인지과학, 제26권 제4호
Korean Journal of Cognitive Science
2015, Vol. 26, No. 4, 377～392.

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

배 장 성 이 창 기
강원대학교 컴퓨터과학과

한국어 의미역 결정(Semantic Role Labeling)은 주로 기계학습에 의해 이루어지며 많은 말뭉치 자원을 필요로 한다. 그러나 한국어 의미역 결정 시스템에 사용되는 Korean PropBank는 의미역이 부착된 용언과 용언 격틀이 PropBank에 비해 각각 1/5, 1/2 수준에 불과하다. 따라서 본 논문에서는 한국어 의미역 결정 시스템을 위해 의미역이 부착된 용언과 용언 격틀을 확장하여 Korean PropBank를 확장 시키고자 한다. 대부분의 의미역 결정 시스템은 학습 도메인에 의존적이기 때문에 적용 도메인 변경에 따른 성능 하락이 나타날 수 있다. 본 논문에서는 기존의 학습 말뭉치와 적은 양의 새로운 학습 말뭉치를 활용하여 새로운 도메인에 대해 의미역 결정 시스템의 성능 하락을 최소화 할 수 있는 도메인 적응 기술을 Structural SVM(S-SVM)과 Deep Neural Network(DNN) 기반 한국어 의미역 결정 시스템에 적용하여 그 실효성을 알아보고자 한다.

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

† 교신저자: 이창기, 강원대학교 컴퓨터과학과, 강원대학교 한빛관 301호
연구분야: 자연어처리, 정보추출, 정보검색, 기계학습
Tel: +82-33-250-8443, E-mail: leeck@kangwon.ac.kr

Ⅰ. 서 론

의미역 결정(Semantic Role Labeling)은 문장에 존재하는 각 술어의 의미와 그 논항들의 의미적인 관계를 결정하는 자연어처리의 한 단계이다. 의미역 결정 연구는 크게 격틀사전에 기반을 둔 방법과 말뭉치에 기반을 둔 방법으로 나눌 수 있으며 최근에는 의미역 말뭉치와 기계학습 알고리즘을 이용한 연구가 활발하게 진행되고 있다[1-5].

일반적인 기계학습 기반의 의미역 결정 시스템은 해당 문장의 술어들을 식별하고 각 술어에 대한 논항들의 의미역을 결정하여 "누가, 무엇을, 누구에게, 어떻게, 왜" 등의 의미관계를 찾아내는 시스템이다. 예를 들면 그림 1의 입력 '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. 'NR'은 '속하.01'의 논항이 아님을 뜻하고 'ARG1', 'ARG2'는 술어 '속하.01'의 논항이 된다.

그림 1. 한국어 의미역 결정 시스템

기계학습 기반의 의미역 결정 시스템은 많은 양의 말뭉치를 필요로 한다. 의미역 결정 시스템에 널리 사용되는 말뭉치로는 영어 의미역 결정을 위한 PropBank[6]와 한국어 의미역 결정을 위한 Korean PropBank[7]가 있다. Korean PropBank는 Virginia 말뭉치와 Newswire 말뭉치로 구성되어있으나, Virginia 말뭉치가 군대 용어로 이루어져 있기 때문에 한국어 의미역 결정에는 주로 Newswire 말뭉치를 사용한다. Newswire 말뭉치는 2,749개의 용언 격틀과 23,707개의 의미역이 부착된 용언을 가지고 있다. 하지만 이는 4,659개의 용언 격틀과 112,917개의 의미역이 부착된 용언을 가지고 있는 PropBank보다 현저히 적은 수준이다. 따라서 본 논문에서는 한

국어 위키피디아에서 추출한 데이터와 한국어로 이루어진 동화 말뭉치를 이용하여 Korean PropBank를 확장한다. Korean PropBank의 확장은 말뭉치 양의 증가라는 의미보다, 새로운 도메인의 추가에 그 의미를 둔다.

일반적인 의미역 결정 시스템은 학습에 사용한 데이터와 평가에 사용하는 데이터가 같은 도메인으로 이루어져 있다. 반면 학습에 사용한 도메인이 아닌 다른 도메인으로 시스템을 평가할 경우 시스템 성능이 10% 이상 하락됨을 볼 수 있다[8]. 이는 기존 시스템을 다른 도메인에 적용할 때 큰 문제점이 된다. 도메인 변경 시 성능 하락을 극복하기 위해서는 변경되는 새로운 도메인에 대해 매번 충분한 양의 학습 데이터를 구축해 주면 된다. 하지만 의미역 결정 시스템에 필요한 말뭉치를 만드는 일은 사람이 손수 만들기 때문에 많은 시간과 비용을 필요로 한다.

이러한 문제를 해결하기 위한 방법 중 하나가 도메인 적응 기술이다. 도메인 적응 기술의 목적은 최소한의 신규 데이터 생성만으로 시스템의 적용 도메인 변경에 따른 성능 하락을 최소화 하는 것이다.

도메인 적응 기술은 의미역 결정 시스템 입력 데이터의 부족한 말뭉치 양을 보완할 수 있는 하나의 방법으로써 신규 도메인에 대해 많은 양의 말뭉치를 확보하지 못 했을 때 유용하게 사용될 수 있다. 본 논문에서는 도메인 적응 기술을 S-SVM과 DNN 두 종류의 한국어 의미역 결정 시스템에 적용하여 실효성을 알아보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 Korean PropBank 확장에 대해 설명한다. 4장에서는 도메인 적응 기술을 적용한 S-SVM과 DNN 두 종류의 한국어 의미역 결정 시스템의 실험 결과에 대해 설명한다. 5장에서는 결론에 대해 기술한다.

Ⅱ. 관련 연구

기존 연구 방법에 따르면, 의미역 결정 시스템은 크게 격틀 사전 기반의 시스템과 의미역 말뭉치에 기반을 둔 시스템으로 나눌 수 있다. 격틀 사전에 기반을 둔 시스템은 입력 문장과 격틀 사이의 유사도 계산 과정을 통해 의미역이 결정되기

때문에 처리속도가 빠르고 높은 정확률을 보이지만, 격틀 사전의 구축이 어렵고 격틀 사전에 기술되지 않은 임의격을 처리하지 못하는 문제가 있다. 의미역 말뭉치에 기반을 둔 시스템은 의미역 말뭉치와 기계학습 알고리즘을 이용하여 의미역을 결정하는 방법이다. 이 방법은 격틀 사전에 기반을 둔 방법에 비해 적용률이 높은 장점이 있으나, 의미역 말뭉치 구축이 어렵다는 단점이 있다[9].

격틀 사전 기반의 시스템과 의미역 말뭉치 기반의 시스템 모두 격틀과 의미역 말뭉치라는 언어 자원을 필요로 하므로 의미역 결정을 연구함에 있어 말뭉치는 필수불가결한 자료라 말할 수 있다.

PropBank는 의미역 결정에 필요한 자원으로써 사람이 수작업으로 만든 영어 말뭉치이다. PropBank는 4,659개의 용언 격틀과 112,917개의 의미역이 부착된 용언을 가지고 있다. Korean PropBank는 PropBank를 기반으로 만들어진 한국어 말뭉치이며 용언 격틀 2,749개, 의미역이 부착된 용언이 23,707개 (Virginia 말뭉치 제외)로 PropBank에 비해 말뭉치의 양이 적어 연구에 어려움이 따른다.

최근 의미역 결정은 의미역 말뭉치와 기계학습 알고리즘을 이용한 방법이 영어 및 한국어에서 많이 연구되고 있다[3,4,8,10]. [3,4]는 Korean PropBank를 학습 말뭉치로 사용하는 시스템으로 [3]은 한국어 의미역 결정을 sequence labeling(순차적 레이블링) 문제로 바꾸어, sequence labeling에 좋은 성능을 보이는 S-SVM을 이용한다. [4]는 최근 여러 분야에서 연구가 이루어지고 있는 DNN을 한국어 의미역 결정에 적용하였다. 한국어 의미역 말뭉치 구축은 [3,4]와 같이 말뭉치 기반의 시스템들을 위해 필요한 일이다. 따라서 본 논문에서는 Korean PropBank를 확장하고, 확장 작업을 보다 빨리 하기 위해 [5]에서 개발한 의미역 반자동 태깅 도구를 사용하였다.

도메인 적응 기술은 새로운 도메인에 대해 부족한 말뭉치 자료의 양을 보완할 수 있는 기술이다. 따라서 본 논문에서는 도메인 적응 기술을 한국어 의미역 결정에 적용하고 그 실효성을 알아본다.

도메인 적응 기술에서는 이미 보유하고 있는 도메인을 '소스(source)', 만들어진 의미역 결정 시스템을 적용 할 도메인에 대해서 '타겟(target)' 으로 정의하여 구분짓고 있다. 대부분의 의미역 결정 시스템이 학습에 사용한 도메인과 실제 적용 할 도메인을 같은 도메인으로 두고 있으나, 이와 달리 학습에 사용한 도메인과 실제 적용할 도메인이 다른 경우 시스템의 성능 하락이 있을 수 있다[8].

[11,12]에서는 도메인 적응 기술을 다음과 같이 분류하고 있다.

- **Source-only(SRC-ONLY)**: 도메인 적응의 가장 기본이 되는 방법으로, 소스 도메인의 학습 데이터만을 의미역 결정 시스템의 학습에 사용하며, 도메인 적응 기술의 베이스라인으로 사용된다.
- **Target-only(TGT-ONLY)**: 새롭게 적용할 도메인, 즉 새로 구축한 타겟 도메인의 학습 데이터만을 의미역 결정 시스템 학습에 사용하며, 도메인 적응 기술의 베이스라인으로 사용된다.
- **ALL**: SRC-ONLY와 TGT-ONLY 방법이 더해진 것으로, 소스와 새롭게 적용할 타겟 도메인의 학습 데이터 모두를 의미역 결정 시스템의 학습에 사용하고 이를 통해 모델을 구축한다.
- **Weighted**: ALL모델에 사용하는 소스, 타겟 도메인의 비율에 맞게 가중치를 적용하여 모델을 구축한다. 이때의 가중치는 사용자가 여러 실험을 통해 최적의 값을 찾아내게 되는 파라미터(parameter)가 된다.
- **Prior**: SRC-ONLY 모델의 가중치 벡터(weight vector)를 타겟 도메인 학습에 이용한다. 타겟 도메인 학습 시 SRC-ONLY 모델로 생성된 가중치를 타겟 도메인 가중치 벡터의 시작점으로 초기화하여 학습[13]하게 되기 때문에 빠른 학습속도를 보인다.
- **Linear Interpolation(LIN-INT)**: SRC-ONLY, TGT-ONLY 방법으로 각각 모델을 구축하고, 선형 보간법을 이용하여 구축된 각 모델을 하나의 모델로 통합하며 수식은 다음과 같다.

$$A*SRC\text{-}ONLY.MODEL+(1-A)TGT\text{-}ONLY.MODEL$$

- **PRED**: SRC-ONLY 방법으로 구축된 모델을 이용하여 새롭게 구축한 타겟 도메인을 분석하고, 그 결과를 TGT-ONLY 모델의 자질로 사용한다.
- **Feature Augmentation(FA)**: 의미역 결정 학습에 필요한 자질을 소스 도메인에 적합한 자질, 타겟 도메인에 적합한 자질, 소스와 타겟 도메인에서 공통적으로 사용 가능한 자질로 분류하고 각각의 자질 모두를 이용하여 독립된 모델을 구축한다.

Ⅲ. Korean PropBank 확장

Korean PropBank는 2,749개의 격틀 정보를 저장하고 있는 프레임 파일과 23,707개의 의미역이 부착된 용언으로 이루어진 말뭉치이다. 이는 영어 격틀 정보의 1/2, 의미역이 부착된 용언이 1/5 수준에 불과하여 한국어 의미역 결정 연구에 어려움이 있다. 본 논문에서는 한국어 위키피디아에서 가져온 질문과 정답 쌍으로 이루어진 문장과 한국어 동화 말뭉치에 대해 의미역 정보를 추가 하여 Korean PropBank를 확장하였다.

의미역 정보 추가를 위해 새로운 학습 데이터에 대해 기존 기계학습 기반 의미역 결정 시스템을 적용하여 일차 결과물을 얻고, [5]에서 개발한 반자동 의미역 태깅 도구를 이용하여 일차 결과물의 틀린 부분을 수정하는 방식으로 진행하였다.

말뭉치 구축 시 A0(행위주), A1(피동자)과 같은 논항에 대한 태그셋은 PropBank annotation guidelines[14]에 정의되어 있는 규칙을 따랐으며, 기본적인 의미역과 별도

표 1. 의미역 Modifier

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

로 문장을 더 상세하게 서술하기 위해 PropBank annotation guidelines의 Modifier 태그 중 일부를 사용하였다. 표 1은 본 논문에서 Korean PropBank 확장에 사용한 의미역 Modifier의 정의를 보여주고 있다.

새로운 학습 데이터를 이용하여 말뭉치를 확장 하면 기존 격틀 사전에 존재하지 않는 용언을 정의해야 하는 일이 발생한다. 이때에는 유의어 관계에 있는 프레임 파일을 참조하여 프레임 파일을 새로 추가하였으며, 유의어 판단은 네이버 유의어 사전 및 세종 사전에 정의되어 있는 정보를 사용하였다.

표 2는 새로 추가 하거나, 의미를 추가한 프레임 파일 목록의 일부이다.

표 2. 추가한 프레임 파일 목록의 일부

새로 추가한 프레임 파일	가로지르, 각색, 간행, 강력, 결합, 고단, 고통, 구제, 그릇되, 기구, 단단하, 대중화, 등용, 반납, 발병, 배양, 유래, 익히, 작곡, 장엄, 저작, 점프, 정의, 추모, 출소, 친숙, 탐험, 토벌, 팽창, 편찬, 평등, 폐기, …
의미를 추가한 프레임 파일	거두, 꺾, 꾸미, 나, 대비, 드리, 물들, 수상, 쓰, 이루, 이르, 찍, 켜, 포장

그림 2는 영어의 ‘give’에 해당하는 의미를 가진 ‘teu-ri’ 라는 프레임 파일에 ‘땋은 머리끝에 댕기를 물리다’ 라는 새로운 의미를 추가한 부분이다. 모든 프레임 파일은 xml 형식으로 구성된다.

그림 3은 확장한 Korean PropBank의 의미역 말뭉치에 대한 예제이다. “SRL”로 표시하여 의미역 정보를 저장하고 있음을 알 수 있으며, ‘열리’와 ‘오래되’ 라는 술어가 “verb”로 표시되어 있고 각각 “argument”라고 속성이 표시된 논항들을 가지고 있는 것을 볼 수 있다. 각 논항들은 “type” 속성을 가져 자신이 어떤 의미역 정보를 가지고 있는지 알려주게 된다. 그림 3에 나타난 나머지 속성은 서술어와 각 논항들의 문장 정보에 해당된다. 확장한 의미역 말뭉치는 JSON(JavaScript Object Notation)표기 방법을 따른다.

```
<frameset>
        <id>드리.02</id>
        <edef>땋은 머리끝에 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 머리에는
                    갑사댕기를 드렸다.</text>
              <parse>
              </parse>
              <relation>
              </relation>
            </example>
      </frame>
</frameset>
```

그림 2. 프레임 파일에 의미를 추가한 부분

```
"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"}
] }
]
```

그림 3. 확장한 의미역 말뭉치

확장한 Korean PropBank는 한국어 위키피디아에서 가져온 905개 문장과 동화 말뭉치에서 가져온 631개의 문장에 각각 2,399 / 1,261개의 의미역이 부착된 용언으로 구성되었다. 새로 추가한 프레임 파일은 156개이며 기존의 격틀 정보에 의미를 덧붙인 프레임 파일은 26개다.

확장한 Korean PropBank의 신뢰성 검증을 위해 2명의 숙련된 평가자를 통해 의미역 태깅 일치도를 구하였다. 평가는 새로 확장한 의미역 말뭉치 중 임의로 추출한 200개의 말뭉치를 대상으로 실시하였다. 평가 기준은 각 평가자가 태깅한 의미역이 서로 일치하는지를 확인하는 것으로 94.3%의 일치도를 보였다.

Ⅳ. 도메인 적응 기술 적용

본 논문에서는 기존 Korean PropBank 말뭉치를 소스 도메인으로, Korean PropBank를 확장하기 위해 만든 새로운 말뭉치를 타겟 도메인으로 정의해 사용하였다. 타겟 도메인은 한국어 위키피디아 도메인과 동화 도메인으로 나누어 사용하고 이를 각 도메인 적응 기술 별로 S-SVM과 DNN 두 종류의 한국어 의미역 결정 시스템에 적용해 본다.

한국어 위키피디아 말뭉치는 총 2,399개의 의미역이 부착된 용언으로 구성되며 학습에 1659개, 평가에 740개를 사용하였다. 동화 도메인의 경우 총 10개의 동화로 이루어져 있으며, 9개의 동화 말뭉치로 학습을 수행하고 남은 1개의 동화 말뭉치로 평가를 진행하여 10개의 동화 모두에 대하여 성능 평가를 수행하였다. 학습에 사용한 시스템은 의미역 결정에 좋은 성능을 보이는 것으로 알려져 있는 S-SVM과 DNN의 한 종류인 FFNN(Feed-Forward Neural Network)을 이용하였다.

S-SVM을 이용한 실험에 사용한 도메인 적응 기술은 다음과 같다. SRC-ONLY, TGT-ONLY, Prior, LIN-INT, ALL. FFNN을 이용한 실험은 SRC-ONLY와 TGT-ONLY, pre-training, ALL 기술을 사용하였다. Pre-training 기술은 SRC-ONLY 모델로 학습한 가중치를 TGT-ONLY 모델의 학습 초기 가중치로 주어 Prior 모델의 개념과 유사하게 학습하였다.

실험은 Intel i5-2400 CPU(3.1GHz), 16GB RAM, Windows7 64-bit OS에서 수행되었고, 한국어 의미역 결정 시스템의 성능을 평가하기 위한 척도로 정확도(Precision), 재현율(Recall)의 조화 평균인 F1 값을 사용하였다.

적용 도메인의 변경이 없는 기존 연구 방법, 즉 소스 도메인의 한국어 의미역 결정 시스템 성능은 S-SVM 시스템이 76.96%, FFNN 시스템이 75.45%의 성능을 보였다. 그러나 위의 시스템을 다른 도메인(한국어 위키피디아)에 적용할 경우 S-SVM 시스템에서 62.8%로 약 14%, FFNN에서 59.36%으로 약 16% 정도 성능이 하락하였다. 이 결과로부터 적용 도메인의 변경 시 한국어 의미역 결정에서 성능하락이 나타남을 알 수 있다.

표 3. 기계학습 시스템의 도메인 적응 기술 별 성능

한국어 위키피디아	S-SVM	FFNN
SRC-ONLY	62.80	59.36
TGT-ONLY	57.62	56.65
Prior/pre-training	64.92	62.43
ALL	64.75	62.14

표 3은 한국어 위키피디아를 타겟 도메인으로 하여 두 종류의 기계학습 시스템에 도메인 적응 실험을 수행한 결과이다. 실험 결과를 보면 새로 구축한 타겟 도메인의 데이터만을 사용하는 TGT-ONLY 모델이 가장 낮은 성능을 보인다. 이 결과는 의미역 말뭉치를 사용하는 기계학습 기반 의미역 결정 시스템에서 의미역 말뭉치의 크기가 의미역 결정 성능에 큰 영향을 미친다는 것을 알 수 있다. 또한 SRC-ONLY 모델과 TGT-ONLY 모델의 성능 비교를 통해 잘 갖추어진 소스 도메인이 이미 존재 한다면, 새로운 도메인에 대해 타겟 도메인으로만 학습 했을 때 보다 높은 성능을 기대할 수 있음을 알 수 있다. 또한 Prior/pre-training 방법의 성능이 ALL 모델에 비해 높은 성능을 보이는 것을 알 수 있다.

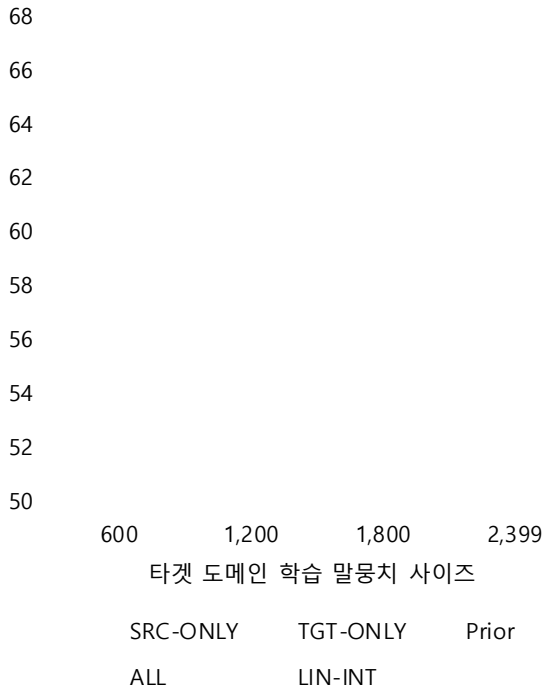

그림 4. 학습 말뭉치 사이즈에 따른 도메인 적응 기술 별 성능
(S-SVM/한국어 위키피디아)

그림 4는 기계학습(S-SVM) 시스템의 도메인 적응 기술 별 성능을 타겟 도메인의 의미역 말뭉치의 크기에 따라 나타낸 그래프이다. SRC-ONLY를 제외한 도메인 적응 기술의 성능 지표가 학습 말뭉치의 사이즈가 증가함에 따라 함께 증가하는 추세를 보인다. ALL과 Prior 모델은 전 구간에서 SRC-ONLY 모델 보다 높은 성능을 보이고 있다. LIN-INT는 타겟 도메인의 학습 말뭉치가 적을 때 성능이 크게 떨어지는데, 불충분한 데이터의 양에 대한 선형 보간법의 한계로 볼 수 있다.

그림 5, 6은 동화 도메인에 대한 도메인 적응 기술 별 성능 평가 그래프이다. 표 3의 결과와 마찬가지로 Prior/pre-training 도메인 적응 기술이 전체적으로 가장 높은 성능을 보였고 ALL 모델도 비슷한 성능을 보여주었다.

마지막으로 도메인 적응 실험을 수행하며 각 도메인 적응 기술 별 학습 시간 속도를 측정한 결과 S-SVM의 Prior와 FFNN의 pre-training이 가장 빠른 학습 속도를 보였다. 여러 실험에서 비슷한 성능을 보였던 ALL 모델은 Prior/pre-training 모델보다 약 80배 이상 느린 학습 속도를 보였다. 따라서 Prior와 FFNN의 pre-training 방법이 ALL 모델보다 성능과 학습 속도에 있어 대부분 우수한 것을 알 수 있다.

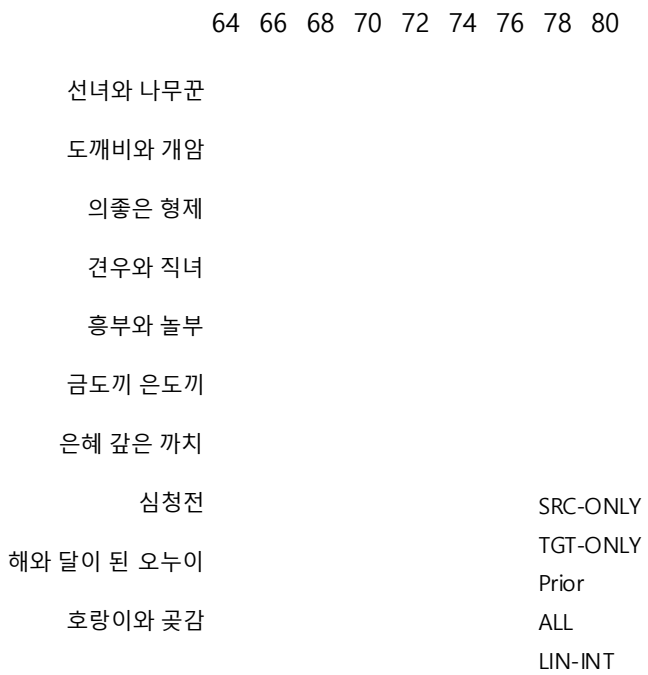

그림 5. 도메인 적응 기술 별 성능 (동화/S-SVM)

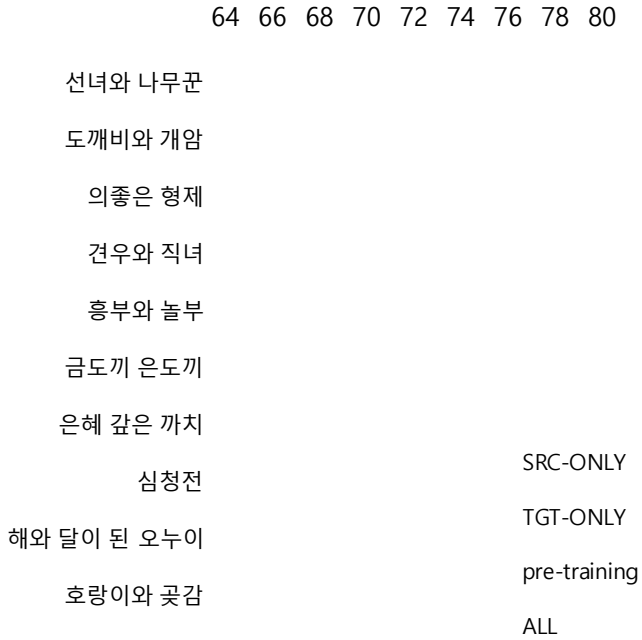

그림 6. 도메인 적응 기술 별 성능 (동화/FFNN)

V. 결 론

본 논문에서는 한국어 의미역 결정 시스템에 주로 사용되는 언어 자원인 Korean PropBank를 확장하기 위하여 한국어 위키피디아로부터 가져온 문장과 동화 말뭉치에 의미역 정보를 추가하여 격틀 정보 183개와 의미역이 부착된 용언 4,660개를 추가하였다.

추가한 두 개의 타겟 도메인에 대해 다양한 도메인 적응 기술을 적용한 결과, 학습데이터와 다른 도메인을 적용 시 큰 폭의 성능 하락이 나타남을 알 수 있었다. 성능 하락의 폭은 도메인마다 다르며, 도메인 적응 기술마다 하락폭이 다른 것을 알 수 있었다. 따라서 신규 도메인에 대해 새로운 시스템을 적용 할 때 신규

도메인의 말뭉치 자원이 많지 않다면, 적절한 도메인 적응 기술을 선택하여 성능 하락을 막을 수 있을 것으로 기대된다.

향후 연구로는 한국어 의미역 말뭉치를 기존과 다른 도메인으로 확장하여 여러 도메인에 적용할 예정이며 기존의 도메인 적응 기술 알고리즘의 문제점을 보완하고 높은 성능을 보였던 기술들의 장점을 합쳐 새로운 방법의 도메인 적응 기술을 개발할 예정이다.

감사의 글

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

참고문헌

[1] 정현기, 김유섭 (2011). 확장된 격틀 사전을 이용한 한국어 부사격 논항의 의미역 결정. **한국정보기술학회논문지,** 167-176.

[2] 김완수, 옥철영 (2015). 한국어 격틀 사전과 의미역 빈도 정보를 사용한 한국어 의미역 결정. **한국정보과학회 학술발표논문집,** 651-653.

[3] 이창기, 임수종, 김현기 (2014). Structural SVM 기반의 한국어 의미역 결정. **한국정보과학회 학술발표논문집,** 574-576.

[4] 배장성, 이창기, 임수종 (2015). 딥 러닝을 이용한 한국어 의미역 결정. **한국정보과학회 학술발표논문집,** 690-692.

[5] 배장성, 오준호, 박천음, 최경호, 이창기 (2014). 한국어 의미역 말뭉치 구축을 위한 반자동 태깅 도구 개발. **한국정보과학회 학술발표논문집,** 592-594

[6] Palmer Martha, Daniel Gildea, Paul Kingsbury (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31-1, 71-106.

[7] Palmer Martha, Ryu Shijong, Choi Jinyoung, Yoon Sinwon, Jeon Yeongmi (2006).

Korean PropBank. LDC Catalog No: LDC2006T03 ISBN, 1-58563.

[8] X. Carreras, L. Marquez (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling, Proceedings of the Ninth Conference on Computational Natural Language Learning. *Association for Computational Linguistics*, 152-164.

[9] 김병수, 이용훈, 이종혁 (2007). 비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정. **정보과학회논문지: 소프트웨어 및 응용,** 34-2, 112-122.

[10] Zhou Jie, Wei Xu.(2015). End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks. *Association for Computational Linguistics*, 1127-1137.

[11] Blitzer John, Daumé III Hal (2010). Domain Adaptation. International Conference on Machine Learning tutorial.

[12] Daumé III Hal (2007). Frustratingly easy domain adaptation. *Association for Computational Linguistics*, 256-263.

[13] Soojong Lim, Changki Lee, Pum-Mo Ryu, Hyunki Kim, Sang Kyu Park, Dongyul Ra (2014). Domain-Adaptation Technique for Semantic Role Labeling with Structural Learning. *ETRI Journal,* 36-3, 429-438.

[14] Babko-Malaya, Olga (2005). Propbank annotation guidelines. URL:http://verbs. colorado.edu.

1차원고접수 : 2015. 04. 06
1차심사완료 : 2015. 06. 02
2차원고접수 : 2015. 09. 01
2차심사완료 : 2015. 10. 15
3차원고접수 : 2015. 10. 26
최종게재승인 : 2015. 11. 05

(*Abstract*)

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

Jangseong Bae Changki Lee

Kangwon National University

Korean semantic role labeling (SRL) is usually performed by a machine learning and requires a lot of corpus. However, the Korean PropBank used in Korean SRL system is less than PropBank. It leads to a low performance. Therefore, we expand the annotated corpus and verb frames for Korean SRL system to expand the Korean PropBank corpus. Most of the SRL system have a domain-dependent performance so, the performance may decrease if domain was changed. In this paper, we use the domain adaptation technique to reduce decreasing performance with the existing corpus and the small size of new domain corpus. We apply the domain adaptation technique to Structural SVM and Deep Neural Network. The experimental result show the effectiveness of the domain adaptation technique.

Key words : Domain adaptation technique, Korean semantic role labeling, Koran PropBank