ISSN(Print): 2287-5905

http://dx.doi.org/10.3745/KTSDE.2012.1.1.031

Disambiguation of Homograph Suffixes using Lexical Semantic Network(U-WIN)

Bae Young-Jun† · Ock Cheol-Young††

ABSTRACT

In order to process the suffix derived nouns of Korean, most of Korean processing systems have been registering the suffix derived nouns in dictionary. However, this approach is limited because the suffix is very high productive. Therefore, it is necessary to analyze semantically the unregistered suffix derived nouns.

In this paper, we propose a method to disambiguate homograph suffixes using Korean lexical semantic network(U-WIN) for the purpose of semantic analysis of the suffix derived nouns. 33,104 suffix derived nouns including the homograph suffixes in the morphological and semantic tagged Sejong Corpus were used for experiments. For the experiments first of all we semantically tagged the homograph suffixes and extracted root of the suffix derived nouns and mapped the root to nodes in the U-WIN. And we assigned the distance weight to the nodes in U-WIN that could combine with each homograph suffix and we used the distance weight for disambiguating the homograph suffixes.

The experiments for 35 homograph suffixes occurred in the Sejong corpus among 49 homograph suffixes in a Korean dictionary result in 91.01% accuracy.

Keywords : Suffix Derived Nouns, Word Sense Disambiguation, Lexical Semantic Network, U-WIN

1. 서 론

빅데이터 개념이 도입되면서 대용량의 문서 및 자료에서 사용자에게 필요한 정보만을 추출하거나 정리해 주는 시스템 또는 서비스에 대한 연구가 최근 활발히 진행되고 있다. 그러나 이러한 시스템 및 서비스들은 아직 의미적 중의성을 처리하지 못하고 있다. 특정 단어가 가지는 의미의 수가 다양할수록 제공되는 정보의 신뢰성이 떨어지는 경향이 있다. 이러한 문제를 해결하기 위해 형태적 처리부터 구문·의미적 처리까지 다양한 방법이 동원되고 있다. 하지만 형태적 처리 기술이 95%이상의 정확률을 보이는 반면, 의미적 처리 기술은 의미적 중의성을 가지는 특정 소수 단어만을 대상으로 처리하여도 70~90%의 정확률을 보이는 등 현재까지 많은 한계를 가지고 있다[1,2,3].

의미 처리 연구는 어휘, 구문, 문장의 의미를 분석하기 위해 이루어 졌으며, 어휘 의미 중의성 해소(Word Sense Disambiguation)에 관한 연구가 많이 진행되어 왔다. 이들은 사용하는 데이터의 형태에 따라서 지식베이스(사전, 의미망, 시소러스, 온톨로지 등)를 이용하는 방법과 말뭉치를 이용하는 방법으로 분류할 수 있고, 방법론에 따라서는 크게 규칙을 이용한 방법, 확률 통계를 이용한 방법 등으로 분류할 수 있다. 지식베이스를 이용한 방법은 기계 가독형 사전(Machine Readable Dictionary)과 어휘의미망과 같은 자원들을 사용하는데, 주로 사전의 뜻풀이나 예문, 개념들의 관계 등을 이용한다[4,5,6,7,8].

말뭉치를 이용한 방법은 1990년대부터 연구가 활발히 진행되어 왔으며, 베이지안 분류기, 결정트리, 신경망 등을 이용한 기계학습 기법을 활용한 연구가 주로 이루어졌다. 말뭉치 의미태그 부착 여부에 따라 비감독 중의성 해소(unsupervised disambiguation)와 감독 중의성 해소(supervised disambiguation)로 나누어진다. 원시말뭉치를 이용할 경우 즉, 의미태그가 부착되지 않은 말뭉치를 사용할 경우 비감독 중의성 해소라고 하며 반대로 의미태그가 부착된 말뭉치를 사용할 경우 감독 중의성 해소라고 한다. 대부분의 경우 감독 중의성 해소의 정확률이 비감독 중의성 해소보다 높게 나타난다. 하지만 감독 중의성 해소를 위해서는 말뭉치에 의미태그 부착 작업이 필요한데, 여기에는 많은 시간, 자원, 노력이 들어간다는 단점이 있다[9].

그 밖에 구조적인 방법을 이용한 연구들이 있는데, 구조적인 방법을 제안한 연구들은 대부분 어휘 사슬(lexical chain)을 기반으로 한 연구들이다. 어휘 사슬은 문맥이나 문장 상에서 의미적으로 관련된 단어들의 연속 또는 연결로, 담화의 일관성과 의미의 연속성 등을 분석하는데 기여한다[10,11].

지금까지 어휘 또는 단어에 대한 의미 중의성 해소 연구는 많이 진행되었지만, 접사파생명사[1]의 접미사 의미 중의

※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No.2012R1A1A2006906).

† 준 회 원 : 울산대학교 정보통신공학과 박사수료
†† 종신회원 : 울산대학교 컴퓨터정보통신공학부 교수

논문접수 : 2012년 8월 6일
심사완료 : 2012년 9월 10일

* Corresponding Author : Cheol-Young Ock(okcy@ulsan.ac.kr)

1) 접사파생명사는 파생어 중 어근에 접사(접두사, 접미사)가 붙어 생성된 명사를 뜻하며, 이중 접미파생명사는 어근에 접미사가 붙어서 생성된 명사를 뜻한다.

성 해소를 위한 연구는 거의 진행된 적이 없다. 현재까지 주로 가급적 많은 접미파생명사를 사전에 등재시켜 처리하였으나, 근본적인 동형이의어 접미사에 대한 언어처리 방법은 거의 연구되지 않았다.

접미사는 단어의 의미를 제약해주는 기능을 하기 때문에 접미사의 의미를 설정해주는 것만으로 접미사가 포함된 단어의 속성과 의미를 부분적으로 파악할 수 있으며, 그 단어가 주어 또는 목적어에 해당한다면 하위범주화 정보의 의미 제약으로 문장의 의미 분석에 활용할 수 있다. 이는 특정 명사 하나에도 문맥의 다양한 의미적 변화를 가져올 수 있는 한자변환 및 기계번역에서 발생할 수 있는 단어의 모호성을 줄여주어 보다 정확한 의미처리를 가능하게 한다. 예를 들어 접미사 '자'의 경우 {'미립자(子)', '유도자(子)', '참석자(者)'} 같은 단어들이 문장에 나타났을 때 접미사의 의미만 다음의 순서대로 {"크기가 매우 작은 요소", "기계장치 또는 도구", "사람"} 파악할 수 있다면 하위범주화 정보, 선택제약 정보 등으로 이용해 문장 또는 문맥의 의미를 보다 정확하게 분석하여 의미처리가 필요한 응용분야에 활용할 수 있을 것이다.

현재 형태소 분석 단계 또는 구문 분석 단계에서 접미사를 따로 처리하지 않고 결합된 어근과 접미사를 하나의 단어로 처리하는 경우가 있다. 그렇지만 이러한 경우 이 단어가 실제 사전에 등재되지 않은 단어이면 미등록어로 처리되거나, 단순 명사로 처리되어 의미 분석 성능을 저하시킨다. 정확한 의미 분석을 위해서 접미사의 의미 중의성 해소가 필요하지만, 접미사를 따로 분리하여 처리하는 시스템이라도 접미사에 대한 의미 중의성 해소는 하지 않는 경우가 대부분이다. 이처럼 자연어처리에서 접미사 의미 중의성 해소를 의미처리에 활용할 수 있지만, 이에 대한 연구는 상대적으로 많은 단어 중의성 해소에 초점이 맞추어져 거의 연구되지 않았다.

접미사는 일반적으로 생산성이 뛰어나 많은 명사들과 결합할 수 있으나, 그럴 경우도 의미적인 제약이 있어 아무 명사와는 결합하지 않는다. 예를 들어 '제(劑)'는 병명이나 약품 등과만 결합한다. 따라서 본 논문에서는 의미태깅된 1,100만 어절의 세종말뭉치에서 접미파생명사를 의미 태깅하고, 의미 결합 제약을 어휘 의미망(U-WIN)에 기반한 모델을 개발하여 접미파생명사 분석을 통한 접미사 중의성 해소를 하고자 한다.

본 논문의 2장에서는 어휘 의미망과 WSD 관련 연구를 살펴보고, 3장에서 U-WIN과 기타 규칙을 이용한 접미파생명사 중의성 해소 방법을 제시한다. 4장에서는 방법론에 대한 실험을 진행한 후 5장에서 결론 및 향후 연구에 대해 논의한다.

2. 어휘 의미망과 접미사 중의성 해소 관련 연구

현재 국내외적으로 의미적 언어 자원 구축에 대한 연구가 다양하게 이루어지고 있다. 국외에서는 WordNet[2], EuroWordNet[3], HowNet[4], UMLS[5], EDR[6], Lexical FreeNet 등이 대표적이다. 이 중 WordNet은 1985년 미국 Princeton 대학교의 심리학자 G. A. Miller를 중심으로 언어학자 Ch. Fellbaum 등 학제간 전문가 및 작업자 그룹에 의해 구축이 시작되었으며, 현재까지 구축되고 있다. 언어학을 비롯하여 자연언어처리, 정보검색, 기계번역 등 여러 분야에서 국제적으로 가장 많이 활용되는 영어 어휘 데이터베이스로 어휘망의 표준으로 인정받고 있다.

국내에서는 부산대의 KorLex[12], 한국전자통신연구원(ETRI)의 ETRINET[1], 국립국어원의 세종의미부류체계[13], 카이스트의 CoreNet[7], 울산대의 U-WIN 등이 대표적이라 할 수 있다. 울산대의 한국어 사용자 어휘지능망 (User-Word Intelligent Network, 이하 U-WIN)은 한국어의 공통적이고 개별적인 속성을 바탕으로 한국인의 보편적인 인지 체계와 개념관계를 파악하여 이를 어휘의 의미적·개념적 네트워크로 형성한 온톨로지적 어휘망이라 할 수 있다. U-WIN은 표준국어대사전을 기반으로 현재 48만 여 어휘가 구축된 상태이다. U-WIN은 전문용어 사전 및 백과사전의 어휘들도 포함하기 때문에 필요시 각 전문 분야별 용어들의 분포도 및 어휘망 형태를 확인할 수 있다. U-WIN의 핵심적 구축 대상은 명사, 동사, 형용사이며, 용언의 하위범주화 정보를 명사 어휘망에 연결하고 있으며, 다른 품사 및 방언, 옛말, 전문용어 등 한국어 어휘 전체를 대상으로 구축 중이다[14].

현재까지 U-WIN은 부족한 어휘를 확보하기 위한 어휘 확장의 방안으로 사용되거나, 어휘간의 유사도 측정을 위한 망으로 사용되었으며, 어휘의 모호성 해결을 위한 방안으로 활용되었다. 임지희(2006)는 U-WIN의 관계정보를 이용하여, 질의어의 모호성을 해결하는 의미적 정보검색의 기반이 되는 기술을 제안하였다. 실험에서 질의어는 전문분야에 주로 사용되는 동형이의어와 보편적으로 사용하는 동형이의어로 구분하고, '질의어+상위어' 형태의 확장 질의어를 설정하였다. 포털사이트의 웹 문서만을 대상으로 한 실험 정확률은 73.5%, 통합검색의 정확률은 68.7%로 나타났다. 이것은 U-WIN이 분류적 상하관계가 아닌 계층적 상하관계를 기반으로 하여 구축됨으로써, 상위어가 질의어의 모호성을 해결하는 유용한 정보로 사용됨을 알 수 있다[15].

조민희(2006)은 U-WIN의 동의어, 유의어, 관련어, 하위어, 상위어 관계 정보를 활용하여 검색어 자동 추천, 관련 단어 제시, 질의어 확장 등을 서비스에 반영하는 사용자 중심의 검색 서비스 요소를 제안하였다. 이러한 어휘지능망의 의미 관계 정보를 활용한 서비스 요소를 통해 현재의 과학기술정보서비스의 검색 만족도를 향상시키는 동시에 사용자가 요구하는 정보를 빠르고 정확하게 검색할 수 있는 서비스 환경으로 개선시키고자 하였다[16].

2) http://wordnet.princeton.edu
3) http://www.illc.uva.nl/EuroWordNet
4) http://www.keenage.com
5) http://www.nlm.nih.gov/research/umls/
6) http://www.wtec.org/loyola/kb/c5_s2.htm
7) http://www.koterm.or.kr

임지희(2007)는 기존의 의미 유사도 측정 방법을 바탕으로 U-WIN 및 정보량을 활용 및 링크 타입·노드의 깊이·최단경로·정보량 등의 요소를 이용한 새로운 의미 유사성 측정방법을 제안하고, 이것을 명사-용언 네트워크 구축을 위해 명사-용언의 연계성을 확보하는 데 활용하였다. 제안한 의미 유사도 측정방법은 링크 기반 방법과 정보량 기반 방법의 단점을 보완하였으며, 한국어 의미망을 활용한 개념적인 클러스터링을 가능하게 하였다[14].

이용훈(2012)은 통계기반의 복합명사 분해 방법과 어휘의미망(U-WIN)과 사전 뜻풀이에서 추출한 의미관계 정보를 이용하는 한국어 복합명사 의미 태깅 시스템을 제안하였다. 전체 시스템은 크게 복합명사 분해, 의미제약, 그리고 의미태깅의 세 가지 부분으로 나누었으며, 분해과정은 세종말뭉치에서 추출한 위치별 명사 빈도를 사용하여 최적의 구성 명사 분해 후보를 선정하고 의미제약을 위한 구성 명사 재분해와 외래어 복원의 과정을 수행하였다. 의미범위 제약과정은 유사도 비교의 계산량을 줄이고 정확도를 높이기 위해 원어 정보와 Naive Bayes Classifier를 이용해 가능한 경우 구성 명사의 의미를 앞서 제약하였다. 의미 분석 및 태깅 과정에서는 bigram 구성 명사의 각 의미 유사도를 구하고 하나의 체인을 만들어가며 태깅을 수행하였다. 성능 평가를 위해 표준국어대사전에서 추출한 3음절 이상의 40,717개의 복합명사를 대상으로 99.26%의 분해 정확도를 보였으며, 95.38%의 의미 분석 정확도를 보였다[17].

접미파생명사와 관련한 논문은 주로 국어학 계열에서 소규모로 연구되었으며 특정 접미사와 관련된 규칙을 형태학적으로 설명하였다. 전산학 계열에서는 접미파생명사를 처리하기 위한 명사파생접미사를 사전정보로 구축한 연구가 있었다. 남윤진(1996)은 말뭉치 분석에 기반한 명사파생접미사의 사전정보 구축에 대한 연구를 진행하였다. 접미사의 의미처리는 아니지만 의미처리에 활용할 수 있도록 접미사를 특성에 맞게 분류하여 접미사 사전을 구축하였다. 접미사의 판별 기준과 그에 따른 파생접미사 목록을 밝히고 각 접미사의 빈도와 다양도 등의 분포상을 50만여 어절의 텍스트를 분석하여 제시하였다. 그리고 각 접미사와 어기의 결합제약 등 접미사가 어절내의 다른 요소들과 맺는 관계들을 설명하였다[18]. 이 논문의 방법론은 실제 시스템으로 구축되지 않고 간단한 실험으로 끝났기 때문에 실제 시스템으로 구축하여 의미처리에 활용해 볼 필요가 있다.

김양진(2008)은 접미사 '-이'의 동음이의 구별 방안으로 형태론적 특성을 중심으로 명사 파생 접미사와 부사 파생 접미사를 구별하는 방법을 제시하였다. 접미사 '-이' 앞의 어기가 동사 또는 형용사일 경우 이는 구별을 위해 실제 패턴을 살펴보고 이를 규칙으로 설정하였다. 특정 대상에 대한 정확한 구별은 가능하지만 대상 확장 시 규칙 설정에 많은 시간이 소모 되는 단점이 있다[19].

3. 접미파생명사 중의성 해소 방법

3.1 접미파생명사의 정의 및 종류

접사파생명사는 독립적인 용법을 지닌 하나의 말(語根 또는 語基)에 접사(접두사, 접미사)가 결합된 단어로, 둘 이상의 형태소로 구성되는 점은 합성어와 같으나 접두사나 접미사가 어근(語根)에 종속적으로 결합되는 점이 합성어와 다르다. 접사는 그 자체로는 자립성을 가지지 못하지만 어근에 붙어 새로운 단어를 만들며, 접사에 따라서는 새로운 단어를 만들어내는 생산성 높은 것과 생상성이 낮은 것이 있다.

대체로 접미사는 새로운 단어를 만드는 측면에서 생산성이 높은 편이다. 그렇지만 접미사는 어근과의 결합 시에 품사적 제약(숫자 혹은 단위성 의존명사와만 결합 "말-들이", "원-어치", "4강", "3중"), 어휘적 제약(고유명사, 고유어, 혹은 한자어와만 결합, "김-씨", "얼음-장", "사진-사"), 의미적 제약(추상명사, 행위성 명사, 상태성 명사, 시간성 명사, 장소성 명사, 구체 명사, 인성 명사)에 따라 제한적인 어근과 결합한다.

표준국어대사전에 등재된 접미사 중 문법정보 "~명사 뒤에 붙어"를 포함하고 방언 및 북한말과 옛말을 제외한 접미사는 339종이며, 이 중 Table 1의 접미사 목록과 같이 49종은 중의성 해소가 필요한 동형이의어이다. Table 1의 개수는 해당 접미사의 의미수이다.

Table 1. Homograph suffixes

no	접미사	개수	no	접미사	개수	no	접미사	개수
1	가	7	18	류	2	35	장	9
2	거리	2	19	률	2	36	전	3
3	경	2	20	보	3	37	정	5
4	계	5	21	부	4	38	제	4
5	공	3	22	사	8	39	조	2
6	관	3	23	상	3	40	주	3
7	광	2	24	생	2	41	증	2
8	구	4	25	선	6	42	지	3
9	국	2	26	성	2	43	지기	2
10	권	3	27	수	4	44	집	3
11	기	7	28	순	2	45	째	2
12	대	5	29	씨	2	46	책	2
13	도	5	30	암	3	47	판	2
14	력	2	31	양	4	48	형	2
15	령	3	32	원	4	49	화	2
16	로	2	33	율	2	총 계		161
17	록	2	34	자	3			

3.2 접미파생명사의 어근과 U-WIN 매핑

동형이의어 접미사의 경우 결합되는 어근과의 의미적 제약에 따라 각기 다른 접미사로 분석된다. 예를 들어, 접미사 '제'는 {制(제도/방법), 祭(제사/축제), 製(만들어진 물건), 劑(약)} 등의 4가지 접미사가 있으며, "추첨제(抽籤制)", "추모제(追慕祭)". "미국제(美國製)", "위염제(胃炎劑)"와 같이 각기 다른 접미사로 분석된다.

본 논문에서는 어휘의미망에 49종의 동형이의어 접미사와 결합할 수 있는 의미제약 속성값을 부여하여 동형이의어 접미사를 분별하고자 한다. 예를 들어, {추첨, 추모, 미국, 위염} 등은 U-WIN에서 다음 Fig. 1과 같은 계층을 가진다.

Fig. 1. Roots of four suffixes '제'

U-WIN에서 각 어근을 포함하고 있는 노드에서 개별 '제'와 결합할 수 있는 최소상계노드를 지정함으로써, 최소상계노드 하위의 모든 노드들이 개별 '제'와 결합할 수 있다. 위의 Fig. 1[B]에서 '위염'의 상위 노드 '질병' 혹은 '병'은 '제(劑)'와 결합할 수 있음을 지정해 둠으로써 '질병'의 모든 하위 명사(340개)들은 '제(劑)'와 결합한다.

현재 형태 의미 주석 세종말뭉치는 접미사에 대해 동형이의어 태깅이 되어 있지 않다. 세종말뭉치에서 접미사 '제'는 2,585회 사용되었으며, 실험적으로 접미사 '제'를 동형이의어 태깅한 결과 중복된 어근(명사)을 포함하여 '제(制)'는 1,946회, '제(祭)'는 445회, '제(製)'는 90회, '제(劑)'는 104회 사용되었다. 또한, '제(制)'와 결합한 어근은 {이사, 교수, 부총리, 신분, 자치구, 선거구…} 등이며, '제(劑)'와 결합한 어근은 {설사, 염, 회복, 증가, 이완, …} 등이다.

Fig. 2와 같이 '제(制)'와 '제(劑)'가 결합한 어근들의 어휘의미망에서의 분포를 살펴보면 '제(制)'는 직위, 위치, 자리 또는 구역 등과 같은 특정 '공간'의 의미를 가지는 단어와 결합하며, '제(劑)'는 현상, 상태와 같은 '모양'의 의미를 가지는 단어 또는 변화, 경과와 같은 '과정'의 의미를 가지는 단어와 결합하는 것을 볼 수 있다.

세종말뭉치에 나타난 어근을 U-WIN에 매핑하기 위해 고려해야 될 점은 세종말뭉치의 어근은 동형이의어 단위로 의미태깅이 되어 있고, U-WIN은 다의어 단위로 구축되어 있다는 점이다. 동형이의어는 형태 즉, 철자는 동일하나 의미들이 전혀 다른 말을 뜻한다. 예를 들면 '배_01(과일)', '배_02(기계, 운송수단)', '배_03(신체부위)'와 같다. 다의어는 하나의 말이 둘 이상의 다르면서 어원적(語源的)으로 관련 있

[a] 제(制) [b] 제(劑)

Fig. 2. Distribution of roots of a suffix '제' over U-WIN

는 의미를 가지고 있는 말을 가리킨다. 즉, '배_03(신체부위)' 중에서도 "사람이나 동물의 몸에서 위장, 창자, 콩팥 따위의 내장이 들어 있는 곳으로 가슴과 엉덩이 사이의 부위."라는 의미의 '배'가 있으며, "아이가 드는 여성의 태내(胎內)."와 같은 의미의 '배'가 있다. 이 둘은 신체부위를 나타내지만, 의미적으로 다른데 이를 다의어라 한다.

위와 같이 세종말뭉치에는 '배_03(신체부위)' 하나로 의미태깅 되어있지만, 실제 U-WIN에는 '배_03(신체부위)_01', '배_03(신체부위)_02' 등 다수가 존재할 수 있고, 이를 자동으로는 정확하게 구분하기 어렵기 때문에 '배_03(신체부위)_01', '배_03(신체부위)_02' 둘 다 어근으로 간주하고 매핑 작업을 진행하였다.

3.3 점층적 누적 가중치 적용

어근의 분포를 바탕으로 최소상계노드를 설정한다. 그러나 상위 노드로 갈수록 의미의 분별력이 없어지고, 하위 노드로 갈수록 의미가 협소해지기 때문에 적절한 최소상계노드를 설정하는 작업이 필요하다. 어근과 매칭되는 U-WIN 노드와 그 상위 노드에 적절한 가중치를 주어 최소상계노드를 설정한다.

U-WIN의 각 노드는 두 가지 가중치를 가진다. 하나는 말뭉치 상의 어근과 일치되는 노드인지 여부에 따른 가중치이며, 이 가중치는 매칭이 될 경우 100의 값을 가지며 일치되지 않을 경우 0의 값을 가진다. 다른 하나는 해당 노드의 하위어에서 거리만큼 감소된 가중치이며, 하위어 중 다른 어근과 매칭이 되는 하위어가 있을 경우 해당 하위어에서 현재 노드까지 거리분의 1만큼의 값들의 합을 가중치로 가진다.

$$f(s_n, x) = A_x + \sum_{i=1}^{M}\left(\frac{1}{d_i}\right) \quad (1)$$

가중치를 부여하는 수식은 <수식 1>과 같다. s는 중의성을 가진 접미사 중 하나이며, x는 어근이다. 함수 $f(s_n, x)$는 어근 x에 대한 특정 의미의 접미사 s의 가중치를 나타낸다. M은 U-WIN에서 x 노드 하위어 중 말뭉치에서 나타난 하위어의 총 개수, d는 해당 노드로부터 하위 노드 사이의 거리이다. A_x는 U-WIN에서의 노드 x가 말뭉치에 출현여부를 확인하는 값으로, 말뭉치에 출현했을 경우 100 값을 가지고 아니면 0값을 가진다. 그리고 $\sum_{i=1}^{M}\left(\frac{1}{d_i}\right)$은 노드 x의 하위노드 중 말뭉치에 출현한 노드로부터 할당된 가중치의 합이다. 즉, 하위 노드들의 개별 가중치에 거리를 나눈 후 모두 합한 값이다.

어근에 해당하는 U-WIN 노드에 가중치 100을 할당하고 그 노드의 상위 노드를 따라 가면서 거리분의 1의 가중치를 더해 준다. 예를 들면 Fig. 2[a]의 '교수'는 접미사 제(制)에 대해 100의 가중치를 가지며, 그 상위어인 '직위'는 1/2의 가중치를, 상위어 '위치'는 1/3의 가중치를 가진다. 이러한 방식으로 U-WIN의 각 노드에 각각의 접미사에 해당하는 모든 가중치를 더해 주었다. 실제 학습 말뭉치에 나타난 Fig. 2[a]의 '교수'는 100의 값을 가지고, 나타나지 않은 '직위'는 3/2(교수:1/2+이사:1/2+부총리:1/2)의 값을 가진다. 이렇게 가중치가 부여된 노드 중 가장 큰 가중치를 가지는 노드를 최소상계노드로 설정한다.

U-WIN의 노드에 가중치를 부여함으로써 실제 학습 말뭉치에 나타나지 않은 어근도 U-WIN의 노드와 매핑 후 상위 계층을 따라가다 보면 특정 접미사의 가중치를 가지게 되어 동형이의접미사의 의미 분별이 가능하게 된다.

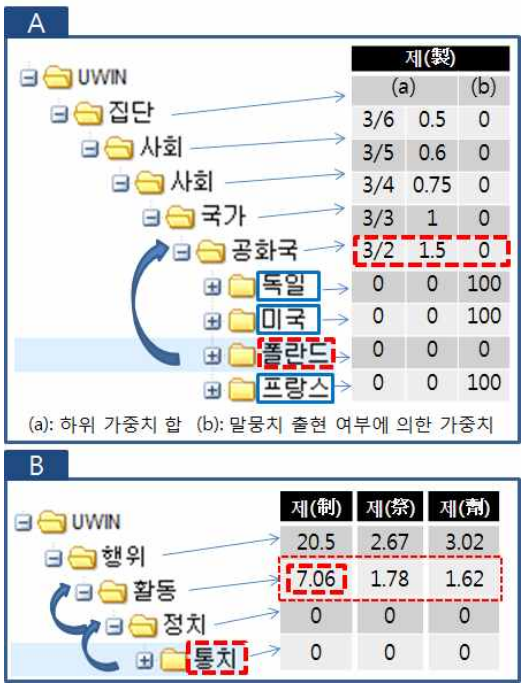

Fig. 3. Disambiguation through weights of hyperonym

Fig. 3은 상위 노드의 가중치 합 또는 비교를 통한 동형이의접미사 의미 분별에 대한 예이다. Fig. 3[A]는 학습말뭉치에 나타난 어근('독일', '미국', '프랑스')이 상위 노드의 '공화국'에 가중치를 부여해, 실제 학습 말뭉치에 나타나지 않은 '폴란드'가 그 상위노드 '공화국'의 가중치 1.5를 가지게 되어 제(製)로 의미가 할당되는 것을 보여준다. 즉, 어근이 '공화국'의 하위 노드일 경우 접미사 '제'는 모두 '제(製)'로 의미 분별된다. Fig. 3[B]는 Fig. 3[A]와 달리 세 가지의 동형이의접미사 '제'의 가중치가 부여되어 있고, 이중 가장 값이 큰 '제(制)'로 의미 분별이 되는 것을 보여준다.

3.4 기타 규칙 적용

본 논문에서는 말뭉치에서 접미사 앞에 나타나는 어근만을 대상으로 의미 중의성 해소를 시도한다. 그러나 어근이 U-WIN 내에 없는 경우 접미사의 의미 중의성을 해소할 수 없다. 그래서 이러한 경우들 중 특정 접미사 앞에 나타나는 어근의 패턴을 분석하여 규칙을 설정하였다.

어근의 형태정보 즉, 품사정보를 분석한 결과 중의성을 가지는 특정 접미사 앞의 일부 어근은 숫자(SN) 및 수사(NR) 또는 기호(SW)로 나타나는 경향이 있다. 예를 들면, 접미사 '경'은 총 두 가지의 의미를 가지지만 이 중 뜻풀이가 "'그 시간 또는 날짜에 가까운 때'의 뜻을 더하는 접미사."인 '경(頃)'만 어근으로 숫자를 가진다는 사실을 말뭉치 상에서 확인할 수 있었다. Table 2는 접미사 '경(頃)'이 말뭉치에서 숫자를 어근으로 가지는 예를 보여준다.

Table 2. Example of roots of a suffix '경'

접미사	말뭉치에 나타난 접미사와 어근의 예
경_25 (頃)	오후_02/NNG 8/SN+:/SP+**00/SN**+경_25/XSN
	BC/SL 427/SN+-/SO+**347/SN**+경_25/XSN

Table 2와 같이 숫자를 포함한 어근이 중의성을 가지는 접미사 중 하나만 사용되는 경우도 있지만, Table 3과 같이 둘 이상의 접미사와 같이 사용되는 경우도 있다. 그래서 이러한 경우는 숫자 앞의 한 단어를 어근으로 선택하여 U-WIN 상의 가중치를 획득하여 중의성 해소를 한다.

Table 3. Example of roots of a suffix '가'

접미사	뜻풀이
	말뭉치에 나타난 접미사와 어근의 예
가_12 (街)	'거리' 또는 '지역'의 뜻을 더하는 접미사.
	예) 을지로/NNP 3/SN+가_12/XSN
가_13 (價)	'원자가(原子價)'를 나타내는 접미사.
	예) 3/SN+가_13/XSN 알코올.

이와 마찬가지로 품사태그가 수사(NR)인 경우도 비슷한 성향을 보였다. 이렇게 어근에 숫자(SN)와 수사(NR)을 포함하는 접미사는 {가, 경, 계, 관, 권, 기, 대, 로, 류, 사, 선, 순,

자, 전} 14종이다. 이는 말뭉치 상에서 빈도 두 번 이상 나타난 접미사를 대상으로 추출한 것이다.

어근에 나타난 기호(SW) 중 특히 단위성의존명사에 해당하는 단어를 어근으로 가지는 접미사가 있다. Table 4의 {대(臺), 순(順)}은 %(퍼센트), kg(킬로그램)의 단위성의존명사에 해당하는 기호를 어근으로 가진다. 이런 형태적 정보를 규칙으로 설정하면 중의접미사 개수가 2개씩인 '대'와 '순'의 의미 분별이 가능해진다.

Table 4. Example of suffixes with symbols

접미사	뜻풀이
	말뭉치에 나타난 접미사와 어근의 예
대_18 (臺)	'그 값 또는 수를 넘어선 대강의 범위'의 뜻을 더하는 접미사.
	예1) 14/SN+%/SW+대_18/XSN 예2) 90/SN+kg/SW+대_18/XSN
순_17 (順)	'차례'의 뜻을 더하는 접미사.
	예) 6.9/SN+%/SW 순_17/XSN

U-WIN 노드에 없는 숫자(SN), 수사(NR), 특정기호(SW)에 해당하는 어근이 나타날 경우 본 절에서 살펴본 것과 같이 기구축한 어근의 형태적 규칙을 통해 접미사의 의미 중의성을 해소하였다.

3.5 접미파생명사 의미 분별 알고리즘

접미파생명사의 의미 중의성 해소를 위한 전체적인 실행 과정은 Fig. 4와 같다. 중의접미사 리스트, 표준국어대사전, U-WIN의 자원을 활용하고, 앞 절에서 살펴본 것처럼 점층적으로 누적된 가중치, 패턴 규칙, 확장 방법 등을 이용하여 의미 분별을 진행한다.

예를 들어 "5·16을 주도한 핵심세력인 육사 8기생들이 바로…"라는 문장의 형태소 분석 및 의미태그 된 말뭉치 "5/SN+·/SP+16/SN+을/JKO 주도하_01/VV+ㄴ/ETM 핵심/NNG+세력/NNG+이/VCP +ㄴ/ETM 육사_06/NNP 8/SN+기_21/NNG+생/XSN+들/XSN+이/JKS 바로_02/MAG"가 입력으로 들어오면 이 중 접미사(XSN)가 있는 '8/SN+기_21/NNG+생/XSN+들/XSN+이/JKS' 어절을 모두 추출한다. 접미사 '생/XSN'은 동형이의접미사이지만 '들/XSN'은 단일 접미사이기 때문에 '들/XSN'은 제외하고 '생/XSN'을 중심으로 앞의 어근 '기_21/NNG'를 확보한다.

Fig. 5와 같이 어근 '기_21/NNG'는 다의어이기 때문에 '기_21_01'과 '기_21_02'를 다의어 리스트에 저장한 뒤 하나씩 순차적으로 처리한다. '기_21_01'과 '기_21_02'를 U-WIN과 매칭시켜 일치하는 노드의 가중치를 확인한 뒤 가중치가 노드에 할당되어 있지 않으면 그 노드의 상위 노드로 옮겨가며 가중치를 확보한다. 만약 해당 어근이 U-WIN에 없을 경우 '규칙 처리' 부분으로 넘겨 특정 패턴이 가지는 접미사별 가중치를 확보한다. 두 방법으로 가중치를 확보한 경우 최고 가중치를 선정해 가장 가능성이 높은 접미사 의미를 선택한 후 의미 분별을 종료한다.

4. 실험 및 평가

실험에 사용된 말뭉치는 세종말뭉치로 동형이의어 단계까지 태깅이 되어 있으나, 접미사는 동형이의어 태깅이 되어

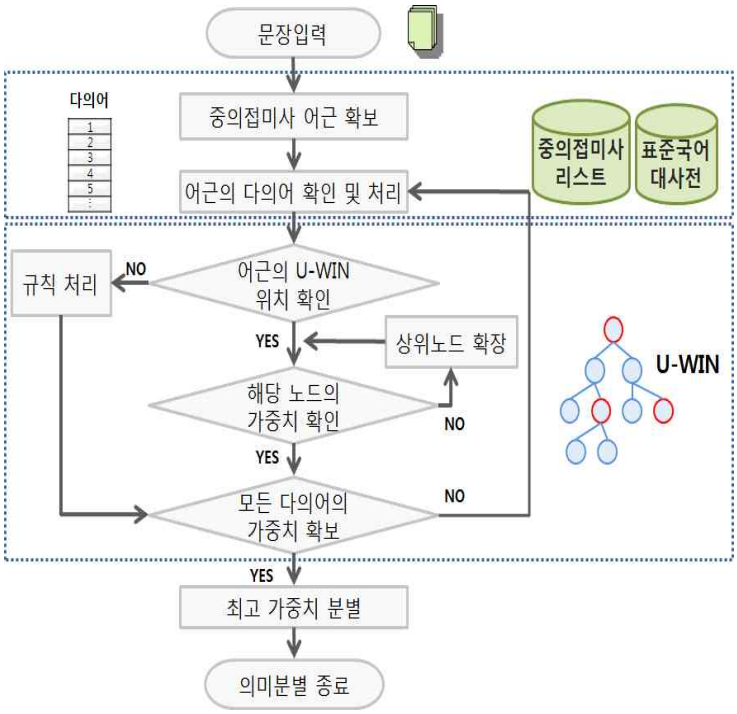

Fig. 4. Algorithm flowchart of suffix derived noun sense disambiguation

Fig. 5. Distribution of polysemy '기' over U-WIN

Table 5. Homograph suffixes for test

A:접미사의 동형이의어 수, B:말뭉치에 나타난 동형이의어 수, C:실험에 사용된 접미사 구분

no	접미사	A	B	C	no	접미사	A	B	C	no	접미사	A	B	C
1	가	7	6	O	18	류	2	2	O	35	장	9	6	O
2	거리	2	1	X	19	률	2	1	X	36	전	3	3	O
3	경	2	2	O	20	보	3	1	X	37	정	5	1	X
4	계	5	5	O	21	부	4	2	O	38	제	4	4	O
5	공	3	2	O	22	사	8	8	O	39	조	2	2	O
6	관	3	3	O	23	상	3	3	O	40	주	3	2	O
7	광	2	1	X	24	생	2	2	O	41	증	2	2	O
8	구	4	2	O	25	선	6	3	O	42	지	3	3	O
9	국	2	2	O	26	성	2	1	X	43	지기	2	2	O
10	권	3	3	O	27	수	4	2	O	44	집	3	1	X
11	기	7	5	O	28	순	2	2	O	45	째	2	2	O
12	대	5	3	O	29	씨	2	1	X	46	책	2	2	O
13	도	5	4	O	30	암	3	0	X	47	판	2	2	O
14	력	2	1	X	31	양	4	1	X	48	형	2	2	O
15	령	3	1	X	32	원	4	3	O	49	화	2	2	O
16	로	2	2	O	33	율	2	1	X	**총 계**		161	112	35
17	록	2	0	X	34	자	3	2	O					

있지 않다. 그래서 접미파생명사의 접미사를 대상으로 동형이의어 태깅을 수작업으로 진행하였다. 세종말뭉치에 포함된 동형이의접미사의 개수는 33,104개였으며, 중복을 제거한 '어근+동형이의접미사' 쌍의 개수는 7,772개였다. 실험에 어근 외에 주변 문맥 정보를 사용하지 않기 때문에 중복을 제거한 '어근+동형이의접미사'를 사용하였다. 실제 말뭉치에 나타난 중의성 접미사의 종수는 Table 5의 49종 중 {록, 암} 2종을 제외한 47종이었다. 이 중 {거리, 광, 양, 정, 씨, 령} 6종은 말뭉치에 나타난 횟수가 5이하로 빈도수 부족으로 제외하였고, {력, 률, 보, 성, 율, 집} 6종은 말뭉치에서 하나의 의미로만 사용되어 의미 중의성 해소가 필요 없기 때문에 제외하였다. 그래서 실험에 사용한 동형이의접미사의 종류는 총 {가, 경, 계, 공, 관, 구, 국, 권, 기, 대, 도, 로, 류, 부, 사, 상, 생, 선, 수, 순, 원, 자, 장, 전, 제, 조, 주, 증, 지, 지기, 째, 책, 판, 형, 화} 35종이었다.

실험의 베이스라인(base line)은 중의접미사의 세종말뭉치에 나타난 빈도를 기준으로 설정하였다. 세종말뭉치에서 개별 동형이의접미사 중 가장 빈도가 높은 접미사를 선정하여 각 접미사별 정확률로 설정하였고, 그 전체 정확률의 평균인 74.90%로 나타났다(Table 7 참조).

실험은 각 노드에 가중치의 비율로 의미 분별하는 방법과 기타 규칙을 결합하여 진행하였다. 기본적으로 비율에 의한 의미 분별 방법은 3.3절에 설명한 것처럼 해당 어근에 대한 가중치가 있을 경우 탐색 없이 가중치를 확보하지만, 설정된 가중치가 없을 경우 상위탐색을 통해서 가중치를 확보한다. 그 뒤, 확보된 가중치들의 비율 계산을 통한 최대값을 선택하도록 하였다. Fig. 6처럼 U-WIN은 다의어 기반으로 구축되었기 때문에 문맥상에 나타난 접미사와 결합한 어근이 U-WIN 상에 다수 분포할 수 있다. 이러한 분포에서 가중치 중 가장 큰 값을 선택하는 것보다 분포의 비율을 고려한 가중치 중 가장 큰 값을 선택하는 것이 모든 가중치 값들을 결과 선택에 보다 잘 반영할 수 있다. 예를 들어, Fig. 6에서 어근 '운영_03_01'를 'A', '운영_03_02'를 'B'라 하고, 'A'와 'B'에 할당된 동형이의 접미사 '자'의 가중치가 존재할 때, 동형이의 접미사 중 하나를 선택하기 위해 가중치 MAX 값을 선택할 경우 '자_01(子)'의 값이 0.233으로 가장

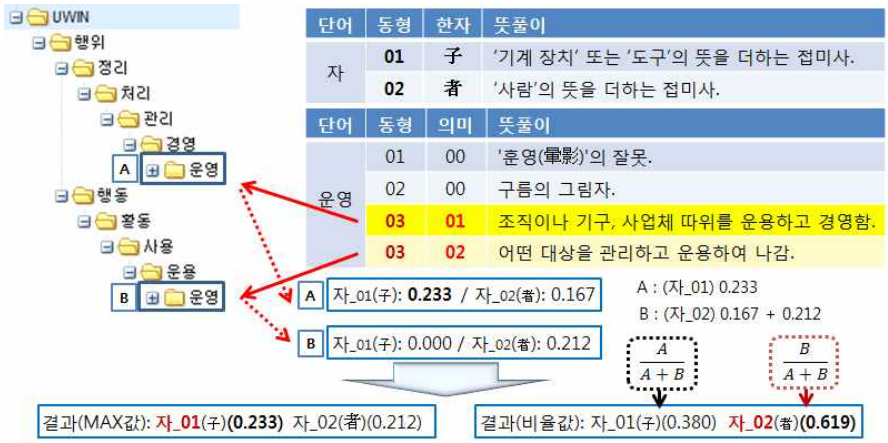

Fig. 6. Sense selection of homograph suffix '자'

크기 때문에 '자_01(子)'를 결과 값으로 출력된다. 하지만 이 값은 다의어 '운영'의 'A' 쪽 값만 고려하기 때문에 잘못된 결과로 출력할 가능성이 높다. 다의어는 의미가 비슷하기 때문에 하나의 다의어를 선택하여 가중치를 반영하는 것보다 모든 다의어들을 포함시켜 가중치를 반영하는 것이 정확률 향상에 도움이 된다. 그래서 비율을 통한 값으로 계산을 하면 'A' 와 'B' 다의어의 가중치 값들을 모두 고려해 보다 정확한 결과 값을 출력할 수 있다.

어근에 대한 실험을 위한 문장의 수가 충분하지 않기 때문에 10-묶음 교차 검증법(10-fold Cross Validation)을 사용하였다. 전체 집합을 10개로 나눈 뒤 9개 집합을 사용해 학습을 진행한 후 나머지 1개의 집합으로 결과를 도출하는 작업을 10번 반복하였다. 그 결과는 Table 6와 같다.

실험결과 평균 정확률은 91.01%로 나타났다. 비율을 계산

Table 6. Results of test

실험 회수	실험 (비율+규칙)	
	정답수/총개수	정확률
1	965 / 1,074	89.85
2	999 / 1,080	**92.50**
3	930 / 1,016	**91.53**
4	957 / 1,018	**94.00**
5	953 / 1,037	**91.89**
6	855 / 918	**93.13**
7	1,053 / 1,158	**90.93**
8	867 / 986	**87.93**
9	943 / 1,059	**89.04**
10	928 / 1,039	**89.31**
평균	**91.01**	

Table 7. Precision comparison of disambiguation of homograph suffixes

A: base line 정확률, B: 비율값+규칙 실험 정확률

no	접미사	개수	A	B	no	접미사	개수	A	B
1	가	388	29.89	**83.50**	19	수	10	**70.00**	40.00
2	경	92	97.82	97.82	20	순	64	93.75	**95.31**
3	계	595	57.31	**84.03**	21	원	329	48.63	**92.09**
4	공	12	91.66	91.66	22	자	1,211	**99.58**	99.33
5	관	372	43.01	**71.77**	23	장	787	65.81	**92.50**
6	구	13	**84.61**	76.92	24	전	488	55.73	**87.29**
7	국	290	94.82	94.82	25	제	807	76.70	**92.06**
8	권	93	53.76	**81.72**	26	조	16	75.00	**81.25**
9	기	413	36.80	**75.78**	27	주	53	52.83	**81.13**
10	대	101	49.50	**85.14**	28	증	102	82.35	**92.15**
11	도	64	64.06	**92.18**	29	지	246	58.13	**80.08**
12	로	51	**94.11**	84.31	30	지기	19	73.68	**84.21**
13	류	26	80.76	**88.46**	31	째	32	**90.62**	87.50
14	부	175	97.14	**99.42**	32	책	16	**62.50**	50.00
15	사	591	42.64	**80.87**	33	판	110	**99.09**	98.18
16	상	1,287	95.18	**97.43**	34	형	71	87.32	**90.14**
17	생	186	82.79	**95.69**	35	화	1,201	98.66	**98.91**
18	선	74	54.05	**83.78**	전체		10,385	74.90	**91.01**

Table 8. Precision of homograph suffixes

no1	접미사	어깨 번호	정답/전체 개수 (부분)	개별 정확률	정답/전체 개수(전체)	전체정확률 [base line]
1	가	12	31 / 34	91.17	324 / 388	83.50 [29.89]
		13	99 / 116	85.34		
		14	17 / 22	77.27		
		16	106 / 115	92.17		
		17	0 / 12	0		
		19	71 / 89	79.77		
2	경	25	88 / 90	97.77	90 / 92	97.82 [97.82]
		27	2 / 2	100.00		
3	계	16	180 / 210	85.71	500 / 595	84.03 [57.31]
		17	3 / 5	60.00		
		18	16 / 33	48.48		
		19	299 / 341	87.68		
		20	2 / 6	33.33		
4	기	39	19 / 34	55.88	313 / 413	75.78 [36.80]
		40	28 / 55	50.90		
		41	131 / 152	86.18		
		42	35 / 49	71.42		
		43	100 / 123	81.30		
5	대	18	6 / 7	85.71	86 / 101	85.14 [49.50]
		19	35 / 44	79.54		
		21	45 / 50	90.00		
6	부	21	5 / 5	100.00	174 / 175	99.42 [97.14]
		23	169 / 170	99.41		
7	사	37	31 / 44	70.45	478 / 591	80.87 [42.63]
		38	162 / 198	81.81		
		39	0 / 2	0		
		40	16 / 23	69.56		
		41	219 / 252	86.90		
		42	38 / 50	76.00		
		43	0 / 3	0		
		44	12 / 19	63.15		
8	상	26	1,216 / 1,225	99.26	1254 / 1287	97.43 [95.18]
		27	2 / 8	25.00		
		28	36 / 54	66.66		
9	생	07	29 / 32	90.62	178 / 186	95.69 [82.79]
		08	149 / 154	96.75		
10	선	19	34 / 40	85.00	62 / 74	83.78 [54.05]
		21	14 / 15	93.33		
		22	14 / 19	73.68		
11	순	17	4 / 4	100.00	61 / 64	95.31 [93.75]
		18	57 / 60	95.00		
12	원	17	134 / 147	91.15	303 / 329	92.09 [48.63]
		18	154 / 160	96.25		
		19	15 / 22	68.18		
13	제	19	602 / 619	97.25	743 / 807	92.06 [76.70]
		20	57 / 77	74.02		
		22	50 / 53	94.33		
		23	34 / 58	58.62		
14	지기	13	5 / 5	100.00	16 / 19	84.21 [73.68]
		14	11 / 14	78.57		
15	책	08	4 / 10	40.00	8 / 16	50.00 [62.50]
		09	4 / 6	66.66		

해서 최대값만 사용하는 방법(90.87%)보다 기타 규칙을 포함한 방법이 0.14% 더 나은 성능을 보였다. 단순히 빈도로 정확률을 측정한 베이스라인(74.9%)보다 최대 16.11% 성능 향상을 보였다.

학습 말뭉치의 어근들과 실험 말뭉치의 어근들의 분포의 차이가 클수록 정확률이 낮아지는데, 8번 실험은 특히 이 차이가 커 다른 실험 말뭉치보다 정확률이 낮게 측정되었다. 그리고 기타규칙 포함 시 0.14%의 성능향상이 있었지만, 이 성능향상은 '어근+접미사'의 중복된 데이터를 제거한 실험 말뭉치에서 나온 수치이다. 중복을 제거하지 않은 말뭉치에서는 100개 중 2.3개 정도의 빈도로 출현하므로 실제 성능의 2~3%의 정확률을 좌우하기 때문에 기타규칙을 고려할 필요가 있다.

Table 7은 실험에 사용된 35가지의 동형이의접미사별 정확률을 베이스라인과 비교한 값을 보여준다. 동형이의접미사 중 {가, 계, 관, 권, 기, 대, 도, 류, 부, 사, 상, 생, 선, 순, 원, 장, 전, 제, 조, 주, 증, 지, 지기, 형, 화} 25가지는 베이스라인보다 높은 정확률을 보였고, {경, 공, 국} 3가지는 동등한 정확률을 보였으며, {구, 로, 수, 자, 째, 책, 판} 7가지는 베이스라인보다 낮은 정확률을 보였다.

U-WIN을 이용한 방법의 장점은 출현 빈도가 낮은 동형이의접미사도 어근으로 확장을 통한 동일 상위어를 가지거나 어근이 같은 부류에 속할 때 적절히 의미 분류를 할 수 있다는 점이다. 일반적으로 동형이의접미사 중 한 쪽 의미에 대한 빈도수가 높으면 중의성해소 결과로 빈도수가 높은 의미로 의미 분별될 가능성이 높다. 하지만 본 논문의 방법을 이용하면 Table 8에 있는 동형이의접미사 {경_27, 부_21, 순_17, 지기_13} 등과 같이 10개 이하의 적은 수의 의미도 의미 분류를 정확히 하는 것을 볼 수 있다. 뿐만 아니라 최고 빈도의 의미와 빈도수가 많이 차이나는 {대_18, 생_07, 선_21} 등의 의미들도 적절히 분류하는 것을 볼 수 있다. 이는 동형이의접미사마다 앞의 어근이 무분별하게 붙는 것이 아니라 의미가 일부 제약되어 특정한 의미를 가진 어근들이 주로 해당 동형이의접미사 앞에 붙는다고 볼 수 있다. 그리고 U-WIN을 바탕으로 하여 해당 어근과 의미적으로 유사한 단어들의 확장을 통해 자료 부족현상을 일부 해결할 수 있다.

오류가 나타나는 유형으로 사용된 빈도수가 부족할 경우, U-WIN 확장 시 가중치를 확보하지 못할 경우, 동형이의접미사가 같은 어근을 가지는 경우로 크게 3가지로 구분된다. 베이스라인보다 낮은 정확률을 보인 동형이의접미사 중 {구, 수, 책} 3가지는 말뭉치에 나타난 전체 개수가 16개 이하로, 학습말뭉치와 실험말뭉치를 구분할 때 학습말뭉치 보다 실험말뭉치 쪽으로 해당 동형이의접미사가 편중될 시 데이터 부족으로 인한 오류가 나타났다.

U-WIN을 통한 동형이의접미사별 가중치를 확보할 때 현재 노드에 가중치 값이 없을 경우 상위어로 확장해 가중치를 확보한다. 그러나 ROOT 전 단계까지 확장해도 가중치를 확보하지 못할 경우 의미 분별을 하지 못한다. {경_25, 계_19, 공_16, 권_05, 권_06, 대_19, 도_21, 로_10, 류_03, 부_23, 사_41, 사_42, 선_19, 수_33, 순_8, 원_19, 자_31, 장_39, 장_46, 조_28, 주_31, 주_32, 증_11, 지_26, 지기_14 째_02, 책_08}의 27종 동형이의접미사에서 총 37개가 실험말뭉치에서 나타났다.

마지막으로 아래의 Table 9와 같이 접미사 앞의 어근이 여러 동형이의접미사와 결합되거나, 어근이 많은 뜻을 가지는 다의어일 경우 또는 상위 노드를 4단계 이상 따라가서 가중치를 확보할 경우 오류로 분별될 확률이 큰 것을 확인할 수 있었다.

Table 9. Errors of homograph suffix '기' and stem '인식(認識)'

기	어근	인식(認識)		
	의미	가중치	정답	분별
氣	기운, 느낌, 성분	0.12		
記	기록	0.49		O
期	기간, 시기	0.43		
器	도구, 기구	0.21		
機	기계 장비	0.20	O	

이러한 오류들은 대부분 자료의 부족으로 인한 오류이지만, 오류들을 해결하기 위해서는 말뭉치에 나타난 어근의 속성 및 주변 문맥 정보 등을 활용하고 동의어, 유의어, 관련어 등의 확장을 통한 가중치 값을 확보하거나, U-WIN 상에서 의미의 핵심을 중심으로 적절한 가중치가 부여되도록 보다 적절한 가중치 부여방법 적용 등을 통해 부분적으로 해결할 수 있다.

5. 결 론

본 논문에서는 동형이의접미사의 중의성을 해소하기 위해 기존의 연구에서처럼 형태적 특성, 전후 문맥 또는 공기정보를 사용하는 대신 U-WIN과 간단한 어근의 형태적 규칙을 사용하였다. 접미파생명사의 어근과 접미사를 분리하여, 어근을 U-WIN과 매핑 시켜 상위어를 따라가며 거리비율로 각 접미사의 가중치를 U-WIN 노드에 매핑한 뒤, 이를 동형이의접미사 의미 중의성 해소에 사용하였다.

기존의 어휘 의미 중의성 해소 연구에서는 중의성 해소 대상을 소수의 어휘 몇 가지만을 대상으로 연구를 진행하였지만, 본 논문에서는 대규모 말뭉치에 나타난 대부분의 중의성 접미사들을 대상으로 중의성 해소를 시도하였다.

실험을 위한 말뭉치는 세종말뭉치를 이용하였으며, 접미사의 동형이의어 단계까지 태그가 부착되어 있지 않기 때문에 수작업으로 동형이의어 태그를 부착한 말뭉치를 이용하였다. 세종말뭉치에 나타나는 중의성 접미사 목록 중 35종을 대상으로 실험한 결과 91.01%의 정확률을 보였으며, 베이스라인으로 설정한 빈도 기반의 방법보다 16.11%의 성능 향상을 보였다.

향후 U-WIN 가중치 부여 방법을 밀도 또는 계층별로 달리 적용하거나, 동형이의어 접미사의 사전 뜻풀이 정보 및

문법 정보를 이용하여 U-WIN의 매핑 정보를 확장한 후 접미파생명사의 중의성 해소 실험을 진행할 것이다.

참 고 문 헌

[1] J. Heo, H. C. Seo and M. G. Jang, "Homonym Disambiguation based on Mutual Information and Sense-Tagged Compound Noun Dictionary", Journal of KIISE, Vol.33, No.12. pp.1073-1089, 2006.

[2] M. H. Kim and H. C. Kwon, "Word Sense Disambiguation using Semantic Relations", Journal of KIISE, Vol.38, No.10, pp.554-564, 2011.

[3] S. J. Kang, "Ontology Construction and Its Application to Disambiguate Word Senses", The KIPS transactions: Part B, Vol.11, No.4, pp.491-500, 2004.

[4] M. Lesk , "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone", In Proceedings of the 5th SIGDOC (New York, NY), pp.24-26, 1986.

[5] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness", In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI, Acapulco, Mexico), pp.805-810, 2003.

[6] P. Resnik, "Selectional preference and sense disambiguation", In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington, D.C.), pp.52-57, 1997.

[7] D. Yarowsky , "Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora", In Proceedings of Coling-92, 1992.

[8] R. Navigli and P. Velardi, "Structural semantic interconnections: A knowledge-based approach to word sense disambiguation", IEEE Trans. Patt. Anal. Mach. Intell. Vol.27, No.7, pp.1075-1088, 2005.

[9] R. Navigli, "Word sense disambiguation: A survey", ACM Computing Surveys, Vol.41, Issue 2, No.10, 2009.

[10] M. Galley and K. Mckeown, "Improving word sense disambiguation in lexical chaining", In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI, Acapulco, Mexico). pp.1486-1488, 2003.

[11] R. Mihalcea, P. Tarau and E. Figa, "Pagerank on semantic networks, with application to word sense disambiguation", In Proceedings of the 20th International Conference on Computational Linguistics (COLING, Geneva, Switzerland), pp.1126-1132, 2004.

[12] A. S. Yoon, S. H. Hwang, E. R. Lee and H. C. Kwon, "Construction of Korean Wordnet 「KorLex 1.5」", Journal of KIISE: Software and Applications, Vol.36, No.1, pp.92-108, 2009.

[13] S. H. Lee, "세종 전자 사전의 어휘 의미 부류 체계", 새국어생활 Vol.17, No.3, pp.51-67, 2007.

[14] J. H. Im, Y. J. Bae, H. S. Choe and C. Y. Ock, "A Measure of Semantic Similarity and its Application in User-Word Intelligent Network", in Proceedings of the KCC, Vol.34, No.1, pp.189-193, 2007.

[15] J. H. Im, H. S. Choe and C. Y. Ock, "Semantic Information Retrieval Based on User-Word Intelligent Network", in Proceedings of KCA, Vol.4, No.2, pp.547-550, 2006.

[16] M. H. Cho, S. F. Choi, H. S. Choi and H. M. Yoon, "Improvement of Science and Technology Information Retrieval Service using Semantic Language Resource", in Proceedings of KCA, Vol.4, No.2, pp.570-574, 2006.

[17] Y. H. Lee, C. Y. Ock and E. B. Lee, "Korean Compound Noun Decomposition and Semantic Tagging System using User-Word Intelligent Network", The KIPS transactions: Part B, Vol.19, No.1, pp.63-76, 2012.

[18] Y. J. Nam and C. Y. Ock, "Constructing Dictionary Information for the Processing of Derivational Suffixes of Nouns based on corpus Analysis", Journal of KIISE, Vol.23, No.4, pp.389-401, 1996.

[19] R. J. Kim and Y. J. Jeong, "A Device for Distinguishing Homonym Relationship of Suffix '-i'", Journal of Korealex, Vol.12, pp.185-207, 2008.

배 영 준

e-mail : young4862@nate.com

2004년 울산대학교 컴퓨터·정보통신공학부 (학사)

2006년 울산대학교 정보통신공학과(석사)

2012년 울산대학교 정보통신공학과 박사수료

관심분야 : 한국어정보처리, 전문용어 인식, 정보검색, 의미분별

옥 철 영

e-mail : okcy@ulsan.ac.kr

1982년 서울대학교 컴퓨터공학과(학사)

1984년 서울대학교 컴퓨터공학과(석사)

1993년 서울대학교 컴퓨터공학과(박사)

1994년 러시아 TOMSK 공과대학 교환교수

1996년 영국 GLASGOW 대학교 객원교수

2007년~2008년 한국정보과학회 언어공학연구회 위원장

2007년 몽골국립대학교 IT대학 명예박사학위

2008년 국립국어원 객원연구원

1984년~현 재 울산대학교 컴퓨터정보통신공학부 교수

관심분야 : 한국어정보처리, 온톨로지, 지식베이스, 기계학습, 문서분류

어휘의미망(U-WIN)을 이용한 동형이의어 접미사의 의미 중의성 해소

배 영 준[†] · 옥 철 영[††]

요 약

현재까지 대부분의 한국어처리시스템에서는 가급적 많은 접미파생명사를 사전에 등재하여 처리하였다. 그러나 접미사는 생산성이 높기 때문에 모든 접미파생명사를 사전에 등재하는 것은 한계가 있다. 따라서 접미파생명사의 의미 분석을 통해서 미등재 접미파생명사를 분석할 필요가 있다.

본 논문에서는 접미파생명사의 의미 분석의 일환으로 한국어 어휘의미망(U-WIN)을 이용한 동형이의어 접미사의 중의성 해소 방법을 제시한다. 형태 의미 주석 세종 말뭉치에서 동형이의어 접미사를 포함한 33,104개의 접미파생명사를 대상으로 실험하였다. 실험을 위해 먼저 동형이의어 접미사를 의미 태깅하였으며, 접미사 앞의 어근을 추출하여 U-WIN의 노드에 매핑시켰다. 또한 동형이의어 접미사와 결합되는 U-WIN 상의 노드들에 대해 거리 가중치를 부여하여 이를 동형이의어 접미사 중의성 해소에 사용하였다.

동형이의어 접미사 49종 중 세종말뭉치에 나타난 35개의 동형이의어 접미사를 대상으로 실험한 결과 91.01%의 정확률을 보였다.

키워드 : 접미파생명사, 의미 중의성 해소, 어휘의미망, U-WIN