CRFs 기반의 한국어 의미역 부착 성능 향상을 위한 자질 선택

창원대학교 ▌박태호·차정원*

1. 서 론

의미역 결정은 서술어와 논항 사이의 의미 관계를 이해하고, 의미 논항의 역할을 분류하는 작업이다. 의미역을 인식하고 분류하는 작업은 지속적인 연구가 진행되고 있지만 형태소나 구문 정보를 이해하는 것보다 부족한 성능을 보이고 있다. 의미 논항의 역할은 문장의 구조나 구문 기능이 변하더라도 변하지 않는 특성을 지니고 있다. 따라서 문장 구조가 다르더라도 동일한 의미 정보를 지니는 경우가 있기 때문에 단순히 구문 정보만으로는 해결하기 어렵다. 예를 들어 서술어의 형태가 능동태인 문장이 서술어가 수동태가 될 때, 문장에서의 주어와 목적어는 서로 바뀌게 된다. 하지만 의미역인 '행위주'와 '피동작주'는 바뀌지 않는다. 이 외에도 한국어 의미역 결정에는 다양한 문제가 있다. 그 중 하나는 의미역을 발생시키는 서술어가 관계 체언보다 선행되어 나타날 경우, 기존의 의존 구문 분석으로 관계를 파악하기 힘든 문제점이 있다. 다음은 이러한 문제점을 나타낸 예문들이다.

(가) 경찰은 범인을 체포했다.
(나) 범인은 경찰에게 체포당했다.
(다) 나는 새로 산 옷을 세탁했다.
(라) 나는 학생이다.
(마) 내가 본 것은 역사책이다.

위의 예문에서 (가)와 (나)의 문장은 의미가 서로 동일하다. 하지만 (가)에서의 주어는 '경찰'이고, (나)에서의 주어는 '범인'이다. 하지만 '체포하다'라는 행위의 행위주는 '경찰'로 동일하다. (다)의 문장의 경우는 '나는 옷을 샀다'와 '나는 옷을 세탁했다.'로 분석할 수 있다. 따라서 '사다'라는 행위의 대상은 '옷'이며, '세탁하다'의 행위의 대상 또한 '옷'이다. 하지만 의존 구문 분석의 결과로는 '사다'와 '옷'의 관계를 알 수 없다. (라)와 (마)의 경우는 긍정지정사인 '이(VCP)'가 결합된 어절에 대한 의미역 결정의 문제점이다. 긍정지정사는 구문 분석에서 동사처럼 활용한다. 하지만 긍정지정사는 행위성이 나타나지 않는다. 따라서 서술어로써 자질을 지닐 수 없으며, 의미역을 가질 수 없다. 단, 긍정지정사가 나타나는 어절이 의미역이 되는 경우는 나타난다. (마)의 문장은 "나는 역사책을 보았다."와 같은 의미를 지닌다. 따라서 (마)의 '역사책이다.'어절은 '보다'행위의 '대상'역할의 의미역이 된다.

이러한 문제점을 토대로 본 논문에서는 능동태와 수동태의 문장 형태 변환에 따른 의미역 결정에 대한 문제와 관계 체언보다 선행되는 서술어에 대한 문제를 해결하는데 도움이 되는 자질에 대한 실험을 진행하였다.

본 논문에서는 울산대 한국어 의미역 말뭉치를 사용하여 CRFs(Conditional Random Fields) 기반의 한국어 의미역 결정 시스템을 구축하였다. 본 논문은 2장에서 의미역에 대한 관련 연구를 소개하고, 3장에서는 의미역 결정에 사용되는 기존의 자질에 대한 소개와 새롭게 제안하는 자질에 대해 소개한다. 4장에서는 제안한 자질을 통한 실험 결과와 결과 성능을 분석하고, 5장에서는 한국어 의미역 결정에서 해결해야 할 문제점에 대해 설명한다. 마지막 6장에서는 결론에 대해 기술한다.

2. 관련 연구

영어권에서는 이미 오래 전부터 CoNLL Shared Task를 진행하여 2004년부터 꾸준히 의미역 결정에 대해 연구를 진행하였다[1]. 초기의 의미역 결정은 형태소나 구문 정보만을 이용하였다. 구구조를 기반으로 하여, 이후로는 의존구조나 구묶음 정보를 활용한 의미역 결정 연구를 진행하였다[2-6]. 또한 형태소나 구문 정보 외에 의미역 결정에 도움이 되는 다양한 자질에 대해서 꾸준히 연구를 진행하였다. 이러한 연구

* 종신회원

를 통해 서술어의 형태나 개체명 정보, 기존에 사용하던 자질들의 조합 자질 등이 의미역 결정 성능 향상에 도움이 된다는 것을 증명하였다[7].

또한 새로운 자질을 찾기보다는 기존에 사용하던 자질의 조합 자질 중 최적의 조합을 찾는 방법에 대한 연구도 진행되었다[8,9]. 이 실험을 통해 조합 자질 최적화만으로도 성능을 향상 시킬 수 있음을 보였다.

한국어 의미역 결정에는 격틀 사전 정보를 이용하여 의미역을 해결하려는 연구가 있었다[10,11]. 격틀 사전에는 서술어가 가지는 다양한 의미에 따라 활용할 수 있는 격틀 정보가 기술되어 있다. 격틀 정보에는 용언과 함께 나타날 수 있는 논항과 논항의 역할이 나타나있다. 또한 각 논항이 갖는 조사 정보와 의미 표현 정보 또한 나타나있다. [10]도 격틀 사전 정보를 이용한 연구로 비지도 학습 중 하나인 self-training 알고리즘을 사용하여 의미역 결정을 진행하였다. [11]은 의미역 결정에 애매성이 큰 부사격 조사 중 몇 개를 선택하여 애매성을 해소하는 방법을 연구하였다. 부사격 조사 중 '-에','-로','-에서','-에게'를 선택하여 해당 조사의 문제점을 해결하고, 의미역 결정을 진행하였다. 기계 학습 중 Structural SVM을 이용한 연구가 있었다[12,13]. [12]는 연속되는 레이블이 독립적이지 않고 영향을 미친다는 가정 하에 i-1번째 레이블이 i번째 레이블에 정보를 전달할 수 있도록 설계되었다. 이는 순차적 레이블링 기반으로 의미역을 해결한 방법이다. [13] 역시 순차적 레이블링 기반으로 격틀 사전 정보와 Korean Propbank 말뭉치에서 추출한 서술어 정보를 통해 서술어 인식 및 분류와 논항 인식 및 분류를 동시에 진행하였다. 제안하는 모델은 문장에서의 서술어를 인식하고, 후에 인식된 서술어와 관계 논항을 찾아 의미역을 분류한다. 학습 자질 중엔 서술어에 포함된 형태소의 군집 정보를 활용하기도 하였다. [14]는 한국어에서의 다양한 형태소 정보를 활용한 논문으로 조사 정보를 다양하게 분류하여 조사의 유무 정보, 조사의 형태소 정보, 조사의 형태소 품사 정보, 조사와 구문 정보의 조합 등을 사용하였고, 이와 함께 어미 정보를 사용하여 CRFs 학습을 하였다. 그리고 본 논문과 마찬가지로 한국어 의미역 결정에 도움이 되는 자질에 대한 연구가 있었다[15]. [15]는 개체명 정보와 WordVector로 생성한 군집 정보, 동사파생접미사 정보가 한국어 의미역 결정에 도움이 됨을 증명하였다. 최근에는 기계 학습에서 어려운 부분인 자질 선택과 그 조합에 대한 문제를 해결한 딥 러닝(Deep Learning)방법을 이용한 연구도 진행되었다[16].

3. 의미역 결정 자질 분석

3.1 기준 시스템

CRFs를 이용한 한국어 의미역 결정 성능을 알아보기 위해 다음의 자질로 기준 성능을 측정하였다. 기준 실험에서 사용되는 자질은 한국어 의미역 결정과 관련하여 자질에 대한 연구에서 검증된 자질을 함께 사용하였다[15].

- 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
- 개체명 정보
- WordVector를 이용한 군집 정보
- 동사파생접미사 정보

3.2 검증 후보 자질

한국어 의미역 결정의 성능 향상을 위해 서론에서 서술한 예문이 가지는 문제를 해결하는데 도움이 될 것으로 판단되는 자질을 연구하였다. 본 논문에서 한국어 의미역 결정에서 나타나는 문제점을 해결하기 위해 제안하는 자질은 다음과 같다.

- 자동사와 타동사 정보
- 능동태와 수동태 정보
- 관계 체언보다 선행되는 서술어 정보
- 세종 격틀 사전에서 나타난 의미 그룹 정보

제안하는 자질은 다음과 같은 근거를 통해 결정하였다. 먼저 자동사와 타동사 정보는 세종 격틀 사전에 포함되어 있는 정보를 사용하였다. 세종 격틀 사전은 자동사와 타동사를 일반자동사, 일반타동사, 일반자타동사, 기능자동사, 기능타동사, 기능자타동사, 숙어자동사, 숙어타동사, 숙어자타동사로 총 9가지로 분류하고 있다. 자동사는 동작이나 작용이 주어에만 영향을 미치는 동사이며, 타동사는 동작의 대상인 목적어를 필요로 하는 동사이다. 자동사와 타동사 자질을 사용한다면 목적어의 필요성에 대해 알 수 있으며, 목적어 위치에 올 수 있는 의미역에 대해 의미역 결정에 도움을 줄 것으로 예상하였다.

두 번째로 능동태와 수동태 정보는 표준대국어사전에 나타난 용언의 능동태와 수동태 정보를 추출하여 사용하였다. 능동태와 수동태는 서론에서 언급하였듯이 같은 '체포하다'라는 의미의 서술어 일지라도 그

형태가 능동인지 수동인지에 따라 주어가 바뀌게 된다. 따라서 서술어가 능동태인지 수동태인지 알 수 있다면 의미역 위치의 변화에 대응할 수 있을 것으로 예상하였다. 영어권에서는 이미 Voice라는 자질로 서술어의 수동태와 능동태를 구분하는 자질을 사용하고 있다.

세 번째로 관계 체언보다 선행되는 서술어 정보는 이미 분석이 완료된 구문 분석 정보를 다시 검색하여 체언 중 관계가 서술어로 연결되어 있는 체언에겐 기준 실험에서 부여하는 서술어 자질을 동일하게 부여하였다.

마지막으로 세종 격틀의 의미 그룹 정보는 세종 격틀에 단어의 의미에 따라 해당 단어가 속하는 의미 그룹 정보를 추출하여 사용하였다. 세종 격틀 사전에는 총 645개의 체언 의미 그룹과 631개의 용언 의미 그룹이 구축되어 있다. 격틀 사전에는 용언과 함께 나타나는 의미역 정보와 그 의미역의 의미 그룹 정보가 포함되어 있다. <그림1> 은 용언 격틀 사전의 일부이다. <그림1>에서 '체포하다'는 '행위자'와 '대상'이 함께 나타난다는 것을 알 수 있으며, <sel_rst> 태그에서 '행위자'에 올 수 있는 단어의 의미 그룹과 '대상'에 올 수 있는 단어의 의미 그룹을 알 수 있다. 본 논문에서는 이 정보를 기계 학습 자질로 활용하기 위해 이진 정보를 생성하였다. 이진 정보 생성 방법은 문장에서 서술어를 인식한 후, 해당 서술어와 의존 관계가 있는 체언의 의미 그룹이 서술어와 함께 나타날 수 있는 의미 그룹이라면 1을 주고, 아닐 경우엔 0을 주었다.

4. 실험 및 토의

실험 말뭉치는 울산대 한국어 의미역 말뭉치를 사용하였으며, 학습 모델로는 CRFs를 선택하였다. 학습에 사용된 논항의 역할 분류는 다음 <표 1>과 같다.

실험에 사용한 말뭉치의 양은 총 35,000문장이며 이 중 28,000문장을 학습에 사용하고, 7,000문장을 평가에 사용하였다. 말뭉치 전체에서 나타난 의미역 127,785개 중 ARG0는 15.96%, ARG1는 59.44%, ARG2는 4.23%, ARG3은 9.44%으로 나타났다. 학습 말뭉치에는 102,246개의 의미역이 있고, 평가 말뭉치에는 25,539개의 의미역이 있다.

3.1에서 언급한 자질을 사용하여 기준 성능을 측정하였다. 이후 제안하는 자질을 추가하여 학습하였고, 자질별로 성능을 측정하였다. 측정된 기준 성능과 각 자질을 추가하여 구한 성능은 다음의 <표 2>와 같다.

표 1 한국어 논항의 분류와 의미

분 류	의 미	분 류	의 미
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		

표 2 기준 성능표

	Precision	Recall	F1
기준 실험	78.23	75.21	76.69
+자/타동사	78.42	75.31	76.83
+능/수동태	78.53	75.18	76.82
+선행 용언	78.26	75.41	76.81
+격틀 사전/의미 그룹	78.43	75.22	76.79
+모든 자질	78.52	75.31	76.88

* 표에서 각각의 자질은 기준 실험에 해당 자질 하나만 추가하여 독립적으로 사용하였다. '모든 자질'에서는 제안하는 4개의 자질을 전부 사용하였다.

• **자동사와 타동사:** 자동사와 타동사 자질을 통해 해당 서술어가 목적어가 필요한지 불필요한지 알 수 있게 되었다. 또한 한국어 문장에서 주격조사를 사용하는 목적어가 종종 나타나는데 이 자질을 사용할 경우 주격조사와 결합된 체언이 행위 대상이 되는 의미역이 될 수 있음을 학습할 수 있었다. 서술어 "좋아하다 "는 일반타동사의 자질을 지니고 있다. "나는 빵은 좋아한다."라는 문장에서 "좋아하다"라는 행위의 대상은

```
<frame_grp type="FTR">↓
    <frame>X=NO-이 Y=N1-을 V</frame>↓
    <subsense>↓
        <sel_rst arg="X" tht="AGT">인간(경찰관|형사|수사관|군인)|인간집단(경찰|검찰)</sel_rst>↓
        <sel_rst arg="Y" tht="THM">인간(주동자|범인|탈옥수|강도|탈영병)</sel_rst>↓
        <eg>저는 아직 그 사람을 체포할 만한 증거를 찾지 못했습니다.</eg>↓
    </subsense>↓
</frame_grp>↓
```

그림 1 용언 '체포하다'의 격틀 사전 정보

‘빵’이지만 ‘빵’이란 단어가 주격조사인 ‘은’과 결합되어 있어 기존 자질에서는 애매성이 나타나게 된다. 따라서 자동사와 타동사 정보를 사용한다면 이와 같은 문제를 해결하는데 도움이 된다.

• **능동태와 수동태:** 능동태와 수동태는 영어권에서와 마찬가지로 성능 증가에 도움이 되었다. 능동태와 수동태에 대한 자질 사용 이전에는 주어가 행위주와 피동작주 중 무엇이 될 것인지 결정하기 어려웠다. 하지만 능동태 문장에서 주어는 행위주가 되고, 수동태 문장에서의 주어는 피동작주가 된다. 따라서 단순히 구문 기능 태그 정보만 사용할 때 나타나던 자질의 애매성을 해소해주는 기능을 수행하였다.

• **체언보다 선행되는 서술어 정보:** 서술어와 체언 사이에 관계가 존재하나 서술어를 기준으로 의존 관계가 생성되는 기존의 구문 분석 결과로는 체언보다 선행되는 서술어와 체언의 관계 정보를 얻을 수가 없었다. 하지만 간단한 규칙을 적용하여 서술어 자질을 지니지 못하는 체언에게 실제로 관계가 있는 서술어 정보를 부여함으로써 해당 체언이 서술어와 의존 관계가 있다는 것을 알 수 있게 되었다. 의미역은 서술어 없이 발생할 수 없기 때문에 기존에 서술어 자질 정보를 갖지 않는 체언들은 모두 의미역을 부여하지 못하였다. 하지만 체언보다 선행되는 서술어 정보를 추가함으로써 이 문제를 해결할 수 있었다. 단, 이번 실험에서 서술어와 체언 사이에 대명사나 의존명사가 존재하는 문제에 대해서는 해결하지 못하였다.

• **세종 격틀 사전 및 의미 그룹 정보:** 세종 격틀 사전과 의미 그룹 정보를 사용한 실험에서도 성능이 향상되었다. 의미역 결정에서는 주어나 목적어 위치에 존재하나 의미 역할을 지니지 않는 체언이 다수 나타난다. 하지만 기존의 자질만으로는 해당 체언이 의미적으로 서술어와 관계가 있는지를 이해하기 어려웠다. 따라서 의미 그룹 정보를 통해 해당 체언이 서술어 격틀에 적합한 위치에 있는지, 또한 체언이 지니는 의미가 부합하는지를 알 수 있게 되었다.

5. 한국어 의미역의 문제점

현재 한국어 의미역에는 다양한 태그셋이 사용되고 있다. 공개되어 사용할 수 있는 말뭉치는 Linguistic Data Consortium에서 제공하는 Korean Propbank 말뭉치[17]와 세종 의미역 말뭉치 그리고 울산대 의미역 말뭉치가 있다. 이 세 개의 말뭉치는 모두 다른 태그셋을 사용하고 있다. 또한 이 말뭉치 전체가 포함하고 있는 문장 수는 20만 문장 정도로 아직은 기계 학습을 수행하기에는 부족한 양이고, 그 중 세종 말뭉치는 누락된 정보와 오류가 많아 학습에 사용하기에는 아직 부적절하다.

또한 의미역을 분류하는 작업은 문장에서 사용되는 용언과 관계있는 모든 논항에 대해서 진행된다. 따라서 한 어절에 역할 정보가 부착될 수 있고, 구나 절에도 역할 정보가 부착될 수 있다. 그리고 하나의 의미 논항이 여러 개의 용언과 관계가 생성되어 2개 이상의 의미 역할이 부여될 수도 있다. 하지만 현재 연구된 시스템은 단일 정보를 부착하고 있기 때문에 이를 해결할 방법 연구가 필요하다.

6. 결 론

본 논문을 통해 기존에 사용하던 형태소 정보나 구문 분석 정보 외에 의미역 결정에 도움이 되는 자질을 검증하였다. 실험을 통해 자동사와 타동사, 능동태와 수동태, 체언보다 선행되는 서술어 정보, 세종 격틀 사전과 의미 그룹 정보가 의미역 결정에서 성능을 향상 시킬 수 있다는 것을 확인하였다. 제안한 네 가지의 자질을 추가하여 CRFs 모델을 통해 학습을 진행하였으며, 의미역 인식 및 분류 성능이 76.88%로 측정되었다.

본 연구를 통해 기존의 의미역 결정 연구보다 나은 성능을 구할 수 있었으나, 아직 해결하지 못한 오류가 많이 남아있다. 따라서 향후 연구로 긍정지정사가 포함된 어절이 의미역이 되는 경우에 대한 연구를 진행할 예정이다. 또한 이중 주어가 나타나는 문장에서 주어가 의미역이 되지 않는 경우에 대한 연구를 함께 진행할 예정이다. 그리고 현재 한 어절에 하나의 의미역만을 부여하는 시스템에 대해서만 연구를 진행하였으나, 한 어절에 의미역이 여러 개가 나타나는 문제를 해결할 수 있는 방법을 연구하고 있다.

참고문헌

[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] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

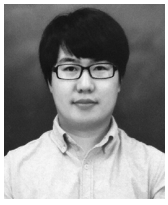
[14] Young-Bum Kim, Heemoon Chae, Benjamin Snyder and Yu-Seop Kim, “Training a Korean SRL System with Rich Morphological Features”, In Proceedings of the Association for Computational Linguistics, Association for Computational Linguistics, pages 637 – 642. 2014

[15] Tae-Ho Park, Jeong-Won Cha, "Korean Semantic Role Labeling Using CRFs", 제27회 한글 및 한국어 정보처리 학술대회, 11-14, 2015.

[16] Jangseong Bae, Changki Lee and Soojong Lim, “Korean Semantic Role Labeling using Deep Learning”, 한국정보과학회 2015 한국컴퓨터종합학술대회 논문집 , 2015.06, 690-692, 2015.

[17] 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.

약 력



박 태 호

2013 창원대학교 컴퓨터공학과 졸업(학사)
2015 창원대학교 컴퓨터공학과 졸업(석사)
2016~현재 창원대학교 친환경해양플랜트FEED 공학과 박사과정
관심분야: 자연어처리, 기계학습, 의미 분석
Email: taehope@changwon.ac.kr

차 정 원

1996 숭실대학교 컴퓨터공학과 졸업(학사).
1999 포항공과대학교 컴퓨터공학과 졸업(석사)
2002 포항공과대학교 컴퓨터공학과 졸업(박사)
2002~2003 USC/ISI(박사후연수)
2004년~현재 창원대학교 컴퓨터공학과 교수
관심분야: 차연어처리, 기계학습, 정보검색
Email: jcha@changwon.ac.kr