http://dx.doi.org/10.5392/JKCA.2016.16.04.418

한국어 명사의 지식기반 의미중의성 해소를 위한 효과적인 품사집합

Efficient Part-of-Speech Set for Knowledge-based Word Sense Disambiguation of Korean Nouns

곽철헌*, 서영훈*, 이충희**
충북대학교 컴퓨터공학*, 한국전자통신연구원**

Chul-Heon Kwak(ho495949@chnu.ac.kr)*, Young-Hoon Seo(yhseo@cbnu.ac.kr)*,
Chung-Hee Lee(forever@etri.re.kr)**

요약

본 논문에서는 지식기반 기법에서 한국어 명사의 의미중의성 해소에 유용한 품사집합을 제시한다. 세종 형태의미분석 말뭉치에서 174,000 문장을 추출하여 테스트 셋으로 이용하고, 표준국어대사전의 뜻풀이와 용례를 이용하여 각 문장의 의미중의성을 해소하였다. 그 결과 전체 테스트 셋의 성능을 가장 좋게하는 15개의 품사집합과 단어별 평균을 가장 높게 하는 17 개의 품사집합이 제시되었다. 실험결과 45 개의 전체 품사집합을 이용하는 것보다 정확도가 최대 12%까지 향상되었다.

■ 중심어 : | 단어중의성해소 | 자연언어처리 | 품사조합 | 지식기반접근법 | 표준국어대사전 |

Abstract

This paper presents the part-of-speech set which is highly efficient at knowledge-based word sense disambiguation for Korean nouns. 174,000 sentences extracted for test set from Sejong semantic tagged corpus whose sense is based on Standard korean dictionary. We disambiguate selected nouns in test set using glosses and examples in Standard Korean dictionary. 15 part-of-speeches which give the best performance for all test set and 17 part-of-speeches which give the best performance for accuracy average of selected nouns are selected. We obtain 12% more performance by those part-of-speech sets than by full 45 part-of-speech set.

■ keyword : | Word Sense Disambiguation | Natural Language Processing | Part-of-Speech Set | Knowledge-based approach | Standard Korean Dictionary |

I. 서 론

단어의 의미 중의성 해소(Word Sense Disambiguation, 이하 WSD)는 하나의 단어가 문맥에 따라 둘 이상의 의미를 가지는 동형이의어(homonym) 또는 다의어(polysemy)의 중의성을 해소하는 자연어 처리 기술 중 하나이다. WSD는 문장에 중의성이 있는 목표 단어의 의미를 해소하여 문맥의 의미를 정확히 파악할 수 있도록 한다. 예를 들어 “배를 타고 가다가 멀미를 하여 배가 아팠다.”라는 문장에서 ‘배’라는 단어가 어떠한 의미

* 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 [10044577, (1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]
접수일자 : 2015년 10월 23일
수정일자 : 2016년 01월 15일
심사완료일 : 2016년 03월 15일
교신저자 : 서영훈, e-mail : yhseo@cbnu.ac.kr

로 사용되었는지 파악하는 연구가 바로 WSD이다. 정보 검색 서비스에 WSD연구를 적용하면 검색어의 정확한 의미파악으로 사용자로부터 요구된 결과 문서의 정확도가 높아질 수 있다. 또한 자동번역 시스템에서도 중의성이 있는 단어들의 의미를 파악함으로써 더욱 정확한 번역을 할 수 있다.

WSD 기술은 기계 가독형 사전(machine readable dictionary)이나 시소러스(thesaurus)와 같은 언어자원을 이용하여 중의성을 해소하는 지식기반 접근법(knowledge based approach)과 대량의 말뭉치를 이용하여 규칙을 추출하고 학습하는 기계학습기반 등의 연구방법으로 나뉜다[1]. 국내의 경우 한국어 명사의 의미 중의성 해소 관련 연구들을 살펴보면 공개되지 않은 사전을 이용하거나, 말뭉치를 기계학습에 적용하여 실험한 연구가 많다. 그러나 기계학습을 위한 말뭉치는 구축할 때 소모되는 비용이 크다. 그리고 말뭉치의 기반이 된 사전의 체계가 변경되면, 말뭉치를 다시 사용하기 위해 추가적인 비용이 소모된다. 그러나 지식기반 접근법의 경우 정의되어있는 지식을 기반으로 하기 때문에 이러한 단점을 보완할 수 있다. 말뭉치가 없거나 도메인이 다양할 때 기본적으로 연구할 수 있는 방법이 바로 지식기반 접근법이다. 현재 한국어의 경우 다양한 도메인의 말뭉치가 존재하지 않기 때문에 이러한 지식기반 접근법에 대한 연구도 필요하다.

본 연구에서는 단어의 의미 중의성 해소를 위해서 사용되는 품사 자질 중 어떠한 품사가 단어 중의성 해소에 있어 긍정적인 혹은 부정적인 기여를 하는지에 대해 실험한다. 실험을 위해서 Lesk[2]가 제안한 알고리즘을 변형하고, 표준국어대사전을 적용하여 사용하기로 하였다. 그리고 세종사전의 형태의미분석말뭉치를 이용하여 여러 가지 품사조합으로 의미 중의성 해소하여 어떤 품사조합이 좋은 성능을 보이는지 분석한다.

본 논문의 구성은 2장에서는 WSD와 관련된 관련 연구에 대해서 소개하며, 3장에서는 기본적인 LESK알고리즘을 어떤 방식으로 한국어에 접목시켜 더 높은 정확률의 품사 조합을 찾을 것인가에 대한 내용에 대해 서술한다. 4장에서는 앞에서 서술한 방법들을 토대로 실험한 후 평가 및 결과를 제시하고 앞으로의 연구 방향에 대해서 논의하는 순서로 구성될 것이다.

II. 관련 연구

해외의 경우 WSD의 연구 초기에는 Lesk의 연구와 같이 구축에 큰 비용이 소요되는 말뭉치나 시소러스 등이 필요하지 않은 지식기반방법으로 많은 연구가 진행되어 왔다. Lesk는 의미분석대상 단어가 포함된 문장에 존재하는 다른 단어들과 대상 단어의 의미별 뜻풀이에 등장하는 단어들의 중복 정도를 파악하여 가장 일치율이 높은 사전적 의미를 대상 단어의 의미로 결정하는 알고리즘이다. 장점은 다른 언어자원이 필요하지 않고 구현이 간단하며 말뭉치로부터 얻기 어려운 저빈도어의 정보를 얻기에는 용이하다는 장점이 있으나, 단점으로는 단어에 따라 정확도의 차이가 크게 나고 사전의 데이터에 민감하게 작용하며 변화하는 언어의 특성을 반영하기 어렵다는 특징이 있다[3]. 이후 Banerjee[4]의 연구에서는 기존의 Lesk의 연구에서 단어별로 정확도의 차이를 보이는 문제를 보완하기 위해 단어 사이의 관계를 연결하여 구축한 워드넷[5]을 이용하였다.

워드넷은 프린스턴 대학의 인지과학연구소에 의해 1985년부터 개발이 시작된 영어권의 의미 어휘 목록으로서 단어를 유의어 집단으로 분류하고 각 단어의 일반적인 정의를 제공할 뿐 아니라 각 어휘 목록 사이의 다양한 의미 관계를 상위어, 하위어, 등위어 등의 형태로 기록하였으며, 이를 통해 영어권 WSD 등의 자연언어처리 관련 연구 발전에 많은 기여를 하였다. 워드넷이 만들어지고 공개된 이후 약 50여개 언어의 어휘의미망이 워드넷을 참조모델로 구축되어 자연언어처리 연구에 활용되고 있으나 국내의 경우 이러한 환경이 제공되고 있지 않다.

국내의 연구로는 강상욱[6]이 한국어 어휘의미망(KorLex)을 이용한 비감독 단어 의미 중의성 해소 방안을 제안하였다. 중의성 어휘의 주변 문맥에 나타나는 공기 어휘 간의 연관성을 카이제곱 통계량으로 계산하여 중의성 해소에 이용하였다. 어휘 의미망으로 중의성 어휘의 의미별 관계어(상위어, 하위어, 등위어 등)를 찾

고, 대규모 말뭉치로부터 이들 어휘 간 공기어 빈도를 추출하여 어휘의 의미를 정의하는 연구가 진행되었다. 세종 전자사전 내에 정의된 어휘별 문형정보와 선택제약정보와 같은 용언의 하위범주화 정보를 이용하여 규칙을 추출하는 연구도 진행하였다.

정한조[1]은 한국어 언어자료 중 공개된 세종 전자사전과 표준국어대사전을 이용하여 연구하였다. 표준국어대사전에 어휘의 뜻풀이에 사용되는 속담과 용례문장들과 세종 전자사전에 포함된 형태분석말뭉치를 결합하여 어휘의 중의성을 해소하는데 이용한 것이다. 문장 내에서 사용된 의미가 어떠한 의미인지 태그된 중의어가 포함된 문장과 표준국어대사전의 용례와 뜻풀이를 결합하여 중의성 해소 목표 어휘의 좌우 5개의 명사, 동사, 형용사를 센스벡터로 만들고 Naïve Bayes Classifier로 어휘 중의성 해소를 시도한 바 있다.

박상근[7] 또한 어휘의미망으로 단어의 의미 중의성을 해소하였으며, 신준철[8]과 박용민[9] 또한 세종 형태분석말뭉치로 단어 공간 벡터를 생성하여 단어의 의미 중의성 해소에 이용하였다.

기계 가독형 사전을 이용한 지식기반의 접근은 사전에 등재된 한정적인 데이터만을 가지고 의미를 결정하기 때문에 무한한 자연언어의 표현과 규칙들을 처리하는데 한계가 있었다. 그러면서 점차 의미정보가 태그된 말뭉치가 등장하게 되고, 말뭉치를 기반으로 하여 규칙을 추출하거나 확률정보를 이용하는 기계학습 기반의 방법들에 대한 연구가 주를 이루게 되었다[10]. 그러나 한국어의 경우 세종사전의 형태분석말뭉치에 오류가 많고 공개된 자료가 부족하기 때문에 제한된 정보만을 이용한 연구가 필요하다.

III. 품사 조합과 LESK알고리즘

본 연구에 이용될 LESK알고리즘은 매우 단순한 방법이다. 사전에 등재된 단어의 뜻풀이와 용례를 이용하여 의미 분석 대상이 포함된 문장과 비교하여, 같은 내용어가 등장하면 이 내용어들이 서로 연관성이 있다고 가정하고, 대상 어휘의 의미를 특정하는 방식이다. 이 방법은 사전에 등재된 내용만 이용하기 때문에 자연어가 가지는 무한한 규칙을 담기에 부족하다. 그러나 현재 한국어의 경우 공개된 자료가 제한적이므로 대규모의 자료가 필요하지 않은 연구가 필요하다.

LESK 알고리즘은 사전에 등재된 단어의 용례와 뜻풀이에 사용된 정보를 이용하여 문장 내 단어 의미 중의성 해소의 대상이 되는 단어의 의미를 구분한다. 더 세부적으로 서술 하자면 문장 내부에 등장하는 키워드들 간에 연관성이 있다고 가정하고 단어의 의미를 특정하는 것이다.

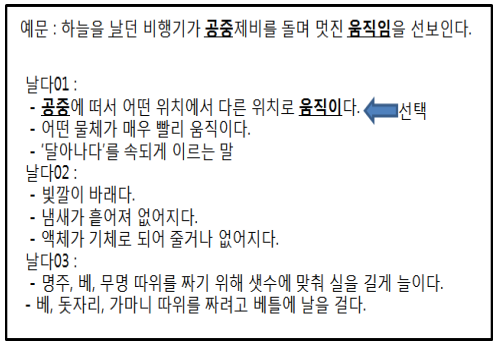

그림 1. LESK 알고리즘 예시

[그림 1]을 보면 날다<동사>의 의미 중의성 해소를 할 때, 사전에 정의된 내용을 형태소 분석을 통해 요소들을 분리하고 주요 내용어인 체언과 용언을 추출하여 의미 중의성 해소 대상 단어가 포함된 문장과 기계 가독형 사전에 등재된 뜻풀이와 용례에 포함된 내용어와 비교하여 가중치를 계산하는 방식으로 “공중(체언)”과 “움직이다(용언)”이 포함되어 있으므로 날다01의 의미와 가장 가깝다고 의미를 정한다.

LESK 알고리즘의 방법상 사전에 등재된 단어의 뜻풀이와 용례만을 사용하여 어휘의 중의성을 판별하기 때문에 이용할 수 있는 정보가 제한적이고 기본적인 알고리즘 이지만 본 논문에서 찾아보고자 하는 품사 조합에 대한 실험을 위해 비교적 구현이 간단한 알고리즘인 LESK 알고리즘을 사용하기로 하였다.

품사 조합은 세종 형태소 품사셋에 존재하는 45개의 모든 품사들에 대해 LESK 알고리즘을 통해 어휘 중의

성 해소 후 정확률을 파악한다. 그리고 품사를 각각 하나씩 빼보면서 같은 실험을 계속하여 정확률의 차이를 구하여 해당 품사가 LESK알고리즘을 통한 WSD에 어떤 영향을 미치는지 알아본다. 간단히 말해 해당 품사가 알고리즘에서 제외되었을 때 정확률이 떨어지게 되면 해당 품사가 긍정적인 영향을 미친 것이고 정확률이 높아지게 되면 부정적인 영향을 미치는 것이다. 이러한 실험을 반복하여 가장 높은 정확률을 보이는 품사조합을 찾는 방식으로 실험을 진행하였다. 실험에는 '다리', '배', '차', '신', '밤', '병', '거리', '눈', '의사', '때', '일', '사회', '속', '문제', '자신', '세계', '사실' 총 17개의 명사를 대상으로 실험하였다. 대상 어휘는 Senseval-2의 한국어 학습데이터의 대상어휘인 '눈', '손', '말', '바람', '자리', '거리', '의사', '목', '점', '밤' 중 세종 형태분석말뭉치에서 고빈도로 등장하는 단어들을 선정하였다. 그리고 말뭉치에서 중의어의 사용빈도가 많은 단어인 '다리', '배', '차', '신', '병', '거리', '의사' 등을 포함하였다. 그리고 지식기반 접근법에 대한 활용성을 검증하기 '밤', '눈', '때', '일', '사회', '속', '문제', '자신', '세계', '사실' 등의 중의어를 추가하여 구성하였다. Senseval은 WSD의 기술을 평가하는 대회로 1998년 시작되어 3년마다 열리는 대회이다. 2004년까지 Senseval-3이 개최되었으며 2007년부터는 SemEval로 이름이 변경되었다.

IV. 실험 및 평가

4.1 실험 데이터

실험 데이터 중 LESK 알고리즘에 사용될 기계가독형 사전의 데이터는 표준국어대사전의 뜻풀이와 용례를 세종 형태소 품사셋에 맞게 형태소 태깅을 한 데이터들을 이용하였다. 이를 검증하기 위해 세종 형태분석 말뭉치를 이용하여 정확률을 측정하였다. 정확률의 측정은 본 연구에 사용된 접근법이 말뭉치를 필요로 하지

표 1. 세종 전체 품사셋 이용시 정확률

정확도 (%)																		
	다리	배	차	신	밤	병	거리	눈	의사	때	일	사회	속	문제	자신	세계	사실	계
P(D)	54.6	6.3	40.8	6.4	5.4	38.8	17.2	86.3	31.7	3.7	71.2	47.5	16.6	0.0	28.6	5.4	76.0	31.6
P(E)	20.9	21.5	66.3	6.3	22.1	46.8	20.1	58.6	23.5	62.6	38.3	74.9	74.9	67.9	77.3	75.0	14.7	41.3
P(D+E)	40.1	15.1	68.0	6.9	13.3	52.1	21.5	86.6	32.5	5.0	78.4	51.7	62.6	4.9	79.5	29.3	67.1	42.0
MFC	76.0	33.5	74.9	71.0	96.9	62.0	55.3	91.4	61.7	99.5	99.7	99.8	99.7	99.5	96.5	94.7	98.4	83.0

표 2. 품사 조합1 이용시 정확률

정확도 (%)																		
	다리	배	차	신	밤	병	거리	눈	의사	때	일	사회	속	문제	자신	세계	사실	계
P(D)	16.1	12.1	22.7	5.0	17.4	34.1	14.2	19.4	11.8	35.3	62.3	23.0	41.9	2.7	40.5	5.6	53.0	32.1
P(E)	22.5	20.2	73.5	6.7	26.7	36.6	16.2	49.1	23.1	37.7	40.2	41.2	71.8	54.2	66.8	74.5	38.4	46.9
P(D+E)	26.3	23.1	73.6	7.1	27.6	47.2	19.6	53.2	21.7	47.9	72.2	34.7	80.5	29.3	75.3	69.2	63.4	54.0

표 3. 품사 조합2 이용시 정확률

정확도 (%)																		
	다리	배	차	신	밤	병	거리	눈	의사	때	일	사회	속	문제	자신	세계	사실	계
P(D)	21.6	13.4	18.6	4.1	18.7	37.2	16.5	27.2	18.3	20.0	66.1	35.7	23.5	0.9	34.6	5.5	71.7	30.4
P(E)	21.4	20.0	74.0	6.1	32.1	40.2	24.6	52.7	19.7	31.0	49.1	22.3	73.1	50.6	74.9	72.4	35.9	45.8
P(D+E)	25.4	23.3	74.1	5.8	30.3	49.5	28.2	55.8	24.9	35.7	75.8	39.1	70.1	15.0	81.2	35.5	77.1	50.0

않기 때문에 세종 말뭉치에서 해당 단어가 등장한 문장 전체가 평가셋이자 정답셋으로 사용되었다. 본 실험의 자질로 사용될 세종 형태소 품사셋은 총 45개로 NNG(일반명사), NNP(고유명사), NNB(의존명사), NP(대명사), NR(수사), VV(동사), VA(형용사), VX(보조용언), VCP(긍정지정사), VCN(부정지정사), MM(관형사), MAG(일반부사), MAJ(접속부사), IC(감탄사), JKS(주격조사), JKC(보격조사), JKG(관형격조사), JKO(목적격조사), JKB(부사격조사), JKV(호격조사), JKQ(인용격조사), JX(보조사), JC(접속조사), EP(선어말어미), EF(종결어미), EC(연결어미), ETN(명사형전성어미), ETM(관형형전성어미), XPN(체언접두사), XSN(명사파생접미사), XSV(명사파생접미사), XSA(형용사파생접미사), XR(어근), SF, SP, SS, SE, SO, SL, SH, SW, NF, NV, SN, NA(기타기호들)로 이루어져 있다.

4.2 실험 방법

세종 형태분석말뭉치에서 대상 어휘가 포함된 문장을 추출한 후, 표준국어대사전의 뜻풀이와 용례를 각각 비교하여 대상 어휘의 의미를 선택하는 방식으로 실험을 진행하였다. 품사셋에 대한 실험을 위해 세종 형태소 품사셋에 포함된 모든 품사를 이용하여 실험을 진행하여 기본 수치를 측정한 후, 각 품사를 품사셋에서 하나씩 배제하고 같은 실험은 반복하여 각 품사가 LESK 알고리즘에 어떠한 영향을 미치는지 분석한다.

4.3 실험 결과 및 평가

실험대상이 되는 어휘 목록에 대해 각 대상 명사가 포함되어있는 문장을 세종 형태분석말뭉치에서 추출하여 표준국어대사전과 비교하여 매칭되는 형태소가 있으면 가중치를 증가한다. 이 가중치의 합이 가장 높은 의미를 대상 어휘의 의미로 정하는 방식이다. 세종 품사셋에 존재하는 모든 형태소 품사를 이용하여 실험한 결과를 베이스라인으로 두고 실험을 한다. 그리고 말뭉치에서 통계적으로 가장 많이 나타난 의미를 해당 단어의 의미로 결정하는 Most Frequent Class 데이터를 참고하기 위해 추가하였다. 위 표의 '계'의 산출방법은 '정답단어수/실험대상단어수' 이다.

[표 1]은 모든 품사셋을 이용하여 실험을 진행하였을 때의 정확도를 나타내는 표이다. P(D)는 표준국어대사전에 등장하는 어휘의 뜻풀이만을 이용하여 판별한 정확률이며, P(E)는 용례만을 이용하여 판별한 정확률이다. P(D+E)는 뜻풀이와 용례를 같이 실험하여 얻은 정확률이다. 결과를 살펴보면 각 어휘별로 정확률의 차이가 매우 큼을 알 수 있다. 이는 각 어휘의 정의나 용례에 등장하는 형태소들이 일상적으로 쓰이는 것들과 차이가 존재하는 경우도 있기 때문이다. '신'이라는 단어의 경우 가장 많이 사용되는 단어 의미의 정의가 "종교의 대상으로 초인간적, 초자연적 위력을 가지고 인간에게 화복을 내린다고 믿어지는 존재." 즉 종교적의미의 신으로 많이 사용되는데, 정의의 내용을 보다시피 일상적으로 많이 사용되지 않는 형태소의 구성으로 이루어져있다. 이로 인해 어휘별로 그 정의나 용례의 형태에 따라 정확률의 차이가 큼을 알 수 있다. 위에서 실험한 내용을 LESK 알고리즘에 사용한 품사 조합 실험에 비교 대상으로 하고 추가적인 실험을 하였다. 모든 품사 조합에서 각 품사를 제외하고 실험을 하였을 때의 성능의 차이를 토대로 품사조합을 추출하기로 하였다. 실험에 사용된 단어 집합을 W라 하고 실험단어 w에 대하여 품사조합 $X(X \subseteq S,\ S$는 세종 전체 품사셋)를 사용하여 실험한 결과 성능을 $E(w, X)$라 했을 때 최적의 품사 조합 X_1을 찾는 수식 1은 다음과 같다.

$$X_1 = \arg max_X \left(\frac{\sum_{w \in W} (E(w, X) - E(w, S))}{|W|} \right) \quad (1)$$

수식 1로 계산하여 얻은 품사조합 X_1로 실험한 결과가 [표 2]이며, 각 단어별 등장 횟수를 고려하기 위한 수식 2는 다음과 같다.

$$X_2 = \arg max_X \left(\sum_{w \in W} \left(\frac{E(w, X) - E(w, S)}{N_w} \right) \times \frac{1}{|W|} \right) \quad (2)$$

여기서 N_w은 실험 말뭉치에서 나타난 단어 w의 출현 횟수이다. 위의 수식으로 얻은 품사조합 X_2로 실험을

한 결과가 [표 3]이다.

[표 2]에 사용된 품사 조합은 [NNG NNB NP NR VV VA VX VCP VCN MAG JKS JKC EP EF ETM] 15개의 품사 조합이며, [표 3]에 사용된 품사 조합은 [NNG NNB NP NR VV VA VX VCP VCN MAG JKS JKC JKG EP EF ETM XSN] 17개의 품사의 조합이다. 두 가지 조합의 차이는 통계 자료를 이용하였을 경우 JKG(관형격조사)와 XSN(명사파생접미사)가 추가 되었다는 것이다. 품사조합1 을 통해 나온 실험 결과를 보면 세종 품사셋의 모든 품사를 이용하여 수행한 실험 결과보다 정확률이 상승하였음을 알 수 있다. 어휘별로 차이가 있기는 하지만 전체 단어 계의 성능이 높아지는 효과를 얻었으며, P(D+E)의 성능이 [표 1]에 비해 12%가 향상되었다. 그러나 품사조합 1을 이용할 경우 성능이 향상되는 단어가 성능이 하락되는 단어보다 많다. 이를 개선하기 위해 품사조합 2를 이용하면 총 단어의 성능향상은 품사조합 1에 비해 4%정도 낮지만 더 많은 단어들의 성능이 향상되는 결과를 보였다. 이러한 결과는 각 어휘별로 특정 품사가 미치는 영향이 다르기 때문에 나타난 영향으로 분석된다. 예를 들어 JKG(관형격조사)의 경우 각 단어에 대해 긍정적인 영향을 주기도 하고 부정적인 영향을 주기도 한다. 이러한 영향들이 수식1에서는 JKG를 품사 조합에서 제외시키는 결과를 보이고, 수식2에서는 품사조합에 포함시킨다. 또한 품사 중에는 LESK 알고리즘을 통해 중의성을 해소할 때에 성능에 전혀 영향을 미치지 않는 요소도 있었다. 어휘의 정의나 용례에 등장하지 않는 품사도 존재하기 때문이다. 보통 SF, SS, SE등과 같은 기호들은 대부분의 경우 사전의 정의나 용례의 문장에 등장하지 않으며, SN(숫자기호)같은 경우 특정 어휘에만 등장하는 경우이다. NR(수사) 또한 비슷한 경우라 할 수 있다. 그 외로 NNG(명사), VV(동사), VA(형용사) 등의 품사는 예상대로 중의성 해소에 긍정적인 영향을 끼쳤다. 본 실험에서 추출한 결과는 본 실험 대상 단어에 해당하는 품사조합이며, 다른 단어를 대상으로 할 경우 상이한 결과를 보일 수 있지만, 본 실험의 결과를 참고 및 활용이 가능하다.

앞서 얻은 결과를 토대로 추가적인 실험을 진행하였다. '손', '이상', '안', '과정', '시', '이해', '사상' 등 총 7개의 추가적인 어휘를 대상으로 같은 실험을 통해 성능 향상이 유사한 결과를 나타내는지 확인을 하였다. 추가 실험 결과 [표 4-표 6]을 보면 앞서 했던 실험결과와 유사하게 나타난 것을 확인 할 수 있다. 그러나 단어마다 사전에 등재된 내용에 따라 형태소 품사의 종류들이 다르고 성능의 차이가 생길 수 있기 때문에 단어에 따라 더 좋은 성능을 보일 수 있는 품사 조합이 존재 할 수 있다.

표 4. 검증 실험 : 세종 전체 품사셋 이용시 정확률

	정확도 (%)							
	손	이상	안	과정	시	이해	사상	계
P(D)	0.0	1.6	2.2	7.7	47.3	73.1	2.0	11.1
P(E)	21.4	34.4	52.3	62.4	4.6	64.2	1.9	33.6
P(D+E)	2.6	4.2	15.1	40.6	46.9	71.9	0.9	19.0

표 5. 검증 실험 : 품사 조합1 이용시 정확률

	정확도 (%)							
	손	이상	안	과정	시	이해	사상	계
P(D)	0.4	7.9	2.4	11.6	47.3	44.6	1.5	11.8
P(E)	14.5	37.9	31.5	48.8	7.5	51.3	8.8	27.5
P(D+E)	9.5	27.3	22.5	45.5	48.8	64.2	4.8	27.5

표 6. 검증 실험 : 품사 조합2 이용시 정확률

	정확도 (%)							
	손	이상	안	과정	시	이해	사상	계
P(D)	2.7	3.7	1.2	8.5	33.7	65.9	0.3	9.9
P(E)	14.8	43.4	37.0	48.8	8.7	54.0	10.4	30.1
P(D+E)	11.0	20.9	17.5	37.5	37.4	69.0	2.1	23.0

III. 결 론

본 논문에서는 한국어 명사 의미 중의성 해소의 실험에서 사용되는 품사가 어떠한 영향을 미치는지에 대해 분석하였다. 분석 결과 사람이 문장 내에 존재하는 중의성 어휘를 판단하는데 중요하게 인식하는 품사들이 포함되었음을 확인할 수 있었다. 그리고 추출된 품사 조합을 통해 지식기반 접근법의 중의성 해소 정확률을

다소 향상시킬 수 있었다. 이러한 품사조합은 분류될 문서의 성격이나 분류에 사용될 사전의 특성에 의해 다소 달라질 수 있으며, 이러한 특성들을 고려한 연구가 추가적으로 필요하다. 체계적인 단어의 의미 구축과 다양한 용례들의 적용을 통한 사전의 구축이 그 예이다.

향후의 연구는 기계학습을 이용한 접근법 등 다른 중의성 해소 방법들에 대해서 품사 조합이 어떠한 영향을 미치는 가에 대해서 추가적으로 연구를 하여 중의성 해소에 유용한 품사 조합의 자질들과 각각의 알고리즘에서 적합한 품사 조합을 찾는 것이다. 또한 각 단어별 정확률의 차이를 줄이고 정확률을 더욱 향상시키려면 현재 사전이 갖는 제한적인 정보를 확장하는 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] 정한조, 박병화, “사전과 말뭉치를 이용한 한국어 단어 중의성 해소,” 한국지능정보시스템학회, 제21권, 제1호, pp.1-13, 2015.

[2] M. Lesk, “Automatic Sense Disambiguation Using Mahine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream cone,” Proceedings of the 5th Annual International Conference on Systems Documentation, pp.24-26, 1986.

[3] 이현아, "가변 크기 문맥과 거리가중치를 이용한 동형이의어 중의성 해소," 한국마린엔지니어링학회, 제38권, 제4호, pp.444-450, 2014.

[4] S. Banerjee and T. Pedersen, “Extended Gloss Overlaps as a Measure of Semantic Relatedness,” International Joint Conferences on Artificial Intelligence, pp.805-810, 2003.

[5] http://wordnet.princeton.edu

[6] 강상욱, 김민호, 권혁철, 전성규, 오주현, “세종 전자사전과 한국어 어휘의미망을 이용한 용언의 어의 중의성 해소,” 동계학술발표회 논문집, pp.414-416, 2014.

[7] 박상근, 최지연, 최기선, “가변길이 윈도우와 빈도 가중치를 이용한 단어 의미 중의성 해소,” 동계학술발표회 논문집, pp.441-443, 2014.

[8] 신준철, 옥철영, “한국어 품사 및 동형이의어 태깅을 위한 단계별 전이모델,” 정보과학회논문지 : 소프트웨어 및 응용, 제39권, 제11호, pp.889-901, 2012.

[9] 박용민, 이재성, “한국어 단어 공간 모델을 이용한 단어 의미 중의성 해소,” 한국콘텐츠학회논문지, 제12권, 제6호, pp.41-47, 2012.

[10] 이용구, “단어 중의성 해소를 위한 지도학습 방법의 통계적 자질선정에 관한 연구,” 한국비블리아학회지, 제22권, 제2호, pp.5-25, 2011.

저 자 소 개

곽 철 헌(Chul-Heon Kwak) 정회원

- 2015년 2월 : 충북대학교 컴퓨터공학과(학사)
- 2015년 3월 ~ 현재 : 충북대학교 컴퓨터공학과(석사 과정)

<관심분야> : 정보검색, 자연언어처리

서 영 훈(Young-Hoon Seo) 종신회원

- 1983년 : 서울대학교 컴퓨터공학과(학사)
- 1985년 : 서울대학교 컴퓨터공학과(석사)
- 1991년 : 서울대학교 컴퓨터공학과(박사)
- 1994년 ~ 1995년 : 미국 Carnegie Mellon 대학 기계번역센터 객원교수
- 1988년 ~ 현재 : 충북대학교 전자정보대학 컴퓨터공학과(교수)

<관심분야> : 자연언어처리, 한영기계번역, 정보검색, 질의응답시스템

이 충 희(Chung-Hee Lee) 정회원

- 1995년 : 한양대학교 전자계산학과(학사)
- 2001년 : 연세대학교 컴퓨터과학과(석사)
- 2014년 : 충북대학교 컴퓨터공학과(박사)
- 2001년 ~ 현재 : 한국전자통신연구원 선임연구원

<관심분야> : 자연언어처리, 기계학습, 정보검색, 질의응답