

# CRFs 기반의 한국어 의미역 결정

박태호, 차정원  
창원대학교, 적응지능연구실  
taehope@changwon.ac.kr, jcha@chagwon.ac.kr

## Korean Sematic Role Labeling Using CRFs

Tae-Ho Park, Jeong-Won Cha  
Changwon National University, Adaptive Intelligent Research Lab.

### 요약

의미역 결정은 서술어와 논항들 사이의 의미 관계를 결정하는 문제이다. 의미역 결정을 위해 구구조 정보와 의존 구조 정보 등의 다양한 자질에 대한 실험이 있었다. 논항은 구문 구조에서 얻을 수 있는 서술어와 논항 관계에 많은 영향을 받지만 구문 구조가 변경되어도 변하지 않는 논항의 의미로 인해 의미역 결정에 어려운 점이 있다. 본 논문에서는 한국어 의미역 결정 문제를 위해 Korean Propbank 말뭉치와 직접 구축한 의미역 말뭉치를 학습 말뭉치로 사용하였다. 본 논문에서는 이전에 연구된 구문 정보와 그 외의 자질들에 대한 성능을 검증하였다. 본 논문에서 제시하는 자질들의 성능을 검증하기 위해 CRF를 사용하였고, 제시된 새로운 자질을 사용하여 논항의 인식 및 분류에서 76.25%(F1)의 성능을 보였다.

주제어: 의미 분석, 한국어 의미역 결정, Korean Propbank, CRF

### 1. 서론

의미 분석은 서술어와 논항 사이의 의미 관계를 결정하는 것이다. 의미역 결정은 구문 분석과 비슷하지만 구문 분석은 ‘주어’ 나 ‘목적어’ 등의 문법 관계를 분석하는 반면 의미역은 논항이 문장에서 지니는 의미에 더 초점이 맞추어져 있다. 의미역은 논항의 의미에 따라 ‘행위주’, ‘피동작주’, ‘경험주’ 등으로 분석할 수 있다. <표1>은 본 논문에서 사용하는 한국어 논항의 분류와 의미를 설명한다. 의미역에서 논항들은 서술어와 관계가 밀접하며 구문 분석에서 분석되는 ‘주어’ 나 ‘목적어’ 등과 유사하다. 이미 구문 분석 정보를 이용한 의미역 결정 실험이 진행되었다[1-5]. 하지만 구문 정보를 이용하더라도 의미역 결정에는 어려운 점이 많다. 그 이유는 서술어의 의미에 따라 문장의 형태가 다를 경우, 구문 정보는 변하지만 논항 정보는 변하지 않는다. 아래는 이를 설명하기 위한 예이다.

- (가) 해커는 서버를 공격했다.
- (나) 서버는 해커에게 공격받았다.

위 문장은 서술어 ‘공격.01’ 이 나타난 문장으로 (가)는 ‘공격하다’, (나)는 ‘공격받다’ 의 형태로 사용되었다. 두 문장에서 주어(SBJ)는 (가)에서는 ‘해커’ 이고, (나)에서는 ‘서버’ 로 서로 다르다. 하지만 ‘공격’ 을 행한 행위주(Agent)인 ‘해커’ 와 ‘공격’ 을 받는 피동작주(Patient)는 ‘서버’ 로 두 문장이 동일하다. 따라서 구문 정보 외의 자질로 격틀 사전 정보 등의 새로운 자질을 사용한 연구가 진행되었다[6,7].

의미역 결정은 정보추출, 의미분석, 질의응답 등 다양한 자연어처리 분야에서 활용할 수 있다. 최근 정보추출

과 질의응답 시스템 분야에 대한 연구가 활발히 진행되고 있고, 의미역 결정 연구 또한 꾸준히 진행되고 있다.

본 논문에서는 Korean Propbank[8]를 사용하여 CRF 기반의 한국어 의미역 결정 시스템을 구축했다. 본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장에서는 의미역 결정에서 사용하는 자질과 본 연구에서 검증하고자 한 자질에 대해 소개한다. 4장에서는 새로운 자질을 통한 학습 결과를 분석하고, 마지막 5장에서 결론에 대해 기술한다.

표 1. 한국어 논항의 분류와 의미

분류	의미	분류	의미
ARG0	Agent	ARGM-EXT	Extent
ARG1	Patient	ARGM-INS	Instrument
ARG2	Start point/ Benefactive	ARGM-LOC	Locative
ARG3	Ending point	ARGM-MNR	Manner
ARGM-ADV	Adverbial	ARGM-NEG	Negation
ARGM-CAU	Cause	ARGM-PRD	Predication
ARGM-CND	Condition	ARGM-PRP	Purpose
ARGM-DIR	Direction	ARGM-TMP	Temporal
ARGM-DIS	Discourse		

### 2. 관련 연구

영어권에서는 의미역 결정을 위해 CoNLL shared Task 의 의미역 말뭉치를 이용하여 2004년부터 꾸준히 연구되었다[9]. 초기에는 구구조를 기반으로 하여 의미역을 결

정하였고, 이후로 의존구조나 구문 정보 등을 이용한 의미역 결정 연구가 진행되었다[1-5]. 또한 의미역 결정에 도움이 되는 다양한 자질들을 분석한 연구로 서술어의 형태나 개체명 정보 등의 자질이 의미역 결정 성능 향상에 도움을 준다는 것이 증명되었다[10].

하지만 구구조 정보와 의존 구조 정보만으로는 성능을 올리기에 부족하여 이전에 사용한 자질들의 최적의 조합을 찾는 연구가 진행되었다[11,12].

한국어 의미역 결정에는 격틀 사전 정보를 이용한 연구가 있었다[6,7]. 격틀 사전에는 서술어의 각 의미에 따른 논항과의 관계가 나타나 있다. 이때 각 논항의 역할과 문장에서 사용하는 형태가 구문 분석의 기능 태그와 조사 정보가 함께 나타나 있다. [6]은 격틀 사전을 이용한 연구로 비지도 학습의 하나인 self-training 알고리즘을 이용하였다. [7]은 의미역 결정에 애매성이 큰 부사격 조사 ‘-에’, ‘-로’, ‘-에서’, ‘-에게’ 를 대상으로 실험하였다. [13,14]는 Structural SVM을 이용한 연구로 [13]은  $i-1$ 번째 레이블과  $i$ 번째 레이블이 독립적이지 않고 영향을 미친다는 것을 기반으로 순차적 레이블링 기반으로 의미역을 결정하였으며, [14] 또한 순차적 레이블링 기반으로 격틀 사전과 Korean Propbank에서 추출한 서술어 정보를 통해 서술어 인식/분류와 논항 인식/분류를 함께 진행하였다. [15]은 다양한 형태소 정보를 활용한 논문으로 조사 정보를 세분화하여 조사의 유무, 조사의 형태소, 조사의 형태소 품사, 조사와 구문 정보의 조합을 사용하였고, 이와 함께 어미 정보를 사용하여 CRF 학습을 하였다. 최근에는 자질 선택과 조합에 대한 어려움을 해결한 딥 러닝(Deep Learning)방법을 이용한 연구도 진행되었다[16].

### 3. 의미역 결정 자질 분석

#### 3.1. 기준 시스템

CRF를 이용한 한국어 의미역 결정 성능을 알아보기 위해 다음의 자질로 기준 성능을 측정하였다.

- 구문 기능 복합 레이블 정보
- 의존 구문 트리에서 서술어의 부모 노드 정보
- 서술어의 형태소, 형태소 품사 정보
- 서술어와 현재 어절의 형태소, 형태소 품사, 의존 관계 조합 정보
- 서술어와의 거리 정보

#### 3.2. 검증 후보 자질

본 논문에서는 기존의 한국어 의미역 결정에 도움을 줄 수 있는 새로운 자질을 찾고 이를 검증하기 위해서 다음의 자질을 추가한 실험을 진행하였다.

- 개체명 정보
- 구문 체크 정보
- 구문 레이블과 기능 레이블을 분리한 정보
- Word Vector 정보
- Word Vector를 이용한 군집(cluster) 정보
- 동사과생접미사 정보

이미 영어권에서 의미역 결정 방법에서 개체명과 시제 단어 사전, 장소 단어 사전을 사용하고 있다. 개체명 정보는 수량 또는 날짜 등의 정보를 통해 시간 의미의 ‘ARGM-TMP’에 도움이 될 것으로 생각되었고, 장소를 나타내는 개체명은 장소 의미의 ‘ARGM-LOC’을 결정하는데 도움이 될 것으로 예상하였다.

기준 실험에서 사용한 구문 분석 정보는 구문과 기능 정보가 결합된 복합 레이블을 사용하였는데, 이를 분리하였을 때의 성능을 알아보기 위해 구문과 기능 정보를 분리하여 자질 정보로 사용하였다.

또한 최근 Word Embedding이나 Feature Embedding을 통해 구한 정보를 이용한 실험이 활발히 진행되고 있어 SKIP GRAM과 CBOW를 이용한 Word Vector 정보와 이를 K-means 알고리즘을 이용하여 군집 정보를 구했다. K-means 알고리즘에서 K는 CBOW와 SKIP GRAM에서 각각 100, 200을 사용하였다. Word Vector 값을 학습 자료로 사용할 때에는 소숫점 넷째자리에서 반올림한 값을 사용하였다.

마지막으로 동사과생접미사 정보를 사용하였다. 영어권에서 사용하는 자질 중 ‘Voice’는 동사의 수동태와 능동태를 구분하는 자질로 이미 성능 향상에 도움이 되는 것이 증명되었다[10]. 한국어 역시 이와 다르지 않을 것이라고 생각되지만 아직 서술어의 수동태와 능동태 정보가 구축되지 않은 상황에서 이를 활용하기 위해 동사과생접미사 중 ‘-당하’와 ‘-시키’가 자질로써 사용 가능하다고 판단하였다. 1장에서 보인 예문 중 (나)의 서술어 ‘공격받다’는 ‘공격당하다’와 같은 의미로 쓸 수 있다. 따라서 서술어에 포함된 동사과생접미사를 자질로 사용하였다.

## 4. 실험 및 토의

실험 말뭉치는 Korean Propbank 말뭉치를 사용하였으며, 학습 모델로는 CRF를 선택하였다.

새로운 자질에 대한 검증에 앞서 이전에 연구된 자질만을 사용하여 CRF로 성능을 확인하였다. 기준 성능은 다음의 <표2>와 같다.

표 2. 기준 성능표

	Precision	Recall	F1
Baseline	72.53	67.48	69.99

이후 새로운 자질을 추가하여 학습하였으며, 각 자질 별로 다음과 같은 결과를 얻었다.

**개체명** 개체명 정보를 추가했을 시, 인명(PER)과 장소(LOC), 수량(QT), 시간(TM), 수(NUM) 정보로 ARG0와 ARGM-LOC, ARGM-TMP의 인식 및 분류 성능이 상승하였다. 인명 정보를 지닌 어절이 행위주 또는 피동작주로 자주 나타났다. 장소 정보는 장소 의미역에서 주로 나타났다. 특히 시제 의미역에서는 수량과 수의 개체명이 시간을 나타내는 정보와 구별할 수 있는 정보로 활용되어 시제

의미역 분류 오류를 크게 줄였다.

**구문과 기능 레이블** 본 연구에서는 기존 실험에서 구문과 기능 레이블을 하나로 묶어 복합 레이블을 사용한다. 하지만 의미역은 기능 레이블과 밀접한 관계가 있기 때문에 구문과 기능 레이블을 분리하여 학습해 보았다. 하지만 성능이 오를 것이라고 생각한 처음 예상과는 다르게 오히려 성능이 떨어졌다. 기능 레이블이 주는 긍정적인 영향보다 분리되어 혼자 남은 구문 레이블 정보가 노이즈로 작용하였기 때문에 성능이 떨어진 것으로 분석된다.

**Word Vector** Word Vector는 CBOW와 SKIP GRAM 두 가지를 사용하여 추출했다. Vector 값은 각각 총 50개로 소수점 넷째자리에서 반올림하여 사용하였다. 하지만 입력 자질을 하나의 고유한 정보로 인식하는 CRF에서는 Vector값에서 유사함을 찾지 못하고 소숫점 한 자리만 달라도 서로 다른 정보로 인식하여 성능은 떨어졌다.

**군집 정보** Word Vector값은 CRF에서 학습하기 위한 자질로는 부적합하여 Word Vector값으로 K-means 알고리즘을 통해 형태소의 군집을 구했다. K 값은 CBOW와 SKIP GRAM 각각 100, 200으로 하였다. 모든 군집 정보가 성능 향상에 도움이 되었으며, CBOW는 K가 200일 때 가장 성능이 많이 올랐고 SKIP GRAM도 K가 200일 때 가장 성능이 좋았다. 최종적으로는 CBOW보다 SKIP GRAM이 조금 더 성능이 좋았다.

**동사파생접미사** 동사파생접미사는 명사와 동사파생접미사가 연결되어 서술어로 사용될 경우 서술어에 존재하는 동사파생접미사를 추출하여 자질 정보로 사용하였다. 학습에 사용된 동사파생접미사는 총 4종류로 ‘-당하’, ‘-되’, ‘-시키’, ‘-하’가 있다. 학습 결과로 ‘-당하’ 또는 ‘-시키’가 연결된 서술어에서 주어를 피동작주로 분류하는데 도움을 주어 성능이 향상되었다.

학습은 Korean Propbank의 newwise 의미역 말뭉치 4,151문장과 직접 레이블링한 의미역 말뭉치 1,237 문장을 함께 사용하였다. 말뭉치 전체에서 나타난 논항 중 ARG0는 19.6%, ARG1는 34.92%, ARG2는 12.17%, ARG3은 1.23%로 나타났다. 전체 문장을 5등분하여 교차 검증을 하였으며 성능은 다음 <표3>과 같다. 교차검증에서 사용한 평가 문서에 포함된 논항은 실험1부터 실험5까지

각각 6,554개, 6,368개, 6,589개, 6,719개, 6,618개이다.

표 3. 한국어 논항 인식 및 분류의 5묶음 교차 검증

실험	Precision	Recall	F1
실험 1	78.32	74.96	76.60
실험 2	78.35	75.75	<b>77.02</b>
실험 3	76.82	73.60	75.18
실험 4	77.88	74.31	76.05
실험 5	78.33	74.58	76.41
평균	77.94	74.64	<b>76.25</b>

## 5. 결론

본 논문을 통해 구문 분석 정보나 격틀 사전 정보 외에 개체명과 Word Vector를 기반으로 한 K-means 알고리즘으로 분류한 군집 정보, 동사파생접미사가 한국어 의미역 결정에 도움이 된다는 것을 확인하였다. 새롭게 추가한 자질로 학습한 CRF 모델을 통해 5묶음 교차 검증 결과 논항 인식 및 분류 성능에서 76.25(%)의 성능을 보였다. 또한 의미역 결정 연구에 관한 이전 연구와 비교하였을 때, 영어권에서의 성능보다는 낮은 성능을 보였지만 한국어 의미역 결정에서는 이전 연구보다 나은 성능을 보였다<표4>. 또한 Deep Learning으로 학습한 의미역 결정 모델보다 성능이 높게 나왔는데, 이는 Deep Learning에서 가중치 조정을 통해 스스로 자질의 조합을 결정하지만 CRF에서는 관계가 분명한 조합을 명확히 해 줌으로써 차이가 나타난다. 하지만 Word Embedding이나 Feature Embedding을 통해 아직 활용하지 못하는 자질을 Deep Learning에서 활용할 수 있도록 하는 연구가 꾸준히 진행되고 있어 어느 모델이 더 우수하다고 결론을 지을 수는 없다고 생각된다.

향후 연구로 이전 연구에서 이미 활용한 자질 중 본 실험에서 누락된 격틀 사전 정보를 이용하여 성능을 검증할 예정이다. 격틀 사전 이용시 명사파생접미사 정보로 찾지 못한 서술어 의미에 따른 논항 구조를 보다 명확히 하여 성능이 향상될 것으로 예상된다. 또한 본 논문에서 사용한 Korean Propbank 말뭉치는 경제 신문 말뭉치로 다양한 도메인에 적용하기 위한 말뭉치 구축을 진행하고 있다.

표 4. 한국어 의미역 결정 성능 비교

	Precision	Recall	F1
English SRL[5]	82.95	78.40	80.61
self-training + Frame Dictionary(case : ‘-에’, ‘-로’, ‘-에서’, ‘-에게’)[6]	83.00	-	-
Sequence Labeling + Structural SVM[13]	-	-	74.77
Structural SVM[14]	81.16	71.53	76.04
Deep Learning[16]	-	-	75.44
Our CRF model(baseline)	72.53	67.48	69.99
Our CRF model(with new feature)	77.94	74.63	76.25

## 참고문헌

- [1] Daniel Gildea and Daniel Jurafsky. "Automatic labeling of semantic roles", Association for Computational Linguistics, Computational Linguistics, 28(3):245-288. 2002.
- [2] Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James Martin, and Dan Jurafsky. "Shallow semantic parsing using support vector machines. Technical", Report TR-CSLR-2003-1, Center for Spoken Language Research, Boulder, Colorado. 2003.
- [3] Kadri Hacioglu. "Semantic role labeling using dependency trees", In Proceedings of COLING, Geneva, Switzerland. 2004.
- [4] Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James Martin, and Daniel Jurafsky. "Semantic role labeling by tagging syntactic chunks", In Proceedings of CoNLL-2004, Shared Task - Semantic Role Labeling. 2004.
- [5] Richard Johansson and Pierre Nugues, "Dependency-based semantic role labeling of PropBank", EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing Pages 69-78, 2008.
- [6] 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, 2007.
- [7] 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, 2011.
- [8] Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon, Korean Propbank, [Online].  
Available: <http://catalog.ldc.upenn.edu/LDC2006T03>
- [9] Xavier Carreras and Lluís Màrquez, "Introduction to the CoNLL-2004 shared task: semantic role labeling", CONLL '04 Proceedings of the Ninth Conference on Computational Natural Language Learning, 2004.
- [10] Sameer Pradhan, Wayne Ward and Daniel Jurafsky, "Semantic role labeling using different syntactic views", Proceeding ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics Pages 581-588, 2005.
- [11] Weiwei Sun, "Improving Chinese semantic role labeling with rich syntactic features", ACLShort '10 Proceedings of the ACL 2010 Conference Short Papers Pages 168-172, 2010.
- [12] Shiqi Li, Qin Lu and Hanjing Li, "Combining constituent and dependency syntactic views for Chinese semantic role labeling", COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters Pages 665-673, 2010.
- [13] Soojong Lim and Hyunki Kim, "Korean Semantic Role Labeling using Sequence Labeling", 한국정보과학회 학술발표논문집, 2014.6, 595-597, 2014.
- [14] Changki Lee, Soojong Lim and Hyunki Kim, "Korean Semantic Role Labeling Using Structured SVM", Journal of KIISE 42(2), 2015.2, 220-226, 2015.
- [15] Young-Bum Kim, Heemoon Chae, Benjamin Snyder and Yu-Seop Kim, "Training a Korean SRL System with Rich Morphological Features", In Proceedings of the Association for Computational Linguistics, Association for Computational Linguistics, pages 637-642. 2014
- [16] Jangseong Bae, Changki Lee and Soojong Lim, "Korean Semantic Role Labeling using Deep Learning", 한국정보과학회 2015 한국컴퓨터종합 학술대회 논문집, 2015.06, 690-692, 2015.