한국어 의미역 결정을 위한 자질 정보 확장

조병철○, 석미란, 김유섭
한림대학교, 융합소프트웨어학과
max91128@naver.com, smr4880@hanmail.net, yskim01@hallym.ac.kr

Expansion of Feature Information for Korean Semantic Role Labeling

Byeong-Cheol Jo○, Mi-Ran Seok, Yu-Seop Kim
Dept. of Convergence Software, Hallym University

요 약

의미역 결정은 주어진 술어와 의존 관계에 있는 여러 논항들과 그 술어간의 의미 관계를 결정하는 것이다. 의미역 결정은 보통 대량의 말뭉치를 이용하여 분류의 관점에서 문제를 해결하고자 한다. 본 논문에서는 한국어 구문 표지 부착된 말뭉치에 구축한 의미역 표지 부착 말뭉치 10,000 문장을 이용한 자동 의미역 결정 방법을 제안한다. 특히, 한국어는 그 특성상 조사와 어미가 문법 관계뿐만 아니라 의미 관계 설정에도 매우 중요한 역할을 하기 때문에 기존의 의미역 결정 연구에서 미비했던 부분인 조사와 어미 정보를 개선하여 새로운 자질 (features) 로 설계하여 의미역 결정을 시도하였다. 기존의 다른 언어에서의 의미역 결정 연구에서 사용된 자질에 본 논문에서 제시된 접사 정보에 기반한 자질을 추가하게 되면 약 77.9%의 F1 점수를 얻을 수 있었는데, 이는 기존 연구에 비하여 약 10% 포인트 향상된 결과이다.

주제어: 자질(Feature), 한국어 접사, 조사, 어미

1. 서론

의미역 결정은 문장의 서술어와 그 서술어와 의존관계에 놓여 있는 논항들 사이의 의미관계를 결정하는 문제이다. 의미역 결정은 문장의 서술어와 논항들 사이의 '주어', '목적어'와 같은 문법 관계를 '행동주', '경험주', '대상' 등과 같은 의미 관계로 사상(mapping)하는 문제로 볼 수 있으며, 일반적으로 본 과정은 구문 분석을 수행한 후에 수행된다[1,2]. 즉, 의미역 결정이란 서술어-논항의 적합한 관계를 결정하는 것이다. 의미역 결정은 기계 번역, 정보 추출, 질의응답과 같은 다양한 자연언어 처리 응용에서 각각의 성능 향상을 위해 사용될 수 있다.

문서 내에 존재하는 중요한 개체 간의 관계를 자동으로 추출하는 작업은 핵심적인 작업으로 꼽히면서도 가장 어려운 작업으로 알려져 있다[3]. 이러한 개체 간의 관계를 추출함에 있어서 문서 내에 존재하는 다양한 자질들을 활용 할 수 있다.

본 논문에서는 의미역 표지 부착 말뭉치로 Proposition Bank[4]체계를 따르는 한국어 PropBank를 이용하여 의미 표지부착 말뭉치를 자동 구축할 수 있는 한국어 의미 역 결정 시스템을 구현하였다. 한국어는 하나의 어근에 다양한 접사가 결합되어 다양한 문법적,의미적인 변이를 하는 특징을 가지고 있다. 이에, 본 연구에서는 한국어의 접사 정보를 자질로 표현하여 의미 역 결정의 성능을 향상시키고자 하였다. 접사 중 조사가 의미역 결정에 있어서 큰 역할을 하기 때문에 새로운 자질(features)로 만들었다. 기존 연구에서 제안된 자질에 본 연구에서 제안한 접사 자질들을 결합한 결과 약 10% 포인트 정도의 성능 향상을 보일 수 있었다.

2. 관련 연구

의미역 결정 연구는 말뭉치 기반 방법(corpus based method)[6]과 격틀사전 기반방법(case frame based method)[7]이 있다.

격틀사전 기반 방법은 술어-논항 관계를 기술한 격틀을 이용하여 의미역이 결정되기 때문에 높은 정확률을 보인다. 하지만 격틀 사전 구축이 어렵고 격틀에 기술되지 않은 문장 형태에는 적용이 불가능 하다는 단점이 있다[1.2].

말뭉치 기반 방법은 말뭉치에 의미역을 부착한 후 기계학습 (Machine Learning), 특히 지도 학습 (Supervised learning) 을 통해 의미역을 결정하는 방법이다. 영어권에서는 PropBank 등의 의미역 부착 말뭉치를 이용하는 방법이 많이 연구 되어 있지만, 한국어의 경우에는 의미역이 태깅된 말뭉치를 구하거나 구축하기 어려워서 격틀 사전을 이용하는 방법이 주로 연구되어 왔다[1,2].

본 논문에서는 말뭉치 기반 방법을 사용한 자동 의미역 결정 방법을 제안한다.

3. 자질 (Feature)

한국어의 경우에는 구문 또는 의미 분석을 하는데 있어서 조사나 어미와 같은 접사가 매우 중요한 역할을 한다. 특히 한국어는 어순에 별다른 제약이 없기 때문에 기존 영어권 언어와 같이 어휘의 위치 정보를 의미 역 결정에 사용할 수 없다. 반면에 조사와 어미가 다양한 형태로 결합되고 이를 통하여 어휘들의 구문 및 의미가 결정되는 경우가 많아 접사 정보의 활용이 의미 역 결정의 성능에 큰 영향을 미칠 것으로 추정된다.

	자질(Feature)	설명(Desription)
기본 자질 (General Features)	A_stem/P_stem	논항과 술어의 어근
	A_POS_LV1/P_POS_LV1	A_stem과 P_stem 품사의 큰 단위
	A_POS_LV2/P_POS_LV2	A_stem과 P_stem 품사의 작은 단위
	A_CASE/P_CASE	A_stem과 P_stem의 격조사
	A-LeftSiblingStem	논항의 왼쪽 형제의 어근
	A-LeftSiblingPOS_LV1	왼쪽 형제 어근의 큰 품사
	A-LeftSiblingPOS Lv2	왼쪽 형제 어근의 작은 품사
	A-RightSiblingPOS Lv1	오른쪽 형제 논항의 큰 품사
	A-RightSiblingPOS Lv2	오른쪽 형제 논항의 작은 품사
	P-ParentStem	서술어의 부모의 어근
	P-ChildStemSet	서술어의 자식의 어근 집합
	P-ChildPOSSet Lv1	P-ChildStemSet의 큰 품사 집합
	P-ChildCaseSet	P-ChildStemSet의 격조사 집합
새로운 자질 (New Features)	조사	조사의 단어
	조사_80	조사 의미역 빈도 80%이상
	한단어_여부	논항이 하나의 단어이면 1
	한단어_어근	접사가 없을때의 단어
	어미_여부	어미가 존재하면 1 아니면 0
	어미_어근	어미가 존재하면 어미의 어근
	어미_클래스	어미의 분류

[표 1] 기존 자질과 새로운 자질

조사는 품사 중 하나로 문장 내에서 주로 체언에 연결되어 뒤에 오는 다른 단어에 대하여 가지는 문법적 관계를 표시하거나, 특별한 의미 요소를 첨가 하여 주는 기능을 지닌 형태이다. 대표적으로는 '~이(가)', '~을(를)', '~에게' 등이 있다.

조사	의미역	확률
를	ARG1	97.5%
에선	ARGM-LOC	100%
처럼	ARGM-EXT	85.6%
보다	ARGM-EXT	97.7%
에_대하_어	ARGM-ADV	85.7%

[표 2] 조사 빈도 80% 이상 의미역. 수동 의미표지 부착한 말뭉치에서 조사가 특정 의미역으로 태깅된 확률이 80%이상 인 것

어미란 어형변화를 가지게 되는 단어에서 어근을 제외하고 어말 위치에 오는 형태소이다.

본 논문에서는 조사와 어미, 그리고 한 단어로 이루어진 논항에 대한 자질을 설계하였다. [표 1]에서는 본 연구에서 사용된 모든 자질이 나열되어 있다. 자질들은 기존 연구에서 활용되었던 기본 자질과 본 연구에서 제시된 새로운 자질들로 구성된다.

기본 자질(General Features): 기본 자질로는 [5]에서 사용한 자질들이 사용되었다.

● Stem: 논항과 술어의 어근이다.

● Pos_LV1: 명사, 동사, 형용사, 부사 등의 기본 품사를 나타낸다.

● Pos_LV2: 용언불가능보통명사, 성상상태부사,문장양태부사,문장접속부사,지시시간부사,양수사,부정부사,지시처소부사,성상정도부사,기타의존명사,성상의태부사,일반동사,성상형용사와 같은 하위 범주를 나타낸다.

● CASE: 격조사인 주격조사, 보격조사, 목적격조사, 부사격조사, 관형격조사를 나타낸다.

새로운 자질(New Features): 한국어에서는 접사가 문장의 의미를 파악하는데 중요하기 때문에 조사, 어미에 초점을 맞추었고, 기존 연구에서 미비했던 부분을 개선하여 아래와 같은 자질(Features)로 설계하였다.

● 조사: 조사 단어 자체를 자질로 사용하였다. 예를 들어 '~이(가)', '~을(를)', '~에게'를 나타낸다.

● 조사_80 : 조사가 특정 의미역으로 매핑 되는 확률이 80% 이상인 것을 나타낸다[표 2]. 예를 들어 조사 '를'이 'ARG1'으로 의미역 부착되는 빈도는 '97.5%'이다.

● 한단어_여부: 논항이 접사가 없이 하나의 단어로 이루어졌는지를 나타낸다.

● 한단어_어근: 논항이 한단어로 이루어진 경우 단어

자체를 자질로 사용한다. 예를 들면, '그러나', '그리고', '반면', '또는' 등이 포함된다.

● 어미_여부: 어미가 존재하는지 여부를 나타낸다.

● 어미_어근: 어미가 존재하면 어미의 어근은 어떤 것인지를 나타낸다. 예를 들어, 먹(다), 걷(다), 하(다) 등이 있다.

● 어미_클래스: 과거시제선어말어미, 종속연결어미, 대등연결어미, 부사형전성어미, 의문형종결어미, 명사형전성어미, 높임선어말어미, 평서형종결어미를 나타낸다.

본 논문에서는 더 좋은 결과를 위해 기본 자질과 개선된 새로운 자질을 결합하였다.

4. CRF

CRF (Conditional Random Field) 는 통계적 모델링 방법 중 하나로 패턴 인식과 기계 학습과 같은 구조적 예측에 사용된다. 일반적인 분류자 (Classifier) 가 이웃하는 표본을 고려하지 않고 단일 표본 라벨을 예측하는 반면 CRF (Conditional Random Field) 는 이웃하는 표본을 고려하여 예측한다. CRF는 자연언어로 된 글 또는 생물학적 서열정보 일련의 데이터에 대한 라벨 예측, 분석에 사용 되기도 한다.

본 논문에서는 crf suite[1]를 사용하였고, CRF 알고리즘 중 평균 퍼셉트론 (average perceptron) 을 사용하여 의미역을 예측하였다.

5. 실험 결과

	의미역	Precision	Recall	F1
General	ARG0	74.63%	80.95%	77.66%
	ARG1	82.52%	84.58%	83.54%
	ARG2	38.72%	35.81%	37.21%
	ARG3	27.59%	19.42%	22.79%
	ARGM-DIS	67.65%	83.98%	74.93%
	…	…	…	…
	total	**67.27%**	**70.5%**	**68.5%**
General + new feature	ARG0	81.42%	85.43%	83.38%
	ARG1	88.02%	89.02%	88.52%
	ARG2	62.51%	59.88%	61.17%
	ARG3	73.68%	61.17%	66.84%
	ARGM-DIS	77.55%	87.97%	82.43%
	…	…	…	…
	total	**77.60%**	**78.70%**	**77.90%**

[표 3] CRF 실험 결과

의미역 태깅 말뭉치 10,000개 중 8,000개를 학습 데이터(train data)로, 실험 데이터(test data)로는 2,000개를 평가를 위해 사용하였다.

실험은 기본 자질과 새로운 자질을 합하여 실험을 진행하였다. 동일한 데이터를 가지고 기본 자질로만 의미역 결정 실험을 했을 경우에는 약 68.5%의 F1 score가 나왔으나 새로운 자질들을 추가하였을 경우에는 약 77.9%의 F1 score를 보였다. 본 연구에서는 한국어에 특화된 자질을 사용함으로써 약 10%의 향상된 결과를 보여줄 수 있었다.

6. 결론

본 논문에서는 지금까지 태깅된 의미역 말뭉치 데이터를 이용하였고, 말뭉치 기반방법으로 CRF 알고리즘을 사용하여 자동으로 의미역을 부착하였다.

한국어에는 의미 분석을 하는데 있어 조사와 어미가 중요해서 기존 자질보다 약 10% 향상된 결과를 얻을 수 있었다.

향후에는 보다 더 좋은 자질을 적용시켜 본 문제에 가장 적합한 방법을 찾고자 한다.

참고문헌

[1] 김병수, 이용훈, 나승훈, 김준기, 이종혁, "부트스트래핑 알고리즘을 이용한 한국어 격조사의 의미역 결정", 한국컴퓨터종합학술대회,vol.33, no.1,pp.4-6, 2006.

[2] 이창기, 임수종, 김현기, "Structural SVM 기반의 한국어 의미역 결정", 정보과학회 논문지, vol.42, no.2,pp.220-226, 2015.

[3] Bunescu, R. C. and Mooney, R. J., "A Shortest Path Dependency Kernel for Relation Extraction," Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp.724-731, 2005.

[4] M. Palmer, D. Gildea, and Paul Kingsbury, "The Proposition Bank: An Annotated Corpus of Semantic Rules", Computational linguistics, 31(1), 71-106, 2005.

[5] Young-Bum Kim, Heemoon Chae, Benjamin Snyder and Yu-Seop Kim*,"Training a Korean SRL System with Rich Morphological Features",The 52nd Annual Meeting of the Association for Computational Linguistics, 637-642, 2014.

[6]Kadri Hacioglu, Sampeer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky, "Semantic role labeling by tagging syntactic chunks," In Proceedings of CoNLL 2004 Shared Task, 2004.

[7] Kurohashi, S, and Nagao, M. "A Method of Case Structure Analysis for Japanese Based on Examples in Case Frame Dictionary," IEICE Transaction Information and System, Vol.E77-D, No.2, pp. 227-239, 1994.

1) http://www.chokkan.org/software/crfsuite