가중치를 이용한 통계 기반 한국어 동형이의어 분별 모델

(A Korean Homonym Disambiguation Model Based on Statistics Using Weights)

김 준 수 † 최 호 섭 † 옥 철 영 ††
(Jun-Su Kim) (Ho-Seop Choe) (Cheol-Young Ock)

요 약 본 논문은 한국어 정보처리에서 발생하는 어휘 중의성 문제 중 한국어에서 그 심각성이 큰 동형이의어 중의성을 해결하기 위하여, 사전 뜻풀이 말뭉치에서 구축한 의미정보(Semantic Information)와 이를 이용한 기존의 통계기반 동형이의어 분별 모델에 대한 실험 결과를 분석하여, 정확률 향상을 위한 새로운 동형이의어 NPH(New Prior Probability of Homonym sense) 가중치 및 인접 어절에 대한 거리 가중치 적용 모델을 제안한다.

사전 뜻풀이 말뭉치의 상위 고빈도 동형이의어 200개 중 중의성이 높은 46개(명사 30개, 동사 16개)를 선별하고, 21세기 세종 계획에서 제공하는 350만 어절 품사 부착 말뭉치에서 이들 동형이의어를 포함하는 47,977개의 문장을 추출하여 실험을 하였다. 기존의 통계기반 동형이의어 분별 모델에서는 72.08%(명사 78.12%, 동사 62.45%)의 정확률을 나타냈으나, NPH 가중치를 부여한 실험 결과 정확률이 평균 1.70% 향상되었으며, NPH와 거리 가중치를 함께 이용한 결과 평균 2.01% 정확률이 향상되었다.

키워드 : 의미중의성해결, 의미분별, 의미정보, 동형이의어

Abstract WSD(word sense disambiguation) is one of the most difficult problems in Korean information processing. The Bayesian model that used semantic information, extracted from definition corpus(1 million POS-tagged eojeol, Korean dictionary definitions), resulted in accuracy of 72.08% (nouns 78.12%, verbs 62.45%). This paper proposes the statistical WSD model using NPH(New Prior Probability of Homonym sense) and distance weights. We select 46 homonyms(30 nouns, 16 verbs) occurred high frequency in definition corpus, and then we experiment the model on 47,977 contexts from '21C Sejong Corpus'(3.5 million POS-tagged eojeol). The WSD model using NPH improves on accuracy to average 1.70% and the one using NPH and distance weights improves to 2.01%.

Key words : Word Sense Disambiguation, WSD, Semantic Information, Homograph

1. 서 론

자연언어처리(NLP)는 형태소 분석, 구문 분석, 의미 분석, 담화 분석 등의 과정을 통해서 자연언어를 이해하고 이에 적합한 반응을 생성하여 "컴퓨터가 자연언어로 인간과 의사 소통이 가능하게 하는 효율성 있는 기술 개발"을 목표로 한다. 이러한 NLP 기술들을 바탕으로 기계 번역, 정보 검색, 자연언어 인터페이스, 문서 교정 및 퇴고, 문서 분류, 자동 요약, 음성 인식, 음성 대화, 언어 학습 등과 같은 다양한 응용 분야에 활용하게 되는 것이다[1]. 그러나 NLP 분석 과정에서 발생하는 여러 가지의 중의성(ambiguity) 문제는 NLP 관련 기술 연구뿐만 아니라 응용 분야에 이르기까지 두루 영향을 미쳐 분석의 정확성과 응용 시스템의 효율성을 저하시키는 중요한 원인이 되고 있다. 따라서, 자연언어처리의 중점 연구 과제는 각 분석 단계에서 발생하는 중의성을 최소화하는 것이다.

자연언어처리의 의미 분석 단계에서 발생하는 중의성 문제는 한 형태가 여러 가지의 의미를 가지는 동형이의어(homograph, 異義語)나 한 단어가 두 가지 이상의 의미를 가지는 다의어(polysemous, 多義語)와 같은 어휘

· 이 논문은 정보통신부 정보통신연구진흥원에서 지원하고 있는 2001년 정보통신기초기술연구지원사업의 지원에 의하여 연구되었음
† 정 회 원 : 울산대학교 컴퓨터정보통신학과
kimjunsu@mail.ulsan.ac.kr
hoseop@mail.ulsan.ac.kr
†† 종신회원 : 울산대학교 컴퓨터정보통신공학부
okcy@mail.ulsan.ac.kr
논문접수 : 2002년 6월 11일
심사완료 : 2003년 7월 30일

적 중의성에서 시작된다. 한국어는 많은 단어들이 한자어에서 유래되어 동형이의어가 많이 발생하게 되어, 한국어 단어 중의성 해소(word sense disambiguation: WSD) 문제는 다의어 수준의 중의성 해결에 앞서 동형이의어에 대한 정확한 분별이 선행되어야 한다.

본 논문의 목적은 한국어 동형이의어 분별에 필요한 대량의 언어 자원 확보 및 선행 동형이의어 분별 모델의 성능 개선에 있다. 동형이의어 분별에 필요한 언어 자원의 확보를 위해 본 논문에서는 [2]에서 제안한 방법을 바탕으로 15만 여 개의 표제어로 구성된 국어사전 뜻풀이를 말뭉치[1]로 활용하여 품사(part of speech: POS) 태그 및 의미(sense) 태그를 부착한 뜻풀이 말뭉치를 구축하고, 동형이의어와 함께 사용된 체언(일반명사)과 용언(동사, 형용사)의 공기빈도(co-occurrence frequency)를 동형이의어 분별에 필요한 의미정보로 구축한다. 또한 [3]에서 제안한 통계적 동형이의어 분별 모델의 분석 결과를 토대로 NPH(New Prior Probability of Homonym sense) 가중치 및 인접 어절에 대한 거리 가중치를 적용한 동형이의어 WSD 모델을 제안한다.

2. 관련 연구

본 장에서는 단어 중의성 해결에 일반적으로 활용되는 언어 자원들과 그 자원을 이용한 다양한 WSD 방법론들에 대하여 살펴보고, 본 연구에 이용되는 언어 자원의 확보 방안 및 동형이의어 분별 모델에 관한 선행 연구들에 대한 이해 및 문제점을 알아본다.

컴퓨터를 이용한 대량의 자료 저장, 검색 및 분석이 가능해지면서 WSD를 위한 지식을 자동으로 획득하려는 연구가 활발히 진행되어 기계 가독형 사전(machine readable dictionary: MRD)과 말뭉치(corpus)가 적극 활용되고 있다. 사전에 바탕을 둔 지식 획득은 현재 언어사용 실태를 잘 반영할 수 없다는 단점을 가지고 있지만 대량의 어휘에 대한 형태소 정보, 품사 정보, 구문 정보, 의미 정보 등 다양한 지식을 제공하고, 또한 사용빈도가 높지 않은 단어들에 대한 정보도 제공하는 장점으로 WSD를 위한 어휘 지식 획득의 자원으로 빈번하게 사용된다[4-7]. 말뭉치에 의한 지식 획득은 현재 사용되는 언어 현상들을 잘 반영하는 장점과 컴퓨터 관련 기술 및 인터넷의 발전으로 대량의 말뭉치를 손쉽게 구축 가능하다는 장점으로 1990년대 이후 말뭉치에 기반한 통계적 WSD 모델들이 급격히 증가하였다[8-10]. 그러나 말뭉치의 가공(품사, 구문, 의미 분석 작업)에 많은 노력과 시간이 소요되고, 자주 사용되지 않는 단어에 대한 자료 부족 문제(data sparseness problem)가 발생하는 단점이 있다.

다양한 언어 자원의 활용과 이를 이용한 단어 중의성 해결 모델들이 제시되고 있으며, 대표적인 WSD 방법론들은 다음과 같다. 첫째는 통합된 규칙 대 규칙에 의한 접근 방식이며, 둘째는 독립적 접근 방식으로서, WSD를 복합적인 의미 분석과는 구분하여 독립적으로 수행하는 방법이다. 전자에 속하는 대표적인 WSD 방법이 선택 제약에 기반한 WSD 방법이며[11,12], 후자에 속하는 것은 통계적 분석에 기반 WSD 분석 방법이다. 통계적 분석에 기반한 WSD는 다시 학습시킬 데이타의 형태에 따라 세 가지 방법으로 구분할 수 있다. 첫째, 정보 이론에 기반한 중의성 해결이나[13] 베이지안 분류(Bayesian classification) 등과 같은 레이블 처리된 학습 집합에 기반한 지도 학습 중의성 해결(supervised disambiguation) 방법[2,3,14,15]. 둘째, 아무런 가공이 되지 않은 원시 말뭉치만을 학습시키는 비지도 학습 중의성 해결(unsupervised disambiguation) 방법[16], 마지막으로 사전이나 시소러스와 같은 자원에 기반한 사전 기반 중의성 해결(dictionary-based disambiguation) 방법[17,18]으로 나눌 수 있다. 최근 국내에서도 이러한 다양한 방법을 이용하여 단어 중의성을 해결하고자 하고 있으나, 소량의 동형이의어를 대상으로 평가하는 데 그치고 있는 실정이다. 다음은 본 논문과 관련한 선행 연구들에 대하여 알아보자.

[2]는 사전의 뜻풀이에서 추출한 공기정보를 바탕으로 통계적 분석 방법을 도입한 동형이의어 중의성 해결 시스템을 제안하고, 9개의 동형이의어 명사(기관, 기구, 눈, 다리, 병, 배, 비, 신, 차)를 대상으로 실험하였다[2]. [2]에서 제안하고있는 동형이의어 분별용 의미정보는 정제된 언어 표현으로 이루어진 사전 뜻풀이를 적극 활용한 것으로 의미 결정성이 높은 공기 단어의 추출 및 의미계층 구조가 완벽하게 구축되지 않은 상황에서 [19]에서 분석된 뜻풀이의 패턴을 중심으로 계층적 구조를 유추할 수 있는 정보의 획득 방안을 제시하였다. 또한 [20]

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 자원이며 자체적으로 훌륭한 말뭉치라 할 수 있다.

2) 9개 동형이의어를 포함하는 사전 뜻풀이 5,246 문장을 학습 말뭉치로 활용하여 대상 동형이의어와 함께 쓰인 체언(일반명사) 및 용언(동사, 형용사)을 의미정보로 추출하였다. 이 의미정보를 바탕으로 통계적 WSD 모델을 제안하고, 학습 말뭉치를 대상으로 체언과 용언의 가중치를 가변적으로 실험한 결과 체언 대 용언(0.9 vs 0.1)의 가중치에서 가장 높은 평균 96.11%의 정확률을 나타내었으며, 학습에 이용되지 않은 실험 말뭉치(국어 정보 베이스 ver1.0과 ETRI의 품사 부착 말뭉치에서 추출한 1,796문장)를 대상으로 실험한 결과 평균 80.73%의 정확률을 보였다.

의 연구에서 발생하는 자료 부족 현상을 극복하기 위해 체언과 용언의 정보를 함께 이용하는 방안을 제시하기도 하였다. 하지만 의미 분별된 뜻풀이 말뭉치의 부재로 9개 실험 대상 동형이의어에 대한 의미정보만을 구축하는 한계를 보이고 있다.

[3]은 [2]에서 제안한 동형이의어 분별 모델의 문제점을 보완하고자 WSD 분야의 대표적인 통계적 모델인 베이지안 분류에 기초한 통계적 분석 모델을 제안하였고, 의미정보의 확장 방안으로 부사류에 대한 정보를 추가하여 실험하였다[3]. 그러나 [3]에서 제안한 통계적 분별 모델은 다음과 같은 문제점을 가지게 된다. 첫째, 밀도[4]에 기반한 확률 계산에서 사용 빈도가 낮은 의미(의미정보의 밀도가 상대적으로 매우 낮은 의미)의 공기단어들이 상대적으로 높은 확률값을 가지게 되어 과분석되는 문제가 종종 발생하게 된다. 둘째, 문장의 구문구조를 고려하지 않고 단순히 동형이의어와 공기하는 의미정보에 의해 분석을 시도하게 되어 문장의 구조가 단순한 단문의 분석에는 효과적이지만 복문이나 중문과 같이 문장의 구조가 복잡한 경우 불필요한 의미정보에 의한 오분석 가능성이 높아지게 되는 문제점을 가지고 있다.

3. 의미정보 구축 및 기존 동형이의어 분별 모델

3.1 동형이의어 중의성 해결을 위한 의미정보 구축

동형이의어 중의성은 동형이의어가 사용된 문맥에서의 다른 단어와의 의미적 관계에 의해서 그 의미가 결정된다[10]. 그리고, 통계에 기반한 동형이의어 중의성 해결 시스템의 정확률은 획득한 의미정보의 정확도와 충실도에 큰 영향을 받게 된다. 의미정보 추출에 사용되는 온라인 자원은, 크게 사전과 말뭉치로 나눌 수 있는데, 현재 국내에는 한국어 동형이의어에 대한 충분한 규모의 의미 태그 부착 말뭉치[2,10]가 없으므로, [2]의 의미정보 구축 방안(9개 동형이의어를 포함하는 사전 뜻풀이 5,246문장을 이용하여 의미정보 추출)을 바탕으로 15만 여 개의 표제어로 구성된 중·소규모의 국어사전 뜻풀이 전체를 말뭉치로 이용하여 품사 태그 및 의미(sense) 태그를 부착한 뜻풀이 말뭉치를 구축하고, 동형이의어와 함께 사용된 체언(일반명사)과 용언(동사, 형용사)의 공기빈도를 동형이의어 분별에 필요한 의미정보로 추출한다. 다음 그림 1은 의미 태그 부착 말뭉치의 구축 및 의미정보를 구축하는 과정을 보여주고 있으며, 각 과정에 대한 설명은 다음과 같다.

첫째, 품사 태그 부착 뜻풀이 말뭉치 구축

품사 태그 부착 뜻풀이 말뭉치를 구축하기 위한 일반적인 선행 작업으로 품사 태거(part-of-speech tagger)를 이용한 작업을 수행한다. 본 논문에서는 문화관광부의 [21세기 세종 계획] 프로젝트에서 제공하는 지능형 형태소 분석기 내의 품사 태거를 이용하여 국어사전의 뜻풀이와 용례에 품사 태그를 부착하였다.

품사 태거를 이용한 자동 품사 태그 부착 시 발생되는 오류를 확인·수정함으로써 고품질의 말뭉치를 구축할 수 있다. 이를 위해 자동 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 논문에서는 명사와 용언 중심의 의미정보 추출이므로 명사(일반명사, 고유명사)와 용언(동사, 형용사)을 중심으로 품사 태그를 확인·수정하였다.

둘째, 사전 기반 의미 태그 부착 뜻풀이 말뭉치 구축

의미 태그 부착 뜻풀이 말뭉치란 품사 태그 부착 말뭉치에 의미 태그를 부착한 말뭉치를 말한다. 본 논문에 필요한 기초 자료 확보를 위해서 반드시 사람이 직접 의미 태그를 부착하여 구축한 고품질의 의미 태그 부착 말뭉치가 필요하다. 본 논문에서는 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 구축하였다.

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.

셋째, 의미정보 추출 및 DB 구축

국어사전에서 추출 가능한 의미정보는 표제어의 뜻풀이/개념, 동의어, 유의어, 반의어, 전문용어, 특수어, 관련어, 방언, 어원, 관용구(idiom) 등 다양하다[19]. 본 논문에서는 동형이의어 중의성 해결에 필요한 의미정보를 추출하기 위하여 뜻풀이에 쓰인 단어들을 이용하고자 하였다. 그 이유는 뜻풀이가 앞에서 서술한 의미정보를

3) [2]의 실험 결과와 비교하기 위해 동일한 9개의 동형이의어 명사와 실험 말뭉치를 채택하였고, 용언류 동형이의어 분별 정확률을 측정하기 위해 7개의 동형이의어 용언(끼다, 말다, 붓다, 지다, 지르다, 바르다, 세다)을 추가하여 실험하였다. [2]의 실험 말뭉치를 대상으로 한 실험 결과 평균 84.42%의 정확률로 [2]의 결과보다 3.69%의 정확률이 향상되었으며, 7개의 동형이의어 용언은 실험 말뭉치(21세기 세종 계획에서 제공하는 350만 어절 품사 부착 말뭉치에서 추출한 3,114문장)에서 평균 70.81%의 정확률을 보였다.

4) 본 논문에서 밀도란 동형이의어의 의미별(사전에서 어깨번호로 구분되는 개별어휘)로 구별된 의미정보 집합에 포함된 공기단어의 비율을 의미한다. 또한, 밀도에 의한 정규화란 확률 계산을 위해 단순 공기 빈도를 바로 이용하는 것이 아니라 밀도에 의해 정규화된 값을 확률 계산에 이용하는 방법을 의미한다. [21](이호, pp. 16-19, 1999) 참조.

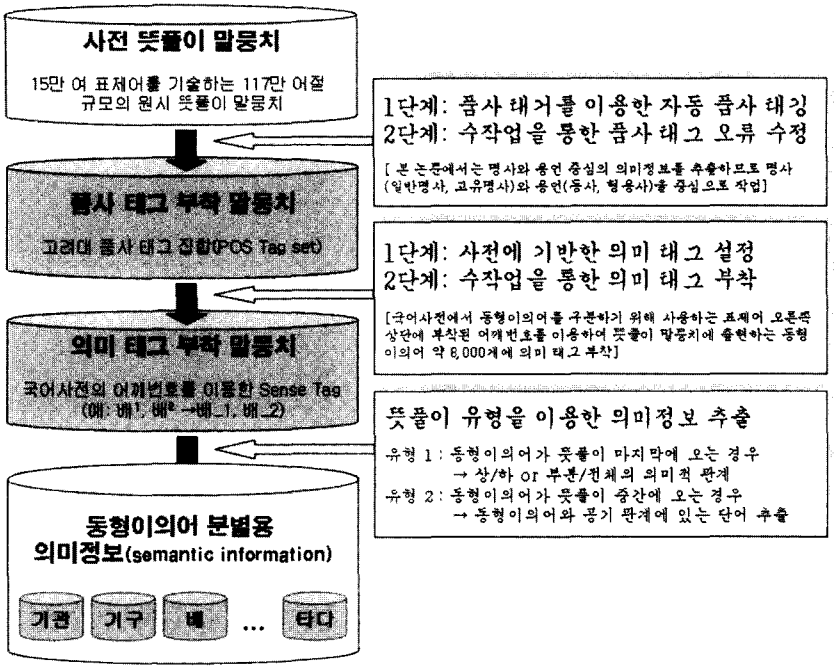
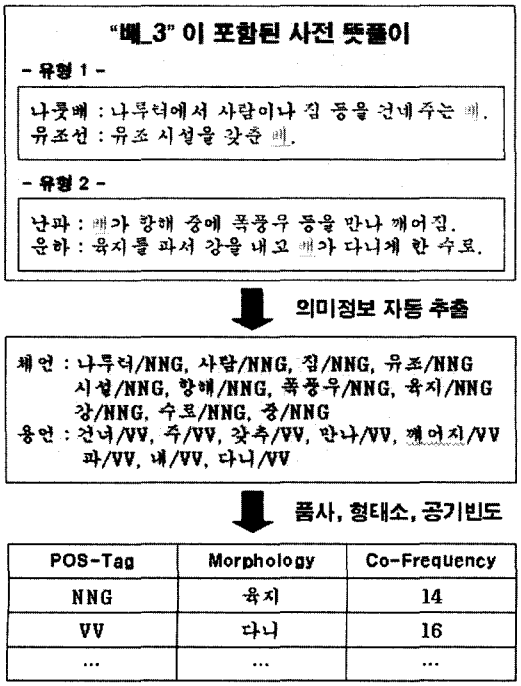


그림 1 의미 태그 부착 말뭉치 및 의미정보 DB 구축 개요

상당수 담고 있기 때문이다. 그렇지만 뜻풀이에 있는 모든 단어를 의미정보로 설정할 수는 없다. 그래서 본 논문에서는 [2]와 [3]의 의미정보 수집 방안을 기반으로 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 성질을 가지는 조사, 어미 등과 같은 형식형태소나 문법형태소는 한 단어의 의미를 분별하는 의미적 역할이 부족함으로 의미정보로서의 자격을 부여하기 어렵기 때문이다. 그러므로 본 논문에서는 동형이의어 중의성 해결에 핵심적인 역할을 하는 명사, 동사, 형용사 중심의 의미정보 수집 및 추출에 역점을 두었다. 그림 2는 뜻풀이 속에 포함된 동형이의어 "배_3(운송수단)"의 의미정보 추출 과정을 보여주는 예이다.



POS-Tag	Morphology	Co-Frequency
NNG	육지	14
VV	다니	16
...	...	...

그림 2 "배_3(운송수단)"의 의미정보 추출 예

3.2 기존 동형이의어 분별 모델과 문제점

3.2.1 기존 동형이의어 분별 모델

[3]에서 제안하고 있는 동형이의어 분별 모델은 WSD 분야의 대표적인 통계적 모델인 베이지안 분류 를 기초로 동형이의어 분별 의미정보를 효율적으로 이용하기 위해 도입된 모델이다. Gale과 Yarowsky는 베이지안 접근(Bayesian approach) 방식을 이용하여 주어진 문장(Context: C)에서 중의성을 가진 단어 $(s_1, s_2, \ldots, s_n)$에 대하여 각각의 의미별 확률 $P(s_i|C)$을 계산하고 이 중 가장 확률이 가장 높은 의미를 선택하는 방법을 이용한다. [3]에서 제안하는 WSD 모델은 베이지안 분류 모델

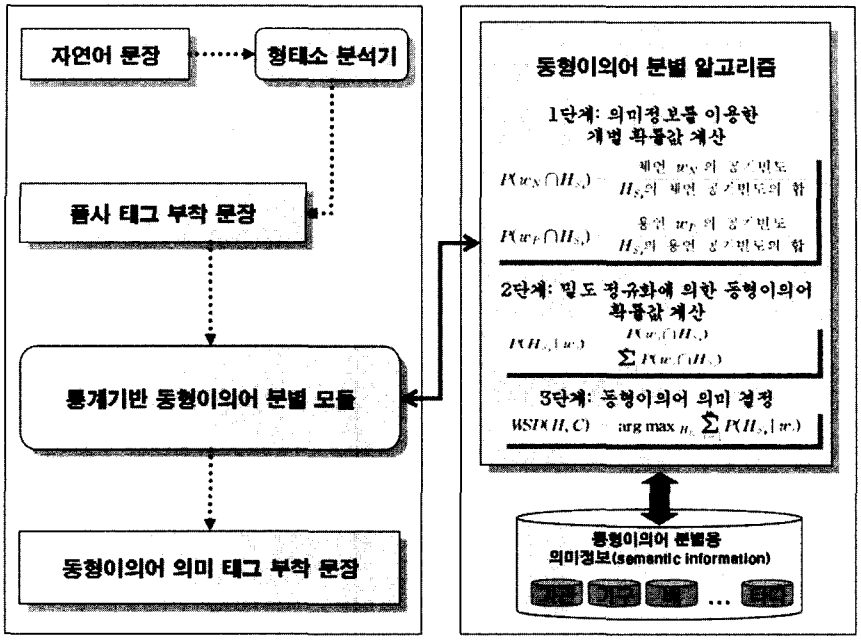


그림 3 동형이의어 분별 과정 및 알고리즘 요약

과 두 가지의 큰 차이점을 보이고 있다. 첫째, 일반적인 베이지안 분류 모델에서는 중의성 단어와 공기하는 단어들에 대한 총 공기 빈도(the total number of co-occurrences)를 이용한 확률 계산법으로써 빈도에 의한 정규화라고 볼 수 있다. 하지만 불균등한 학습 자료를 이용할 때는 밀도에 기반한 확률계산법[21]을 적용하여야 한다. 따라서 [3]에서는 다음의 수식 (2), 수식 (3) 그리고 수식 (4)와 같이 밀도를 고려하여 학습 자료의 개수가 동형이의어 의미에 따라 불균등하더라도 일관된 결과를 얻게 된다. 둘째, 베이지안 분류 모델에서는 $P(C|S_i)$를 독립 가정을 이용하여 $P(C|s_i) = \prod_{w \in C} P(w|s_i)$로 계산하며 분석의 편리 및 분별력 향상을 위하여 로그(log)를 일반적으로 적용하고 있다. 이 경우 공기 빈도에 따라 확률값의 편차가 매우 커지게 되어 소수의 단어에 의한 오분석 가능성이 높아지게 된다. 그러므로 [3]에서는 이러한 문제를 줄이기 위하여 로그의 이용 없이 확률값을 합하는 방법을 이용하고 있다.

동형이의어가 포함된 임의의 문장 C에서 나타나는 동형이의어 H의 의미(즉, 사전에서 어깨번호로 구분되는 개별어휘)는 $H_{s_1}, H_{s_2}, \cdots, H_{s_n}$로 표현할 수 있다. 다음 수식들에 의하여 $H_{s_1}, H_{s_2}, \cdots, H_{s_n}$ 중 하나로 분별하게 되는 것이다. H_{S_k}는 동형이의어 H의 k-번째 의미를 나타내며, $w_j (\in C)$는 H_{S_k}의 의미정보 집합에 속하며 문장 C에서 동형이의어 H와 공기하여 나타나는 단어를 나타낸다. 또한, w_j는 $H_{s_1}, H_{s_2}, \cdots, H_{s_n}$ 각각의 의미정보 집합들에 중복되어 나타나기도 한다. 수식 (1)은 수식 (2)를 통해 얻은 확률들을 점수로 판단하고 합하여 가장 높은 값을 가지는 동형이의어 의미를 선택하게 된다.

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

동형이의어 H의 $H_{s_1}, H_{s_2}, \cdots, H_{s_n}$들이 상호배타적이며 완비적인(mutually exclusive and exhaustive) 사상들로 분할되어 있다고 가정하고, 사상 w_j는 사상 H와 교집합을 형성할 경우 $H_{s_1}, H_{s_2}, \cdots, H_{s_n}$는 상호배타적이므로 이들간에는 겹침(intersection)이 없게 된다. 따라서 수식(2)를 이용하여 공기하는 의미정보 w_j가 의미 H_{S_k}로 결정지을 확률을 구하게 된다.

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

수식 (3)과 수식 (4)에서는 의미정보를 체언, 용언의 집합으로 구분하고 동형이의어 의미정보 집합의 체언 집합과 용언 집합으로 구분하고 밀도를 고려하여 확률을 구하게 된다.

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

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

3.2.2 기존 모델의 동형이의어 분별 과정 및 오류 분석

[예문 1] 원문 : 그 바람에 **배**보다 배꼽이 더 커버린 과자값이 들었지만 그뒤에도 카메라가 쥐어진 할애비만 보면 외손녀는 울 듯 비죽거렸고, 그때마다 아내는 카메라를 동댕이치겠다고 위협했다.
(세종계획 품사 부착 말뭉치에서 추출한 문장, POS Tag 생략)

[예문 1]에는 두 개의 동형이의어 '배[5]'와 '들다'가 포함되어 있다. 이 중 '배'는 문맥상 신체 부위를 나타내는 '배_1'로 분별하는 것이 올바르다. [예문 1]에 나타난 '배' 에 대한 동형이의어 분별 정보 즉, 각 의미정보 집합들에 들어 있는 단어들과 이들의 공기빈도는 표 1과 같다.

표 1 예문 1에서 추출한 의미정보(앞/ 뒤 5어절 내)

배_1 (신체)	체언	배꼽(2)
	용언	크다(26), 들다(9)
배_3 (선박)	체언	바람(9)
	용언	크다(25), 들다(1)
배_4 (과일)	체언	바람(1), 과자(2)
	용언	크다(1), 들다(1)
배_6 (곱절)	체언	값(1)
	용언	크다(5)

추출된 단어(의미정보)의 공기빈도를 수식(3)과 수식(4)로 계산하고 수식 (2)에 적용한 확률의 결과는 표 2와 같다. 결과적으로 [3]에서 제안한 통계적 모델은 '배_4(과일)'로 결정하여 동형이의어 분별에 실패하게된다. 분별이 실패하게 된 주요 원인을 다음과 같이 분석할 수 있다.

표 2 통계적 방법에 의해 구한 확률값

	배_1 (신체)	배_3 (선박)	배_4 (과일)	배_6 (곱절)
배꼽	1.00			
바람		0.48	0.52	
과자		0.00	1.00	
값				1.00
크다	0.31	0.26	0.11	0.32
들다	0.39	0.05	0.56	
합계	1.70	0.79	2.19	1.32
의미 분별 결과 (실패) : 배_4(과일)				

5) 동형이의어 '배'의 사전 뜻풀이(분별에 사용된 동형이의어 의미별)
배_1(신체부위) : ① **사람이나 동물의 몸에서 위장 따위의 내장이 들어 있는 부분.**
② 긴 물건 가운데의 볼록한 부분.
③ 정상 진동 또는 정상파에서 진폭이 가장 큰 부분. 복.
배_3(운송수단) : 사람, 물건을 싣고 물 위로 떠다니는 물건. 선박.
배_4(과일) : 배나무의 열매.
배_6(갑절, 곱절) : 갑절, 곱절

첫째, 밀도에 기반한 정규화 확률 계산에서 사용 비율이 낮은 의미(의미정보의 밀도가 상대적으로 매우 낮은 의미)의 공기단어들이 상대적으로 매우 높은 확률을 가지게 되어 중첩되는 의미정보들에서 과분석 되는 경우가 발생하게 된다. 예를 들어, 표 2에서 사전뜻풀이에서 추출한 의미정보 '들다'의 경우, '배_3'과 '배_4'에서 동일하게 1회씩 출현한다. '배_3'과 '배_4'가 출현한 뜻풀이의 문장 개수는 각각 513회와 24회이며, 출현 비율은 동형이의어 개별 의미 중 각각 45.39%와 2.62%를 차지하게 되며 '배_4'로 해석될 확률이 동일한 공기빈도에서 17배 높아지게 되는 문제가 있다. 둘째, [3]의 모델은 구문구조를 이용한 분석을 하지 않고 단순히 동형이의어와 공기하는 의미정보가 무엇인가에 따라 분석하게 된다. 따라서 동형이의어가 단문에서 사용되었다면 구문구조를 분석하지 않더라도 대부분 분별이 가능하다. 그러나, 복문 혹은 중문에서는 추출된 의미정보가 해당 동형이의어의 의미 분별에 결정적인 역할을 하지 못할 수도 있다. 따라서, 본 논문에서는 이 두 가지 큰 문제점을 해결하기 위한 방법을 제시하고자 한다.

4. 가중치를 적용한 통계기반 확률 모델

4.1 새로운 동형이의어 사전 확률 가중치

임의의 문장 C 에 동형이의어와 공기하는 단어 w_j ($\in H_{S_1} \cap H_{S_2} \cap \cdots \cap H_{S_k}$)가 각각의 의미정보 집합에 동시에 나타날 경우 3장의 수식 (2)를 통해 동형이의어 의미별 확률을 계산하게 된다. 각각의 의미정보 집합에서의 공기빈도가 같다면 밀도 정규화에 의해 빈도 총합이 작은 의미에 높은 확률을 부여하게 된다. 공기하는 단어 w_j가 '배_1(신체)'과 '배_4(과일)'의 의미정보 집합에 동시에 포함된 경우 '배_1(신체)'의 공기빈도가 '배_4(과일)'의 공기빈도보다 적어도 15배 이상 되어야만 '배_1(신체)'로 분별될 확률이 높게 된다. 그러나 사전 뜻풀이 말뭉치에서 획득한 의미정보에서 공기빈도 15회는 매우 높은 수치로 그 단어 하나만으로도 통계적인 분석 없이도 동형이의어를 분별할 수 있는 결정적인 정보가 될 수 있다.

일반적인 확률 계산에서는 빈도에 의한 사전 확률(prior probability)을 가중치로 적용한다. 하지만 일반 말뭉치와 사전 뜻풀이 말뭉치는 많은 빈도 차이로 인해 본 논문에서 얻어진 빈도를 손쉽게 이용하기에는 다소 문제가 있다. 하지만 공기 가능한 단어들의 부류는 일정할 것이라는 가설 아래 본 논문에서는 의미정보에 분포하는 단어들을 해결의 실마리로 보고 의미정보 집합별 단어종수를 적극 이용하고자 한다. 그림 4는 사전 뜻풀이 말뭉치에서 얻어진 동형이의어 '배'의 사전 확률이다.

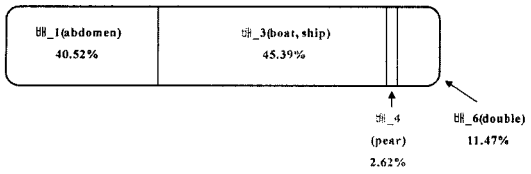


그림 4 임의의 문장에 동형이의어 '배'가 나타날 사전 확률

동형이의어 의미정보 집합 $H_{s_1}, H_{s_2}, \cdots, H_{s_n}$ 각각에 포함된 체언과 용언의 단어종수를 이용하여 수식 (5)와 같은 새로운 동형이의어 의미별 사전확률(new prior probability of homonym senses: NPH)을 얻게 된다. 그림 4의 빈도에 의한 사전 확률이 그림 5와 같은 단어종수에 의한 사전 확률로 변하게 된다.

$$P_{NPH}(H_{S_k}) = \frac{\text{sum of frequency in } H_{S_k}}{\sum_{j=1}^{n} \text{sum of frequency in } H_{S_j}} \quad (5)$$

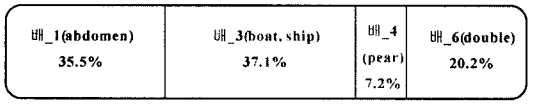


그림 5 동형이의어의 새로운 사전 확률
(예 : 동형이의어 '배')

개별의미 H_{S_k}의 의미정보들은 $P(w_j \cap H_{S_k})$을 확률로 가지고 있다. 새로운 사전 확률 $P_{NPH}(H_{S_k})$을 3장의 수식 (3)과 수식 (4)에 적용하여 $P(w_j \cap H_{S_k}) \times P_{NPH}(H_{S_k})$을 새로운 확률로 이용한다. 통계적 분별 모델에 새로운 확률을 적용하면 수식 (2)는 수식 (2)'와 같이 표현할 수 있다.

$$P_{NPH}(H_{S_k} \mid w_j) = \frac{P(w_j \cap H_{S_k}) \times P_{NPH}(H_{S_k})}{\sum_{i=1}^{n} P(w_j \cap H_{S_i}) \times P_{NPH}(H_{S_i})} \quad (2)'$$

동형이의어 분별에 실패한 [예문 1]에 대하여 표 1에 추출한 의미정보들에 대해 NPH 가중치 적용하여 확률을 계산하게 되면 올바르게 분별함을 표 4에서 알 수 있다.

표 4 새로운 사전확률 가중치의 적용 결과

	배_1 (신체)	배_3 (선박)	배_4 (과일)	배_6 (곱절)
배꼽	1.00			
바람		0.83	0.17	
과자			1.00	
갚				1.00
크다	0.47	0.36	0.04	0.13
들다	0.70	0.08	0.22	
합계	2.17	1.43	1.27	1.13
의미 분별 결과 (성공) : 배_1(신체)				

표 2와 표 4를 비교하면 '배_1'의 의미정보에 속하는 '들다'는 빈도 9로 '배_4'에서의 빈도 1보다 높다. 기존 [3]의 모델에서는 '배_1'로 0.39 확률을 가져 '배_4'의 확률 0.56 보다 낮았다. NPH 가중치를 고려한 표 4의 결과를 보면 0.70으로 '배_1'의 확률이 높아짐을 알 수 있다.

4.2 어절간 거리에 대한 가중치

동형이의어 의미 분별에서 구문구조를 이용할 수 있다면 동형이의어 분별 시에 양질의 의미정보를 선별하여 불필요한 요소를 줄일 수 있을 것이다. 따라서 인접 어절에 대한 정보를 효율적으로 이용하는 방안을 마련해야 한다.

분별하고자 하는 동형이의어의 앞/뒤 5어절로 제한된 문맥을 이용할 때의 분별 정확률이 전체 어절을 이용한 경우와 크게 차이가 없다는 것은 동형이의어 분별을 의미정보가 동형이의어와 인접한 어절에 많이 분포한다는 것이다[10,21]. 특히, 앞/뒤 5어절 범위에서도 동형이의어와 더욱 인접한 의미정보가 상대적으로 분별에 큰 영향을 준다는 점은 자명하다. 따라서 인접 거리에 대한 가중치를 적절히 적용해 보고자 한다.

[21세기 세종 계획]에서 제공하는 150만 어절 규모의 의미 태그 부착 말뭉치를 대상으로, 동형이의어 의미 결정에 필요한 명사(일반명사, 고유명사)와 용언(동사, 형용사)을 수작업을 통해 분석한 결과, 대상 동형이의어를 중심으로 앞/뒤 5어절 범위 내에 97.8%가 분포함을 알 수 있다. 특히, 앞/뒤 5어절 범위에서도 대상 동형이의어와 더욱 인접한 어절에 많이 분포함을 다음의 그림 6

표 3 동형이의어 '배'의 의미정보 집합별 단어종수 및 총빈도합

어휘	의미	어휘종수		총빈도합	
		체언	용언	체언	용언
배	배_1	639	307	2,323	1,313
	배_3	668	283	1,593	1,114
	배_4	130	67	164	102
	배_6	363	82	676	178
소 계		1,800	739	4,756	2,707

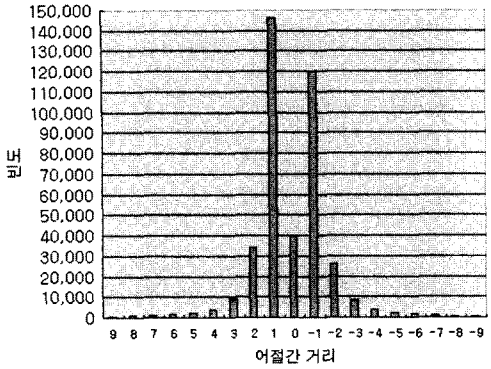


그림 6 세종 150만 의미 태그 부착 말뭉치에서 수작업으로 추출한 의미정보 분포도

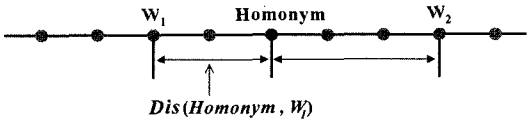


그림 7 동형이의어와 의미정보 어휘간의 인접거리

을 통해 알 수 있다. 따라서 어절간 거리를 고려하지 않고 추출된 의미정보를 효율적으로 사용하기 위해 인접 어절 거리에 대한 가중치를 적용해 보고자 한다.

동형이의어와 의미정보로 사용되는 단어간의 거리를 고려하여 수식 (6)과 같은 거리 가중치를 만들고 3장의 수식 (2)에 적용한 수식 (2)''을 얻게 된다.

$$Dis(H, w_j) = \frac{1}{\sqrt{|d(H) - d(w_j)|}} \quad (6)$$

$$P(H_{S_k}, C) = P(H_{S_k} \mid w_j) \times Dis(H, w_j) \quad (2)''$$

거리 가중치는 어절간 거리를 반영하게 되어 거리가 멀어질수록 의미 분별에 미치는 영향은 감소하게 된다. 따라서 인접 어절에서 단어가 발견되면 높은 확률을 유지하게 되며 멀어질수록 감소하는 방법을 적용하게 된다.

[예문 2] 원문 : 우리를 기다리는 것은 잘 익은 배였습니다.
(우리/NP+를/JKO 기다리/VV+ㄴ/ETM 것/NNB+은/JX 잘/MAG 익/VV+은/ETM 배/NNG+이/VCP+었/EP+습니다/EF+./SF)
(세종계획 품사 부착 말뭉치내 문장)

[예문 2]에서 발견되는 의미정보를 보면 '배_3(운송수단)', '배_4(과일)'의 의미정보로 각각 '기다리다'와 '익다' 한 개씩만 가지며 의미정보가 중첩되지 않아 동일한 확률값 1을 가지고 있다. [3]에서 제안한 통계적 모델을 통해서는 의미를 분별할 수 없다. 또한 NPH 가중치만을 적용하면 높은 사전 확률을 가지는 '배_3(운송수단)'에 의해 '기다리다'의 확률이 높은 값을 가지게 되어 분별에 실패하게 된다. 이러한 경우 거리 가중치 적용을 통해 '기다리다'와 '익다'는 각각 0.5와 1.0의 확률값을 가지게 되어 올바른 분석을 할 수 있다. 하지만 모든 경우에 거리 가중치가 효율적으로 적용되지는 않는다. 거리 가중치의 효율성을 높이기 위해 NPH 가중치를 적용한 결과에서 분별의 차이가 근소한 경우(예문3의 경우: 최고 확률값을 가지는 의미가 36.17% 다음 값이 23.83%로 편차가 20%이내)에 한해서 적용하는 방법을 택하도록 한다.

5. 실험 및 분석

본 장에서는 [2]와 [3]의 연구에서 각각 제안하고 있는 동형이의어 분별 모델과 본 논문에서 제안하는 NPH 및 거리 가중치 적용 방법을 다음과 같은 비교 실험을 수행하였다. 첫째, [2]와 [3]의 비교 실험에 공통으로 이용된 9개(기관, 기구, 눈, 다리, 병, 배, 비, 신, 차) 단어와 외부 정확률 실험에 이용된 1,796문장(국어 정보 베이스 ver1.0과 ETRI 품사 부착 말뭉치에서 추출)을 대상으로 '실험 1'을 수행한다. 둘째, 가중치를 적용한 모델의 일반화 실험을 위해 분별 대상 동형이의어 선정 및 실험 문장의 확장을 위해 본 논문에서는 사전 뜻풀이 말뭉치에 나타나는 상위 고빈도 동형이의어 200여 개를 1차 대상으로 중의성이 높은(하나의 의미로 90% 이상 사용되는 단어 제외) 동형이의어 명사 30개, 동사 16개를 다음 표 5와 같이 선정한다. 그리고 이들 동형이의어를 포함하는 47,977개의 문장을 [21세기 세종 계획]에서 제공하는 350만 어절 품사 부착 말뭉치에서 실험 대상 문장을 추출하고 '실험 2'를 수행한다.

표 5 '실험 2'에 사용된 동형이의어

명사(30개)	결정, 경기, 국, 기구, 기원, 날, 눈, 대, 독, 등, 못, 배, 부정, 비, 상, 성, 의사, 의지, 이상, 장기, 장수, 절, 주장, 중, 지도, 차, 창, 철, 판, 표
동사(16개)	갈다, 고르다, 괴다, 끼다, 닫다, 들다, 묻다, 붓다, 쉬다, 싸다, 쓰다, 이르다, 지다, 차다, 켜다, 타다

'실험 1'과 '실험 2'는 [2], [3] 그리고 본 논문에서 제안하고 있는 NPH 가중치, 거리 가중치 그리고 이들 가중치를 상황에 따라 병행한 모델에 대한 비교 실험을 한다. '실험 1'의 결과 [3]과 본 논문에서 제안한 모델들이 9개 대상 동형이의어 중 7개에서 [2]의 실험 결과보다 정확률이 향상됨을 표 6을 통해 알 수 있다.

'실험 1'에서 본 논문의 가중치를 적용한 결과 NPH 가중치를 적용한 후 계산된 확률값을 비율로 환산하여 그 편차가 20% 이내의 경우 거리 가중치를 적용(가중치 1+2)한 실험에서 최고 1.5% 정확률이 향상되었다.

표 6 '실험 1'의 9개 동형이의어 분별 정확률 비교

(가중치 1: NPH 적용 결과, 가중치 2: 거리 적용 결과, 가중치 1+2: 가중치 1과 2를 병행한 결과, 정확률 단위: %)

동형이의어	의미	문장수	[2] 모델	[3] 모델	가중치1	가중치2	가중치 1+2
기관	몸	17	88.24	88.23	88.24	82.35	82.35
	장치	2	100.00	50.00	50.00	50.00	50.00
	조직	185	92.43	94.05	94.59	94.59	94.59
	정확률	204	92.16	93.14	93.63	93.14	93.14
기구	장치	24	75.00	83.33	87.50	79.17	83.33
	조직	98	89.8	72.45	61.22	70.41	64.29
	정확률	122	86.89	74.59	66.39	72.13	68.03
눈	신체부위	431	79.81	89.10	95.59	95.13	95.59
	식물	1	100.00	100.00	0	100.00	10.000
	기상현상	79	81.01	64.56	48.10	62.03	59.49
	정확률	511	80.04	85.32	88.06	90.02	90.02
다리	교각	21	71.43	33.33	14.29	23.81	23.81
	신체부위	58	84.48	94.83	100	96.55	98.28
	정확률	79	81.01	78.48	77.22	77.22	78.48
병	그릇	12	16.67	16.70	0.00	8.33	8.33
	병사	0	0	0	0	0	0
	질병	151	86.75	98.01	98.68	98.68	98.68
	정확률	163	81.60	92.02	91.41	92.02	92.02
배	과일	6	33.33	16.67	0.00	16.67	16.67
	운송수단	92	75.00	79.35	85.87	86.96	86.96
	신체부위	50	62.00	62.00	64.00	56.00	62.00
	정확률	148	68.92	70.95	75.00	73.65	75.68
비	청소도구	1	0	0	0	0	0
	기상현상	86	80.23	86.05	91.86	91.86	91.86
	비석	10	30	20.00	20.00	20.00	20.00
	비율	1	0	0	0	0	0
	정확률	98	73.47	77.55	82.65	82.65	82.65
신	신발	2	100.00	50.00	50.00	50.00	50.00
	종교	372	86.29	88.44	90.32	83.87	87.63
	정확률	374	86.36	88.24	90.11	83.69	87.43
차	운송수단	46	50.00	54.35	56.52	60.87	60.87
	음료	39	48.72	46.15	43.59	46.15	43.59
	차이	12	83.33	83.33	83.33	83.33	83.33
	정확률	97	43.30	54.64	54.64	57.73	56.70
총 문장수 / 평균정확률		1,796	80.73	83.13	84.30	83.96	84.63

표 7 '실험 2'의 동형이의어 분별 정확률 비교

(가중치 1: NPH 적용 결과, 가중치 2: 거리 적용 결과, 가중치 1+2: 가중치 1과 2를 병행한 결과, 정확률 단위: %)

동형이의어	문장수	[2] 모델	[3] 모델	가중치1	가중치2	가중치 1+2
명사	29,479	72.75	78.12	80.72	81.04	81.55
동사	18,498	60.50	62.45	62.72	56.77	62.20
합계/평균	47,977	68.03	72.08	73.78	71.68	74.09

'실험 1'은 [2]와 [3]의 비교 실험에 사용된 문장을 대상으로 본 논문이 제안하는 가중치의 적용 효과를 실험하기 위한 한 방법이다. 하지만, 9개의 대상 단어가 모두 명사 동형이의어라는 문제와 1,796개의 소규모 문장을 대상으로 한 실험으로써 일반화에는 다소 부족한 점이 많다. 따라서, 본 논문에서는 분별 대상 동형이의어와 실험 문장을 확장한 '실험 2'를 수행한 결과는 다음 표 7과 같다. 기존의 통계 모델[3]에 비해 본 논문에서 제안한 NPH 가중치를 부여한 '가중치 1'에서 정확률이 평균 1.70% 향상되었으며, NPH와 거리 가중치를 함께 이용한 '가중치 1+2'에서 정확률이 평균 2.01% 향상되었다.

표 8 '실험 2'의 30개 명사 동형이의어 분별 정확률 비교
(가중치 1: NPH 적용 결과, 가중치 2: 거리 적용 결과, 가중치 1+2: 가중치 1과 2를 병행한 결과, 정확률 단위: %)

동형이의어	의미 개수	문장수	[2] 모델	[3] 모델	가중치1	가중치2	가중치 1+2
결정	2	1,528	92.21	94.63	94.83	94.31	94.96
경기	3	653	83.15	79.33	77.95	81.32	81.16
국	2	234	90.60	88.03	88.89	91.45	91.03
기구	3	505	82.57	81.58	78.42	79.21	79.80
기원	3	282	64.54	58.16	58.51	60.28	60.64
날	2	4,215	88.42	93.52	95.73	95.07	96.06
눈	2	4,041	78.64	84.58	87.68	88.02	88.39
대	4	94	58.51	64.89	56.38	58.51	57.45
독	2	139	86.33	89.93	91.37	91.37	91.37
등	3	459	72.11	83.66	87.15	86.49	86.71
못	3	2,446	86.75	88.59	91.50	92.48	92.11
배	4	2,029	60.67	63.38	64.17	67.08	66.14
부정	3	996	40.16	39.66	37.05	38.96	39.06
비	4	635	83.94	87.09	90.08	91.97	91.97
상	4	186	69.89	62.37	61.83	73.12	70.97
성	4	635	66.46	70.71	71.97	73.86	73.86
의사	2	733	75.31	69.44	65.48	67.67	66.98
의지	2	752	72.87	75.80	76.20	75.93	76.20
이상	3	2,843	46.92	60.39	74.36	72.88	74.71
장기	3	302	55.63	54.64	49.34	47.02	49.34
장수	3	145	51.72	49.66	48.97	48.97	48.97
절	3	295	62.37	68.81	60.34	64.41	74.24
주장	3	2,038	76.35	94.65	98.09	96.96	97.89
중	3	957	47.02	55.49	64.16	59.67	60.19
지도	2	497	82.70	82.90	83.10	81.49	81.89
차	3	855	57.89	70.41	70.53	71.35	71.58
창	2	285	79.30	78.25	78.60	86.67	85.26
천	3	163	64.42	65.64	62.58	66.26	65.64
판	3	164	65.85	70.12	69.51	67.07	69.51
표	3	363	60.61	63.09	60.88	60.88	60.61
총 문장수 / 평균 정확률		29,479	72.75	78.12	80.72	81.04	81.55

'실험 2'의 결과 중 명사 동형이의어에 대한 실험인 표 8을 보면 '가중치 1+2' 적용에서 [2]와 [3]의 모델에 비해 각각 23개, 21개의 동형이의어에서 각각 8.8%와 3.4%의 정확률 향상이 있었다. 동사 동형이의어에 대한 실험인 표 9에서는 '가중치 1'의 적용에서 가장 높은 62.72%의 정확률을 보이고 있으나, [2]와 [3]의 모델에 비해 크게 향상되지는 않았다. 이는 가중치 적용 방법들이 동사 동형이의어의 분별보다는 명사 동형이의어의 분별에 효과적으로 적용됨을 알 수 있다.

6. 결론 및 향후 과제

본 논문에서는 [2]에서 제안한 방법을 바탕으로 15만여 개의 표제어로 구성된 국어사전 뜻풀이를 이용하여 동형이의어 분별에 필요한 의미정보를 구축하였으며, [3]에서 제안한 통계적 WSD 분석 결과를 토대로 NPH 가중치 및 인접 어절에 대한 거리 가중치를 적용한 동형이의어 분별 모델을 제안하였다.

'실험 1'과 '실험 2'를 통해 기존의 WSD 모델과 본 논문이 제안하는 가중치 적용 모델과의 비교 실험 결과 '실험 1'에서는 '가중치 1+2'를 적용할 때 최고 1.5%의 정확률이 향상되었다. '실험 2'의 명사 동형이의어 대해 '가중치 1+2' 적용 모델에서 [2]와 [3]의 모델에 비해 각각 23개, 21개에서 8.8%, 3.4% 정확률이 향상되었다. 동사 동형이의어에 대해서는 '가중치 1'적용 모델이 가장 높은 62.72%의 정확률 나타내었다.

본 논문이 제안하고 있는 가중치 적용 방법의 문제점은 다음과 같이 요약할 수 있으며, 이는 향후 개선해야 할 연구 과제이기도 하다. 첫째, 동형이의어 분별에 사

표 9 '실험 2'의 16개 동사 동형이의어 분별 정확률 비교
(가중치 1: NPH 적용 결과, 가중치 2: 거리 적용 결과, 가중치 1+2: 가중치 1과 2를 병행한 결과, 정확률 단위: %)

동형이의어	의미 개수	문장수	[2] 모델	[3] 모델	가중치1	가중치2	가중치 1+2
갈다	3	179	70.39	64.25	60.89	60.34	60.89
고르다	3	318	58.18	57.86	52.83	43.71	48.74
괴다	2	59	74.58	77.97	79.66	72.88	76.27
끼다	2	403	74.19	73.20	74.19	70.22	72.70
달다	4	383	52.48	64.75	67.36	67.62	67.62
들다	3	5,494	56.46	61.92	63.00	58.05	65.47
묻다	3	1,662	66.06	64.86	66.85	58.00	68.29
붓다	2	182	79.67	76.92	76.92	73.63	75.27
쉬다	3	519	84.78	79.19	80.92	74.95	75.92
싸다	3	453	73.73	67.77	63.13	58.72	60.04
쓰다	3	4,245	54.04	53.78	56.68	49.54	55.05
이르다	3	1,704	52.76	60.50	49.24	42.25	44.31
지다	4	675	78.22	70.22	64.00	60.00	64.74
차다	4	754	57.43	60.74	62.20	55.84	57.43
켜다	2	136	91.18	83.82	87.50	85.29	87.50
타다	5	1,332	70.50	72.52	77.93	72.30	77.25
총문장수 / 정확률		18,498	60.50	62.45	62.72	56.77	62.20

용되는 의미정보를 정제·분류할 필요가 있다. 특정 문장에 출현한 동형이의어를 분별하기 위해 필요한 문장 내의 의미정보들 중 동형이의어와의 의미적 관계의 중요도에 따른 의미 정보 가중치를 고려해야 한다. 다시 말하면, 동형이의어 분별에 직접적인 영향을 주는 의미정보와 간접적인 영향을 주는 의미정보에 서로 다른 가중치를 부여한다는 것이다. 이것은 본 논문의 '실험 2'에서와 같이 동형이의어 중 용언과 관련된 WSD 정확률이 떨어지는 부분을 보안할 수 있는 방법이기도 하다. 둘째, 하위 범주화나 선택 제약 등 구문 패턴을 이용한 WSD 방법을 고려할 필요가 있다. 이것은 인접 어절에 대한 거리 가중치를 보안할 수 있는 방법으로, 인접 어절 제한에서 고려되지 못하는 구문적인 정보를 활용하는 것이다. 즉 명사 동형이의어인 경우에는 용언과 관련된 의미정보를 먼저 고려하고, 용언 동형이의어인 경우에는 명사와 관련된 의미정보를 먼저 고려함으로써, 동형이의어와 의미정보간의 밀접한 의미적 상관성을 부여할 수 있다.

참 고 문 헌

[1] 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, pp. 1-40, 1998.

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

[3] 이왕우 외, "Bayes 정리에 기반한 개선된 동형이의어 분별 모델", 제13회 한글 및 한국어 정보처리 학술대회, pp. 465-471, 2001.

[4] J. Markowitz, T. Ahlswede, and M. Evens, "Semantically significant patterns in dictionary definitions," Proceedings of the 24th Annual Conference of the Association for Computational Linguistics, New York, pp. 112-119, 1986.

[5] A. Alonge, "Analysing dictionary definitions of motion verbs," Proceedings of the 15th International Conference on Computational Linguistics, pp. 1315-1319, 1992.

[6] Alpha k, Luk, "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions," Proceeding on 33rd Annual Meeting of the Association for Computational Linguistics, pp. 181-188, 1995.

[7] E. Brill, *A Corpus-Based Approach to Language Learning*. Ph.D. thesis Computer and Information Science, University of Pennsylvania, 1993.

[8] Rebecca Bruce and Janyce Wiebe, "Word-Sense Disambiguation Using Decomposable Models," Proceeding on 32rd Annual Meeting of the Association for Computational Linguistics, pp. 139-146, 1994.

[9] Hwee Tou Ng and Hian Beng Lee, "Integrating multiple Knowledge sources to disambiguate word sense: An examplar-based approach," Proceeding on 34th Annual Meeting of the Association for Computational Linguistics, pp. 40-47, 1996.

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

[11] E. Kelly and P. Stone, "Computer Recognition of English Word Senses," Amsterdam, The Netherlands: North-Holland, 1975.
[12] S. F. Weiss, "Learning to disambiguate," Information Storage and Retrieval, Vol. 9, pp. 33-41, 1973.
[13] P. Brown, V. Della Pietra, S. Della Pietra and R. Mercer, "Word sense disambiguation using statistical methods," Proceedings of the 29th Annual Conference of the Association for Computational Linguistics, pp. 264-270, 1991.
[14] D. Yarowsky, "Word-Sense Disambiguation Using Statical Model of Roget's Corpora," Proceedings of the 14th International Conference on Computational Linguistics, pp. 454-460, 1992.
[15] Jun-Su Kim, Wang-Woo Lee, Chang-Hwan Kim and Cheol-Young Ock, "A Korean Homonym Disambiguation System Based on Statistical Model Using Weights," Proceedings of the 16th Pacific Asia conference, pp. 166-176, 2002.
[16] 박성배, 장병탁, 김영택, "의미 부착이 없는 데이타로부터의 학습을 통한 의미 중의성 해소", 한국 정보과학회 '2000 봄 학술 발표 논문집 B, 제27권 1호, pp. 330-332, 2000.
[17] 송영빈, 최기선, "동사의 애매성 해소를 위한 시소러스의 이용과 한계", 제12회 한글 및 한국어 정보처리 학술대회 발표논문, pp. 255-261, 2000.
[18] 이창기, 이근배, "의미 애매성 해소를 이용한 WordNet 자동 매핑", 제12회 한글 및 한국어 정보처리 학술대회 발표논문, pp. 262-268, 2000.
[19] P.O. Cho, C.Y. Ock, "A Korean Noun Semantic Hirarchy based on Semantic Features," Proceedings of the 18th ICCPOL Vol.1, 1999.
[20] 박영자, *사전을 이용한 단어 의미 자동 클러스터링 유전자 알고리즘 접근법*, 박사 학위 논문, 연세대학교 대학원, 1998.
[21] 이호, *단어 의미 중의성 해결을 위한 분류 정보 모형*, 박사 학위 논문, 고려대학교 대학원, 1999.

김 준 수

1998년 울산대학교 수학과 이학사. 2000년 울산대학교 수학과 이학석사. 2001년~2003년 (주)시리울산 한국어시소러스 연구소 선임연구원. 2000년~현재 울산대학교 컴퓨터정보통신학과 박사과정(수료). 관심분야는 한국어정보처리, 의미분별, 정보검색, 통계처리

최 호 섭

1998년 경남대학교 국어국문학과 문학사 2000년 경남대학교 국어국문학과 문학석사. 2000년~2001년 한국전자통신연구원 지식정보검색연구팀 파견연구원. 2002년~2003년 (주)시리울산 한국어시소러스 연구소 선임연구원. 2002년~현재 울산대학교 컴퓨터정보통신공학과 석박사통합과정. 관심분야는 한국어정보처리, 온톨로지, 지식베이스, 언어이해, 지식처리

옥 철 영

1982년 서울대학교 컴퓨터공학과 학사 1984년 서울대학교 컴퓨터공학과 석사 1993년 서울대학교 컴퓨터공학과 박사 1984년~현재 울산대학교 컴퓨터정보통신공학부 교수. 1994년 러시아 TOMSK 공과대학 교환교수. 1996년 영국 GLASGOW 대학교 객원교수. 관심분야는 한국어정보처리(의미분별, 온톨로지), 지식베이스, 기계학습, 문서분류