비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정

(Unsupervised Semantic Role Labeling for Korean Adverbial Case)

김 병 수 [†] 이 용 훈 [††] 이 종 혁 [†††]

(Byoung-Soo Kim) (Yong-Hun Lee) (Jong-Hyeok Lee)

요 약 말뭉치를 이용하여 통계적으로 의미역 결정(semantic role labeling)을 하기 위해서는, 의미역을 태깅하는 작업이 필수적이다. 그러나 한국어의 경우 의미역이 태깅된 대량의 말뭉치를 구하기 힘들며, 이를 직접 구축하기 위해서는 많은 시간과 노력이 필요한 문제점이 있다. 본 논문에서는 비지도 학습의 하나인 self-training 알고리즘을 적용하여, 의미역이 태깅되지 않은 말뭉치로부터 의미역을 결정하는 방법을 제안한다. 이를 위해, 세종 용언 전자사전의 격틀정보를 이용하여 자동으로 학습 말뭉치를 구축하였으며, 확률 모델을 적용하여 점진적으로 학습하였다. 그 결과, 4개의 부사격 조사에 대해 평균적으로 83.00%의 정확률을 보였다.

키워드 : 의미 분석, 의미역 결정, 부사격, 비지도 학습

Abstract Training a statistical model for semantic role labeling requires a large amount of manually tagged corpus. However, such corpus does not exist for Korean and constructing one from scratch is a very long and tedious job. This paper suggests a modified algorithm of self-training, an unsupervised algorithm, which trains a semantic role labeling model from any raw corpora. For initial training, a small tagged corpus is automatically constructed from case frames in Sejong Electronic Dictionary. Using the corpus, a probabilistic model is trained incrementally, which achieves 83.00% of accuracy in 4 selected adverbial cases.

Key words : semantic analysis, semantic role labeling, adverbial case, unsupervised learning

1. 서 론

의미 분석은 일반적으로 형태소 분석과 구문 분석의 과정을 거쳐 이루어지는 자연언어처리의 상위 단계로 크게 단어의 의미 중의성을 해소하는 단계(word sense disambiguation)와 문장 내의 서술어와 논항들 사이의 의미 관계를 결정하는 단계(semantic role labeling)로 나누어진다. 이러한 의미 분석 단계를 처리하는데 가지는 난점은 단어의 다의성과 '주어', '목적어', '간접목적어'와 같은 문법 관계(grammatical relation)를 '행위주', '대상' 등과 같은 의미 관계(semantic relation)로 사상(mapping)하는데 발생하는 애매성이라고 할 수 있다. 본 논문에서는 의미 분석 단계 중 후자에 해당하는 의미역 결정 문제에 대해 다루고자 한다.

의미역 결정이란, 문장 내의 서술어-논항(predicate-argument) 관계에 적합한 의미 관계를 결정하는 과정이다. 즉, 그림 1과 같이 문장의 표층격(surface case)에 해당하는 문법 관계를 심층격(deep case)에 해당하는 의미 관계로 사상하는 과정으로 볼 수 있다. 이러한 의미 관계는 전통적으로 격 관계(case role), 의미역(thematic role, θ role)으로 불리며 오랜 기간에 걸쳐 언어학자들 사이에서 연구되어 왔던 어려운 주제 중 하나이다.

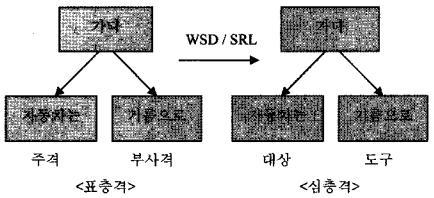

그림 1 의미역 결정의 예

· 이 논문은 첨단정보기술연구센터를 통한 과학재단 및 2006년도 두뇌한국21사업에 의하여 지원되었음

† 비 회 원 : 포항공과대학교 정보처리학과
akirus82@postech.ac.kr
†† 정 회 원 : 포항공과대학교 컴퓨터공학과
yhlee95@postech.ac.kr
††† 종신회원 : 포항공과대학교 컴퓨터공학과
jhlee@postech.ac.kr

논문접수 : 2006년 8월 13일
심사완료 : 2006년 8월 30일

의미역 결정은 표층격과 심층격 사이의 규칙적인 연결성의 문제를 포함한 여러 가지 흥미로운 순수언어학적 연구 주제를 포함하고 있을 뿐만 아니라 자연언어처리 측면에서도 유용한 정보로 활용되고 있다. 예를 들어, 기계번역(MT)에서는 동일한 의미 관계를 가지지만 서로 다른 문법 관계를 지닌 두 언어의 구문을 연결하는 중간 단계(interlingual representation)로 사용된다. 이외에도 대량의 의미 관계 정보를 필요로 하는 정보추출(IE), 질의응답(QA)과 같은 다양한 자연언어처리 응용에서 성능을 향상시키는데 중요한 역할을 한다. 이에 따라 최근 들어 자동으로 의미역을 결정하는 방법론에 대한 연구들이 활발히 진행되고 있다.

본 논문의 구성은 다음과 같다. 2장에서는 기존에 제안된 의미역 결정 방법론, 관련 연구, 문제점, 그리고 연구 동기에 대해 언급하고, 3장에서는 본 논문에서 제안하는 의미역 결정 시스템의 구조를 소개한다. 4~6장에서는 시스템의 각 부분에 대해 살펴보고, 7장에서는 수정된 self-training 알고리즘을 설명한다. 8장에서는 실험 및 결과를 분석하고, 마지막으로 9장에서는 결론에 대해 기술한다.

2. 기존연구

기존의 의미역 결정 연구는 크게 격틀사전에 기반한 방법(case frame based method)[1,2]과 말뭉치에 기반한 방법(corpus based method)[3-11], 그리고 이들 방법을 통합한 하이브리드 방법(hybrid method)[12-16]으로 나눌 수 있다.

격틀사전에 기반한 방법은 서술어와 논항들의 쓰임을 기술한 격틀사전을 이용하는 방법으로, 서술어와 논항에 대한 문법 관계를 기술한 문틀(frame)과 논항들의 정보를 기술한 선택제약(selectional restriction) 등을 이용하여 서술어-논항 관계에 대해 적합한 격틀을 선택하여 의미역을 결정하는 방법이다. 격틀사전에 기반한 방법은 입력 문장과 격틀 사이의 간단한 유사도 계산 과정을 통해 의미역이 결정되기 때문에 처리속도가 빠르고 높은 정확률을 가지는 장점이 있지만, 격틀사전과 같은 고비용의 언어자원이 필요하고 격틀사전에 기술되지 않은 임의격[1]에 대해서는 처리하지 못하여 적용률이 낮은 단점이 있다.

말뭉치에 기반한 방법은 의미역이 태깅된 대량의 말뭉치를 이용하여 통계적 혹은 기계적 학습 방법으로 의미역 결정을 하는 방법이다. 즉, 말뭉치로부터 의미역 결정에 도움이 되는 자질들(features)을 추출하고 이들을 다양한 학습 알고리즘에 따라 모델을 학습하여 의미역 결정을 하는 방법이라고 할 수 있다. 지금까지 지지벡터기계(support vector machine), 결정트리(decision tree), 최대 엔트로피(maximum entropy) 모델 등 다양한 학습 알고리즘이 의미역 결정에 적용되었다. 이 방법은 적용률이 높고 격틀사전에 기반한 방법에 비해 견고하다는 장점이 있다. 그러나 의미역을 태깅하여 학습에 필요한 말뭉치를 구축하는 작업은 많은 시간과 노력이 필요하고, 의미역 결정 시스템의 성능이 구축된 학습 말뭉치의 양과 질에 크게 의존할 수밖에 없는 단점이 있다.

하이브리드 방법은 둘 이상의 방법을 통합하여 각 방법의 단점을 서로 보완하는 방법이다. 예를 들어, [13]은 정확률이 높은 격틀사전에 기반한 방법과 적용률이 높은 말뭉치에 기반한 방법을 통합하여 보다 정교한 의미역 결정 모델을 제시하였다. 이 외에도 'CoNLL-2005 Shared Task Semantic Role Labeling'에서 평가한 의미역 결정 시스템 중 상위의 시스템들이 다양한 기계학습 모델을 통합하여 좀 더 정확한 의미역 결정을 하였다[16]. 그러나 하이브리드 방법은 기본적으로 말뭉치에 기반한 방법의 비중이 크기 때문에 의미역을 태깅하는 작업이 불가피한 단점이 있다.

최근 들어 말뭉치를 이용한 통계적인 학습 방법이 유행하면서 다수의 의미역 결정 연구들이 진행되었고, 성능을 통해서 통계적 학습 방법의 효과를 입증하였다. 하지만 한국어의 경우 영어권의 FrameNet이나 PropBank에서 제공하는 의미역이 태깅된 말뭉치가 없어 지도 학습(supervised learning)을 적용하기 어렵다. 일반적으로 의미역이 태깅된 말뭉치를 구축하는 작업은 의미역의 수가 많고[2], 각 의미역 사이에 미세한 차이가 존재하여 의미역을 결정하기 어려운 문제점이 있다. 따라서, 본 논문에서는 의미역이 태깅되지 않은 말뭉치로부터 의미역을 결정하는 비지도 학습(unsupervised learning)을 기반으로 한 의미역 결정 시스템을 제안한다. 이를 위해, 세종 용언 전자사전으로부터 의미역을 결정하는데 필요한 격틀정보를 추출하고, 이를 격틀사전에 기반한 방법을 통해서 학습에 필요한 말뭉치를 자동으로 구축한 후 학습 말뭉치로부터 확률정보를 추출하여 확률 모델을 수정된 self-training 알고리즘에 따라 점진적(incrementally)으로 학습하며 의미역을 결정하였다. 본 논문에서는 의미역 결정을 하는데 애매성이 큰 부사격 조사 '에', '로', '에서', '에게'를 대상으로 실험하였다.

1) 언어학자들 사이에 필수격(obligatory case)과 임의격(optional case)을 구별하는 방법에 대해서는 여러 가지 이견이 있지만, 본 논문에서는 세종 용언 전자사전에 기술된 경우 필수격으로, 그렇지 않은 경우 임의격으로 간주하겠다.

2) 세종 용언 전자사전의 의미역 기술 지침에 따르면 행위주, 경험주, 심리경험주, 동반주, 대상, 장소, 도착점, 방향, 결과상태, 출발점, 도구, 영향주, 기준치, 내용, 목적 등 총 15개의 의미역에 대해서 정의하고 있다.

3. 의미역 결정 시스템 구조

비지도 학습을 기반으로 한 의미역 결정 시스템은 그림 2와 같이 크게 세 부분으로 나누어진다. 첫째, 서술어-논항 관계 추출기는 격틀 모델의 전처리 단계로 입력 문장에서 서술어-논항 관계를 찾아 추출한다. 둘째, 격틀 모델은 격틀사전을 이용하여 서술어-논항 관계에 적합한 격틀을 할당하여 확률 모델을 학습하는데 필요한 말뭉치를 자동으로 구축한다. 셋째, 확률 모델은 서술어, 논항, 격조사 등의 자질들을 선택하여 여러 단계로 구성된 확률 모델로 의미역을 결정한다. 이때 확률 모델을 학습하기 위해서 기존의 self-training 알고리즘을 수정하여 적용하였다.

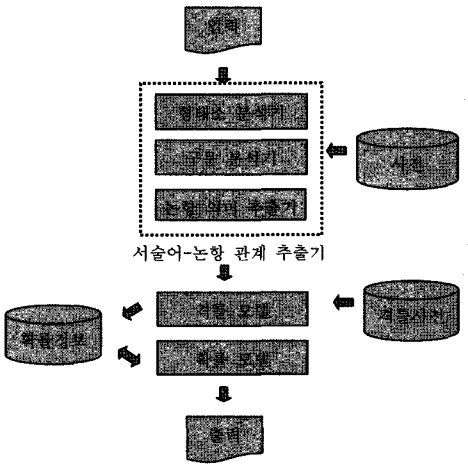

그림 2 의미역 결정 시스템 구조

4. 서술어-논항 관계 추출기

서술어-논항 관계 추출기는 형태소 분석기와 구문 분석기를 통해 서술어와 그 서술어가 취하는 논항들을 의존트리(dependency tree) 형태로 추출한다. 예를 들어, 그림 3과 같은 구문 분석의 결과에서 서술어의 원형을 복원한 뒤 '하다'에 대해 주어, 목적어, 부사어로 쓰인 논항들을 추출하고, 논항 의미 추출기를 통해서 논항들이 가질 수 있는 의미들을 추출한다. 본래는 의미역 결정 단계 전에 입력 문장의 논항들은 단어의 의미 중의성 해소 과정을 거쳐 하나의 의미가 정해져야 하나 이를 위해서는 별도의 시스템이 필요하고, 단어의 의미 중의성 해소는 자연언어처리의 한 연구 분야로 쉽게 해결할 수 있는 문제가 아니기 때문에 본 연구에서는 이는 고려하지 않았다. 대신 논항의 의미들과 격틀사전의 선택제약 사이의 유사도 계산을 통해서 최대 유사도를 가지는 의미를 논항의 의미로 선택하는 방법을 통해서 단어의 의미 중의성 해소 단계를 대신하였다. 서술어-논항 관계 추출기에 사용된 형태소 분석기 및 구문 분석기는 포항공과대학교 지식 및 언어공학 연구실에서 개발한 KoMA와 KoPA를 사용하였다.

```
| 했다. (평서형종결)
=> 하(YBDO)+ 였(fmbtp)+ 다(fmofd)+ .(g)
    | 야영을 (목적어) ----> 했다.
    => 야영(CMCPA)+ 을(fjco)
    | 습지에서 (부사어) ----> 했다.
    => 습지(CMCN)+ 에서(fjcao)
    | 어제 (부사어) ----> 했다.
    => 어제(SBO)
```

그림 3 형태소 분석 및 구문 분석 결과의 예

5. 격틀사전을 이용한 의미역 결정

5.1 용언사전으로부터 격틀사전 구축

세종 용언 전자사전은 표제어에 대해 다양한 통사적, 의미적 정보가 XML형태로 수록되어 있다. 따라서 이들 중 의미역 결정에 필요한 정보들을 선별하여 전산적 처리가 용이하도록 격틀사전을 구축해야 할 필요가 있다. 예를 들어, 그림 4와 같이 표제어 '가다'의 경우, 'X=N0-이 Y=N1-로 V'와 같은 문틀(frame), '교통기관(버스|기차|비행기)', '장소'와 같은 선택제약(selectional restriction), 'THM', 'GOL'과 같은 의미역(semantic role)들을 추출하여 격틀사전을 구축하였다.

격틀사전의 선택제약은 세종전자사전의 명사 의미부류를 기준으로 기술되어있다. 세종 명사 의미부류는 최상위부류들에 대해 단계적으로 의미영역을 분할하여 구축된 의미부류들의 위계적 체계로 총 582개의 의미부류로 분류된다[17]. 최상위 부류로는 그림 5와 같이 <구체물>, <집단>, <장소>, <추상적대상>, <사태> 등 5개의 부류가 설정되었는데 이중 처음 4개는 논항명사의

```
<orth>가다</orth>
  <sense n="01">
    <frame>X=N0-이 Y=N1-로 V</frame>
      <sel_rst arg="X" tht="THM">교통기관(버스|기차|비행기)</sel_rst>
      <sel_rst arg="Y" tht="GOL">장소</sel_rst>
      <eg>철수는 무작정 부산으로 가는 버스를 탔다.</eg>
  </sense>
  <sense n="04">
    <frame>X=N0-이 Y=N1-로 V</frame>
      <sel_rst arg="X" tht="THM">교통기관(자동차|전차)</sel_rst>
      <sel_rst arg="Y" tht="INS">연료(가스)|에너지원(전기)</sel_rst>
      <eg>이 차는 기름이 아니라 전기로 간다.</eg>
  </sense>
```

그림 4 표제어 '가다'의 격틀정보

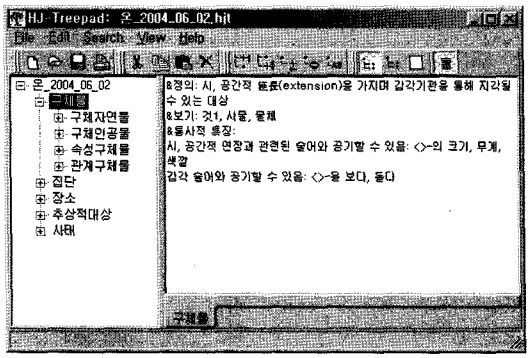

그림 5 세종 명사 의미부류

의미부류이고, 나머지 <사태> 부류는 술어명사의 의미부류이다. <사태>는 술어명사에 해당하는 의미부류로 다음에 설명할 술어명사 및 기능동사 구문 처리에 있어서 유용하게 사용된다.

5.2 술어명사 및 기능동사 구문 처리

술어명사(predicate noun)란 사건이나 행위, 개체들의 관계 등을 나타내는 의미적 실체인 술어가 명사의 형태로 실현된 것을 의미하며, 기능동사[3](support verb)란 술어명사를 중심으로 문장 구성이 가능하도록 술어의 위치를 채우고 술어명사의 현동화(actualization)를 뒷받침해주는 동사를 말한다. 이런 술어명사와 술어명사의 고유 논항, 기능동사로 구성하는 기본구문을 기능동사 구문이라고 한다[18].

세종 용언 전자사전에 기술된 대표적인 기능동사인 '하다', '되다', '시키다'는 각각 79개, 50개, 18개의 격틀을 가지고 있다. 다양한 술어명사와 사용될 수 있는 기능동사의 특징은 격틀을 선택하는 유사도 계산 수식이 정교하지 않다면 입력 문장의 유사도를 계산하여 적합한 격틀을 선택하는데 변별력이 없기 때문에 잘못된 격틀이 선택될 수 있다. 따라서 그림 6과 같이 'Npr[4]-을 V' 형태로 나타나는 기능동사 구문에 대해서 'Npr+V'의 형태와 같이 변형해서 처리해주면 더 정확한 격틀을 선택할 수 있고 처리속도도 빨라진다. 이는 한국어의 기능동사 구문에서만 나타나는 구문적인 특징으로 [12,18]에서 필요성을 언급하기도 하였다.

(예 1 기능동사 구문 처리)

> 우리 부대는 어제 습지에서 **야영을 했다.**
> => 우리 부대는 어제 습지에서 **야영했다.**

실제 '야영하다' 의 경우 1개의 문틀을 가지고 있기 때문에 '야영을 하다'로 격틀을 선택하는 것보다 유사도를 계산하는 횟수가 79번에서 1번으로 줄어들었고 올바른 의미역으로 결정되었다. 또한 6장에서 설명할 확률 모델의 경우 서술어를 중심으로 확률 수식이 구성되어 있기 때문에 술어명사와 기능동사의 특징을 반영하기 어려운 문제점이 있다. 따라서 확률 모델을 적용하는데 있어서도 '하다'가 아닌 '야영하다'와 같이 기능동사 구문의 특징을 반영한 개별적인 확률을 가지기 때문에 올바른 의미역을 결정하는데 도움이 된다.

5.3 격틀 선택

격틀사전을 이용하여 입력 문장에 대해서 적합한 격틀을 선택하는 과정은 서술어-논항 관계 추출기에서 추출한 의존트리와 격틀에 기술된 선택제약 사이의 유사도를 계산하여 전체 유사도가 가장 높은 격틀을 선택하는 과정으로 생각할 수 있다. 이를 위해서 의존트리에서 '주어', '목적어' 등에 해당하는 구문 관계와 일치하는 격요소(case component)를 격틀사전에서 찾아야 한다. 그 다음 각 논항들에 대해서 유사도를 계산하고, 이들의 합을 통해서 각 격틀마다 전체 유사도를 계산한다.

$$sim(c_i,c_j)=\frac{2\times \min_{pths(mscs(c_i,c_j),rt)} len_e p}{\min_{pths(c_i,c_j)} len_e p+2\times \min_{pths(mscs(c_i,c_j),rt)} len_e p}$$

그림 6 두 의미 사이의 유사도 계산

본 논문에서는 그림 6의 수식을 이용하여 논항과 선택제약 사이의 유사도를 계산한다[19]. 위의 수식은 세종 명사 의미부류와 같이 트리 형태의 계층구조에서 두 자식과 공통 조상 사이의 관계를 유사도로 나타낸 것이다. 이때 입력 문장의 논항과 그림 4에서 알 수 있듯이 선택제약 모두 하나 이상의 의미를 가질 수 있으므로 이들 사이의 유사도 중에서 최대가 되는 의미를 논항의 의미로 선택하는 방법을 사용하여 단어의 의미 중의성 해소 과정을 대신하였다.

$$total=(\sum_{N} weight\times arg\,max_{C_i,C_j} sim(c_i,c_j))\times\sqrt{\frac{L}{N}}$$

그림 7 문장 전체의 유사도 계산

위의 수식은 그림 6의 수식을 이용하여 각 논항들의 유사도를 계산한 후 가중치(weight)에 따라 더하여 입력 문장과 격틀 사이의 전체 유사도를 계산한다. 격틀사전의 한 표제어 안에서는 격틀 사이에 주어와 목적어가 동일한 경우가 많고 실제 격틀을 선택하는데 부사격 논항의 비중이 큰 경우가 많다. 따라서 각 격에 해당하는 가중치의 최적화된 값을 찾아 할당하면 보다 정확한 의

3) 기능동사는 'Light verb'라고도 한다.
4) Npr은 술어명사를 나타낸다.

미역 결정을 할 수 있다. 본 논문에서는 실험적으로 각 격에 해당하는 가중치를 구하여 주격, 목적격보다 부사격에 대해서 약 두 배의 가중치를 주었다. 또 주격이나 목적격과 같은 필수격 논항이 생략되어 입력 문장에 필수격 논항의 수(L)가 격틀에 기술된 필수격 논항의 수(N)보다 작다면, 유사도를 낮게 해야 하므로 'L/N'을 곱하여 전체 유사도를 보정(normalize)하였다. 그러나 입력 문장의 필수격 논항의 수가 격틀에 기술된 필수격 논항의 수보다 많다면, 현재 유사도를 계산하는 격틀이 잘못된 격틀일 가능성이 높으므로 다음 격틀에 대해서 유사도 계산을 하였다.

6. 확률 모델을 이용한 의미역 결정

6.1 확률정보 구축

기존의 다른 자연언어처리 분야에 비지도 학습을 적용한 경우는 사람이 직접 학습 말뭉치를 구축하거나 몇 가지 규칙을 설정하여 구축하였다. 그러나 의미역은 서술어와 논항, 격조사의 관계에 따라 다양한 결정되기 때문에 소량의 학습 말뭉치나 몇 가지 규칙만으로는 부사격의 다양한 의미를 나타내기에 한계가 있다. 또 본 논문과 같이 확률 모델을 사용하여 의미역을 결정하는 경우 충분한 학습 말뭉치가 있어야 실제 언어의 사용을 반영하는 확률을 얻을 수 있고, 확률이 존재하지 않아 의미역을 결정할 수 없는 문제가 발생하지 않는다.

따라서 본 논문에서는 격틀 모델의 결과로 의미역이 결정된 말뭉치에서 확률 모델의 수식에 따라 확률정보를 추출하여 학습에 필요한 확률정보를 구축하였다. 이와 같이 격틀사전을 이용하여 자동으로 학습 말뭉치를 구축하면 일반적인 규칙에 비해 서술어의 특성에 따라 기술된 다양한 문틀과 선택제약을 통해 비교적 정확하고 충분한 학습 말뭉치를 얻을 수 있고, 추후 격틀을 추가하여 적용률을 높여 추가적으로 학습 말뭉치를 확보할 수 있다는 장점이 있다.

6.2 확률 모델

본 연구에서는 [5,12]에서 제시한 linear interpolation 방법과 Backoff 방법을 결합한 확률 모델을 형태소 분석, 구문 분석 결과 및 격틀사전에서 얻을 수 있는 정보에 맞게 수정하였다.

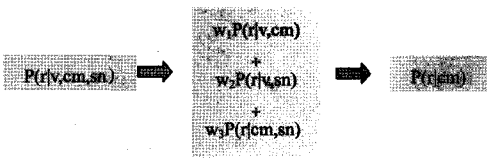

그림 8 확률 모델

일반적으로 의미역 결정은 서술어와 논항, 격조사 등의 조건이 만족될 때 특정한 의미역(r)으로 결정되는 레벨 1의 확률 모델로 생각할 수 있다. 그러나 논항은 논항의 어휘 정보를 그대로 이용하면 학습 예제 부족 문제(data sparseness problem)가 생길 수 있으므로 세종 명사 의미부류 정보를 이용하였다. 레벨 1은 세 가지 조건이 모두 만족하지 않으면 확률이 존재하지 않아 의미역을 결정하지 못하는 문제가 발생한다. 따라서 좀 더 일반적인 확률을 얻기 위해서 레벨 2와 같이 세 가지 조건 중 한가지 조건을 제거한 형태로 서술어(v)와 논항의 의미부류(sn), 격조사(cm)의 조합으로 확률 모델을 분해하였고, 각 확률에 대한 가중치의 곱의 합을 결합한 *linear interpolation* 형태로 구성하였다. 이때 각 확률 모델의 가중치는 평가 말뭉치 중 일부를 선택하여 성능을 평가한 후 모델의 성능에 비례하여 결정하였으며, 각 가중치의 합이 1이 되도록 하였다. 레벨 3은 격조사가 주어졌을 때 의미역이 결정되는 확률 모델로 8장의 실험에서 기본 모델처럼 각 부사격 조사에 대해서 최다 빈도를 가지는 의미역으로 결정하는 모델을 의미한다.

각 레벨은 그림 9에서 알 수 있듯이 의미역 결정 결과가 신뢰할 수 있는 임계값(Θ_1)에 도달하지 못하거나 확률이 없는 경우 다음 레벨로 진행하면서 Backoff 형태로 의미역 결정을 한다. 임계값(Θ_1)은 1부터 0까지 점차적으로 줄여나가면서 신뢰할 수 있는 의미역 결정 결과를 학습 말뭉치에 추가하였다. 그러나 레벨 2에서 레벨 3이 바로 적용될 경우 대부분의 의미역이 기본 모델로 결정되는 문제가 있으므로 레벨 3은 마지막 학습 단계에서 적용하여 확률이 작아 신뢰할 수 없거나 확률이 없는 대상만 의미역을 결정하였다.

7. 비지도 학습을 기반으로 한 의미역 결정

통계적 학습 방법이 널리 사용됨에 따라 대량의 학습 말뭉치의 필요성이 증가되고 있다. 그러나 이러한 말뭉치를 구축하는 작업은 많은 시간과 노력을 필요로 하는 문제점이 있다. 따라서 최근에는 대량의 학습 말뭉치를 사용하지 않고 소량의 학습 말뭉치만 가지고 학습을 하는 비지도 학습 방법의 중요성이 증가하였다. 지금까지, 정보 추출(IE), 단어의 중의성 해소(WSD), 개체명 인식(NER), 통계적 구문분석(statistical parsing) 등 다양한 자연언어처리 응용에 적용되었고[20-22], 의미역 결정에 비지도 학습 방법이 사용된 연구로는 [14,15]이 있다.

7.1 Self-training 알고리즘

본 논문은 비지도 학습 방법의 하나인 self-training 알고리즘을 수정하여 점진적으로 확률 모델을 학습하였다. 비지도 학습 방법은 특성에 따라 크게 co-training, co-EM, self-training, EM 등과 같은 방법들로 나눠

볼 수 있다[21,23,24]. Self-training, EM 방법의 경우 자질들을 한 가지 관점에서 문제를 해결하는데 비해 co-training, co-EM 방법의 경우 서로 다른 두 가지 관점에서 문제를 바라보는데 그 특징이 있다. Co-training이 적용된 대표적인 예로, 인터넷 문서를 분류하는 문제의 경우 문서 자체의 내용 즉, 문서에 나타난 단어들을 중심으로 문제를 보는 관점과 인터넷 문서의 특징인 하이퍼링크를 중심으로 문제를 보는 관점으로 자질들을 분리하여 두 개의 분류기로 문제를 해결하였다. 일반적으로 co-training, co-EM의 경우가 self-training 방법에 비해 성능이 높지만 위의 예와 같이 자연스럽게 두 자질 집합으로 분리되어야 하며, 각각의 자질들이 독립적이어야 한다는 가정이 있어야 한다. [24]는 두 개의 자질 집합으로 나누기 힘든 문제에 대해서 임의로 자질들을 나눠서 co-training과 co-EM에 적용하여 self-training과 성능을 비교하였다. 그 결과 문제를 정확히 두 자질들로 나누기 힘든 경우 위의 두 방법을 적용하여도 self-training과 성능 차이가 크게 나지 않았으며, 평가하는 말뭉치에 따라서는 self-training이 높은 성능을 보였다.

의미역 결정의 경우 co-training과 co-EM과 같이 다양한 자질들을 두 자질 집합으로 나누는 명확한 기준이 없으며, 임의로 자질들을 나누어도 각 자질 집합들이 의미역을 결정하는데 충분한지 여부도 확인하기 어려운 문제점이 있다. 또한 각각의 자질 집합 사이에 독립성을 가정하기도 힘들기 때문에 본 연구에서는 self-training 알고리즘을 적용하였다.

일반적인 self-training 알고리즘의 경우 그림 9에서와 같이 학습 말뭉치를 구축하는 부분(initialization)과 이 말뭉치를 이용하여 반복적으로 모델을 학습하는 부분(iteration)으로 구성된다. 이때 일반적인 경우 학습을 반복하는 과정에서 새로 의미역이 결정된 대상은 일반적으로 한번의 학습 단계가 끝나고 나서 일괄적으로 학습 말뭉치에 추가된다. 그러나 본 논문에서 제안하는 수정된 self-training 알고리즘의 경우 의미역이 결정된 대상을 바로 학습 말뭉치에 추가하여 확률 모델을 수정한다. 즉, 그림 10과 같이 기존의 self-training 알고리즘의 경우 학습 말뭉치를 통해 구축된 모델 M_1으로 의미역이 태깅되지 않은 말뭉치를 평가하고 새로 학습 말뭉치에 추가할 대상을 선택한다. 그리고 한번의 학습 단계가 끝나게 되면 이 대상들을 학습 말뭉치에 추가하게 되고 다음 학습 단계에서는 새로운 모델 M_2로 학습을 하게 된다. 하지만 수정된 self-training 알고리즘은 학습 말뭉치에 추가할 대상이 선택되면 바로 모델을 수정

CF-based Model (Initialization) :

Select case frame to determine the arguments to be labeled, along with their roles.

- Let A be the set of annotated arguments. ; A = Ø
- Let U be the set of unannotated arguments, initially all arguments.
- Let N be the newly annotated arguments. ; N = Ø

Add to N each argument whose role assignment is unambiguous.

Set U to U – N and set A to A + N.

Probability-based Model (Iteration) :

Let n be the newly annotated argument. ; n = Ø

repeat

Select N from all arguments in U for which ;

- the highest probability candidate meets the threshold (Θ_1).
- Set U to U – n and set A to A + n.
- Compute the probability model, using counts over the arguments in A.
  - if the candidate doesn't meet the threshold (Θ_1) than move to next level.
  - if the probability is lower than the specific threshold (Θ_2), then apply level 3.

Set U to U – N and set A to A + N.

If number of N is 0 than decrease the threshold (Θ_1).

until Θ_1=0

그림 9 수정된 Self-training 알고리즘

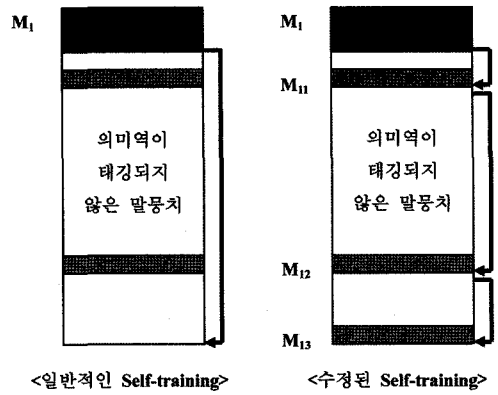

그림 10 Self-training 알고리즘 비교

하여 모델 M_{11}, M_{12}, M_{13} 등으로 학습하게 된다[5]. 즉, 한번의 학습 단계가 끝나기 전에 모델을 지속적으로 수정함으로써 최선의 모델로 의미역이 태깅되지 않은 말뭉치를 평가할 수 있게 된다.

논 본문과 같이 확률 모델을 사용하는 경우, 초기 학습 과정에 신뢰할 수 있는 대상(의미역이 결정된 대상)을 추가하면 전체적인 확률을 안정시킬 수 있고 이전에 없었던 확률들이 생길 수 있기 때문에 의미역 결정을 하는데 도움이 된다. 또한 M_{13}와 같이 다음 학습 단계에서 학습 말뭉치에 추가되어야 할 대상들이 미리 추가됨에 따라 전체적으로 학습 단계의 수가 줄어들어 학습 시간 측면에서도 효율적이게 된다. 일반적인 self-training 알고리즘과 제안된 self-training 알고리즘의 성능 및 실행시간 비교는 이후 8장에서 자세히 살펴보도록 하겠다.

8. 실험 및 평가

8.1 실험 말뭉치

본 논문에서는 세종 전자사전에 기술되어 있는 예문들을 추출하여, 부사격 조사를 포함하지 않은 문장, 문법적 오류가 있는 문장들을 제거한 후 실험에 사용하였다. 총 34,371개의 문장들 중 임의로 1,225개의 문장을 선택하여 평가 말뭉치로, 나머지 문장들은 학습 말뭉치로 구축하였다. 평가 말뭉치는 단문 외에도 복문 및 중문을 포함한 문장들로 구성되어 있고, 문장 하나에 평균 1.2개의 해당 부사격 조사를 포함하고 있었다.

평가 말뭉치는 세종 전자사전에서 정의한 15개의 의미역에 대해 격틀사전의 정보를 이용하여 의미역을 결정하였으며, 격틀에 기술되지 않은 부사격 조사에 대해서는 세종 전자사전의 의미역 기술 지침서를 참고하여 의미역을 결정하였다.

의미역 결정의 정확한 실험 결과를 위해 형태소 분석과 구문 분석 과정에서 발생하는 오류를 수정하고, 단어의 의미 중의성 해소 과정을 거쳐 올바른 논항의 의미를 결정할 수 있으나 본 연구에서는 의미역 결정에 초점을 맞추고 있기 때문에 이전 단계의 오류에 대해서는 수정 없이 그대로 결과를 이용하였다.

8.2 실험 결과

표 1은 의미역 결정 모델에 따른 부사격의 성능을 보여준다. 기본 모델(baseline)은 각 부사격에 대해서 최다 빈도를 가지는 의미역으로 결정하는 모델로 '에'와 '에서'는 장소를, '로'와 '에게'는 도착점을 할당하였다. '로'의 경우 상대적으로 다른 부사격에 비해 다양한 의미를 가지기 때문에 기본 모델의 정확률이 낮게 나온 것을 알 수 있다. 격틀 모델은 격틀사전에 기술된 격틀에 대해서 일부만 의미역이 결정되므로 정확률(precision)과 재현률(recall)로 나누어 평가하였고, 나머지 모델들은 정확률로만 평가하였다. '에게'의 경우 간접목적어로, 필수 부사격에 해당하여 격틀사전에 많이 기술되었기 때문에 다른 부사격에 비해 격틀 모델을 통해서 높은 재현률를 얻을 수 있었다. 제안 모델은 본 논문에서 격틀 모델의 결과를 이용하여 확률 모델을 통해 점진적으로 학습한 모델로 평균 83.00%의 정확률을 보였고 기본 모델보다 약 37%정도의 성능 향상을 보였다.

표 1 의미역 결정 모델에 따른 성능

	에	로	에서	에게	전체
기본 모델	49.22	29.38	52.75	62.50	45.72
격틀 모델	90.29	90.45	96.92	95.76	91.54
	62.43	51.16	34.62	70.63	56.82
제안 모델	84.57	79.69	80.77	86.63	83.00

(단위 : %)

* 정확률 : 정확하게 의미역이 결정된 개수 / 의미역이 결정된 개수
* 재현률 : 정확하게 의미역이 결정된 개수 / 의미역을 결정해야 하는 개수

표 2는 의미역 별 세부 성능을 보여준다. 표 2에서 보여지는 것처럼, '로'의 경우 총 11가지의 의미역으로 사용되기 때문에 다른 부사격보다 애매성이 높아 성능이 낮게 나온 것으로 보여진다. 그러나 '에서'의 경우 총 3가지 의미역으로 쓰이는데 비해서 예상보다 성능이 낮게 나온 이유는 격틀 모델의 재현률이 낮아 충분한 학습 말뭉치를 확보하지 못했고, 또한 '에서'의 특징을 확률 모델에서 제대로 반영하지 못했기 때문이라고 할 수 있다. '에서'는 확률 모델에 반영된 자질(서술어, 격조사,

5) 그림 10에서 화살표는 각 모델의 평가 범위를 나타내며, 밝은 회색 부분은 새로 의미역이 결정된 대상을 나타낸다. 그리고 M_{ij}에서 i는 i번째 학습 단계를, j는 i번째 학습 단계에서 j번째 모델을 나타낸다.

표 2 의미역 별 세부 성능

의미역	에		로		에서		에게	
	정확률	빈도수	정확률	빈도수	정확률	빈도수	정확률	빈도수
행위주	100	2	50	2	100	5	90.00	20
경험주	100	2	0	0	0	0	57.14	14
심리경험주	0	0	0	0	0	0	100	2
동반주	0	0	100	1	0	0	0	0
대상	83.33	49	83.33	6	0	1	0	1
장소	85.79	346	71.43	7	89.58	96	87.50	8
방향	0	0	71.43	14	0	0	0	0
도착점	86.69	219	91.50	114	0	0	92.00	100
결과상태	0	0	85.15	101	0	0	0	0
출발점	0	2	0	0	68.75	80	81.81	11
도구	70.00	10	67.71	96	0	0	50	2
영향주	79.59	49	69.23	39	0	0	50	2
기준치	70.83	24	57.14	7	0	0	0	0
목적	0	0	0	0	0	0	0	0
내용	0	0	100	2	0	0	0	0

(단위 : %, 개)

논항의 의미부류) 외에 동사의 이동이나 변화의 의미에 따라 장소와 출발점으로 구별되는 특징이 있다.

아래의 예에서 알 수 있듯이 '나오다'의 경우 '겸허하다'와 다르게 물리적 이동의 의미를 가지기 때문에 출발점으로 의미역이 결정되었다. 그러나 형태소 분석 및 구분 분석의 결과만으로는 동사의 이동이나 변화의 의미를 파악할 수 없기 때문에 '에서'의 의미역을 결정하는데 어려움이 있다. 추후 이러한 특징을 확률 모델에 반영한다면 성능을 개선할 수 있으리라고 생각된다.

(예 2 '에서'의 의미역 결정)

> 사람들은 죽음 앞에서|**장소** 겸허하다.
> 수도꼭지에서|**출발점** 물이 잘 나온다.

표 3은 학습 말뭉치의 양과 질 따른 확률 모델의 성능을 보여준다. 격틀 모델 1(F_1-measure = 0.6090)은 격틀을 선택하는데 임계값을 두어 정확률은 높지만 재현률이 낮은 학습 말뭉치를 구축하는 모델이고, 격틀 모델 2(F_1-measure = 07012)는 임계값을 두지 않고 정확률은 다소 낮지만 재현률이 높은 학습 말뭉치를 구축하는 모델이다. 표 3의 성능에서 알 수 있듯이 격틀 모델 1과 같이 학습 말뭉치가 부족한 경우 학습 예제 부족 문제가 발생할 수 있기 때문에 모든 부사격에서 학습 말뭉치가 충분한 격틀 모델 2에서 좋은 성능을 나타내었다. 따라서 본 논문에서는 격틀 모델 2를 통해서 학습 말뭉치를 구축하였다.

표 3 학습 말뭉치 양과 질에 따른 확률 모델 성능

	에	로	에서	에게
격틀 모델 1	69.23	62.10	68.80	64.28
격틀 모델 2	71.30	65.48	72.41	66.67

(단위 : %)

확률 모델의 성능이 다소 낮은 이유 두 가지 이유로 생각된다. 첫째로 확률 모델이 의미역 결정에 영향을 미치는 복잡한 언어 현상을 반영하지 못하였기 때문이다. 그리고 기존의 비지도 학습 방법과 달리 본 논문에서는 자동으로 학습 말뭉치를 구축하였기 때문에 부분적으로 오류가 포함되어 있기 때문이다. 이 오류는 확률 모델을 학습하는 초기 단계에 의미역이 잘못 결정된 대상들이 선택되면 전체적으로 시스템의 성능을 감소시키는 문제를 발생시킨다. 이는 격틀을 선택하는데 적절히 임계값을 설정하여 학습 말뭉치의 양과 질을 조절하는 과정을 통해서 오류가 미치는 영향을 줄이도록 해야겠다.

표 4는 일반적인 self-training 알고리즘과 수정된 self-training 알고리즘의 성능과 실행시간을 비교하였다. 수정된 self-training 알고리즘의 경우 신뢰할 수 있는 대상이 선택되면 학습 말뭉치에 추가하여 순간순간 모델을 수정하기 때문에 학습 말뭉치의 문장 순서에 따라서 모델이 달라지며, 학습을 하는 반복 횟수도 달라지게 된다. 따라서 학습 말뭉치의 문장 순서에 영향을 받는지 여부를 확인하기 위해서 동일한 학습 말뭉치를 문장의 순서를 임의로 바꾸어 5회 실험을 하였다. 그 결과 평균적으로 '에'는 2%, '로'는 0.5%, '에서'는 4.5%, '에게'는 0.5%의 성능 향상을 보였으며, 약 80초 정도의 실행시간을 개선하였다. 특히 수정된 self-training 알고리즘은 최악의 성능에서도 일반적인 self-training 알고리즘 이상의 성능을 보였다.

표 4 Self-training 알고리즘 성능 비교

		일반적인 Self-Training	수정된 Self-training				
			1회	2회	3회	4회	5회
정확률	에	82.57%	83.71%	84.57%	84.29%	84.86%	85.26%
	로	79.12%	80.15%	79.64%	79.64%	79.64%	79.38%
	에서	76.23%	79.86%	80.66%	81.22%	80.66%	81.77%
	에게	86.25%	86.88%	86.25%	87.50%	86.25%	86.25%
실행시간		5m 12s	3m 46s	3m 48s	4m 5s	3m 57s	3m 45s

수정된 self-training 알고리즘의 효과는 크게 두 가지로 살펴볼 수 있다. 첫째로 '에서' 와 같이 장소와 출발점의 빈도수가 비슷한 경우, 격틀 모델을 통해 구축된 학습 말뭉치에 한 의미역의 많이 존재하게 되면 확률 모델로 학습을 하는 과정 내내 한 의미역에 치우친 방향으로 학습을 하게 된다. 하지만 신뢰할 수 있는 대상을 한번의 학습이 끝나기 전에 바로 확률정보에 추가함에 따라 수정된 모델로 학습을 할 수 있기 때문에 이후에 학습을 하는 과정에서 개선 방향으로 학습을 할 수 있다. 둘째로 한번의 학습 과정에서 수정된 모델로 순간 순간 학습을 하기 때문에 새로 확률정보가 추가되어 의미역을 결정할 수 있는 대상들이 선택되는데, 이는 전체적으로 학습 과정을 반복하게 되면 학습의 횟수를 감소시켜 실행시간을 단축시키는 장점을 가진다.

그림 11은 수정된 self-training 알고리즘의 학습곡선을 나타낸다. 제안된 시스템은 확률 모델을 반복적으로 학습함에 따라 격틀 모델로 의미역이 결정되지 않은 대상들에 대해 의미역을 결정한다. 격틀 모델의 결과 재현률이 가장 낮았던 '에서'의 경우 성능 향상의 폭이 가장 컸으며, 재현률이 가장 높았던 '에게'의 경우 성능 향상의 폭이 가장 낮았다. '에게'를 제외한 나머지 부사격은 약 10% 이상의 성능 향상을 보였고 학습곡선이 증가하는 방향으로 올바르게 학습이 이루어지고 있었다.

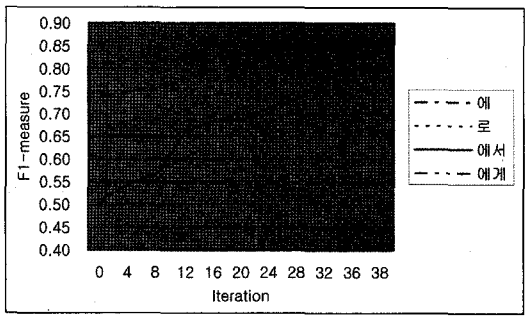

그림 11 수정된 Self-training 알고리즘의 학습 곡선

9. 결 론

본 논문에서는 비지도 학습에 기반한 의미역 결정을 하기 위해 의미역이 태깅되지 않은 말뭉치로부터 신뢰할 수 있는 대상들을 선별하여 학습 말뭉치에 추가하고 이를 확률 모델에 따라 점진적으로 학습하는 의미역 결정 시스템을 제안하였다. 기존에 학습 말뭉치를 구축하는 방법과 다르게 세종 용언 전자사전이라는 유용한 언어자원을 이용하여 자동으로 학습 말뭉치를 구축하였으며, 술어명사 및 기능동사 구문 처리, 수정된 self-training 알고리즘 적용을 통해서 시스템의 성능 및 효과를 입증하였다. 그 결과, 부사격에 대해 평균적으로 83.00%의 정확률을 보였으며, 제안된 self-training 알고리즘은 학습 시작 단계(격틀 모델)보다 평균적으로 13%정도의 성능 향상을 보였다. 그리고 직접적인 비교는 어렵지만 같은 세종 용언 전자사전을 이용한 [13]의 결과와 비교했을 때 '에'의 경우 약 2%. '로'의 경우 5%의 성능 차이를 보였다. 그러나 [13]의 경우 기계학습 방법으로 지지벡터기계를 선택하여 지도 학습을 하였다는 점과 형태소 분석, 구문 분석, 단어의 의미 중의성 해소 등 이전 단계에서 발생하는 오류를 수정하여 제어하였다는 점을 생각해 볼 때 본 논문에서 제안한 방법의 성능을 예상해 볼 수 있다.

본 논문에서는 간단한 확률 모델을 통해서 의미역 결정을 하였으나 구분 분석 결과에서 얻을 수 있는 서술어와 논항의 정보(형태소, 파생접사, 보조사 등), 주어와 목적어의 정보들을 확률 모델에 반영하여 실험을 보완할 필요가 있다. 또 형태소 분석, 구문 분석, 단어의 의미 중의성 해소 단계에서 발생하는 오류를 수정하여 본 연구에서 제안하는 시스템의 정확한 성능을 평가할 필요가 있다. 마지막으로 제안된 self-training 알고리즘을 의미역 결정 외에 정보추출, 개체명 인식 등 다양한 자연언어처리 분야에 적용해서 그 효과를 입증하는 것도 향후 과제로 남아있다.

참 고 문 헌

[1] Kurohashi, S, and Nagao, M. "A Method of Case Structure Analysis for Japanese Based on Examples in Case Frame Dictionary," IEICE Transaction Information and System, Vol.E77-D, No.2, pp. 227-239, 1994.

[2] Stephen Beale, Serei Nirenburg, and Kavi Mahesh,

"Semantic Analysis in The Mikrokosmos Machine Translation," In Proceeding of Symposium on NLP, 1995.

[3] Aria Haghighi, Kristina Toutanova, and Christopher Manning, "A Joint Model for Semantic Role Labeling," In Proceedings of CoNLL 2005 Shared Task, 2005.

[4] Daniel Gildea and Daniel Jurafsky, "Automatic Labeling of Semantic Roles," In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 2000.

[5] Daniel Gildea and Daniel Jurafsky, "Automatic Labeling of Semantic Roles," Computational Linguistics, Vol.28, No.3, pp. 245-288, 2002.

[6] Daniel Gildea and Martha Palmer, "The Necessity of Parsing for Predicate Argument Recognition," In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 239-246, 2002.

[7] Kadri Hacioglu, Sampeer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky, "Semantic role labeling by tagging syntactic chunks," In Proceedings of CoNLL 2004 Shared Task, 2004.

[8] Nianwen Xue and Martha Palmer, "Calibrating Features for Semantic Role Labeling," In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004.

[9] Sameer Pradhan, Kadri Hacioglu, Wayne Ward, James H. Martin, and Daniel Jurafsky, "Semantic Role Chunking Combining Complementary Syntactic Views," In Proceedings of CoNLL 2005 Shared Task, 2005.

[10] S.B. Park, "Decision Tree Based Disambiguation of Semantic Roles for Korean Adverbial Postposition," IEICE Transaction Information and System, Vol.E86-D, No.8, 2003.

[11] Vasin Punyakanok, Peter Koomen, Dan Roth, and Wentau Yih, "Generalized Inference with Multiple Semantic Role Labeling Systems," In Proceedings of CoNLL 2005 Shared Task, 2005.

[12] Jung-Hye Park, "Determination of Thematic Roles according to Syntactic Relations Using Rules and Statistical Models," MS Thesis, Pohang University of Science and Technology, 2002.

[13] Myung-Chul Shin, "Integration of Case-Frame Dictionary into Machine Learning Techniques for Semantic Role Assignment of Korean Adverbial Cases," MS Thesis, Pohang University of Science and Technology, 2006.

[14] Robert S. Swier and Suzanne Stevenson, "Unsupervised Semantic Role Labelling," In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 95-102, 2004.

[15] Robert S. Swier and Suzanne Stevenson, "Exploiting a Verb Lexicon in Automatic Semantic Role Labelling," HLT/EMNLP, 2005.

[16] Xavier Carreras et al, "Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling," In Proceeding of CoNLL-2005, 2005.

[17] 이성헌, "전자사전 구축과 의미부류 - 세종 명사 의미부류 체계의 예", 한국사전학, 2005.

[18] 이성헌, "전자사전에서의 기능동사 구문 처리문제 - 세종 체언사전의 경우", 한국사전학, 2004.

[19] Emmanuel Blanchard, et al. "A typology of ontology-based semantic measures," EMOI - INTEROP, 2005.

[20] Rada Mihalcea, "Co-training and Self-training for Word Sense Disambiguation," In Proceedings of CoNLL 2004, pp. 33-40, 2004.

[21] Rayid Ghani and Rosie Jones, "A Comparison Of Efficacy And Assumptions Of Bootstrapping Algorithms For Training Information Extraction Systems," Proceedings of the LREC 2002 Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data, 2002.

[22] Stephen Clark, James R Curran, and Miles Osborne, "Bootstrapping POS tagger using Unlabelled Data," In Proceedings of CoNLL 2003, pp. 49-55, 2003.

[23] Avrim Blum and Tom Mitchell, "Combining Labeled and Unlabeled Data with Co-training," In Proceedings of the Workshop on Computational Learning Theory, pp. 92-100, 1998.

[24] Kaml Nigam and Rayid Ghani, "Analyzing the Effectiveness and Applicability of Co-training," In CIKM, pp. 86-93, 2000.

[25] 이희자, 이종희, 한국어 학습용 어미·조사사전, 한국문화사, 2001.

[26] 홍재성 외, 21세기 세종계획 전자사전 개발 연구보고서, 국립국어원, pp. 62-66, 2005.

[27] Rosie Jones et al, "Bootstrapping for Text Learning Tasks," In IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications, pp. 52-63, 1999.

[28] Scott Yih and Kristina Toutanova, 2006. "Automatic Semantic Role Labeling," HLT-NAACL 2006 tutorial.

[29] Steven Abney, "Bootstrapping," In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 360-357, 2002.

[30] Xavier Carreras et al, "Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling,' In Proceeding of CoNLL-2004, 2004.

김 병 수

2005년 서울시립대학교 컴퓨터통계학과 학사. 2007년 포항공과대학교 정보처리학과 석사. 2007년~현재 티맥스소프트 전임연구원. 관심분야는 자연언어처리, 의미분석, 의미역 결정 등

이 용 훈

2002년 숭실대학교 컴퓨터학부 학사. 2004년 포항공과대학교 컴퓨터공학과 석사 2004년~현재 포항공과대학교 컴퓨터공학과 박사과정. 관심분야는 자연언어처리, 구문분석, 의미분석 등

이 종 혁

1980년 서울대학교 수학교육학과 학사 1982년 한국과학기술원 전산학과 석사 1988년 한국과학기술원 전산학과 박사 1989년~1991년 일본전기(NEC) 중앙연구소 초청연구원. 1991년~현재 포항공과대학교 컴퓨터공학과 교수. 1998년~1999년 미국 CRL/NMSU(뉴멕시코주립대학) 방문교수. 관심분야는 자연언어처리, 기계번역, 정보검색 등