CRF를 이용한 복수 의미역 문제 해결

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

Multiple Semantic Role Labeling Problems Solving using CRFs

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

요 약

의미역 결정에서 하나의 의미 논항이 둘 이상의 의미역을 가지는 경우는 복수의 레이블을 할당하기 때문에 어려운 문제이다. 본 논문은 복수의 의미역을 가지는 항의 의미역 결정을 위한 새로운 자질을 제안한다. 복수의 의미역을 결정하기 위해서 체언보다 선행되어 나타나는 용언에 대한 자질을 추가하였다. 또한 문장의 용언에 따라 의미역을 결정하기 위해서 문장 내의 용언 수만큼 각각에 용언에 대한 의미역을 결정할 수 있도록 반복적으로 레이블링하는 방법을 제시하였다. 본 논문의 실험 결과로 제안한 방법은 74.90%의 성능(F1)을 보였다.

주제어 : 한국어 의미역 결정, 울산대 한국어 의미역 말뭉치, CRFs

1. 서 론

의미역 결정은 서술어와 논항 사이의 의미 관계에 따라 논항의 역할을 분류하는 작업이다. 의미역은 논항의 유무와 형태에 따라 문장 의미를 분석하는데 도움을 준다. 의미역 결정이 가지는 다양한 문제점을 해결하기 위해서 꾸준한 연구가 진행되었으나 여전히 많은 문제점이 남아있다.

의미역은 구문 관계와 유사한 형태를 보이지만 서술어의 형태에 따라 논항의 역할을 결정하기 어려운 문제가 있다. 또한 하나의 의미 논항이 여러 개의 용언과 관계를 가지게 되면 역할도 여러 개를 가질 수 있다. 이때의 문제는 하나의 의미 논항이 갖는 여러 개의 역할이 용언에 따라 역할이 달라 질 수 있다는 점이다. 따라서 레이블링 모델 중 단일 분류 모델로는 해결할 수 없다[12,13]. 다음은 하나의 의미 논항이 여러 개의 용언과 관계를 가질 때의 의미역 결정 예문이다.

(가) 나는 밥을 먹고, 잠을 잤다.
(나) 나는 동생과 싸워서 부모님께 혼났다.
(다) 내가 기르는 강아지가 새끼를 낳았다.

위의 예에서 (가)는 용언이 '먹다'와 '잤다'로 2개 있다. 이때 '나'가 두 행위의 행위주(ARG0)가 된다. (가)는 역할이 2개 이상이지만 그 역할이 동일하다. (가)와 같은 문장은 역할을 하나만 레이블링 하더라도 문제가 되지 않는다. (나)는 용언이 '싸우다'와 '혼나다'로 2개 있다. 하지만 (가)와는 다르게 '나'의 역할이 '싸우다'와 관계될 때는 '행위주(ARG0)'이지만, '혼나다'와 관계될 때는 '대상주(ARG1)'가 된다. '혼나다'의 행위주는 '부모님'이 된다. 따라서 하나의 의미 논항에 하나의 역할만을 부여하면 (나)의 문장처럼 하나의 의미 논항이 다양한 역할을 수행할 때의 문제를 해결하지 못한다. (다)의 문장 역시 '강아지'가 '나'로부터 '기르다'라는 용언의 '대상주'이지만 '낳다'의 '행위주'로 2개의 역할이 서로 다르다.

CoNLL 형식의 말뭉치는 문장의 용언 수만큼 역할 레이블링 컬럼이 존재하며, 각 용언에 따라 역할이 레이블링 되어있다. 이에 본 논문은 다양한 역할이 부여된 말뭉치를 적극 활용하고, 용언에 따른 정확한 역할을 부여하기 위한 연구를 진행하였다. 본 논문에서는 울산대학교 한국어 의미역 말뭉치를 한국어 Propbank 말뭉치 형태로 변환하여 사용하였다. 울산대학교 한국어 의미역 말뭉치 중 용언 분석이 잘못된 문장을 제외하고 35,000문장을 선택하여 사용하였다.

본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장에서는 제안 방법에 대해서 설명한다. 4장에서는 실험 방식과 실험 결과를 분석하여 설명한다. 마지막으로 5장에서는 결론에 대해 기술한다.

2. 관련 연구

영어권에서는 2004년부터 CoNLL Shared Task를 진행하여 의미역 결정에 대해 연구를 진행하였다[1]. 초기의 의미역 결정은 형태소나 구문 정보만을 이용하였고, 후에 구구조와 의존구조, 구묶음 정보를 활용한 의미역 결정 연구를 진행하였다[2-6]. 또한 형태소나 구문 정보외의 의미역 결정에 도움이 되는 자질에 대해서 다양한 연구를 진행하였다. 이러한 연구를 통해 기존의 자질 외에 서술어의 형태나 개체명 정보, 조합 자질 등이 의미역 결정 성능 향상에 도움이 된다는 것을 증명하였다[7]. 한국어 의미역 결정에는 세종전자사전에서 추출한 격틀 사전 정보를 사용하여 의미역 결정 문제를 해결하려는

연구가 있었다[8,9]. [8]은 격틀 사전 정보를 이용한 연구로 비지도 학습 중 하나인 self-training 알고리즘을 사용하여 의미역 결정을 해결하였다. [9]는 의미역 결정에 애매성이 큰 부사격 조사 중 4개의 조사를 선택하여 애매성을 해소하는 방법을 연구하였다. 부사격 조사 중 '-에', '-로', '-에서', '-에게'를 선택하였고 해당 조사의 문제점을 해결하고, 의미역 결정을 진행하였다. 기계 학습 중 Structural SVM을 이용한 연구가 있었다[10,11]. [10]는 연속되는 레이블이 독립적이지 않고 영향을 미친다는 가정을 두고 i-1번째 레이블이 i번째 레이블에 정보를 전달할 수 있도록 설계되었다. 이는 순차적 레이블링 기반으로 의미역을 해결한 방법이다. [11]도 순차적 레이블링 기반으로 격틀 사전 정보와 Korean Propbank 말뭉치에서 추출한 서술어 정보를 통해 서술어 인식 및 분류와 논항 인식 및 분류를 동시에 진행하였다. [11]에서 제안하는 모델은 문장에서의 서술어를 인식하는 과정(PI)과 인식된 서술어와 관계 논항을 찾아 의미역을 결정하는 작업(AC)을 차례대로 수행한다. 학습 자질 중엔 서술어에 포함 된 형태소의 군집 정보를 활용하기도 하였다. [12,13]는 개체명 정보와 WordVector로 생성한 군집 정보, 동사파생접미사, 자/타동사, 능동/피동 정보가 의미역 결정에 도움이 되는 것을 증명하였다. 최근에는 기계 학습에서 어려운 부분인 자질 선택과 그 조합에 대한 문제를 해결한 딥 러닝(Deep Learning)방법을 이용한 연구도 진행되었다[14].

3. 제안 방법

본 연구에서 하나의 의미 논항에 관계 서술어에 따라 여러 개의 역할을 부여하기 위해서 한 문장을 문장 내의 서술어 수만큼 반복하여 의미역을 부착하도록 하였다. 한 번 동작할 때, 하나의 서술어에 대한 의미역을 부착하여 모든 서술어에 대한 의미역을 부착하도록 하였다. 말뭉치는 울산대 한국어 의미역 말뭉치를 사용하였다. 울산대학교 한국어 의미역 말뭉치 중 용언에 대한 분석이 잘못된 문장을 제외하고 35,000문장을 선택하여 실험을 진행하였다. 또한 일부 누락되거나 잘못 부착된 의미역에 대한 수정을 추가로 진행하였다. 수정한 말뭉치 오류는 형용사의 수식을 받는 체언에 의미역이 누락된 것과 체언의 나열에서 해당 체언들이 동일한 서술어에 수식되지만 일부 체언에 의미역이 누락된 것이다. 전체 35,000문장 중 학습에는 28,000문장을 사용하고, 평가에는 7,000문장을 사용하였다. 의미역 말뭉치의 문장은 평균적으로 약 2.74개의 용언을 지니고 있다. 울산대 말뭉치에 포함된 의미역 수는 필수격이 총 14,613개이고, ARG0은 2,184개, ARG1은 10,285개, ARG2는 636개, ARG3는 1,508개다. 부사격은 총 2,919개다.

모델 학습에 사용한 자질은 이전 연구를 통해 성능 향상이 검증된 자질의 조합을 사용하였다[12,13]. 학습에 사용된 자질은 다음과 같다.

- 형태소
- 형태소 품사 태그
- 의존 구문 분석 정보
- 관계 용언 형태소
- 관계 용언 형태소 품사 태그
- 개체명
- 용언과의 거리
- Word Vector를 기반으로한 cluster 정보
- 체언보다 선행되는 용언 정보
- 자동사/타동사 정보
- 능동태/수동태 정보

4. 실험 및 토의

학습에는 CRF를 사용하였으며, 제안하는 방법을 통해 실험 결과 표1과 같은 성능을 얻을 수 있었다.

표 1. 실험 결과 성능

	precision	recall	F1
performance	75.81%	74.02%	74.90%

2개 이상의 역할을 지니는 의미 논항은 총 1,893개다. 복수의 의미역을 지니는 의미 논항은 평균적으로 2.63개의 역할을 지니며, 총 4,505개의 역할이 존재한다. 표2는 복수의 의미역을 지니는 논항의 모든 역할에 대한 시스템 성능표이다.

표 2. 복수 의미역에 대한 성능

	precision	recall	F1
performance	77.01%	72.08%	74.47%

실험 결과를 분석해보면 시스템이 의미역으로 인식하지 못한 논항이 2,704개, 의미역으로 인식하였으나 역할을 잘못 부착한 수가 1,850개, 의미역이 아닌 논항을 의미역으로 인식한 수가 2,292개였다. 필수격 논항에서의 오류는 역할을 잘못 부착한 오류보다 미인식한 오류가 더 많이 나타났다. 특히 전체 대상주(ARG1) 10,285개중 1,010개를 미인식했다. 시스템이 의미역이 아닌데 의미역으로로 인식한 필수격 논항 수는 총 1,901개였다. 필수격에서 의미역을 미인식하는 문제는 구문 정보는 주어나 목적어가 되고, 서술어와 관계가 존재하나 의미역이 아닌 체언이 종종 나타나기 때문에 학습 시 노이즈로 작용되었기 때문이다. 그 예로 "낚시대는 초리대가 굵은 것을 사용한다."라는 문장에서 '낚시대'는 주어(SBJ)이면서 '사용하다'라는 서술어와 직접 관계가 생성되어 있다. 하지만 '낚시대'가 '사용하다'라는 행위의 행위주가 될 수 없기 때문에 의미역을 지니지 못한다. 이는 구문 분석에서 나타나는 이중 주어와 비슷한 유형이며, 이러한 문장들로 인해 학습 시 자질간 노이즈가 발생하게 된다. 이를 해결할 수 있는 방법으로 단어의 의미 중의성을 해소(WSD)하거나 세종전자사전의 용언 격틀 정보를 활용하는 방법이 있다. 격틀 정보를 활용한다

면 '사용하다'의 행위주에는 '낚시대'라는 사물이 올 수 없기 때문에 의미역이 될 수 없다고 판단할 수 있다. 다른 필수격의 오류에서 역할을 잘못 부착한 오류 중에서는 행위주(ARG0)와 대상주(ARG1)을 서로 바꾸어 부착한 오류가 355개였다. 능동태와 피동태의 정보를 활용하여 기존보다 오류를 196개 줄였지만 아직까지 역할이 서로 바뀐 오류가 남아있다. 부사격에서는 장소(LOC)와 방법(MNR), 시간(TMP) 표현에서 오류가 두드러졌다. 장소는 필수격인 착점(ARG3)과 역할이 서로 잘못 부착된 오류가 많았다. 이는 장소가 되는 국가명들이 문장의 활용에 따라 실제 장소를 뜻하거나 정부를 뜻하여 두 가지의 역할을 모두 수행하기 때문에 발생하는 문제이다. 이는 개체명 결정에서 발생하는 애매성과 유사하다. 방법은 전체 641개 중 301개의 미인식 오류가 있었다. 시제 역시 미인식 오류가 많은데 전체 386개 중 216개를 미인식하였다. 방법과 시제는 그 역할을 나타낼 수 있는 단어의 활용이 무수히 많기 때문에 학습 말뭉치의 양을 확장할 필요성이 있다. 표3은 실험 결과에 따른 오류 유형과 통계이다.

표 3. 오류 유형에 따른 오류 수와 비율

의미역	오류 유형		오류 수	비율
필수격	미인식	ARG1	1,010	14.7%
	역할 오류	ARG0 & ARG1	335	4.9%
부사격	미인식	MNR	301	4.4%
		TMP	216	3.2%
	역할 오류	ARG3 & LOC	205	3.0%

5. 결론

본 논문에서는 하나의 의미 논항이 여러 개의 서로 다른 역할을 지닐 때, 레이블링할 수 있는 방법에 대한 연구를 진행하였다. 실험에서는 단일 레이블링 모델인 CRF를 사용하였지만, 문장내의 서술어 수에 따라 동작하여 각각의 서술어에 대한 의미역을 결정할 수 있도록 하였다. 울산대 의미역 말뭉치를 사용하여 모델을 생성한 결과 74.90%의 성능을 보였다. 기존에 제안된 한국어 의미역 결정 시스템보다는 낮은 성능을 보이나, 본 연구는 다양한 역할을 모두 고려하여 동작하는데 의의가 있다. 향후연구로는 아직 낮은 성능을 높이기 위한 의존 구조에 나타나지 않는 누락된 관계를 찾는 방법에 대한 연구를 진행할 예정이다. 또한 세종전자사전의 격틀 정보를 활용하기 위한 방법과 단어들의 의미중의성을 해소하여 자질로 활용하는 방법을 함께 연구할 예정이다.

참 고 문 헌

[1] 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.

[2] Daniel Gildea and Daniel Jurafsky. "Automatic labeling of semantic roles", Association for Computational Linguistics, Computational Linguistics, 28(3):245-288. 2002.

[3] 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.

[4] Kadri Hacioglu. "Semantic role labeling using dependency trees", In Proceedings of COLING, Geneva, Switzerland. 2004.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Soojong Lim and Hyunki Kim, "Korean Semantic Role Labeling using Sequence Labeling", 한국정보과학회 학술발표논문집 , 2014.6, 595-597, 2014.

[11] Changki Lee, Soojong Lim and Hyunki Kim, "Korean Semantic Role Labeling Using Structured SVM", Journal of KIISE 42(2), 2015.2, 220-226, 2015.

[12] Tae-Ho Park, Jeong-Won Cha, "Korean Semantic Role Labeling Using CRFs", 제27회 한글 및 한국

어 정보처리 학술대회, 11-14, 2015.
[13] Tae-Ho Park, Jeong-Won Cha, "Feature Selection for Korean Sematic Role Labeling Using CRFs", 정보과학회지, 34권. 8호. 통권 327호. 37-41. 2016.
[14] Jangseong Bae, Changki Lee and Soojong Lim, “Korean Semantic Role Labeling using Deep Learning”, 한국정보과학회 2015 한국컴퓨터종합 학술대회 논문집 , 2015.06, 690-692, 2015.