UWIN을 이용한 접미파생명사 중의성 해소

배영준 옥철영
울산대학교 컴퓨터정보통신공학과

{young4862, okcy}@ulsan.ac.kr

Derived Nouns of Suffixes Disambiguation using User-Word Intelligent Network

Young-Jun Bae Cheol-Young Ock

Dept of Computer Engineering & Information Technology, Ulsan University

요 약

지식정보화 사회로의 진입으로 언어처리의 필요성은 점차 확대되고 있으나, 현재의 언어처리 기술은 의미분석에 기반하지 않음으로써 많은 한계를 가지고 있다. 본 논문에서는 의미분석의 일환으로 접미사의 중의성 해소를 위해 한국어 사용자 어휘지능망(U-WIN)을 이용한 접미파생명사 분석 방법을 제시한다. 세종 말뭉치에서 중의성 접미사를 포함한 32,647개의 문장을 대상으로 접미사 앞의 어근을 추출하여 U-WIN과 매핑되는 노드에 가중치를 부여한 뒤 이를 접미사 중의성 해소에 사용한다. 동형이의 접미사 49종 중 세종말뭉치에 나타난 25개의 동형이의접미사만을 대상으로 실험한 결과 91.83%의 정확률을 보였다.

1. 서론

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 뜻의 수가 다양할수록 제공되는 정보의 신뢰성이 떨어지는 경향이 있다. 이러한 문제를 해결하기 위해 형태적 처리부터 구문·의미적 처리까지 다양한 방법이 동원되고 있지만 의미적 처리 기술은 현재까지 많은 한계를 가지고 있다.

의미 처리 연구는 어휘, 구문, 문장의 의미를 분석하기 위해 이루어 졌으며, 이 중 2음절 이상의 단어 또는 어휘 의미 중의성 해소(Word Sense Disambiguation)에 관한 연구가 많이 진행되어 왔다[1,2]. 그러나 접사파생명사의 접미사 의미 중의성 해소를 위한 연구는 거의 진행된 적이 없다. 그리고 현재까지 가급적 많은 접미파생명사를 사전에 등재시켜 처리하였으나[3], 근본적인 동형이의어 접사에 대한 언어처리 방법은 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)'도 병명이나 약품 등과만 결합한다. 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.

2. 어휘 의미망

현재 국내외로 의미적 언어 자원 구축에 대한 연구가 다양하게 이루어지고 있다. 국외에서는 WordNet, EuroWordNet, HowNet, Lexical FreeNet, EDR 등이 대표적이며, 국내에서는 부산대의 KorLex, ETRI의 어휘개념망, 카이스트의 CoreNet 등이 대표적이라 할 수 있다.

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 현재 47만 여 어휘가 구축된 상태이다. U-WIN의 핵심적 구축 대상은 명사, 동사, 형용사이며, 다른 품사 및 방언, 옛말, 전문용어 등 한국어 어휘 전체를 대상으로 구축 중이다[4].

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 복합어와 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 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 방언 및 북한말과 옛말을 제외한 접미사는 339종이며 이 중 {가, 거리, 경, 계, 공, 관, 광, 구, 국, 권, 기, 대, 도, 력, 령, 로, 록, 류, 률, 보, 부, 사, 상, 생, 선, 성, 수, 순, 씨, 암, 양, 원, 율, 자, 장, 전, 정, 제, 조, 주, 증, 지, 지기, 집, 째, 책, 판, 형, 화} 등 49종은 중의성 해소가 필요한 동형이의어이다.

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.

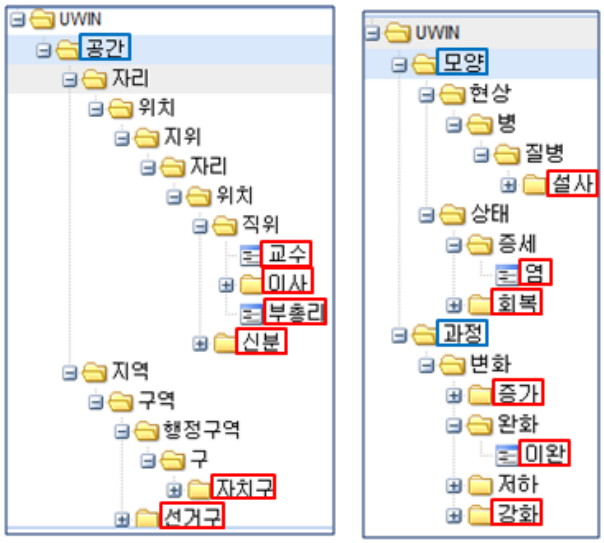
본 논문에서는 어휘어미망에 49종의 동형이의어 한자어 접미사와 결합할 수 있는 의미제약 속성값을 부여하여 동형이의어 접미사를 분별하고자 한다. 예를 들어, {추첨, 추모, 미국, 위염} 등은 U-WIN에서 다음 (그림 1)과 같은 계층을 가진다.

(그림 1) 4 종류의 접미사 '제'의 어근

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 (그림 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 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.

아래의 (그림 2)와 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.



(a) 제(制) (b) 제(劑)

(그림 2) 접미사 '제'의 어근의 어휘의미망 분포

3.3 점층적 가중치 적용

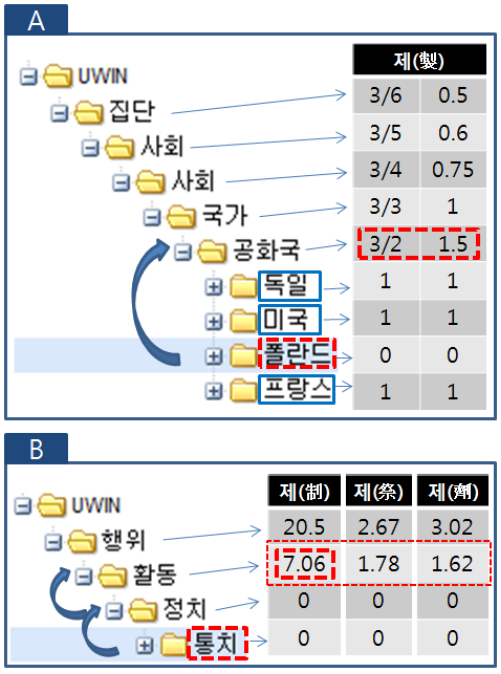
어근의 분포를 바탕으로 최소상계노드를 설정한다. 그러나 상위 노드로 갈수록 의미의 분별력이 없어지고, 하위 노드로 갈수록 의미가 협소해지기 때문에 적절한 최소상계노드를 설정하는 작업이 필요하다.

어근과 매칭되는 U-WIN 노드와 그 상위 노드에 적절한 가중치를 주어 최소상계노드를 설정한다. 어근에 해당하는 U-WIN 노드에 가중치 1을 할당하고, 그 노드의 상위 노드를 따라 가면서 거리분의 1의 가중치를 더해 주었다. 예를 들면 (그림 2)(a)의 '교수'는 접미사 제(制)에 대해 1의 가중치를 가지며, 그 상위어인 '직위'는 1/2의 가중치를, 상위어 '위치'는 1/3의 가중치를 가진다. 이러한 방

식으로 U-WIN의 각 노드에 각각의 접미사에 해당하는 모든 가중치를 더해주었다. 예를 들면, (그림 2)(a)의 '교수'는 1의 값을 가지고, '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 따라가 보면 특정 접미사의 가중치를 가지게 되어 동형이의접미사의 의미분별이 가능하게 된다.

(그림 3)은 상위 노드의 가중치 합 또는 비교를 통한 동형이의접미사 의미분별에 대한 예이다. (그림 3[A])는 학습 말뭉치에 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 의미분별된다. (그림 3[B])는 (그림 3[A])와 달리 동형이의접미사 '제'의 3가지 가중치가 부여되어 있고, 이중 가장 값이 큰 '제(制)'로 의미분별이 되는 것을 보여준다.



(그림 3) 상위노드의 가중치를 통한 의미분별

4. 실험 및 평가

실험에 사용된 말뭉치는 세종말뭉치로 동형이의어 단계까지 태깅이되어 있으나, 접미사는 동형이의어 태깅이 되어 있지 않다. 그래서 접미파생명사의 접미사를 대상으로 동형이의어 태깅을 수작업으로 진행하였다. 세종말뭉치에 포함된 동형이의접미사의 개수는 32,647개였으며, 종수는 29종이었다. 29종 중에 4종은 하나의 뜻으로만 사용되어서 제외하고 25종의 동형이의접미사만을 대상으로 실험하였다.

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 정확률을 더한 값이 70.98%로 나타났다.

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 <표 1>과 같다.

<표 1> 실험 결과

No	정확률	전체 정확률
1	90.90	91.83
2	93.02	
3	93.36	
4	94.20	
5	92.55	
6	94.16	
7	91.31	
8	88.57	
9	90.90	
10	90.30	

실험결과 정확률은 91.83%으로 나타났으며, 단순히 빈도로 정확률을 측정한 베이스라인(70.98%)보다 20.85% 성능향상을 보였다. 실험결과 아래의 <표 2>와 같이 접미사 앞의 어근이 여러 동형이의접미사와 결합되는 어근이 나타나거나, 어근이 많은 뜻을 가지는 다의어일 경우 또는 상위 노드를 4단계 이상 따라가서 가중치를 확보할 경우 오류로 분별될 확률이 큰 것을 확인할 수 있었다.

<표 2> 어근 '인식(認識)'과 접미사 '기'의 분별오류

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	

5. 결론

동형이의접미사의 중의성을 해소하기 위해 기존에 연구에서처럼 형태적 특성, 전후 문맥 또는 공기정보를 사용하는 대신 U-WIN을 사용하였다. 접미파생명사의 어근과 접미사를 분리하여, 어근을 U-WIN과 매핑 시켜 상위어를 따라가며 거리비율로 각 접미사의 가중치를 U-WIN 노드에 매핑

한 뒤, 가장 큰 가중치를 가진 노드를 최소상계노드로 설정하였다.

말뭉치는 세종말뭉치를 이용하였으며, 접미사의 동형이의어 단계까지 태깅이 되어 있기 않기 때문에 수작업으로 동형이의어 태그를 부착하였다. 세종말뭉치에 나타나는 중의성 접미사 목록 중 25종을 대상으로 실험한 결과 91.83%의 정확률을 보였으며, 베이스라인으로 설정한 빈도 기반의 방법보다 20.85%의 성능향상을 보였다.

향후 U-WIN 가중치 부여 방법을 밀도 또는 계층별로 달리 적용하거나, 동형이의접미사의 사전 뜻풀이 정보 및 문법 정보를 이용하여 U-WIN의 매핑 정보를 확장한 후 접미파생명사의 중의성 해소 실험을 진행할 것이다.

참고문헌

[1] 김민호, 권혁철, “한국어 어휘의미망의 의미 관계를 이용한 어의 중의성 해소”, 정보과학회논문지, vol.38, no.10, p554-564, 2011

[2] 허정, 서희철, 장명길, “상호정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해소”, 정보과학회논문지, vol.33, no.12. p1073-1089, 2006

[3] 남윤진, 옥철영, “말뭉치 분석에 기반한 명사파생접미사의 사전정보 구축” 정보과학회논문지, vol.23, no.4, p389-401, 1996

[4] 임지희, 배영준, 최호섭, 옥철영, “U-WIN을 이용한 의미 유사도 측정과 활용”, 2007 한국컴퓨터종합학술대회 논문집, vol.34, no.1, p189-193, 2007