한국어 의미역 결정을 위한 Korean PropBank 확장 및 도메인 적응 기술 적용

배장성○, 오준호, 황현선, 이창기
강원대학교 컴퓨터과학과
{jseffort, jho, hhs4322, leeck}@kangwon.ac.kr

Extending Korean PropBank for Korean Semantic Role Labeling and Applying Domain Adaptation Technique

JangSeong Bae○, JunHo Oh, HyunSun Hwang, Changki Lee
Dept. of computer science, Kangwon National University

요 약

한국어 의미역 결정(Semantic Role Labeling)은 주로 기계 학습에 의해 이루어지며 많은 말뭉치 자원을 필요로 한다. 그러나 한국어 의미역 결정 시스템에서 사용되는 Korean PropBank는 의미역 부착 말뭉치와 동사 격틀이 영어 PropBank의 1/8 수준에 불과하다. 따라서 본 논문에서는 한국어 의미역 결정 시스템을 위해 의미역 부착 말뭉치와 동사 격틀을 확장하여 Korean PropBank를 확장 시키고자 한다. 의미역 부착 말뭉치를 만드는 일은 많은 자원과 시간이 소비되는 작업이다. 본 논문에서는 도메인 적응 기술을 적용해 보고 기존의 학습 데이터를 활용하여, 적은 양의 새로운 학습 말뭉치만을 가지고 성능 하락을 최소화 할 수 있는지 실험을 통해 알아보고자 한다.

주제어: 한국어 의미역 결정, Korean PropBank, 도메인 적응 기술

1. 서론

의미역 결정(Semantic Role Labeling)은 문장의 각 술어의 의미와 그 논항들의 의미적인 관계를 결정하는 자연 언어 처리의 한 단계이다. 의미역 결정은 일반적으로 기계 학습에 의해 이루어지게 되며 현재까지 연구가 활발하게 진행되고 있다.

일반적인 기계 학습 기반의 의미역 결정 시스템은 해당 문장의 술어들을 식별하고 각 술어에 대한 논항들의 의미역을 결정하여 "누가, 무엇을, 누구에게, 어떻게, 왜" 등의 의미 관계를 찾아내는 시스템이다. 예를 들면 그림 1 의 '상어는 연골어류에 속하는 물고기이다.' 와 같은 텍스트로 된 문장이 주어졌을 때 의미역 결정 시스템에 의해 '속하.01' 이라는 술어와 의미역이 달린 술어의 논항들을 얻게 된다. 여기서 'NR' 은 의미역이 달리지 않았음을 뜻하고 'ARG1' 은 술어 '속하.01' 의 논항이 된다.

의미역 결정 시스템은 기계 학습에 필요한 많은 양의 말뭉치를 필요로 한다. 의미역 결정 시스템에서 널리 사용되는 말뭉치로 PropBank[1]가 있으나 이는 영어 의미역 결정을 위한 말뭉치이기 때문에 한국어에 적용할 수 없다. 이를 해결하기 위해 Korean PropBank[2]가 만들어졌으나 의미역 부착 말뭉치와 동사 격틀이 영어 PropBank의 1/8 수준에 불과하다. 따라서 본 논문에서는 한국어 Wikipedia에서 추출한 데이터를 이용하여 Korean PropBank를 확장하고자 한다. 의미역 결정 시스템은 크게 격틀 사전에 기반을 둔 시스템과 말뭉치에 기반을 둔 시스템으로 나눌 수 있으므로[3] Korean PropBank를 확장하기 위해 본 논문에서는 말뭉치를 늘리는 방법과 격틀 사전을 확장하는 방법 모두를 고려한다.

일반적인 의미역 결정 시스템은 학습에 사용한 데이터와 테스트 데이터가 같은 도메인으로 이루어져 있다. 반면 학습에 사용한 도메인과는 다른 도메인으로 테스트를 할 경우 의미역 결정 시스템 성능이 큰 폭으로 하락됨을 볼 수 있다.[4] 의미역 결정 시스템에 필요한 말뭉치를 만드는 일은 사람의 수작업으로 이루어지기 때문에 많은 시간 및 비용을 필요로 한다. 따라서 모든 도메인에 대해 학습 데이터를 가지고 있기란 쉽지 않은 일이다. 학습에 사용한 데이터와 실제 평가, 적용하는 도메인이 다른 경우에 대해 의미역 결정 시스템을 적용하는 것을 도메인 적응 기술이라고 부른다. 도메인 적응 기술은 위에서 언급한 도메인 변경시의 성능 하락 혹은 신규 도메인에 대한 많은 양의 말뭉치를 확보하지 못했을 시 유용하

게 쓰일 수 있는 기술이다. 본 논문에서는 [5]에서 제안한 도메인 적응 기술 및 여러 도메인 적응 기술을 확장한 Korean PropBank 말뭉치에 적용해 본다.

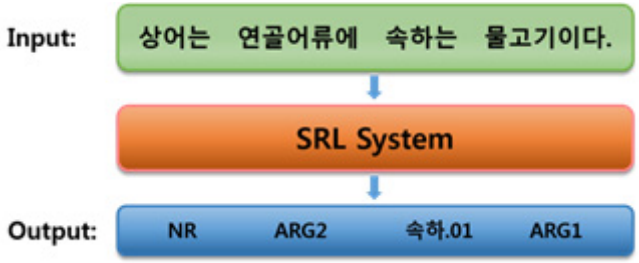

그림 1. 의미역 결정 시스템

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 Korean ProBank 확장에 대해 설명한다. 4장에서는 확장한 Korean PropBank 데이터에 대해 도메인 적응 기술들을 적용하고 그 결과를 보여준다. 5장에서는 결론에 대해 기술한다.

2. 관련 연구

기계 학습 기반 자연어 처리를 연구함에 있어 말뭉치는 필수불가결한 자료이다. PropBank는 의미역 결정에 필요한 자원으로써 사람에 의해 수작업으로 만들어진 의미역 부착 말뭉치이다. Korean PropBank는 PropBank를 기반으로 만들어진 한국어 의미역 부착 말뭉치인데 그 양의 부족함 때문에 연구의 진행에 어려움이 따른다. [6]은 한국어 의미역 결정을 위해 Korean PropBank를 학습 말뭉치로 사용하는 의미역 결정 시스템이다. 의미역 부착 말뭉치 구축은 [6]과 같은 시스템들을 위해 필요한 일이다.

따라서 본 논문에서는 Korean PropBank 확장을 보다 빠른 시간 내에 끝내기 위해 [7]에서 구축한 반자동 태깅 도구를 사용한다.

도메인 적응은 수년간 연구되어온 기술이지만 [5][8-11] 한국어 의미역 결정에 대한 연구 및 실험은 많지 않다. 도메인 적응 기술에서는 보유하고 있다고 가정하는 말뭉치를 'Source', 적용 할 도메인에 대해서 'Target' 이라고 구분 짓고 있다. [10]에서는 도메인 적응 기술들을 분류 및 평가해 놓았다. [10]에 따르면 SRC-ONLY는 가장 기본이 되는 도메인 적응 기술 방법이며 Target 도메인의 데이터는 무시된다. TGT-ONLY 방법은 기존의 Source 데이터를 무시하고 새로운 도메인의 데이터로만 학습하게 된다. Source와 Target 데이터 모두를 학습에 이용하는 ALL 방법도 있으며, SRC-ONLY와 TGT-ONLY에 대해 선형 보간법(Linear Interpolation)을 사용한 도메인 적응 방법도 있다. 또한 학습 모델의 가중치 벡터를 다른 도메인에 적용하는 [5]에서 제안한 Prior 모델도 있다.

3. Korean PropBank 확장

Korean PropBank는 2749개의 격틀 정보를 저장하고 있는 프레임 파일과 4882개의 문장으로 이루어진 말뭉치이다. 영어의 1/8 수준에 불과한 Korean PropBank를 확장시키기 위해 본 논문에서는 Wikipedia에서 가져온 질문과 정답 쌍으로 이루어진 문장에 대해 의미역 정보를 추가 하였다.

의미역 정보를 추가하는 방법은 Wikipedia에서 가져온 문장에 대해 기계 학습을 한 번 한 후에 [7]에서 개발한 반자동 의미역 태깅 도구를 이용하여 말뭉치를 구축하는 방식으로 진행하였다.

말뭉치 구축 시 A0(행동주), A1(피동주)과 같은 논항은 Korean PropBank의 프레임에 정의되어 있는 규칙을 따랐다. 기본적인 의미역과는 별도로 문장을 좀 더 상세하게 서술하기 위해 의미역 Modifier를 사용 하였다. 사용한 의미역 Modifier는 영어 PropBank에서 사용하는 것의 일부를 따르고[12], 표 1 은 의미역 Modifier의 종류 및 정의 이다.

LOC	행동이 일어나는 장소를 나타낸다.
DIR	경로와 방향을 나타낸다.
MNR	행동이 어떻게 일어나는가를 명시한다.
TMP	행동이 언제 일어났는지 명시.
EXT	술어로 인한 양적인 변화를 나타낸다.
PRP	행동의 목적을 나타낸다.
PRD	주어에 대한 서술을 하고 있으나 주 서술어가 아닌 경우에 명시 한다.
CAU	행동이 일어나게 된 원인을 나타낸다.
DIS	술어에 연관 된 접속사.
NEG	부정의 의미를 명시한다.
CND	조건이나, 경우, 만약에 대해 나타낸다.
INS	행동에 사용된 도구나 수단을 나타낸다.
AUX	보조 용언을 나타낸다.
ADV	위의 목록에 해당하지 않으며, 부사적인 역할을 하는 경우에 사용한다.

표 1. 의미역 Modifier

Wikipedia에서 가져온 문장으로 말뭉치를 확장 할 때 Korean PropBank에 존재하지 않는 술어들은 유의어 관계에 있는 프레임 파일을 참조하여 프레임 파일을 새로 추가 하였으며, 이는 네이버 유의어 사전 및 세종 사전을 참조하여 진행하였다. 표 2 는 새로 추가 하거나, 의미를 추가한 프레임 파일 목록의 일부이다.

새로 추가한 프레임 파일	각색, 간행, 강력, 결합, 고단, 고통, 구제, 등용, 반납, 발병, 배양, 변동, 살생, 세우, 순수, 유래, 곡, 장엄, 저작, 점프, 정의, 추리, 추모, 출소, 탐험, …
의미를 추가한 프레임 파일	거두, 꺾, 꾸미, 드리, 물들, 이르, …

표 2. 추가한 프레임 파일 목록

표 3 은 영어의 'give' 에 해당하는 의미만을 가진 'teu-ri' 라는 프레임 파일에 '땋은 머리끝에 댕기를 물리다' 라는 의미를 추가한 부분이다. 모든 프레임 파일은 xml 형식을 따른다.

```
<frameset>
      <id>드리.02</id>
      <edef>땋은 머리 끝에 댕기를 물리다.</edef>
      <roleset>
          <role argnum="1" argrole="happen this action"/>
          <role argnum="2" argrole="where"/>
      </roleset>
     <frame>
          <mapping>
            <rel>드리다</rel>
            <mapitem src="obj" trg="arg1"/>
            <mapitem src="NP_AJT" trg="arg2"/>
          </mapping>
          <example>
             <text>분홍 두루마기에 연두 토시를 끼고 머리에
                는 갑사댕기를 드렸다.</text>
            <parse>
            </parse>
            <relation>
            </relation>
          </example>
     </frame>
</frameset>
```

표 3. 의미를 추가한 프레임 파일의 일부분

```
"SRL" : [
{"verb" : "열리", "sense" : 1, "word_id" : 14, "weight" :
26.6804,
      "argument" : [
      {"type" : "ARG1", "word_id" : 16, "text" : "영화제이며,"},
      {"type" : "ARGM-LOC", "word_id" : 12, "text" : "섬에서"},
      {"type" : "ARGM-TMP", "word_id" : 13, "text" : "매년"}
] },

{"verb" : "오래되", "sense" : 1, "word_id" : 19, "weight" :
26.6804,
     "argument" : [
    {"type" : "ARGM-LOC", "word_id" : 17, "text" : "세계에서"},
    {"type" : "ARGM-EXT", "word_id" : 18, "text" : "가장"},
    {"type" : "ARG1", "word_id" : 20, "text" : "영화제이기도"}
] }
]
```

표 4. 확장한 말뭉치의 의미역 부분

표 4 는 확장한 Korean PropBank중 의미역 부분에 대한 예제이다. '열리' 와 '오래되' 라는 술어가 각각 논항들을 가지고 있는 것을 볼 수 있다. 확장한 모든 말뭉치는 JSON(JavaScript Object Notation) 표기 형식을 따른다.

확장한 Korean PropBank는 의미역 말뭉치 889 문장과 새로 추가한 프레임 파일 89개, 의미를 추가한 프레임 파일 14개이다. 확장한 Korean PropBank의 신뢰성을 위해 2명의 평가자로 의미역 태깅 일치도를 구하였다. 평가 대상은 확장한 Korean PropBank중 임의로 추출한 106개의 술어에 대해 실시하였으며 문장수로는 16문장에 해당된다. 평가 기준은 평가자들이 태깅한 의미역이 서로 일치하는지를 확인하는 것으로 하였으며 94.3%의 일치도를 얻었다.

4. 도메인 적응 기술 적용

기존 Korean PropBank 4882문장을 Srouce 데이터로, 확장한 Korean PropBank 889문장을 Target 도메인으로 정의하고 한국어 의미역 결정 시스템에 대해 도메인 적응 기술 별 성능 평가를 실시한다.

성능을 평가 할 도메인 적응 기술은 다음과 같다. TGT-ONLY(baseline), SRC-ONLY(baseline), Prior, 선형 보간법(Linear Interpolation), ALL. 학습 속도 평가 시 SRC-ONLY와 선형 보간법의 경우 SRC-ONLY 학습 부분의 속도가 고정 되어 있으므로 표기하지 않았다. 모든 실험은 AMD A10-5800K APU (3.80 GHz), 16 GB RAM, Windows7 64-bit OS에서 수행되었다.

성능 평가 척도는 정확도(Precision), 재현율(Recall)의 조화 평균인 F-measure를 사용하였다.

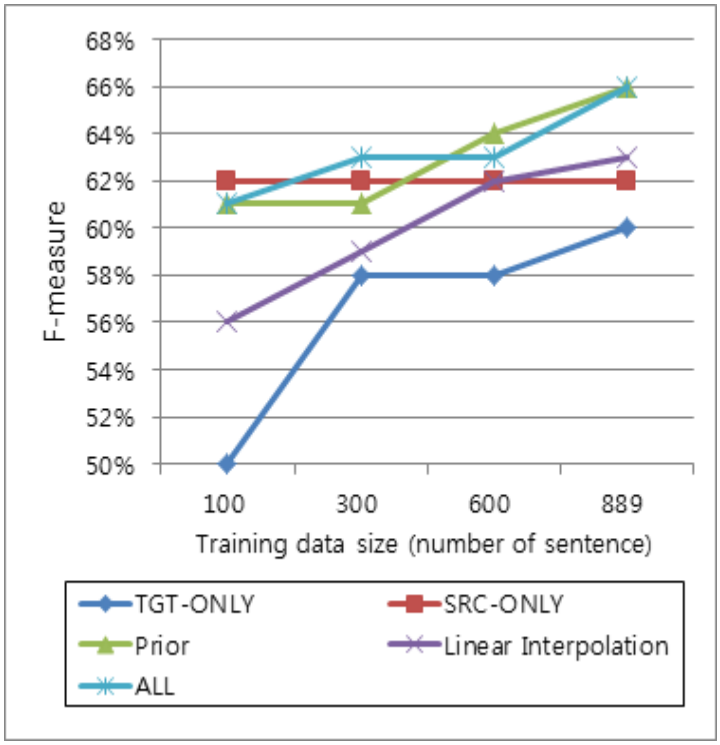

그림 2. 도메인 적응 기술 별 성능

기존 [6]의 실험에서 Source 도메인으로 학습한 의미역 결정 시스템을 Source 도메인에 적용 했을 시 74.2%(F)의 성능을 보였었다. 그러나 다른 도메인 말뭉치인 확장한 Korean PropBank (Wikipedia)에 적용 했을 시 그림 2

에서 보이는 바와 같이 62.2%로 약 12.0% 정도 성능이 하락하였다. 도메인 적응 기술 별 성능 평가 결과는 그림 2 와 같다. Prior 도메인 적응 기술과 ALL 방법이 한국어 의미역 결정 시스템에서 가장 높은 성능을 보였다. TGT-ONLY 방법은 다른 도메인 적응 기술 방법들 보다 낮은 성능을 보이는데, 이는 Target 도메인의 말뭉치가 적기 때문이라고 유추할 수 있다.
또한 그림 2 를 보면 TGT-ONLY가 매우 낮은 성능을 보이다가 학습 말뭉치의 양이 증가할수록 점차 성능이 향상되는 걸 볼 수 있다. 이로부터 학습 말뭉치의 양이 의미역 결정 시스템 성능에 영향을 미친다는 것을 알 수 있다. 그림 2 에서 TGT-ONLY 방법이 가장 낮은 성능을 보인다는 것은 반대로 도메인 적응 기술이 활용 할 만 한 기술이라고 얘기할 수 있다. 그림 3 은 도메인 적응 기술 별 학습 속도를 보여주는데 Prior 방법이 TGT-ONLY 모델과 유사한 속도를 내며, ALL 방법보다 월등히 빠른 학습 속도를 보여준다는 것을 알 수 있다.

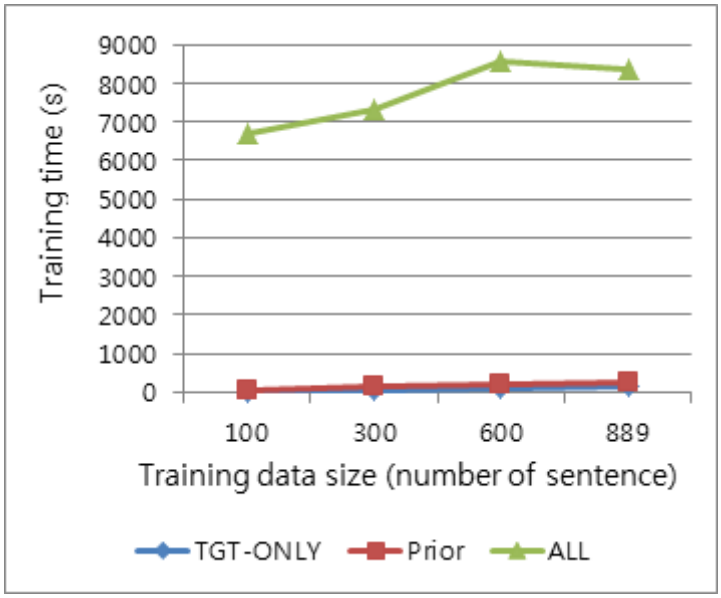

그림 3. 도메인 적응 기술 별 학습 속도

5. 결론

본 논문에서는 한국어 의미역 결정 시스템에 필요한 Korean PropBank를 확장하기 위하여 Wikipedia 로부터 가져온 문장에 의미역을 추가하여 약 1/5의 말뭉치를 확장 시켰다. 또한 Korea PropBank의 격틀 정보에 해당하는 프레임 파일을 추가 하였다.

도메인 적응 기술 적용에 대한 실험 결과로부터 학습에 사용한 도메인과 다른 도메인에 대한 의미역 결정 시스템 적용 시 의미역 결정 성능이 하락함을 확인 하였다. 또한 도메인 적응 기술을 말뭉치가 적은 도메인에 대해 적용 하였을 때 TGT-ONLY(baseline)방법이 60.0%, Prior 방법이 65.6%로 약 5.6%의 성능 향상이 있었다. 이는 기존의 학습 데이터를 활용하여, 적은 양의 새로운 학습 말뭉치만을 가지고 성능 하락을 최소화할 수 있음을 나타낸다. 또한 도메인 적응 기술 중 Prior 모델이 성능 및 학습 시간 면에서 가장 우수하다고 보여 졌다.

향후 연구로는 한국어 의미역 말뭉치를 확장은 물론 도메인 적응 기술들을 다른 도메인들에 대해서도 적용, 비교 평가 할 필요가 있다.

감사의 글

본 연구는 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술 개발사업(정보통신)의 일환으로 수행하였음[10044577, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술개발]

참고문헌

[1] Palmer Martha, Daniel Gildea, Paul Kingsbury. "The proposition bank: An annotated corpus of semantic roles", Computational Linguistics 31, 1, 71-106, 2005.
[2] Palmer Martha, et al. "Korean Propbank", LDC Catalog No.: LDC2006T03 ISBN : 1-58563, (2006)
[3] 김병수, et al. “비지도 학습을 기반으로 한 한국어 부사격의의미 역 결정”, 정보과학회논문지: 소프트웨어 및 응용, 34.2, 2007.
[4] X. Carreras and L. Marquez, “Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling,” Proc. CoNLL, Ann Arbor, Michigan, USA, June 30 , pp. 152-154, 2005.
[5] C. Lee and M. Jang, “A Prior Model of Structural SVMs for Domain Adaptation,” ETRI J., vol. 33, no. 5, pp. 712-719, Oct. 2011.
[6] 이창기, 임수종, 김현기. Structural SVM 기반의 한국어 의미역 결정. 한국정보과학회 학술발표논문집, 574-576, 2014.
[7] 배장성, 오준호, 이창기,et al. 한국어 의미역 말뭉치 구축을 위한 반자동 태깅 도구 개발. 한국 정보과학회 학술발표논문집, 592-594, 2014.
[8] DAHLMEIER, Daniel; NG, Hwee Tou. Domain adaptation for semantic role labeling in the biomedical domain. Bioinformatics, 26.8: 1098-1104, 2010.
[9] NILSSON, Jens; RIEDEL, Sebastian; YURET, Deniz. The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL shared task session of EMNLP-CoNLL. Sn, p. 915-932, 2007.
[10] H. Daumé and D. Marcu, “Domain Adaptation for Statistical Classifiers,” J. Artif. Intell. Res., vol. 26, no. 1, pp. 101-126, May 2006.
[11] Soojong Lim, Changki Lee, Pum-Mo Ryu, Hyunki Kim, Sang Kyu Park, Dongyul Ra. Domain-Adaptation Technique for Semantic Role Labeling with Structural Learning. ETRI Journal, vol.36, no.3, pp.429-438, June 2014.
[12] BABKO-MALAYA, Olga. Propbank annotation guidelines. URL: http://verbs, 2005.