의미 정보를 이용한 한국어 의미역 인식 연구

임수종○, 김현기
한국전자통신연구원, 자동통역인공지능연구센터
{isj, hkk}@etri.re.kr

A Study of Korean Semantic Role Labeling using Word Sense

Soojong Lim○, Hyunki Kim
Automatic Speech Translation and AI research Center, ETRI

요 약

기계학습 기반의 의미역 인식에서 주로 어휘, 구문 정보가 자질로 주로 쓰이지만, 의미 정보를 분석하는 의미역 인식은 단어의 의미 정보 또한 매우 주요한 정보이다. 그러나, 기존 연구에서는 의미 정보를 활용할 수 있는 방법이 제한되어 있기 때문에, 소수의 연구만 진행되었다. 본 논문에서는 동형이의어 수준의 의미 애매성 해소 기술, 고유 명사에 대한 개체명 인식 기술, 의미 정보에 기반한 필터링, 유의어 사전을 이용한 클러스터 및 기존 프레임 정보를 확장하는 방법을 제안한다. 제안하는 방법은 기존 연구 대비 뉴스 도메인인 Korean Propbank는 3.14, 위키피디아 문서 기반의 WiseQA 평가셋인 GS 3.0에서는 6.57의 성능 향상을 보였다.

주제어: 의미역 인식, 의미 정보, 의미 기반 유의어

1. 서론

의미역 인식(Semantic Role Labeling)이란, 자연어 문장에서 'who does what to whom'을 인식하는 기술이다. 문장 내에서 서술어를 중심으로 서술어에 대해 의미적인 역할(예를 들어, 행위자격, 경험자격, 대상격, 도구격 등)을 하는 문장의 부분(units)과 이에 대한 의미 역할을 결정하는 것을 말한다. 문장의 서술어와 논항들 사이의 '주어', '목적어'와 같은 문법 관계를 의미역으로 사상(mapping)하는 문제로 볼 수 있으며, 일반적으로 구문분석을 수행한 후에 의미역 인식이 수행된다[1,2].

기계학습에 기반한 기존 연구는 주로 Johansson and Nugues[3]가 제시한 형태소, 구문 자질을 기반으로 이를 조합하는 등의 방법이 주를 이루었으나, 의미 정보는 의미역 인식에서 형태소, 구문 정보 못지않게 중요한 정보이다. 예를 들어, '타다'라는 서술어가 행위자격과 대상격을 갖는다면, 유사한 의미인 '승차하다'도 마찬가지로 행위자격과 대상격을 갖는다는 것을 유추하여 '승차하다'에 대한 학습 데이터가 존재하지 않더라도 의미역 인식을 학습하는데 도움이 될 것이다. 그러나, 동형이의어가 존재할 경우 '탈것이나 짐승의 등 따위에 몸은 얹다'는 뜻이 아닌 '불씨나 높은 열로 불이 붙어 번지거나 불꽃이 일어나다'라는 뜻으로 잘못 유추될 수도 있기 때문에, 의미 정보를 이용할 때는 글자 그대로의 단어가 아닌 단어가 갖는 의미를 고려해야 한다.

본 논문에서는 기존 한국어 의미역 인식 시스템에, 일반 명사에 대한 동형이의어 정보, 고유 명사에 대한 개체명 인식 정보를 의미 자질로 추가하고, 유의어 사전을 이용하여 단어를 의미 수준으로 클러스터링하며, 유의어 관계를 통해 정교하게 구축된 프레임을 자동으로 확장하여 한국어 의미역 인식 성능을 개선하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 개선하고자 하는 기존 시스템 및 본 논문에서 제안하는 의미 정보에 대해 설명하며, 4장에서는 실험 및 결과를 분석하고, 마지막으로 5장에서 결론을 기술한다.

2. 관련 연구

의미역 인식 연구는 서술어와 이에 대한 논항 관계를 구축한 FrameNet과 같은 언어자원 기반의 의미역 연구와 Proposition Bank 기반 기계학습 기반의 의미역 연구로 나눌 수 있다. 언어자원에 기반한 방법은 사전에 구축된 격틀(frame)과 선택 제약(selectional restriction) 정보를 이용하여 서술어에 대한 논항의 의미역을 결정하는데, 정교하게 구축된 언어자원에 기반하기 때문에 높은 정확률을 보이기는 하지만, 이러한 언어자원 구축이 어렵고, 격틀에 기술할 수 없는 부가역 등은 적용하지 못하는 문제가 있다[3, 4].

의미역 태깅된 말뭉치를 이용한 기계학습 방법은 Proposition Bank를 이용하여 영어권에서 활발하게 진행되고 있으며, Maximum Entropy(ME), Support Vector Machine(SVM), Conditional Random Field(CRF) 등 다양한 기계학습 기법 적용[1,5,6]에 집중되었고, 학습 자질에 대해서는 CoNLL-2008에서 제시한 형태소 및 구문 자질

을 변형 또는 조합하여 적용하는 방법 이외에는 추가 자질 등에 대한 연구는 진행되지 못 했다.

본 논문은 추가 자질로 의미 정보 사용을 제안하며, 또한 기계학습으로 인식된 결과를 의미 언어자원을 이용하여 검증하고, 오류를 수정한다. 또한, 추가된 의미 정보가 성능에 어떠한 영향을 주는지 실험을 통해 살펴보고자 한다.

3. 한국어 의미역 인식을 위한 의미 정보

본 논문에서 채택한 기존 베이스라인 시스템은 Structural SVM을 이용한 순차적 레이블링 방법[6, 7]이다. 이 시스템에서 사용한 자질 중에서 의미 정보로 대체하려는 자질은 다음과 같다.

- 서술어 어휘/품사 및 의미 정보
- 논항 후보 어절의 실질 형태소 어휘/품사 정보
- 서술어의 논항 후보 어절 어휘/품사 조합 정보
- 서술어 및 논항 후보의 의존 구문 트리상 부모 노드/가장 왼쪽 자식 노드/가장 오른쪽 자식 노드의 실질 형태소 어휘/품사 정보
- 서술어와 논항 후보의 의존 구문 트리상 가장 낮은 공통 부모(lowest common ancestor)의 실질 형태소 어휘/품사 정보

위에서 열거한 자질들의 특징은 주로 의미를 갖는 실질 형태소에 대한 어휘 정보로, 표준국어 대사전에서 정의한 동형이의어 수준의 의미 정보(어깨 번호)로 대체를 하여 보다 정확한 뜻을 이용하려 한다.

3.1 동형이의어 수준 의미 정보

어휘의미중의성 해소(Word Sense Disambiguation)은 다른 뜻을 갖는 동일 형태의 어휘에 대해서 의미를 구분해 주는 것을 말한다. 예를 들어, 문장에서 사용된 단어 '눈'에 대해서 '얼음의 결정체'인지, '감각기관' 인지를 결정해 준다. 주로 고유 명사의 의미를 정해주는 개체명 인식과 다르게 어휘의미중의성 해소는 의미 목록(Sense Inventory, 주로 일반적인 사전)에 정의된 일반 명사를 대상으로 의미를 결정해 준다. CoNLL-2008에서는 WordNet 기반의 어휘의미 자질에 대한 사용을 정의하고는 있으나, 널리 사용되지 못 하였다.

본 논문에서는 한국어에 대해 표준국어대사전에 정의된 동형이의어 수준으로 의미를 구분하여 자질로 사용한다. 앞에서 언급한 '타다'의 경우는 기존에 '타/VV'를 자질로 사용하였는데, 이를 '타.02/VV'로 수정하여 사용한다.

3.2 개체명 의미 정보

동형이의어 수준의 의미 정보는 사전에 등재된 단어를 대상으로 하는데, 실제로 위키피디아, 뉴스 등의 문서에서는 사전에 등재되지 않은 다양한 개체명(주로 고유 명사)이 등장한다. 이 또한 의미 정보를 학습해야 하기 때문에 고유 명사에 대해서는 개체명 정보를 의미 정보에 기반한 자질로 채택한다.

개체명(Named Entity; 이하 NE)은 고유한 의미를 가지는 명사를 말하며, 주로 인명, 지명, 기관명과 다양한 숫자 표현(날짜, 시간, 길이, 단위 등)을 말한다. 이러한 개체명은 의미역 중에서 '장소격', '방향격', '시간격'과 연관이 있다. CoNLL-2008 shared task open track 에서는 의미역을 인식하는데 도움이 되는 정보로 판단하여 개체명을 제공하여 의미역을 인식하는데 사용되었으나[5], 다른 기존 연구에서는 다양하게 사용하지 못 했다. 본 논문에서는 인명, 지명, 기관명, 인공물 등 15개의 개체명 범주를 자질로 사용하였고, 개체명에 해당하지 않는 경우는 '개체명 아님'(NE_NONE) 범주로 하였다.

3.3 의미 정보에 기반한 부가격 필터링

Korean Propbank(KPB)[8]에서 제공하는 프레임은 각 서술어의 의미에 대해 필수격에 해당하는 정보만을 기술한다. 시간, 장소, 원인과 같은 모든 서술어에 선택적으로 사용될 수 있는 부가격의 경우에는 프레임에 기술할 수가 없기 때문에 이러한 정보는 포함되어 있지 않다.

이러한 언어자원에 명시되어 있지 않더라도, 개체명이나 일반 단어의 경우도 시간이나 장소를 나타내는 경우 해당 격으로 기계학습 결과가 인식되지 않은 경우에는 시간격이나 장소격으로 인식된 결과를 수정하였다. 또한 부정격이나 부가격은 리스트를 작성하여 잘못 인식된 경우를 수정하였다.

3.4 유의어 사전 기반 의미 클러스터링

고유명사에 대한 개체명 인식 결과는 15개의 개체명 범주로 매핑이 되기 때문에 자료희귀성 문제를 줄일 수 있으나, 사전에 등재된 일반 어휘를 대상으로 하는 동형이의어 수준의 의미 정보는 어휘가 1개 이상의 의미 정보로 분할(표준국어대사전에는 동사 '타다'에 대해 11개 어깨번호가 존재)될 수 있기 때문에 자료 희귀성 문제를 심화시킬 수 있다.

이 문제를 해결하기 위해서 본 논문에서는 유의어 사전[9]을 이용하여 유의어 클러스터를 구성하였다. 유의어 사전은 표준국어대사전의 어깨번호 수준(동형이의어)으로 유의어가 구성되어 있다. 각 표제어에 대해 1차 유의어, 2차 유의어로 구성되어 있고, 2차 유의어는 1차 유의어의 유의어 개념이다. 2차 유의어 확장은 표제어의 유의어 망을 한눈에 볼 수 있을 뿐 아니라, 어휘가 가진 다양한 의미를 확대된 외연 속에서 예측할 수 있게 하는 장점이 있다[9].

표1은 타다02에 대한 유의어 사전이 구성된 예이다.

표 1 유의어 사전 구성 예

표제어	품사	1차 유의어	2차유의어
타다02	동사	등산하다	등반하다01,등정하다01
	동사	승차하다02	타다02
	동사	오르다	높아지다,뛰다01,붙다,상승하다01,서다01,솟다01,올라가다,올라타다,옮다,적히다,치밀다,치솟다,타다02,탑승하다,퍼지다,확산되다
	동사	올라가다	누진하다01,상승하다01,오르다,타다02
	동사	이용하다01	다루다01,부리다01,사용하다03,써먹다,쓰다03,타다02,활용하다

1차 유의어의 경우 표제어의 1차 유의어를 다시 표제어로 위치하여 1차 유의어가 표제어에 존재하지 않거나 이미 클러스터에 속할 경우에는 클러스터 구성을 종료하고 다른 의미 클러스터를 구성하는 방법을 사용하였다.

2차 유의어의 경우에는 1차 유의어를 기반으로 이미 구성된 클러스터에 추가하는 방식을 사용하였다. 표제어는 9,568개, 1차 유의어는 14,623(중복허용 37,157개)개 표제어로 구성이 되고, 2차 유의어는 20,865개(중복허용 307,320개) 표제어로 구성이 된다.

표 2 유의어 기반 클러스터 예

Cluster #	1차 유의어	2차유의어(추가)
0565	타02/VV 승차하02/VV	탑승하/VV
	등산하/VV 오르/VV 올라가/VV	등반하01/VV 등정하01/VV 상승하/VV
	이용하01/VV	활용하/VV

유의어 사전에 기반하여 구성된 클러스터의 예인 표2를 보면 1차 유의어에도 여러 가지 의미가 혼재되어 있다. 이는 유의어 사전이 다의어 수준이 아닌 동형이의어 수준으로 되어 있는 한계이다.

3.5 유의어 사전 기반 프레임 확장

KPB에서는 그림1과 같이 용언 '연주하다'에 대해 프레임 정보를 제공하여 필수 의미역에 대한 정보를 제공한다. 본 논문에서는 위와 같은 프레임 정보를 사용하여, 기계학습에서 잘못 인식된 의미역을 바로잡고자 한다.

FrameSet(XML)

```
<framefile>
  <predicate>
    <lemma>연주(yeon-ju)</lemma>
    <frameset>
      <id>연주.01</id>
      <edef>play</edef>
      <roleset>
        <role argnum = 0, argrole = agent/>
        <role argnum = 1, argrole = thing played/>
      </roleset>
      <mapping>
        <rel>연주하다</rel>
          <mapitem src = sbj, trg = arg0/>
          <mapitem src = obj, trg = arg1/>
      </mapping>
```

그림 1 Korean Propbank Frame 예

그러나 이러한 정보는 한국어 모든 용언에 제공되지 않기 때문에, 한국어에 적용할 경우 모두 용언의 의미에 대해서 프레임 정보를 사용하지는 못한다. 실제로 KPB에서 용언 '타다'에 대해서 프레임을 제공하지만, 'ride', 'take', 'burn'에 해당하는 프레임만을 제공한다. '타다06: 악기의 줄을 퉁기거나 건반을 눌러 소리를 내다.'의 뜻에 해당하는 'play'라는 뜻을 갖는 프레임 정보를 사용할 수 없다.

본 논문에서는 1차 유의어를 사용하여 프레임이 제공하지 않더라도 1차 유의어에 해당하면 프레임을 공유한다고 가정하고 유의어의 프레임을 사용하여 의미역 인식 결과 중 프레임과 배치되는 결과는 배제하였다. '연주하다' 예에서 보면 유의어 '타다06'은 연주하다와 마찬가지로 ARG0, ARG1만을 갖는데, 만약 자동 인식 결과가 프레임에서 존재하지 않는 의미역(이를테면 ARG2)을 인식한 경우에는 차선의 결과를 최종 결과로 선택한다. 이와 더불어, 한국어 형용사의 경우 행위자격이 아닌 대상격만을 갖는 특징이 프레임에서 기술되어 이를 반영하였다.

4. 실험 및 결과

본 논문에서는 KPB와 질의응답 시스템 평가를 위해 한국전자통신연구원(ETRI)에서 구축한 WiseQA[10] 2차 표준평가셋을 학습 및 평가용으로 각각 사용하였다. KPB는 4,882문장으로 구성되는데, 이중 약 1/6인 811문장을 평가용으로 사용하고, 나머지는 학습용으로 사용하였다. WiseQA 2차 표준평가셋 750셋을 학습용으로 사용하고, ExoBrain 언어분석 말뭉치[11] 중에서 의미역 태깅된 117셋, 439문장을 평가용으로 사용하였다. KPB에서 정의한 24개 의미역을 사용하였다.

기계학습을 위한 형태소 분석, 어휘의미 분석, 개체명 인식, 구문 분석 정보는 ETRI 언어분석기[12]를 사용하여

자동으로 분석된 결과를 이용하였다. 따라서 자동으로 추출된 학습 자질에는 오류가 포함되어 있을 수 있다. 한국어 논항 인식/분류(Argument identification and classification)에 대해서만 성능을 측정하였고, 정확율(Precision), 재현율(Recall)에 기반하여 계산하는 F1-score를 성능 척도로 사용하였다.

실험은 앞에서 언급한 기존 한국어 의미역 인식 시스템과 비교하여, 본 논문에서 제안하는 의미 정보 자질을 차례대로 적용하면서 성능 향상 추이를 관찰하였다.

표 3 KPB 실험 결과

	정확율	재현율	F1-score	성능향상
baseline	82.25	71.32	76.40	
+WSD & NE	82.46	72.39	77.10	0.70
+ 부가격 필터링	84.60	72.00	77.79	0.69
+ 유의어 클러스터	85.30	72.60	78.44	0.65
+ 프레임 확장	86.00	73.99	79.54	1.10

표 4 ExoBrain 언어분석 말뭉치 실험 결과

	정확율	재현율	F1-score	성능향상
baseline	74.94	66.64	70.55	
+WSD & NE	75.69	69.64	72.54	1.99
+ 부가격 필터링	78.32	71.44	74.72	2.18
+ 유의어 클러스터	78.72	71.41	74.89	0.17
+ 프레임 확장	80.12	74.33	77.12	2.23

표3과 표4는 본 논문에서 제안된 방법을 이용하여 실험한 결과이다. 한국어 의미역 인식을 위해 본 논문에서 제안된 모든 의미 정보는 한국어 의미역 인식 성능 확장에 기여했으며, 뉴스 도메인인 KPP는 3.14, 위피키디아를 기반으로 한 ExoBrain 언어분석 말뭉치를 이용한 평가셋에서는 6.57의 성능향상을 보였다.

특히 기계학습의 결과 중 명확하게 오류로 볼 수 있어 수정을 한 부가격 필터링 및 프레임 확장의 경우 큰 폭의 성능 향상을 보였다. 이러한 사실은 기계학습에 기반한 방법이 학습 데이터 구축을 통한 확장성 등 여러 가지 장점이 있지만, 의미 정보를 다룰 경우 인간이 구축한 리소스에 기반할 경우 기계학습이 갖고 있는 단점을 보완할 수 있다는 점을 보여준다.

5. 결론

한국어 의미역 인식 기술을 개선하기 위해서 기존에 채택하지 않았던 의미 정보를 자질로 채택하는 방법에 대해서 제안하였다. 제안한 방법은 의미 정보를 사용하지 않는 방법에 비해 뉴스 도메인인 KPB에서는 3.14, ExoBrain 언어분석 말뭉치를 이용한 평가셋에서는 6.57의 성능 향상을 나타냈다. 이러한 실험 결과를 통해 의미 분석을 위해서는 기본적인 형태소, 구문 정보 이외에도 의미 정보와 사람에게 검증된 의미 자원을 사용하는 것이 성능 향상에 도움이 된다는 것을 알 수 있다.

향후 연구로는 서술어와 논항의 의미 경로(path), 의미 거리(sense distance) 등의 정보, 표준국어대사전에서 필수 문틀로 제시하는 정보와 결합하여 프레임을 확장하는 등 한국어 의미 개념망을 사용하는 것을 방향으로 하고 있다.

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.R0101-15-0062, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)

참고문헌

[1] R.Johansson and P.Nugues, "Dependency-based Semantic Role Labeling of PropBank," in Proceedings of the EMNLP-2008, 2008.

[2] 임수종, 배용진, 김현기, 나동렬, "도메인 적응 기술을 이용한 한국어 의미역 인식," 정보과학회 논문지, 제42권, No.4, pp.475-482, 2015

[3] 김병수, 이용훈, 이종혁, "비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정," 정보과학회 논문지:소프트웨어 및 응용, 제34권, No.2, pp.112-122, 2007

[4] A.Giuglea and A.Moschitti, "Semantic Role Labeling via FrameNet, VerbNet and Propbank," in Proceedings of the ACL-2006, 2006.

[5] M. Surdeanu et al., "The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies," Proc. of the CoNLL-2008, pp.159-177, 2008

[6] S.Lim, C.Lee and D.Ra, "Dependency-based Semantic Role Labeling Using Sequence Labeling with a Structural SVM," in Pattern Recognition Letters, vol.34, pp.696-702, 2013

[7] 임수종, 김현기, "한국어 의미역 인식을 위한 다양한 기계 학습 자질 연구", 제9회 정보과학회 빅데이터학회 공동심포지엄, pp.125-128, 2015

[8] M.Palmer et al., "Korean Propbank," Linguistic Data Consortium, Philadelphia, 2006.

[9] 김기형, "단답형 QA를 위한 의미매칭 기반 정답 추출을 위한 유의어/반의어/연관어 수준 어휘지식 DB 시작품 제작," ETRI 용역 과제 결과보고서, ㈜낱말,

2014
[10] http://exobrain.re.kr/onedintro
[11] 임수종, 권민정, 김준수, 김현기, "ExoBrain을 위한 한국어 의미역 가이드라인 및 말뭉치 구축,", 제 27회 한글 및 한국어 정보처리 학술대회 논문집, 2015.
[12] 임준호, 윤여찬, 배용진, 김현기, 이규철, "지배소 후위 제약을 적용한 트렌지션 시스템 기반 한국어 의존 파싱 모델," 제26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.