한국어 의미 분석을 위한 세종의미망 확장

박태호○, 신창욱, 박성재, 박다솔, 신영태, 차정원
창원대학교
{taehope, papower1, tjdwo1289, dasol_p, zzz1421, jcha}@changwon.ac.kr

Extension of the Sejong Semantic Network for Korean Semantic Analysis

Tae-Ho Park, Chang-Uk Shin, Seong-Jae Park, Da-Sol Park, Young-Tae Shin, Jeong-Won Cha
Changwon National University

요 약

자연어처리를 위해 다양한 한국어 어휘의미망이 존재한다. 오랜 기간 동안 한국어 의미망 구축에 대한 연구가 진행되었으나 의미망에 포함되지 못한 단어가 다수 남아 있다. 또한 이미 구축된 단어라 할지라도 그 내용이 부족한 면이 있다. 이에 본 연구는 한국어 의미 분석에 도움을 주기 위하여 새로운 단어들을 의미망에 추가로 구축하였고, 이미 구축된 단어에 대해서는 내용을 보완하였다. 본 연구는 세종전자사전의 체언과 용언을 기반으로 단어를 새롭게 추가하였으며, 총 2,611개의 새로운 단어와 7,238개의 유의어 정보를 추가로 확장하였다.

1. 서 론

의미 분석은 문장의 문법이나 어휘 정보 외에 겉으로 들어나지 않는 정보들을 필요로 한다. 이에 영어권에서는 오래전부터 WordNet을 통해 단어들 사이의 관계와 의미 유사 관계 등을 구체화하여 데이터를 구축해왔다. WordNet은 단어들 사이의 의미 유사 정보를 분석하여 단어 사이의 관계 정보를 체계화하였다. 한국어에서도 다양한 의미망이 구축되어 활용되고 있다. 한국어 어휘의미망 중 구축된 어휘 수가 가장 큰 어휘망은 세종전자사전이다. 세종전자사전은 581개의 의미 분류와 약 54만 개의 어휘가 포함되어 있다[표1][1]. 세종전자사전 외에도 한국어 어휘의미망의 종류가 다양하고, 그 규모가 작지 않다. 하지만 아직 현재까지 구축된 의미망에 포함되지 않은 단어가 많이 남아 있으며, 의미망에 포함된 단어라도 의미 유사 단어 등의 정보가 많이 부족한 상황이다. 이에 본 연구에서는 세종의미망의 의미 분류를 통일하고 부족한 단어를 추가하여 전체 단어수를 확장하고 의미 정보를 보완하였다.. 본 논문은 2장에서 어휘의미망의 관련연구를 소개하고, 3장에서는 의미망을 확장하는 방법과 그 결과를 설명한다. 마지막으로 4장에서 결론에 대해 기술한다.

2. 관련 연구

단어 사이의 관계와 이를 구조화한 어휘의미망 구축은 오래전부터 언어학자들이 꾸준히 연구해왔다. 초기의 어휘의미망은 기계학습이나 자연어처리를 위해서가 아니라 언어학자들이 단어들이 가지는 정보를 체계화하기 위하여 처음 시작되었다[2,3]. 이후 기계학습과 자연어처리 연구의 발전에 따라 의미망의 활용성이 부각되었고, 자연어처리의 자원으로써 좀 더 체계적이고 활용 가능한 다양한 정보를 추가하여 구축하였다. 영어권에서는 WordNet이 단어들 사이의 상·하위 관계와 각 단어들의 유어의 정보를 체계적으로 구축하였다. 한국어 의미망 또한 다양한 연구가 진행되었다. 한국어에는 다양한 의미망이 존재한다. 부산대학교에서 구축한 KorLex[4]와 울산대학교의 UWordMap[5], 서울대학교에서 구축한 세종

표 1 한국어 및 영어 어휘의미망[1]

	구축기관	구축방식/참조모델	의미/개념(n) vs. 어의(w) 수	구축 품사
한국어 명사워드넷	호남대학교	직접	20,000w	명
세종 전자사전	서울대학교	직접	581n vs. 540,000w	모든 품사
U-Win	울산대학교	직접	46,339n vs. 약250,000w	모든 품사
한국어 시소러스	포항공과대학교	참조/PWN	18,362n vs. 21,390w	명
KorLex	부산대학교	참조/PWN	130,639n vs. 147,906w	명,동,형,부,분류사
다국어 어휘 데이터베이스	고려대학교	참조/PWN	5,500w	명
CoreNet	KAIST	참조/NTT어휘대계	2,938n vs. 62,632w	명,동,형

전자사전[6] 등의 한국어 의미망 데이터가 있다. 이러한 단어 어휘의미망 자원은 다양한 자연어 처리 분야에서 활용되고 있다. 특히 형태소 분석과 구문 분석 정보만으로 이해하기 어려운 자연어처리 분야에서 다양한 형태로 활용되고 있다. [7-9]은 어휘의미망 정보를 이용하여 단어들의 중의성을 해소하여 문장의 정확한 의미를 이해하는데 도움을 줄 수 있도록 연구하였다. [7]은 세종전자사전과 KorLex를 이용하여 용언의 어의 중의성을 해소하였다. [8]과 [9]는 울산대학교 어휘의미망을 이용하여 각각 동형이의어 분별과 복합명사의 의미 분석을 연구하였다. 이처럼 한국어 어휘의미망의 구축과 어휘의미망을 이용한 자연어처리 연구가 활발히 진행되고 있다.

3. 실험 및 토의

3.1. 실험 방법

새로운 단어를 의미망에 추가하기 위해서 [10]의 시스템을 활용하였다. [10]은 단어의 의미 그룹을 예측하는 시스템으로써 이를 활용하여 반자동으로 새롭게 추가되는 단어의 의미 그룹을 결정하였다. [10]은 단어들의 word2vec 정보와 유의어 그룹 정보를 사용하여 의미 분류가 되지 않은 단어의 의미 그룹을 예측한다. [10]의 시스템은 5-best에서 78%의 성능을 보인다. 또한 누락된 유의어 정보는 기존에 직접 구축한 유의어 그룹 정보를 사용하여 부족한 정보를 확장하였다.

3.2. 실험 결과

3.1에서 제안하는 시스템을 통해 새롭게 추가된 단어가 체언이 2,563개, 용언이 48개로 총 2,611개의 단어를 추가하였다. 또한 본 연구를 진행하여 기존의 세종의미망에 누락된 유의어 정보를 추가하여 9,045개의 단어에 7,238개의 유의어 정보가 추가되었다. 이 외에도 세종의미망이 지닌 오류를 수정하는 작업을 함께 진행하여 동사가 형용사로 분류되어 있는 등의 잘못된 품사 분류 정보를 수정하였다. 세종전자사전에 포함되어 있던 단어 중 의미 그룹이 미결정되어 있던 98개의 단어에 대해 [10]의 시스템을 이용하여 의미 그룹을 결정해주었다. [표 2]는 본 연구를 통하여 새롭게 추가된 단어들의 일부 예이다.

4 결론

본 연구를 통해 기존 세종의미망에 체언을2,563개, 용언 48개로 총 2,611개의 단어를 추가하였다. 또한 기존에 누락되어 있던 유의어 정보를 새롭게 추가하였다. 유의어가 새롭게 추가된 단어는 9,045개이고, 유의어 수는 7,238개이다. 본 연구를 통하여 기존 의미망에서 부족한 단어와 유의어 정보를 확장함으로써 의미 분석에서 활용할 수 있는 정보가 증가하였다. 하지만 아직 의미망에 미포함된 단어가 많이 남아 있으며, 이미 포함된 단어에도 누락된 정보가 남아 있다. 따라서 향후연구로 유의어 사전에는 존재하나 의미망에는 미포함된 단어에 대한 의미망 확장 작업과 유의어뿐만 아니라 반의어 등의 다양한 정보를 추가할 예정이다. 그리고 본 연구를 통해 새로 추가한 의미 정보나 유의어 정보에 대해 정확한 평가를 위해 이전에 연구된 한국어 의미망을 참조하여 검토할 예정이다. 또한 구축된 데이터를 체계화하여 쉽게 활용 가능하도록 API를 제공하는 시스템을 구축할 예정이다.

표 2 새롭게 추가한 단어 목록

구분	어휘	의미 그룹
체언	부듯가	교통기관관련건물
	구세군	기구
	오존층	기체
	배심원	직업인간
	우동	음식
	병원장	직위인간
	양배추	채소
	...	
용언(동사)	빼내다	단독행위
	펴다	단독행위
	비꼬다	외향적심리상태
	패하다	결과행위
	...	
용언(형용사)	어둡다	속성값
	잘나다	평가속성값
	케케묵다	부정적품격속성값
	복되다	긍정속성값
	...	

사 사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NO. 2015R1A5A7036384).

참 고 문 헌

[1] Aesun Yoon, Soonhee Hwang, Eunryoung Lee, Hyuk-Chul Kwon. "Construction of Korean Wordnet 「KorLex 1.5」." Journal of KISS : Software and Applications, 36.1. 92-108. 2009.
[2] Kilgarriff, Adam, and Christiane Fellbaum. "WordNet: An Electronic Lexical Database." (2000): 706-708
[3] https://wordnet.princeton.edu/
[4] Yoon Ae-sun. "Korean WordNet, KorLex 2.0 —A Language Resource for Semantic Processing and

Knowledge Engineering—." HAN-GEUL, .295 (2012.3): 163-201.
[5] 어휘지도(UWordMap)를 활용한 명사와 용언의 다의어 중의성 해소 / 신준철, 옥철영(울산대)
[6] 이성헌. 세종 전자 사전의 어휘 의미 부류 체계. 2007.
[7] Sangwook Kang, Minho Kim, Hyuk-chul Kwon, SungKyu Jeon, Juhyun Oh. "Word Sense Disambiguation of Predicate using Sejong Electronic Dictionary and KorLex." KIISE Transactions on Computing Practices, 21.7 (2015.07): 500-505.
[8] Joon-Choul Shin, Cheol-Young Ock. "Improvement of Korean Homograph Disambiguation using Korean Lexical Semantic Network (UWordMap)." Journal of KIISE, 43.1 (2016.01): 71-79.
[9] Young-Jun Bae, Cheol-Young Ock. "Semantic Analysis of Korean Compound Noun using Lexical Semantic Network(U-WIN)." Journal of KISS : Software and Applications, 40.12 (2013.12): 833-847.
[10] 박다솔, 차정원. “워드 임베딩을 이용한 세종 전자사전 확장”, 제28회 한글 및 한국어 정보처리 학술대회, pp.75-78. 2016.