도메인 적응 기술을 이용한 한국어 의미역 인식

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

Korean Semantic Role Labeling Using Domain Adaptation Technique

Soojong Lim○, Yongjin Bae, Hyunki Kim
Automatic Speech Translation and AI Research Center, ETRI

요 약

기계학습 방법에 기반한 자연어 분석은 학습 데이터가 필요하다. 학습 데이터가 구축된 소스 도메인이 아닌 다른 도메인에 적용할 경우 한국어 의미역 인식 기술은 15% 정도 성능 하락이 발생한다. 본 논문은 이러한 다른 도메인에 적용시 발생하는 성능 하락 현상을 극복하기 위해서 기존의 소스 도메인 학습 데이터를 활용하여, 소규모의 타겟 도메인 학습 데이터 구축만으로도 성능 하락을 최소화하기 위해 한국어 의미역 인식 기술에 prior 모델을 제안하며 기존의 도메인 적응 알고리즘과 비교 실험하였다. 추가적으로 학습 데이터에 사용되는 자질 중에서, 형태소 태그와 구문 태그의 자질 값을 기존보다 단순하게 적용하여 성능의 변화를 실험하였다.

주제어: 한국어 의미역 인식, 도메인 적응 기술, prior 모델, 자질값 단순화

1. 서론

의미역 인식(Semantic Role Labeling)이란 자연어 문장에서 'Who does what to whom'을 인식하는 기술로, 문장의 서술어를 중심으로 서술어에 대한 의미적인 역할(예를 들어, 행위자, 경험자, 대상격, 도구격 등)을 하는 문장의 부분을 인식하는 것을 말한다. 지식을 처리하는 응용 서비스가 발달함에 따라서 형태소 분석, 개체명 인식 같은 어절 단위 자연어 분석 기술 이외에도 의미역 인식 같은 문장 단위 의미 분석 기술에 대한 수요도 점점 늘어나고 있는 추세이다.

영어권에서는 CoNLL-2004를 시작으로 의미역 인식에 관한 연구가 활발히 진행되고 있는데, Out of Domain을 다루기 시작한 CoNLL-2005에서는 구구조 기반 학습 데이터가 구축된 소스 도메인(WSJ corpus)이 아닌 다른 타겟 도메인(Brown corpus)에 적용할 경우 10% 이상의 성능 하락 현상이 일어났고[1], 의존 구분 분석 결과 기반 의미역 인식을 수행한 CoNLL-2008 Shared Task에서도 마찬가지로 성능 하락 현상이 발생하였다[2].

이러한 성능 하락 현상을 극복하는 방법은 타겟 도메인에 대해서도 소스 도메인만큼의 학습 데이터를 구축하여 타겟 도메인에서도 같은 시스템을 새롭게 구축하는 것이지만, 이는 시간과 비용적인 측면에서 장애가 되는 요소이다. 도메인 적응 기술(Domain Adaptation)은 이러한 문제를 적은 양의 타겟 도메인의 학습 데이터 구축으로도 소스 도메인에 비해 급격한 성능 하락을 방지하기 위해서 제안되었다.

그림1[3]은 소스 도메인의 데이터와 알고리즘(Algorithm 1)을 이용하여 구축된 소스 모델($\mathbf{w}_{src}$)를 입력받아, 타겟 도메인에서 구축된 학습 데이터와 도메인 적응 알고리즘(Algorithm 2)를 이용하여 최종적으로 타겟 도메인에 최적화된 타겟 모델($\mathbf{w}_{tgt}$)을 구축하는 도메인 적응 기술을 적용하는 과정을 보여준다.

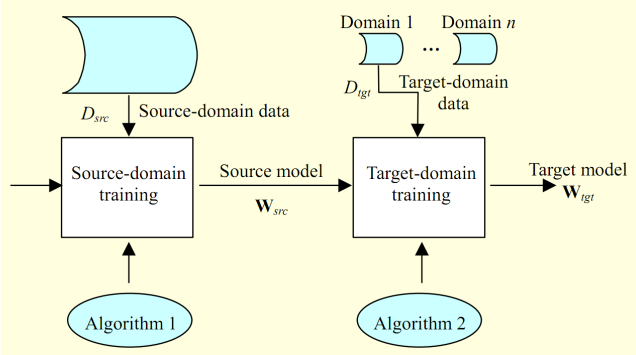

그림 1. 도메인 적응 시스템 구축 과정

본 논문은 뉴스 도메인 Korean Propbank에서 개발된 한국어 의미역 인식 시스템[4]를 위키피디아 기반 질의응답 시스템에서 사용하기 위해, 도메인 적응 기술을 적용하여 백과사전(위키피디아 문서) 도메인에서 효과적으로 작동하는 한국어 의미역 인식 시스템을 구축한다. 본 논문 구성은 다음과 같다. 2장에서는 기존 연구를 소개하고, 3장에서는 본 논문에서 제안하는 도메인 적응 기술을 적용한 한국어 의미역 인식 기술에 대해서 설명하고, 4장에서는 실험 및 결과를 분석하며, 5장에서는 결론에 대해서 기술한다.

2. 관련 연구

도메인 적응 기술은 다양한 분야의 기술에 적용하기 위해서 제안되어 왔다.

Daume and Marcu[5]는 다음과 같이 도메인 적응 기술을 분류하였다.

- source only(SRC-only): 소스 도메인의 학습 데이

터만을 사용하는 것으로 도메인 적응 기술의 베이스라인으로 간주

- target only(TGT-only): 타겟 도메인의 학습 데이터만을 사용
- All and weighted model: 소스, 타겟 도메인의 학습 데이터를 모두 사용하지만 데이터의 비율이 다름을 감안하여 가중치를 적용
- PRED: SRC-only 방법으로 구축된 기술을 이용하여 타겟 도메인의 학습데이터를 분석하고 그 결과를 타겟 도메인의 모델을 구축할 때 자질로 사용
- Linearly interpolation: SRC-only, TGT-only 방법으로 각각 모델을 구축하고 이를 선형보간법(linearly interpolation)을 적용하여 하나의 모델로 통합하는 방법으로 다음의 수식을 이용

  LININT Model
  = λ * Source Model + (1-λ) * Target Model
- Feature Augmentation(FA): 공통적으로 사용 가능한 자질, 소스 도메인에 특화된 자질, 타겟 도메인에 특화된 자질과 같이 3가지로 분류하여 각각의 자질을 이용하여 모델을 독립적으로 구축.
- prior 모델: SRC-only 모델의 가중치 벡터(weight vector)를 타겟 도메인 기술을 구축시 이용. 이를 분류 문제로 변환한 개념은 그림 2[6]과 같다. 타겟 가중치 벡터를 찾기 위해 소스 가중치 벡터를 시작점으로 참조하여 타겟 도메인 학습데이터로 학습.

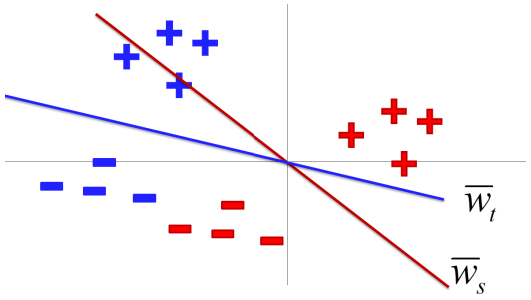

그림 2. prior 모델 적용 예

본 논문에서는 뉴스 도메인에서 구축된 한국어 의미역 인식 시스템을 위키피디아 문서에서 적용할 때 나타나는 성능 하락을 최소화하기 위해서 소규모로 구축된 위키피디아 학습 데이터를 도메인 적응 기술인 prior 모델을 적용하여 백과사전 도메인 한국어 의미역 인식 시스템을 제안한다. 또한 도메인 적응 기술로 기존에 제시된 다른 방법들과 비교 실험을 통해서 제안된 방법의 우수성을 입증하고자 하며, 알고리즘적인 접근 방법 외에도 자질 값을 단순화하는 방법도 실험을 통해 살펴보았다.

3. 도메인 적응 기술 기반 한국어 의미역 인식

일반적으로 의미역을 인식하기 위해서는 대상 문장에서 서술어를 인식하는 단계와 서술어의 의미를 결정하는 단계가 선행된다. 의미가 결정된 서술어 정보를 바탕으로 각각의 서술어와 논항의 관계성을 결정하는 단계와 관계가 있다고 결정된 논항의 의미역을 인식하는 논항 분류 단계를 통해서 최종적으로 의미역이 결정된다.

본 논문에서는 결정된 서술어의 의미를 이용하여 논항을 인식하고 분류하는 하는 문제에 대해서만 도메인 적응 기술을 적용한다.

3.1 소스 도메인 한국어 의미역 인식

도메인 적응 기술은 소스 도메인에서 구축된 시스템을 소규모의 타겟 도메인 학습 데이터를 이용하여 타겟 도메인에 효과적으로 작동하는 시스템을 구축하는 것으로 이 절에서는 소스 도메인에서 구축된 한국어 의미역 인식에 대해서 간략하게 소개한다.

본 논문에서는 소스 도메인 기술로 [4] 연구의 결과를 사용하였는데, 이는 순차적 레이블링 기법을 이용한 한국어 의미역 인식 시스템이다. 이 연구에서는 그림 3과 같은 알고리즘을 사용하였으며, 본 논문에서는 이를 prior 모델을 적용하기 위해서 변경한다.

Input: S, λ, T, k

Initialize: $\mathbf{w}_1 = \mathbf{0}$

For $t = 1,2,\ldots,T$ **do**

  Choose $A_t \subseteq S$, where $|A_t| = k$

  Set $A_t^+ = \{(\mathbf{x}, \mathbf{pr}, \mathbf{y}) \in A_t : l(\mathbf{w}_t; (\mathbf{x}, \mathbf{pr}, \mathbf{y})) > 0\}$

  $\forall (\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) \in A_t^+$:

    $\mathbf{y}_{ij}^* = \text{argmax}\{L(\mathbf{y}_{ij}, \mathbf{y}) - \mathbf{w}_t^T \Psi(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y})\}$

  $\eta_t = 1/\lambda t$

  $\mathbf{w}_{t+1} = (1-\eta_t\lambda)\mathbf{w}_t + \frac{\eta_t}{k} \sum_{(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}) \in A_t^+} \delta \Psi_{ij}(\mathbf{x}_i, \mathbf{pr}_{ij}, \mathbf{y}_{ij}^*)$

Output: $\mathbf{w}_{T+1}$

그림 3. 순차적 레이블링 의미역 인식 알고리즘[4]

3.2 다중 도메인 한국어 의미역 인식

소스 도메인 이외에 다른 도메인에서도 효과적으로 작동하는 한국어 의미역 인식 기술을 개발하기 위해서 본 논문에서는 2가지 방법을 시도하였다. 첫 번째는 타겟 도메인 적용 시에 학습데이터가 충분하지 못 하여 생기는 자료 희귀(data sparseness) 문제를 해결하기 위해 중요 자질 값의 분류 단위를 단순화하는 방법이고, 두 번째 방법은 도메인 적응 기술을 적용하여 한국어 의미역 인식 알고리즘을 수정하는 방법이다.

3.2.1 자질값 분류 단위 단순화

학습 데이터에서 중요 자질로 사용하는 세종 형태소 태그셋을 기존에 세분류로 사용하던 것을 소분류로 변환하여 학습하고, 세종 구문태그 역시 기능 태그만을 사용한다. 세종 형태소 태그셋의 소분류, 세분류 관계는 표 1과 같다. 괄호 안의 숫자는 학습 데이터에서 해당 태그가 출현한 빈도이다. 다양한 격조사(주격, 보격, 목적격 등)은 소분류에서 격조사(JK) 하나로 통일되지만, 의미

역 인식에서 격조사를 세분하여 얻을 수 있는 정보가 더 중요하다고 판단하여 소분류로 변환하지 않았다.

구문 자질 값을 단순화하기 위해서는 argument pruning 기법을 적용하여 세종 구문태그 중에서 기능 태그만을 학습 대상으로 하고, 나머지 태그는 관계없음으로 처리하였다.

표 1. 세종 형태소 태그의 세분류-소분류 매핑

세종 세분류	세종 소분류
NNG(일반명사, 219,707) NNP(고유명사, 32,739) NNB(의존명사, 29,920)	NN(명사)
VCP(긍정지정사, 6,879) VCN(부정지정사, 444)	VC(지정사)
MAG(일반부사, 9,736) MAJ(접속부사, 870)	MA(부사)
EP(선어말어미, 8,923) EF(종결어미, 15,970) EC(연결어미, 48,823) ETN(명사형전성어미, 3,942) ETM(관형형전성어미, 49,571)	E(어미)
XSN(명사파생접미사, 21,741) XSV(동사파생접미사, 24,216) XSA(형용사파생접미사, 2,273)	XS(접미사)

3.2.2 prior 모델 기반 접근 방법

본 논문에서는 자질 값을 단순화하는 방법과 함께 알고리즘을 이용하여 타겟 도메인에서 한국어 의미역 기술을 개발하도록 시도하였다. 여러 가지 도메인 적응 알고리즘 중에서 Chelba and Acero[7]이 제안한 prior 모델을 변형하여 structural SVM에 적용한 연구[8]를 참조하여 한국어 의미역 기술에 적용하였다.

prior 모델은 기계학습의 목적인 최적화된 가중치 벡터 결정을 위해 상대적으로 소규모인 타겟 도메인 학습 데이터만을 이용하기보다, 소스 도메인에서 학습된 가중치 벡터를 참조하여 좀더 효과적으로 최적의 타겟 도메인 가중치 벡터를 탐색하는 개념이다.

소스 도메인 한국어 의미역 인식 기술에 prior 모델을 적용하기 위해서 목적 함수를 아래와 같이 수정한다.

$$f(\mathbf{w};A_t)=\frac{\lambda}{2}\left\|\mathbf{w}-\mathbf{w}_{src}\right\|^2+\frac{1}{k}\sum_{(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y}_{ij})\in A_t} l(\mathbf{w};(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y}_{ij}))$$

$$where \quad l(\mathbf{w};(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y}_{ij}))=\max\{0,\max_{\mathbf{y}}\{L(\mathbf{y}_{ij},\mathbf{y})-\mathbf{w}^T\delta\Psi_{ij}(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y})\}\}$$

$$and \quad \delta\Psi_{ij}(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y})=\Psi(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y}_{ij})-\Psi(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y}).$$

위 식에서 $\mathbf{w}_{src}$는 소스 도메인에서 학습된 가중치 벡터이고, $\mathbf{x}_i$는 학습 데이터의 i번째 문장 벡터, $\mathbf{pr}_{ij}$는 i번째 문장의 j번째 서술어(predicate), $\mathbf{y}_{ij}$는 i번째 문장의 j번째 서술어에 대한 의미역 인식 결과 벡터, A_t는 학습 데이터에서 임의로 선택된 부분집합이다.

타겟 도메인의 최적 가중치 벡터 $\mathbf{w}$를 결정하기 위해서 소스 가중치 벡터 $\mathbf{w}_{src}$를 선행 정보로 이용하여 타겟 도메인의 목적 함수를 위와 같이 정의한다. 목적 함수가 최소값이 되는 최적 가중치 벡터를 구하기 위해서 subgradient 함수는 다음과 같이 정의하였다.

$$\nabla f(\mathbf{w};A_t)=\lambda(\mathbf{w}-\mathbf{w}_{src})-\frac{1}{|A_t|}\sum_{(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y}_{ij})\in A_t^+}\delta\Psi_{ij}(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y}_{ij}^*)$$

$$where \quad \mathbf{y}_{ij}^*=\arg\max_{\mathbf{y}}\{L(\mathbf{y}_{ij},\mathbf{y})-\mathbf{w}^T\delta\Psi_{ij}(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y})\}$$

$$and \quad A_t^+=\{(\mathbf{x},\mathbf{pr},\mathbf{y})\in A_t : l(\mathbf{w};(\mathbf{x},\mathbf{pr},\mathbf{y}))>0\}$$

이 두가지 수식을 이용하여 그림 3의 알고리즘에 prior 모델을 적용하면 그림 4와 같은 알고리즘이 된다.

Inputs: $D_{tgt}, \lambda, T, k, \mathbf{w}_{src}$

1: $\mathbf{w}_1=\mathbf{0}$ // Initialization.

2: For $t=1,2,\ldots,T$ do

3: Choose $A_t\subseteq D_{tgt}$, where $|A_t|=k$

4: Set $A_t^+=\{(\mathbf{x},\mathbf{pr},\mathbf{y})\in A_t : l(\mathbf{w}_t;(\mathbf{x},\mathbf{pr},\mathbf{y}))>0\}$

5: $\forall(\mathbf{x}_{ii},\mathbf{pr}_{ii},\mathbf{y}_{ii})\in A_t^+$:
$\mathbf{y}_{ii}^*=\arg\max_{\mathbf{y}}\{L(\mathbf{y}_{ii},\mathbf{y})-\mathbf{w}_t^T\Psi(\mathbf{x}_i,\mathbf{pr}_{ii},\mathbf{y})\}$

6: $\eta_t=1/\lambda t$

7: $\mathbf{w}_{t+1}=\mathbf{w}_t-\eta_t\lambda(\mathbf{w}_t-\mathbf{w}_{src})+\frac{\eta_t}{k}\sum_{(\mathbf{x}_{ij},\mathbf{pr}_{ij},\mathbf{y}_{ij})\in A_t^+}\delta\Psi_{ij}(\mathbf{x}_i,\mathbf{pr}_{ij},\mathbf{y}_{ij}^*)$

8: Return $\mathbf{w}_{T+1}$ as output

그림 4. prior 모델을 적용한 의미역 인식 알고리즘

위 알고리즘에서 D_{tgt}는 타겟 도메인의 학습데이터 집합, λ는 정규화 상수, T는 반복횟수, k는 sub-gradient를 계산하기 위해 한 번에 사용하는 학습데이터 수, $\mathbf{w}_{src}$는 소스 도메인 최적 가중치 벡터이다.

학습을 위한 자질은 소스 도메인에 적용된 자질을 그대로 사용하였다[4].

4. 실험

본 논문에서는 Korean Propbank[9]에서 군대 도메인인 Virginia 말뭉치를 제외한 뉴스 도메인(Newswire) 말뭉치를 소스 도메인으로 사용하였다. 타겟 도메인은 백과사전(위키피디아) 도메인으로 질의응답 시스템 평가를 위해 구축된 ETRI 표준 평가셋(200셋)에 Korean Propbank 의미역을 사용하여 구축된 말뭉치[10]를 이용하였다. 소스 도메인 말뭉치는 4,882문장으로 구성된다. 타겟 도메인 말뭉치는 총 907문장으로 구성되며, 평가셋 1번부터 160번까지 650문장을 학습데이터로 사용하

고, 161번부터 200번까지 257문장은 테스트 데이터로 사용한다. 타겟 도메인 말뭉치는 소스 도메인 말뭉치에 비해 약 1/7 수준이다.

실험 데이터는 세종 형태소 태그 세분류와 모든 구문 태그를 그대로 사용한 데이터(기존)과 3.2.1에서 언급한 자질 값을 단순화한 데이터(단순)으로 구분하였다. 제안하는 방법과 비교하기 위해서 데이터 측면에서 SRC-only, TGT-only, ALL(소스와 타겟 학습 데이터 모두 사용)로 구분하여 소스 도메인 한국어 의미역 기술을 이용하여 실험하고, 알고리즘 측면에서는 PRED와 FA를 이용한 도메인 적응 시스템을 구축하여 제안하는 prior 모델과 비교하였다.

기계학습을 위해 추출하는 자질을 위한 형태소 분석 및 구문 분석[11]은 ETRI 언어 분석기를 사용하여 자동으로 분석된 결과를 이용하여, 사용된 형태소 분석과 구문 분석 결과에 오류가 포함되어 있다.

실험결과는 표2와 같다. 소스 도메인에서 학습한 한국어 의미역 시스템을 소스 도메인에서 테스트한 성능은 74.77(F1)으로 이를 그대로 타겟 도메인에서 테스트하면 성능이 59.0(F1)으로 약 15% 정도 하락하였다.

도메인 적응 알고리즘을 적용하지 않은 실험 결과에서는 모든 데이터를 사용하며, 자질로 사용된 형태소/구문 분석 태그 단위를 단순화할 때 성능이 향상됨을 실험 결과에서 볼 수 있다. 또한, 도메인 적응 알고리즘을 적용한 실험 결과는 제안한 방법이 64.3(F1)으로 PRED, FA 방법에 비해 나은 결과를 보임을 알 수 있다.

그러나, 자질 값을 단순화한 학습데이터에서는 알고리즘을 적용한 경우 성능이 개선되지 않는 결과를 보인다. 이는 자질 값을 단순화할 경우 데이터만을 사용하면 성능이 개선되는 반면, 도메인 적응 알고리즘을 적용하면 학습되는 효과가 성능에 부정적인 영향을 주는 것으로 유추해 볼 수 있다.

표 2. 실험 결과

실험방법	자질값	Prec.	Rec.	F1
SRC-only	기존	64.0	54.6	59.0
	단순	66.1	55.1	60.1
TGT-only	기존	66.5	54.7	60.0
	단순	65.5	53.2	58.7
All	기존	69.8	58.5	63.6
	단순	69.9	58.5	63.7
PRED	기존	68.6	55.3	61.2
	단순	67.9	55.0	60.8
FA	기존	63.2	60.7	61.9
	단순	63.7	59.3	61.4
제안 방법	기존	69.0	60.2	64.3
	단순	68.4	59.0	63.3

5. 결론

본 논문은 뉴스 도메인 학습 데이터를 사용하여 구축된 한국어 의미역 인식 시스템이 새로운 도메인(백과사전)에서 성능이 하락되는 현상을 최소화하기 위해서 자질 값을 상대적으로 큰 분류로 구분하여 단순화하여, 자료 희귀성 문제를 피하는 방법과 도메인 적응 알고리즘 중 prior 모델을 적용하는 방법을 제안했다.

실험 결과로 자질 값을 단순화하는 방법은 데이터만을 사용한 경우 성능 향상이 있지만, 도메인 적응 알고리즘을 이용하는 경우는 성능 하락 현상이 발생하였다. 도메인 적응 알고리즘을 적용하기 어려운 경우, 자질 값을 좀더 큰 분류로 묶어 단순화하면 소규모 데이터에서는 성능이 개선된다는 사실을 알 수 있다. 알고리즘을 적용한 실험에서는 본 논문에서 제안한 방법이 다른 비교 실험 방법에 비해서 1~3% 정도 성능이 향상됨을 알 수 있다.

향후 연구로는 의미 자질 등 다양한 자질을 의미역 인식에 활용하고, 다른 도메인에도 본 논문에서 제안하는 방법을 적용할 필요가 있다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10044577, (1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]

참고문헌

[1] X. Carreras and L.Marquez, " Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling," Proc. CoNLL-2005, pp.152-154, 2005

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

[3] S. Lim et al., "Domain-Adaptation Technique for Semantic Role Labeling with Structural Learning," ETRI Journal, vol. 36, no. 3, June, pp. 429-438, 2014.

[4] 임수종, 김현기, "순차적 레이블링을 이용한 한국어 의미격 인식," 한국 정보과학회 학술발표 논문집, vol.2014 No.6., pp.595-597, 2014.

[5] J.Blitzzer and Hal Daume, "Domain Adaptation," , ICML tutorial, 2010.

[6] H. Daume and D. Marcu, "Domain Adaptation for Statistical Classifiers," , *J. Artif. Intell. Res.*, vol.26, no.1, pp.101-126, 2006.

[7] C. Chelba and A. Acero, "Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot," , *Comput. Speech Language*, vol.20, no.4, pp.382-399, 2006.

[8] C.Lee and M.Jang, "A Prior Model of

Structural SVMs for Domain Adaptation,", ETRI Journal, vol.33, no.5, pp.712-719, 2011.
[9] Martha Palmer et al., "Korean Propbank." http://catalog.ldc.upenn.edu/LDC2006T03
[10] 배장성, 오준호, 황현선, 이창기, "한국어 의미역 결정을 위한 Korean PropBank 확장 및 도메인 적응 기술," 제26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.
[11] 임준호, 윤여찬, 배용진, 김현기, 이규철, "지배소 후위 제약을 적용한 트렌지션 시스템 기반 한국어 의존 파싱 모델," 제26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.