> 참조). 이로 인해, "신체부위"로 오분석되었다. 그 외의 각 예에서, 오분석을 야기하는 고빈도 단어들은 **밑줄**로 표시하였다.

4. 실험 및 평가

본 논문에서는 비학습 코퍼스(국어 정보 베이스(ver 1.0)와 ETRI 품사 부착 코퍼스)에서 9개의 동형이의어를 포함하고 있는 문장을 추출하여 3장에서 제시한 동형이의어 중의성 해결 모델의 일반성 및 강건성을 측정하였다. 비학습 코퍼스를 대상으로 한 실험에서 명사와 용언의 가중치는 0.9/0.1로 고정하여 실험하였다. 3장과 4장의 실험에서 의미정보 추출을 위해서는 UNIX 환경에서 perl을 이용하였고, 의미 중의성 해결 실험은 Window NT 환경에서 Visual C++를 이용하여 실험하였다.

4.1 비학습 코퍼스 데이타 및 정확률

실험에 사용된 데이타는 총 1,796 문장에 38,266 어절로, 한 문장당 평균 어절 수는 21.3개이다. 동형이의어의 의미별 데이타 수와 실험 결과의 정확률은 <표 11>과 같다.

표 11 비학습 코퍼스의 동형이의 의미별 문장 수와 실험 정확률

단어	의미	실험 데이타 수	실험 정확률(%)
기관	몸(器官)	17	88.2
	조직(機關)	185	92.4
	장치(機關)	2	100.0
기구	장치(機具)	24	75.0
	조직(機構)	98	89.8
눈	신체부위(目)	431	79.8
	식물(木)	1	100.0
	기상현상(雪)	79	81.0
다리	교각(橋脚)	21	71.4
	발(下肢)	58	84.5
병	그릇(甁)	12	16.7
	병사(兵)	0	0
	질병(病)	151	86.8
배	과일(梨)	6	33.3
	운송수단(船)	92	75.0
	신체부위(腹)	50	62.0
비	청소도구	1	0
	기상현상(雨)	86	80.2
	비석(碑)	10	30.0
	비율(比)	1	0
신	신발	2	100
	종교(神)	372	86.3
차	운송수단(車)	46	50.0
	음료(茶)	39	48.7
	차이(差)	12	83.3
총문장 수		1,796	
평균정확률			80.7

4.2 실험의 결과 및 평가

비학습 코퍼서의 문장을 대상으로 한 실험에서는 주로 구문구조를 고려하지 않음으로써 발생한 오분석이 대부분이었다. <표 12>는 구문구조를 고려하지 않아서 발생하는 오분석의 예를 보이고 있으며, 오분석에 영향을 주는 부분을 **밑줄**로 표시하였다.

표 12 구문구조를 고려하지 않아 발생한 오분석의 예

오류 발생 문장	의미분별 대상단어	오분석 결과	올바른 결과
아침에 일어나면 라이온 치약을 쓰고, **일제 커피포트에 물 끓이어 커피를 마신 뒤** 세이코 시계를 차고 ***도요타 차를 타고 출근한다***	차	음료	운송수단
도모데도바 공항에서 집으로 돌아오는 길에 위치한 제르진스키 광장과 크레믈린 입구, 고리끼 공원에는 **비가 내리는 중**에도 사람들로 붐비고 있었으며 아르바뜨 거리와 고리끼 거리 연변에는 여기저기 탱크가 ***눈에 띄었다.***	눈	기상현상	신체부위
여름철, **나무들의 잎이 무성할 때는 푸른 잎에 가리어** ***눈에 잘 띄지 않으나*** 기생하는 나무가 낙엽이 진 겨울철이 되면 잘 보입니다.	눈	식물	신체부위
떨어져 내리었으면, 떨어져 내릴 수 있다면, 하는 생각을 하면서 현주는 ***눈을 감았다.***	눈	기상현상	신체부위
나는 밤새 침대시트를 적시며 고열로 신음하다가 이튿날 아침 한 조각의 빵과 함께 아직 남아 있는 ***병 속의 꼬냑을*** 단숨에 들이켰다.	병	질병	그릇

<표 12>의 첫 번째 예문에서 "차"는 "운송수단"으로 의미분별 되어야 하나, "일제 커피포트에 물을 끓이어 커피를 마신 뒤"라는 구문이 "음료"의 의미정보 집합에 해당하는 의미공기 단어를 많이 내포하고 있어 의미분별에 실패하였다. 이는 현재의 동형이의어 중의성 해결 모델이 단순히 전체 문장에서 명사와 용언을 추출하여 의미정보 집합 내의 의미공기 관계에 있는 단어의 통계적인 정보를 이용함으로써 발생되는 현상이다.

따라서, 동형이의어와 직접적인 의미공기 관계에 있는 구문만을 대상으로 의미분별 할 수 있도록 개선할 필요가 있다. 비학습 코퍼스의 문장당 평균 어절 수는 21.3개로, 3장의 가중치를 결정하기 위한 뜻풀이말(문장당

평균 7.3 어절)과는 달리 중문 및 복문의 형태이다. 따라서 완전한 형태의 구문구조 분석 없이도 동형이의어를 중심으로 인접한 좌우 7개의 어절(뜻풀이말에서의 평균 어절수)에서 명사와 용언을 추출하여 의미분별하는 방법들이 추후 연구되어야 할 것이다.

5. 결론 및 향후 연구

본 논문에서는 동형이의어의 중의성 해결을 위한 의미정보를 동형이의어를 포함하고 있는 사전의 뜻풀이말에서 추출하고, 추출된 의미정보에서 명사와 동사를 모두 고려한 의미 중의성 해결 방법을 제안하였다.

5,246문장의 사전 뜻풀이말을 학습 코퍼스로 하여 의미정보를 구축하였다. 정확한 의미정보를 구축하기 위해서 뜻풀이말을 두 가지 유형으로 구분하였다. 첫 번째는 뜻풀이말의 해당 표제어와 동형이의어가 상-하의어의 관계를 가지는 유형이고, 두 번째는 동형이의어가 뜻풀이말의 중간에 나타나는 유형이다. 구축된 의미정보는 뜻풀이말의 표제어를 포함한 뜻풀이말의 보통명사와 용언으로 구성되어 있으며, 명사와 용언의 출현 빈도 정보를 가지고 있다.

본 논문에서 제안한 동형이의어 중의성 해결 시스템은 의미정보 내의 명사와 용언을 동시에 고려하여 명사만을 이용한 기존 연구의 자료 부족 문제를 크게 완화시켰으며, 비교적 작은 코퍼스인 사전만을 이용하여 제한된 의미 계층 구조를 유추하고 이용함으로써 대용량의 의미 계층 구조가 없는 경우에 적합한 모델이다. 본 논문에서는 명사와 용언이 의미분별에 기여하는 가중치를 결정하기 위해 의미정보 추출에 사용된 학습 코퍼스를 이용하여 실험하였다. 실험 결과 명사와 용언의 가중치가 0.9/0.1일 때 실험 대상의 9개의 동형이의어의 평균 의미분별 정확률이 96.11%로 가장 높았다.

또한 본 논문에서 제안하는 동형이의어 중의성 해결 시스템의 일반성 및 강건성을 측정하기 위하여 국어정보베이스 I과 ETRI 코퍼스에서 추출한 1,796 문장의 비학습 코퍼스를 대상으로 한 실험에서 평균 80.73%의 정확률을 보였다.

실험 과정에서 발생한 여러 유형의 오류를 해결하여 의미분별 정확률을 향상시키기 위해서는 앞으로 다음의 연구들이 더 진행되어야 할 것이다.

첫째, 의미정보 구축 시의 자료부족 문제를 해결하기 위해 첫 번째 유형의 뜻풀이말을 다음 단계로 확장하는 방법의 연구가 필요하다. 둘째, 명사와 용언의 가중치를 결정하는 실험에서 발생한 두 개 이상의 동형이의어로 인한 복합적인 의미공기 유사도 측정 방법 및 소수의 고빈도 단어에 의한 오분석 방지 방법들이 연구되어야 한다. 셋째, 비학습 코퍼스를 대상으로 한 실험에서 구문구조를 고려하지 않아 발생한 오류 해결 방법들이 연구되어야 할 것이다. 마지막으로 사전에서의 전체 동형이의어를 대상으로 의미정보를 구축하고 이를 정보검색이나 기계번역시스템에서 이용하는 방법들도 연구되어야 할 것이다.

참 고 문 헌

[1] 김영택, "자연언어처리", 교학사, 1994.

[2] 박성배, 장병탁, 김영택, "의미 부착이 없는 데이타로부터의 학습을 통한 의미 중의성 해소", 한국 정보과학회 '2000 봄 학술 발표 논문집 B', 제 27 권 1호, pp.330 - 332, 2000.

[3] 박영자, "사전을 이용한 단어 의미 자동 클러스터링 : 유전자 알고리즘 접근법", Ph.D. these, 연세대학교, 1998.

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

[5] 송도규, "인지언어학과 자연언어 자동처리", 홍릉과학출판사, 1997

[6] 송영빈, 최기선, "동사의 애매성 해소를 위한 시소러스의 이용과 한계", 제 12 회 한글 및 한국어 정보처리 학술대회 발표논문, pp.255 - 261, 2000.

[7] 조평옥, 옥철영, "의미속성에 기반한 한국어 명사 의미체계", 정보과학회논문지(B), 26권, 4호, pp.584 -594, 1999.

[8] 이창기, 이근배, "의미 애매성 해소를 이용한 WordNet 자동 매핑", 제 12 회 한글 및 한국어 정보처리 학술대회 발표논문, pp.262 - 168, 2000.

[9] 정보-전자 연구회 편, "자연언어처리입문", 대광서림, 1993.

[10] 조정미, "코퍼스와 사전을 이용한 동사 의미분별", Ph.D. these, 한국과학기술원, 1998.

[11] 조평옥, 안미정, 옥철영, 이수동, "사전 뜻풀이말에서 구축한 한국어 명사 의미계층구조", 한국인지과학회 논문지, 10권 4호, pp.1 - 10, 1999.

[12] Alpha k, Luk, "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions," 33rd Annual Meeting of the ACL, pp.181-188, 1995.

[13] David Yarowsky, "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," Preceedings of COLING 92, pp.454-460, 1992.

[14] James Allen, "Natural Language Understanding," The Benjamin / Cummings Publishing Company,

Inc. 1994
[15] Markoto Nagao저, 황도삼,최기선,김태석 공역, "자연언어처리", 홍릉과학출판사, 1998.
[16] Markoto Nagao저, 황도삼,최기선,김태석 공역, "자연언어이해", 홍릉과학출판사, 1999.
[17] Nancy Ide and Jean Veronis, "Introduction to the Special Issue on Word Sense Disambiguation :The State of the Art," Computational Linguistics, Vol 24, No. 1, pp1 - 40, 1998.
[18] Rebecca Bruce and Janyce Wiebe, "Word-Sense Disambiguation Using Decomposable Models," 32rd Annual Meeting of the ACL, pp.139 - 145, 1994.

허 정
1999년 울산대학교 컴퓨터정보통신공학부 학사. 2001년 울산대학교 컴퓨터정보통신공학부 석사. 2001년 ~ 현재 ETRI 언어공학연구부 연구원. 관심분야는 한국어정보처리, 의미기반 정보검색, 의미분석

옥 철 영
정보학회논문지 : 소프트웨어 및 응용
제 28 권 제 5 호 참조