

형태 의미 정보를 이용한 한국어 의미역 결정

박태호[○], 차정원

창원대학교 친환경해양플랜트FEED공학과[○], 창원대학교 컴퓨터공학과
taehope@changwon.ac.kr, jcha@changwon.ac.kr

Korean Semantic Role Labeling Using Word Sense

Tae-Ho Park[○], Jeong-Won Cha

Changwon National University Department of Eco-friendly Offshore Plant FEED Engineering[○],
Changwon National University Dept. Computer Engineering

요 약

의미역 결정은 서술어와 서술어의 수식을 받는 논항간의 의미 관계를 인식하고, 그 역할을 분류하는 자연어처리의 연구 분야이다. 의미역 결정을 위해 Word Embedding 정보나 단어 그룹화 정보 등 다양한 의미 정보를 활용하는 연구가 진행되었다. 본 연구는 이러한 의미 정보를 활용할 수 있는 정보 중 형태 의미 정보를 이용하여 의미역 분류를 시도하였다. 21세기 세종계획에서 구축한 형태 의미 분석 말뭉치 중 14,335문장을 학습에 사용하였으며, 단어의 의미 번호가 변함에 따라 논항 역할 결정의 변화를 관측하였다. 실험 결과로 77.36%의 의미역 결정 성능을 보였다.

1. 서 론

의미역 결정은 의미 분석을 위한 자연어 처리 작업의 일부이다. 의미역 결정은 서술어와 관계있는 논항의 역할을 결정하는 작업이다. 의미역은 문장의 구조가 바뀌어도 ‘행위주(ARGO)’ 나 ‘피동작주(ARG1)’ 와 같은 의미 논항은 바뀌지 않기 때문에 의미 분석에 필요한 정보를 제공한다.

의미역 결정을 위해서 다양한 의미 정보를 활용한 연구가 있었다[1,2]. [1]에서는 word embedding을 통한 word vector와 k-means 알고리즘을 통한 군집 정보를 사용하여 자질로 사용하였다. 이처럼 word vector나 군집 정보는 한국어 어휘에서 의미 정보를 추출한 새로운 정보이다. [1]의 연구에서는 word vector나 어휘 그룹 정보가 의미역을 결정하는데 도움이 됨을 증명하였다.

하지만 이러한 의미 정보를 사용함에도 해결하지 못한 의미역 결정 오류가 다수 존재한다. 의미역 결정에서 논항의 역할은 서술어의 의미에 따라 결정된다. 서술어는 각각의 의미와 활용에 따라 함께 사용되는 논항이 정해져있다. 이를 정리한 내용이 ‘격틀 사전’ 이다. [3,4]는 ‘격틀 사전’ 정보를 이용하여 의미역 결정을 진행하였다. 예문 (가)는 ‘격틀 사전’ 정보를 이용했을 때 나타날 수 있는 의미역 결정의 예이다.

(가) 이게 이제 내게 배당된 견뎌내야 할 청춘의 무게인지...

예문 (가)에서 ‘배당되다’ 라는 서술어가 세종 격틀 사전에서 ‘수여’ 라는 의미 그룹 정보를 지닌다. 따라서 ‘내게’ 라는 어절은 ‘수여’ 의 의미를 지니는 서술어와 관계 논항이라는 것을 자질로 사용할 수 있다. 의미 그룹 정보를 사용하기 이전에는 사용할 수 있는 정보가

‘-게’ 조사와 같이 형태 정보뿐이므로 ‘내게’ 라는 어절이 ‘ARG3(착점)’ 으로 분석되었지만 의미 그룹 정보를 사용하여 ‘ARG2(수혜자)’ 로 올바르게 분석할 수 있다.

이처럼 논항의 역할은 서술어의 격 정보에 따라 달라지는데 word vector나 군집 정보에서는 정확한 의미를 알기 어렵다. 다음은 동일한 어휘의 서술어가 의미에 따라 논항의 역할이 달라지는 예이다.

- (나) 전등불이(ARG1) 벽에 달려있다.
- (다) 자동차는(ARG0) 도로 위를 달려다.
- (라) 건물 사이로 하늘이(ARG1) 보이다.
- (마) 그는(ARG0) 경찰에게 신분증을(ARG1) 보였다.

(나)와 (다)에서 서술어는 ‘달리다’ 로 동일하지만 (나)의 ‘달리다’ 는 ‘달다03’ 의 피동사로 ‘물건을 일정한 곳에 걸거나 매어 놓는다’ 는 뜻을 지니고, (다)의 ‘달리다’ 는 ‘차, 배 따위가 빨리 움직인다’ 는 뜻을 지닌다. (나)에서 주격은 ‘전등불’ 이고, (다)에서 주격은 ‘자동차’ 이다. 하지만 두 문장에서 사용된 서술어의 의미가 다르고, 그 뜻에 따라 태(態, voice) 또한 다르다. 결과적으로 ‘전등불’ 과 ‘자동차’ 는 모두 주어이지만 ‘전등불’ 은 의미 역할이 ‘피동작주’ 가 되고, ‘자동차’ 는 ‘행위주’ 가 된다.

(라)와 (마) 역시 의미에 따라 피동과 사동으로 태(態, voice)가 바뀌는 서술어이다. (라)의 ‘보이다’ 는 ‘보다01’ 의 피동사이고, (마)의 ‘보이다’ 는 동일하게 ‘보다01’ 의 의미지만 사동사의 태(態, voice)를 지닌다. 이 외에도 위의 예처럼 의미 역할이 반대되는 경우는 아니지만 함께 사용되는 논항이 달라지는 것을 확인할 수 있는 예도 있다.

- (바) 방세가(ARG1) 몇 달을 밀렸다.
- (사) 길동은(ARG1) 철수에게(ARGO) 밀려 넘어졌다.

위의 예에서 (바)의 주격은 ‘방세’ 이고, (사)의 주격은 ‘길동’ 이다. (바)와 (사)의 예는 주격이 모두 ‘대상주(피동작주)’ 이다. 단 (사)의 ‘밀리다’ 의 경우 ‘밀다01’ 의 피동사로 서술어의 ‘행위주’ 가 ‘철수’ 로 나타난다. 이처럼 (바)와 (사)의 ‘밀리다’ 와 같은 서술어는 의미가 달라짐에 따라 함께 사용되는 논항이 달라지는 경우이다. 이러한 서술어는 의미에 따라 함께 사용되는 논항과 논항에 사용되는 조사를 알 수 있기 때문에 단순히 구문 정보와 조사 정보만을 사용할 때보다 정확하게 논항의 역할을 구분할 수 있게 도움을 줄 수 있다.

이처럼 의미의 애매성으로 인해 동일 어휘임에도 분석 결과가 달라지는 경우가 의미역 결정에서 나타난다. 따라서 의미역을 결정하는데 중요한 역할을 하는 술어의 의미 정보를 보다 명확하게 결정지를 필요가 있다. [5]에서는 기존에 사용하던 FrameNet의 의미 그룹 정보와 PropBank의 predicate senses 정보를 함께 사용하여 설계한 신경망 모델에서 predicate senses 정보가 의미역 결정에 도움이 됨을 증명하였다. 이에 본 논문에서는 한국어 의미역 결정에서 형태 의미 분석 정보가 의미역 결정에 미치는 영향을 알아보기 위해 실험을 진행하였다.

본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장에서는 실험 방식과 실험 결과를 분석하여 설명한다. 마지막으로 4장에서는 결론에 대해 기술한다.

2. 관련 연구

한국어 의미역 결정을 위해 의미를 표현할 수 있는 다양한 자질을 사용한 연구가 있었다. [3,4]는 세종전자사전에 수록된 격틀 사전의 격 정보를 활용하여 의미역 결정을 진행하였다. 격틀 사전에는 술어에 따라 함께 사용되는 체언의 종류와 조사 정보가 나타나 있다. [3]은 격틀 사전을 이용하여 self-training 알고리즘을 적용하였다. [4]는 의미역 결정에서 애매성이 크게 나타나는 특정 조사를 선택하여 의미역 결정 성능 향상 실험을 진행하였다. 애매성이 큰 조사로는 부사격 조사인 ‘에’, ‘로’, ‘에서’, ‘에게’ 가 있다. [2]는 한국어 의미역 결정을 위해 딥러닝(Deep Learning) 방법을 적용하였다.

영어권에서도 단어의 의미 표현 정보를 활용하기 위해서 word embedding을 통해 생성한 word vector로 다양한 방식의 딥러닝 알고리즘을 선보였다. [5]는 FrameNet의 정보와 PropBank의 정보를 조합하여 신경망의 학습 정보로 사용하였다. 실험의 결과로 FrameNet의 정보만 사용할 때보다 PropBank의 predicate senses 정보를 함께 사용할 때 성능이 향상됨을 증명하였다. 이는 단어의 의미애매성을 해소할 경우 의미역 결정에 도움이 됨을 증명한 것이다. [6]은 기존에 연구한 One-step 모델을 개선한 모델이다. one-step 모델은 SENNA[7]와 동일한 구조로 IOBES 태그 대신에 IOB 태그를 사용하여 레이블링한 모델이다. 이를 발전시켜 Two-step 모델을 개발하였는데,

IOBES 태그를 사용하며 논항 분류와 분리하여 두 단계에 걸쳐 레이블링을 진행한다. [8]은 다양한 방식의 심층신경망 모델을 이용하여 의미역 결정 성능을 측정하였다. [9]에서는 CNN과 LSTM 두 모델의 성능을 비교 측정하였다. 그러나 이들 방법은 모두 의미 정보를 직접 사용하지 않아 의미 정보가 미치는 영향을 확인하지 못했다.

3. 제안 방법

본 논문에서는 학습과 평가를 위해서 울산대 한국어 의미역 말뭉치 중 형태 의미 분석이 된 말뭉치를 사용하였다. 형태의미분석 말뭉치는 세종계획에서 구축한 말뭉치를 사용하였고, 의미 번호는 표준대국어 사전의 의미 번호를 따른다. 학습에 사용한 자질은 [1]에서 사용한 자질과 형태 의미 분석 말뭉치에서 추출한 어깨번호 정보를 함께 사용하였다. 표 1은 의미 표현 자질의 성능 검증을 위해 baseline 실험에 사용한 자질이다. 학습에는 CRF를 사용하여 모델을 생성하였으며, 5 배수 교차 검증으로 평가하였다. 학습에는 11,468문장을 사용하였고, 평가에는 2,867문장을 사용하였다.

표 1 baseline에 사용된 자질

자 질	내 용
Word	어절
phrase_tag	구문 분석 표지
H.POS	어절의 첫 형태소 품사
L.POS	어절의 마지막 형태소 품사
H.LEX	어절의 첫 형태소
L.LEX	어절의 마지막 형태소
Head_H.POS	서술어 어절의 첫 형태소 품사
Head_L.POS	서술어 어절의 마지막 형태소 품사
Head_H.LEX	서술어 어절의 첫 형태소
XSV XSA LEX	동사파상접미사 또는 형용사파상접미사
Named Entity	개체명
Distance	현재 어절과 서술어와의 거리
Active_Passive	능동/피동태 정보

표 2의 baseline 결과는 표 1의 자질만 사용한 실험 결과이다. p는 정확도를 의미하고, r은 재현율을 의미한다. 표 2는 baseline에서 격틀 사전의 의미 그룹 정보¹⁾, 형태 분석 말뭉치의 의미 번호 그리고 Word2Vector로 생성한 word vector를 이용한 cluster 정보를 각각 추가하여 실험한 결과이다. cluster 정보는 skip-gram으로 word

1) 21세기 세종계획에 포함된 정보로 단어들의 의미 유사에 따라 특정 그룹을 지어 분류한 정보. 예로 ‘반문하다’, ‘문의하다’ 등은 ‘질문’ 의미 그룹에 속하고, ‘격침시키다’, ‘박멸하다’ 등은 ‘과괴행위’ 의미 그룹에 속한다.

vector를 생성 후, k-means 알고리즘을 통해 200개의 cluster를 생성하여 해당 단어가 포함된 그룹 번호를 자질로 사용하였다. k를 100, 200, 300으로 하여 실험했을 때, 200이 가장 성능이 우수하여 k 값을 200으로 정하여 이후 실험에 적용하였다.

표 2 의미 표현 자질을 사용한 실험. 격틀 사전의 의미 그룹 정보(a), 체언 의미 번호(b), 용언 의미 번호(c), cluster(d)를 사용한 실험 결과.

구분	p	r	F1	증감
baseline	78.30	74.01	76.09	-
baseline + a	78.52	74.45	76.43	+0.34
baseline + d	78.54	74.58	76.50	+0.07
baseline + b + c	78.54	74.96	76.72	+0.12
baseline + a + d	78.32	75.29	76.78	+0.06
baseline + a + b + c	78.86	75.28	77.03	+0.25
baseline + b + c + d	78.92	75.45	77.14	+0.11
baseline + a+b+ c+d	78.77	75.79	77.25	+0.11
baseline + a + c + d	78.86	75.92	77.36	+0.11

실험 결과 모든 의미 표현 자질이 의미역 결정에 도움이 됨을 확인하였다. 그 중 형태 의미 분석 정보가 의미역 결정에 가장 큰 도움을 주었다. 모든 의미 표현 자질을 사용한 실험에서는 평균 성능이 77.25%로 나타났다. 추가적으로 서술어에만 의미 번호 정보를 사용했을 때 성능이 향상되어 77.36%로 나타났다.

4. 결론

본 논문에서는 단어의 의미에 따라 논항이 역할이 바뀌는 것을 확인하고, 단어 의미 정보의 활용에 따라 의미역 결정 성능 변화를 확인하기 위해 실험을 진행하였다. 이를 확인하기 위해서 형태 의미 분석 말뭉치를 사용하였으며, CRF의 학습 자질로 단어 의미 번호를 사용하였다. 또한 서술어의 의미에 따라 태(態, voice)의 변화나 격틀의 변화를 구분하여 함께 사용된 논항의 역할이나 사용 유무를 통계를 내어 관측하였다. 학습에는 11,468문장과 평가에는 2,867문장을 사용하여 5배수 교차 검증으로 성능을 측정하였으며, 평균 성능은 77.25%로 나타났다. 또한 서술어만 의미 번호 정보를 사용하였을 때는 77.36%로 성능이 측정되었다.

향후 연구로 서술어에만 의미 번호를 사용했을 때 성능이 향상되는 정확한 이유를 분석하는 작업을 진행할 예정이다. 본 논문에서 의미 번호 정보를 사용하였지만 세중 형태의미분석 말뭉치에서 모든 단어에 의미 번호가

부착되어 있지 않기 때문에 실험에 사용된 단어들 중에 의미중의성이 해소되지 않은 단어도 많이 포함되어 있다. 따라서 의미 번호가 부착되지 않은 단어의 중의성을 해소하는 문제를 함께 진행할 예정이다. 또한 AMR(Abstract Meaning Representation) 모델 생성을 위해 구문 분석, 서술어 의미 결정, 의미역 결정이 동작하는 구조를 설계하기 위한 연구를 진행할 계획이다.

참고 문헌

- [1] Tae-Ho Park, Jeong-Won Cha. "Feature Selection for Korean Sematic Role Labeling Using CRFs." Communications of the Korean Institute of Information Scientists and Engineers, 34.8 (2016.8): 37-41.
- [2] Jangseong Bae, Changki Lee and Soojong Lim, "Korean Semantic Role Labeling using Deep Learning", 한국정보과학회 2015 한국컴퓨터종합학술대회 논문집, 2015.06, 690-692, 2015.
- [3] Byoung-Soo Kim, Yong-Hun Lee and Jong-Hyeok Lee, "Unsupervised Semantic Role Labeling for Korean Adverbial Case", Journal of KISS : Software and Applications - 2007.6 34(2), 2007.2, 112-122.
- [4] Hyun-Ki Jung and Yu-Seop Kim, "Semantic Role Labeling of Korean Adverbial Arguments by using the Expanded Case Frame Dictionary", Journal of Korean Institute of Information Technology 9(10), 2011.10, 167-176.
- [5] FitzGerald, Nicholas, et al. "Semantic Role Labeling with Neural Network Factors." EMNLP. 2015.
- [6] Fonseca, Erick R., and Joao Luis G. Rosa. "A two-step convolutional neural network approach for semantic role labeling." Neural Networks (IJCNN), The 2013 International Joint Conference on. IEEE, 2013.
- [7] Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning." Proceedings of the 25th international conference on Machine learning. ACM, 2008.
- [8] Wang, Zhen, et al. "Chinese Semantic Role Labeling with Bidirectional Recurrent Neural Networks." EMNLP. 2015.
- [9] Zhou, Jie, and Wei Xu. "End-to-end learning of semantic role labeling using recurrent neural networks." ACL (1). 2015.