비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정

김병수°, 이용훈, 나승훈, 김준기, 이종혁
포항공대 정보통신대학원 정보처리학과, 포항공대 컴퓨터공학과, 첨단정보기술 연구센터
{akirus82°, yhlee95, nsh1979, yangpa, jhlee}@postech.ac.kr

Unsupervised Semantic Role Labeling for Korean Adverbial Case

Byoung-Soo Kim°, Yong-Hun Lee, Seung-Hoon Na, JunGi Kim, Jong-Hyeok Lee
Dept. of Graduate School of Information Technology, POSTECH
Dept. of Computer Science & Engineering, POSTECH
Advanced Information Technology Research Center (AITrc)

요 약

본 논문은 한국어정보처리 과정에서 구문 관계를 의미 관계로 사상하는 의미역 결정 문제에 대해 다루고 있다. 한국어의 경우 대량의 학습 말뭉치를 구하기 힘들며, 이를 구축하기 위해서는 많은 시간과 노력이 필요한 문제점이 있다. 따라서 본 논문에서는 학습 말뭉치를 직접 태깅하지 않고 격틀사전을 이용하여 자동으로 학습 말뭉치를 구축하고 간단한 확률모델을 적용하여 점진적으로 모델을 학습하는 수정된 self-training 알고리즘을 사용하였다. 실험 결과, 4개의 부사격 조사에 대해 평균적으로 81.81%의 정확률을 보였으며, 수정된 self-training 방법은 기존의 방법에 비해 성능 및 실행시간에서 개선된 결과를 보였다.

1. 서론

일반적으로 의미 분석은 형태소 분석과 구문 분석의 과정을 거쳐 이루어지는 자연언어처리의 상위 단계로 크게 단어의 의미 중의성을 해소하는 문제(Word Sense Disambiguation)와 문장의 서술어와 논항들 사이의 의미 관계를 결정하는 문제(Semantic Role Labeling)로 나눌 수 있다. 의미 분석 단계에서 문제가 되는 것은 단어의 다의성과 문장의 문법 관계를 의미 관계로 사상하는데 발생하는 애매성이다[13]. 본 논문에서는 이러한 의미 분석 단계 중 후자에 해당하는 의미역 결정 문제에 대해서 다루고자 한다.

의미역 결정이란, 문장의 서술어와 그 서술어가 취하는 명사 논항들 사이에 적합한 의미 관계를 결정하는 것이라고 할 수 있다. 즉, <그림 1>과 같이 문장의 표층격(Surface Case)에 해당하는 문법 관계를 심층격(Deep Case)에 해당하는 의미 관계로 사상하는 문제로 볼 수 있다. 이러한 의미 관계는 격 관계(Case role), 의미역(Thematic role, θ role)으로 불리며 오랜 기간에 걸쳐 언어학자들 사이에서 연구되어 왔던 어려운 주제 중 하나이다.

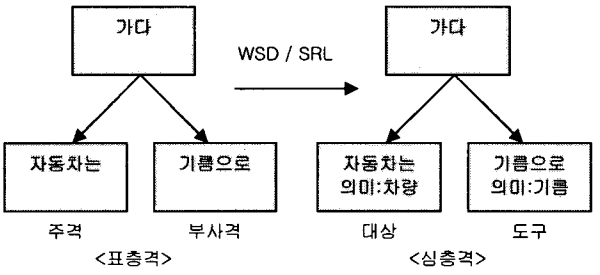


그림 1. 의미역 결정의 예

의미역 결정은 대량의 의미 관계 정보를 필요로 하는 기계번역(MT), 정보추출(IE), 질의응답(QA), 문서요약(Summarization)과 같은 다양한 자연언어처리 응용에서 성능을 향상시키는데 중요한 역할을 한다. 이에 따라 최근 들어 자동으로 의미역을 결정하는 방법에 대한 연구들이 활발히 진행되고 있다.

한국어의 경우 격조사에 의해 논항의 의미역이 부여되고, 하나의 격조사가 서술어의 특징에 따라 다양한 의미를 가지는 특징이 있다. 특히 부사격 조사의 경우 다양한 의미로 사용되고 필수격(Obligatory Case)보다 임

의격(Optional Case)으로 사용되는 경우가 많아 격틀사전에도 기술되지 않기 때문에 의미역을 결정하는데 있어서 문제가 되고 있다. 그 동안 한국어의 의미역 결정에 대한 연구로는 [1,2,10,13,18]가 있고, 일부 부사격 조사에 대해 중점적으로 다룬 연구로는 [13,18]가 있다.

본 논문은 세종전자사전의 용언사전으로부터 의미역을 결정하는데 필요한 격틀사전을 추출하고 이를 이용하여 학습에 필요한 말뭉치를 자동으로 구축한 후 확률 모델을 통해 의미역 결정을 하는 비지도 학습(Unsupervised learning)을 기반으로 한 의미역 결정 시스템을 제안한다. 일반적으로 의미역을 결정하는 작업은 의미역 수가 많고 각 의미역 사이에 미세한 차이가 있어 의미역을 결정하기가 어렵다. 또한 한국어의 경우 영어권의 FrameNet이나 PropBank에서 제공하는 대량의 의미역이 태깅된 말뭉치도 없어 기존의 지도 학습(Supervised learning) 방법을 적용하기 어렵기 때문에 비지도 학습 방법을 선택하였다. 본 연구에서는 의미역을 결정하는데 애매성이 큰 부사격 조사 '에', '로', '에서', '에게'를 대상으로 실험을 하였다.

본 논문의 구성은 다음과 같다. 2절에서는 기존에 제안된 의미역 결정 방법들에 대해 소개하고, 3절에서는 본 논문에서 제안하는 의미역 결정 시스템 구조를 소개한다. 4~6절에서는 의미역 결정 시스템의 각 부분에 대해 설명하고, 7절에서는 수정된 self-training 알고리즘에 대해 설명한다. 8절에서는 실험 및 결과를 분석하고, 마지막으로 9절에서는 결론 및 향후 연구에 대해서 간단히 기술한다.

2. 기존 연구

기존의 의미역 결정 연구는 크게 격틀사전에 기반한 방법(Case frame based method)[12]과 대량의 말뭉치에 기반한 방법(Corpus based method)[2,7,8,18], 그리고 이들 방법을 통합한 하이브리드 방법(Hybrid method)[13,16,17]으로 나눌 수 있다.

격틀사전에 기반한 방법은 서술어와 논항들의 쓰임을 기술한 격틀사전을 이용하는 방법으로, 서술어와 논항들에 대한 문법 관계를 기술한 프레임(Frame)과 논항들의 정보를 기술한 선택제약(Selectional restriction) 등을 이용하여 입력 문장에 대해 적합한 격을 할당하여 의미 관계를 설정하는 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 필요하다는 점과 격틀사전에 기술되지 않은 임의격에 대해서는 처리하지 못한다는 단점이 있다.

말뭉치에 기반한 방법은 의미역이 태깅된 대량의 말뭉치를 이용하여 기계적 혹은 통계적 학습 방법으로 의미역을 결정하는 방법이다. 즉, 말뭉치로부터 의미역 결정에 도움이 되는 자질(Features)들을 추출하고 이들을 다양한 학습 알고리즘에 따라 모델을 학습하여 의미역을 결정하는 방법이라고 할 수 있겠다. 지금까지 지지벡터기계(Support Vector Machine), 결정트리(Decision Tree), 최대 엔트로피(Maximum Entropy) 모델 등 다양한 학습 알고리즘이 의미역 결정에 사용되었다. 이 방법은 적용률이 높고 격틀사전에 기반한 방법에 비해 견고하다는 장점이 있다. 그러나 의미역을 태깅하여 말뭉치를 구축하는 작업은 많은 시간과 노력을 필요로 하는 단점이 있다.

하이브리드 방법은 둘 이상의 방법을 통합하여 두 방법의 단점을 서로 보완하는 방법이다. 예를 들어, [13]은 정확률이 높은 격틀사전에 기반한 방법과 적용률이 높은 말뭉치에 기반한 학습 방법을 통합하여 보다 정확한 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[20]. 하이브리드 방법은 의미역 결정 자체의 특징보다는 모델을 통합하는 측면의 비중이 큰 방법이라고 할 수 있겠다.

3. 의미역 결정 시스템의 구조

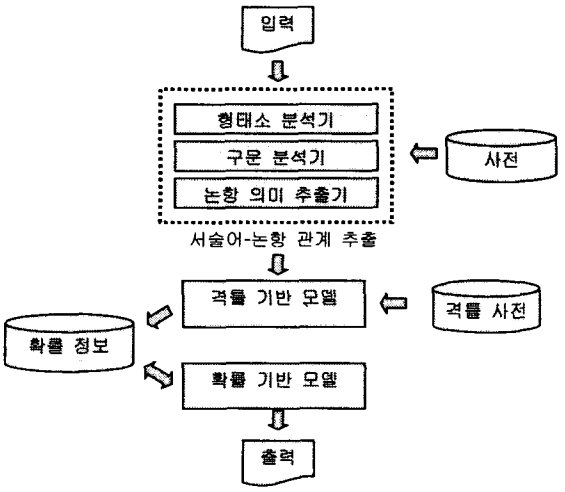


그림 2. 의미역 결정 시스템 구조

비지도 학습을 기반으로 한 의미역 결정 시스템은 <그림 2>와 같이 크게 세 부분으로 나누어진다. 첫째, 서술어-논항 관계 추출기는 입력 문장에서 서술어와 논항 관계를 트리 형태로 추출한다. 둘째, 격틀 기반 모델은 격틀사전을 이용하여 서술어-논항 관계에 적합한 격틀을 할당하여 의미역을 결정한다. 이 결과는 확률 기반 모델을 학습하는 초기 말뭉치로 사용되기 때문에 정확하고 충분한 양을 모으는 것이 중요하다. 셋째, 확률 기

반 모델은 서술어, 논항, 격조사 등의 자질들을 선택하여 Backoff 확률모델로 구성하였다. 이때 점진적으로 확률 기반 모델을 학습하기 위해서 수정된 self-training 알고리즘을 사용하였다.

4. 서술어-논항 관계 추출기

```
| 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)
    | 부대는 (주어) ----> 했다.
    => 부대(CMCN)+ 는(fjb)
        | 우리 (부사어) ----> 부대는
        => 우리(CTP1)
```

그림 3. 형태소 분석 및 구문 분석 결과의 예

서술어-논항 관계 추출기는 형태소 분석기와 구문 분석기를 통해 서술어와 그 서술어가 취하는 논항들을 의존트리(Dependency tree) 형태로 추출한다. 예를 들어, <그림 3>과 같은 구문 분석의 결과에서 동사 '하다'에 대해 주격, 목적격, 부사격으로 쓰인 논항들을 추출한다. 그 다음 논항 의미 추출기를 통해서 논항들이 가질 수 있는 의미들을 추출한다. 원래 입력 문장의 논항들은 단어의 의미 중의성 해소 과정을 거쳐 하나의 의미가 정해져야 하나 이를 위해서는 별도의 모듈이 필요하고, 단어의 의미 중의성 해소는 자연언어처리의 한 연구 분야로 쉽게 해결할 수 있는 문제가 아니기 때문에 본 연구에서는 이는 고려하지 않도록 하겠다. 서술어-논항 관계 추출기에 사용된 형태소 분석기 및 구문 분석기는 포항공과대학교 지식 및 언어공학 연구실에서 개발한 KoMA와 KoPA를 사용하였다.

5. 격틀사전을 이용한 의미역 결정

5.1 용언사전으로부터 격틀사전 구축

세종전자사전의 용언사전에는 표제어에 대해 다양한 통사적, 의미적 정보가 XML 형태로 수록되어 있다. 따라서 이들 중 의미역 결정에 필요한 정보들을 선별하여 전산적 처리가 용이하도록 격틀사전을 구축해야 할 필요가 있다. 예를 들어, <그림 4>에서와 같이 'X=N0-이 Y=N1-로 V'와 같은 프레임(Frame), '교통기관(버스|기차|비행기)'와 같은 선택제약(Selectional restriction), 'THM'과 같은 의미역(Semantic role) 정보들을 추출하여 격틀사전을 구축하였다.

격틀사전의 선택제약은 세종전자사전의 명사의미부류를 기준으로 기술되어있다. 세종 명사의미부류는 트리구조로 최상의 의미부류 <구체물>, <집단>, <장소>, <추상적대상>, <사태> 등에 대해 총 582개의 의미부류로 분류된다. 특히, <사태>는 술어명사에 해당하는 의미부류로 다음 절에서 설명할 기능동사 구문 처리에 있어서 유용하게 사용된다.

```
<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 사전정보

용언사전은 2007년에 완성을 목표로 현재 기술이 미흡한 부분에 대한 보완이 이루어지고 있다. 따라서 이 부분은 격틀 사전에서 제거하였다. 그 결과 표제어 하나당 평균적으로 1.97개의 격틀이 있음을 확인하였다. 그러나 기능동사의 경우 10개 이상의 격틀을 가지고 있는 경우가 많기 때문에 기능동사를 제외한 표제어들이 가지는 격틀의 수는 1.97개 이하로 예상된다.

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[1](Support verb)란 술어명사를 중심으로 문장 구성이 가능하도록 술어위치를 채우고 술어명사의 현동화(Actualization)를 뒷받침해주는 동사를 말한다. 이런 술어명사와 술어명사의 고유논항, 기능동사로 구성하는 기본구문을 기능동사구문이라고 한다[3].

세종전자사전에 기술된 대표적인 '하다', '되다', '시키다'는 각각 79개, 50개, 18개의 격틀을 가지고 있다. 이렇게 다양한 술어명사와 사용될 수 있는 기능동사의 특징은 유사도 계산 수식이 정교하지 않다면 문장의 유사도를 계산하여 하나의 격틀을 선택하는데 변별력이 없기 때문에 잘못된 격틀이 선택될 수 있다. 따라서 <예 1>과 같이 'Npr[2]-을 V' 형태로 나타나는 기능동사 구문에 대해서 'Npr+V'의 형태와 같이 변형해서 처리해주면 처리속도도 빨라지고 더 정확한 격틀을 선택할 수 있다.

[1] 기능동사는 'Light verb'라고도 한다.

[2] Npr은 술어명사를 나타낸다.

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

예 1. 기능동사 구문 처리의 예

실제 '야영하다' 의 경우 1개의 프레임을 가지고 있기 때문에 '야영을 하다'로 의미역을 결정하는 것보다 유사도를 계산하는 횟수도 79번에서 1번으로 줄어들었고 올바른 의미역이 결정되었다. 또한 확률모델을 적용하는데 있어서도 '하다'가 아닌 '야영하다'와 같이 개별적인 확률을 가지기 때문에 정확히 의미역을 결정할 수 있다.

5.3 격틀 선택

격틀사전을 이용하여 입력 문장에 대해서 적합한 격틀을 선택하는 과정은 서술어-논항 관계 추출기에서 추출한 논항과 격틀에 기술된 선택제약 사이의 유사도를 계산하여 문장의 전체 유사도가 가장 높은 격틀을 선택하는 과정으로 생각할 수 있다.

$$\mathrm{sim}(c_i,c_j)=\frac{2*\min_{\mathrm{pths}(\mathrm{mscs}(c_i,c_j),rt)} \mathrm{len}_e p}{\min_{\mathrm{pths}(c_i,c_j)} \mathrm{len}_e p+2*\min_{\mathrm{pths}(\mathrm{mscs}(c_i,c_j),rt)} \mathrm{len}_e p}$$

* pths(x) : x와 y사이의 경로들의 집합
* $\mathrm{len}_e(p)$: 경로 p에 포함된 간선(edge)들의 수
* mscs(x,y) : x와 y사이의 가장 가까운 공통 조상

수식 1. 두 의미 사이의 유사도 계산

본 논문에서는 각 논항에 대해 <수식1>[9]을 이용하여 선택제약에 기술된 의미와 유사도를 계산한다. 입력 문장의 논항과 <그림 4>에서 알 수 있듯이 선택제약 모두 하나 이상의 의미를 가질 수 있으므로 이들 사이의 유사도 중에서 최대가 되는 값을 논항의 의미로 선택하는 방법을 사용하여 단어의 의미 중의성 해소 과정을 대신하였다.

$$\mathrm{total\_score}=\left(\sum_{i,j}\mathrm{weight}*\max \mathrm{sim}(c_i,c_j)\right)*\sqrt{\frac{L}{N}}$$

* weight : 각 필수격 논항의 가중치
* L : 입력 문장에 있는 필수격 논항의 수
* N : 격틀에 있는 필수격 논항의 수

수식 2. 문장 전체의 유사도 계산

두 의미 사이의 유사도 계산 과정을 거쳐 각 논항의 유사도가 계산되면 문장 전체의 점수를 구하여 적합한 격틀을 선택한다. 같은 서술어의 경우 격틀 사이에 주어와 목적어가 동일한 경우가 많고 실제 격틀을 선택하는데 부사격 논항의 비중이 크기 때문에 주격, 목적격 논항보다 부사격 논항에 가중치를 주었다. 또 주격이나 목적격 논항이 생략되어 입력 문장에 논항의 수가 격틀에 기술된 논항의 수보다 작다면(L<N) 점수를 보정해주어야 하기 때문에 $\sqrt{L/N}$ 을 곱하였다.

6. 확률모델을 이용한 의미역 결정

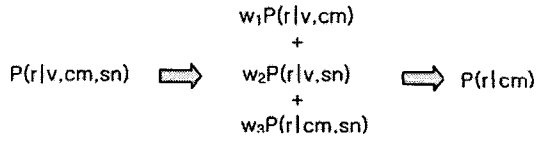
6.1 확률 정보 구축

기존의 다른 자연언어처리 분야에 비지도 학습을 적용한 경우는 사람이 직접 학습 말뭉치를 구축하거나 규칙을 설정하여 구축하였다. 그러나 의미역 결정은 서술어와 논항, 격조사의 관계에 따라 다양한 의미로 사용되기 때문에 소량의 학습 말뭉치나 몇 가지 규칙만으로는 부사격 조사의 다양한 쓰임을 나타내기에 한계가 있다. 또 본 연구와 같이 확률모델을 사용하여 의미역을 결정하는 경우 충분한 학습 말뭉치가 있어야 안정적인 확률을 얻을 수 있고 확률이 존재하지 않아 의미역을 결정할 수 없는 문제가 발생하지 않는다.

따라서 본 시스템에서는 격틀 기반 모델의 결과로 의미역이 결정된 말뭉치에서 확률모델의 수식에 따라 확률 정보를 추출하여 학습에 필요한 확률 정보를 구축하였다. 이와 같이 격틀사전을 이용하여 자동으로 학습 말뭉치를 구축하면 일반적인 규칙에 비해 서술어의 특성에 따라 기술된 다양한 프레임과 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 Backoff 확률모델

본 연구에서는 [9]에서 제시한 Backoff 확률모델을 형태소 분석, 구문 분석 결과 및 격틀사전에서 얻을 수 있는 정보에 맞게 수정하였다.



* 의미역(r), 서술어(v), 격조사(cm), 논항의 의미부류(sn)

그림 5. Backoff 확률모델

일반적으로 의미역 결정은 서술어와 논항, 격조사 등의 조건이 만족될 때 특정한 의미역을 가지는 레벨 1의 확률로 생각할 수 있다. 그러나 논항은 논항의 어휘 정보를 이용하면 학습 예제 부족 문제(Data Sparseness Problem)가 생길 수 있으므로 세종전자사전의 명사의미부류 정보를 이용하였다. 그러나 레벨 1은 세가지 조건이 모두 만족하지 않으면 확률이 존재하지 않아 의미역을 결정하지 못하는 문제가 발생한다. 따라서 좀 더 일

반적인 확률을 얻기 위해서 레벨 2와 같이 세가지 조건 중 한가지 조건을 제거한 형태로 서술어, 논항, 격조사의 조합으로 확률을 분해하였고, 각 확률에 대한 가중치의 곱의 합을 결합한 형태로 구성하였다. 레벨 3은 격조사가 주어졌을 때 의미역이 결정되는 확률로 8절의 실험에서 기본모델처럼 부사격 조사에 대해서 최다빈도를 가지는 의미역을 결정하는 수식을 의미한다.

각 레벨은 의미역 결정 결과가 믿을 수 있는 임계값(Θ_1)에 도달하지 못하거나 확률이 없을 경우 다음 레벨로 진행하면서 보완을 한다. 그러나 레벨 3이 바로 적용될 경우 대부분의 의미역이 기본모델에 해당하는 방식으로 정해지는 문제가 있다. 따라서 레벨 3은 <알고리즘 1>에서 알 수 있듯이 특정 임계값(Θ_2) 이하일 때만 적용하여 확률이 작아 신뢰도가 낮거나 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, 단어의 의미 중의성 해소(Word Sense Disambiguation), 개체명 인식(Named Entity Recognition), 통계적 구문분석(Statistical Parsing) 등 다양한 자연언어처리 응용에 적용되었다[19]. 의미역 결정에 비지도 학습 방법이 사용된 연구로 [16,17]이 있다.

3.2 Self-training 알고리즘

본 논문은 비지도 학습 방법의 하나인 self-training 알고리즘을 사용하여 반복적으로 확률모델을 학습하였다. 비지도 학습 방법은 특성에 따라 co-training, co-EM, self-training, EM 등과 같은 방법들로 나눠 볼 수 있다[6,11,14]. Self-training, EM 방법이 자질들을 한 가지 관점에서 문제를 해결하는데 비해 co-training, co-EM 방법의 경우 서로 다른 두 가지 관점에서 문제를 바라보는 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. Co-training, co-EM의 경우 일반적으로 self-training 방법에 비해 성능이 높지만 위의 예와 같이 자연스럽게 두 가지 자질 집합으로 분리되어야 하며, 각각의 자질들이 독립적이어야 한다는 가정이 있어야 한다. [11]은 두 개의 자질 집합으로 나누기 힘든 문제에 대해서 임의로 자질들을 나눠서 co-training과 co-EM에 적용하여 self-training과 성능을 비교하였다. 그러나 명확히 두 관점으로 나누기 힘든 경우 위의 두 방법을 적용하여도 self-training과 성능 차이가 크게 나지 않았으며 평가하는 말뭉치에 따라서는 self-training이 높은 성능을 보이기도 하였다.

의미역 결정 문제의 경우 co-training과 co-EM과 같이 다양한 자질들을 두 자질 집합으로 나누는 명확한 기준이 없으며 각각의 자질 집합 사이에 독립성을 가정하기도 힘들기 때문에 본 연구에서는 self-training 알고리즘을 사용하였다.

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 items 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

알고리즘 1. 변형된 Self-Training 알고리즘

일반적인 self-training 알고리즘의 경우 <알고리즘 1>에서와 같이 초기 학습 말뭉치를 구축하는 부분(Initialization)과 이 말뭉치를 이용하여 반복적으로 모델을 학습하는 부분(Iteration)으로 구성된다. 이때 반복 학습을 하는 과정에서 새로 의미역이 결정된 대상은 일반적으로 한번 반복이 끝나고 나서 일괄적으로 학습 말뭉치에 추가된다. 그러나 본 논문에서 제안하는 수정된 self-training 알고리즘의 경우 의미역이 결정된 데이터를 바로 학습 말뭉치에 추가하여 확률모델을 수정한다. 특히 확률모델을 사용하는 경우, 초기에 학습 말뭉치에 믿을 수 있는 데이터를 추가하게 되면 확률이 안정되고 이전에 없었던 확률들이 생기기 때문에 보다 나은 모델로 학습을 할 수 있다. 또 수정된 확률모델로 나머지 부

분에 대해 의미역을 결정하기 때문에 더 정확히 결정할 수 있으며 추가적으로 반복 학습을 하지 않고 평가할 수 있기 때문에 전체적으로 반복횟수가 줄어들어 실행 시간에 있어서도 효율적이다. 일반적인 self-training 알고리즘과 제안된 self-training 알고리즘의 성능 및 실행 시간은 이후 8절에서 자세히 분석하도록 하겠다.

8. 실험 및 평가

8.1 실험 말뭉치

본 논문에서는 세종전자사전에 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.

세종전자사전에서는 행위주, 경험주, 심리경험주, 동반주, 대상, 장소, 방향, 도착점, 결과상태, 출발점, 도구, 영향주, 기준치, 목적, 내용 등 총 15개의 의미역을 정의하였다. 평가 말뭉치는 15개의 의미역에 대해 격틀사전의 정보를 이용하여 의미역을 결정하였으며, 격틀에 기술되지 않은 부사격 조사에 대해서는 세종전자사전의 의미역 기술 지침서를 참고하여 의미역을 결정하였다.

보다 정확한 실험 결과를 위해 형태소 분석과 구문분석 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

	에	로	에서	에게	전체
기본모델	49.22	29.38	52.75	62.50	45.72
격틀모델	89.46	91.26	96.43	95.65	91.21
	59.17	48.33	29.67	68.75	53.74
격틀모델+확률모델	83.29	78.09	80.11	86.25	81.81

(단위 : %)

* 정확률 : 의미역 결정이 맞은 개수 / 의미역이 결정된 개수

표 1. 모델에 따른 의미역 결정 결과

<표 1>은 의미역 결정 모델에 따른 의미역 결정 결과를 보여준다. 기본모델은 각 부사격 조사에 대해서 최다 빈도를 가지는 의미역으로 할당하는 모델로 '에'와 '에서'는 장소를, '로'와 '에게'는 도착점을 할당하였다. '에'와 '로'는 다양한 의미로 사용되기 때문에 기본 모델의 정확률이 낮게 나온 것을 알 수 있다. 격틀모델은 격틀사전에 기술된 격틀에 대해서 일부만 의미역이 결정되므로 정확률(Precision)과 재현률(Recall)로 나눠서 평가하였고 나머지 모델은 정확률로만 평가하였다. 격틀모델+확률모델은 본 논문에서 제안한 모델로 평균 81.81%의 정확률을 보였고 기본모델보다 약 35% 정도의 성능 향상을 보였다. '에게'의 경우 필수 부사격으로 격틀에 많이 나타나기 때문에 다른 부사격 조사에 비해 격틀모델을 통해서도 높은 재현률를 얻을 수 있었다.

의미역	에	로	에서	에게
행위주	100	50	100	85.00
경험주	100	0	0	57.14
심리경험주	0	0	0	100
동반주	0	100	0	0
대상	83.67	100	0	0
장소	85.26	71.43	92.71	87.50
방향	0	85.71	0	0
도착점	83.11	91.23	0	94.00
결과상태	0	82.18	0	0
출발점	0	0	63.75	72.73
도구	70.00	64.21	0	50
영향주	79.59	67.50	0	50
기준치	75.00	28.57	0	0
목적	0	0	0	0
내용	0	100	0	0

(단위 : %)

표 2. 부사격 조사 별 의미역 결정 결과

<표 2>는 부사격 조사 별 의미역 결정 결과를 보여준다. 전체적으로 도구, 영향주, 기준치, 출발점과 같은 의미역이 대상, 장소, 도착점 등의 의미역보다 성능이 낮은데 이는 상대적으로 빈도수가 적기 때문에 확률모델로 학습하는 과정에서 낮은 확률을 가지기 때문이다. 이 점은 추후 확률모델의 보완이 이루어져야 할 것이다.

<표 2>의 결과에서 상대적으로 '로'의 성능이 낮은 이유는 총 11가지의 다양한 의미역으로 사용될 수 있기 때문이다. '에서'의 경우는 3가지 의미역으로 쓰이는데 비해서 성능이 낮은 이유는 격틀모델의 적용률이 낮아 충분한 학습 데이터를 확보하지 못했고 결과적으로 '에서'의 특징이 확률모델에 제대로 반영되지 않았기 때문이라고 할 수 있다. 또한 '에서'는 확률모델에 반영된 자질 외에 동사의 이동이나 변화의 의미에 따라 장소와 출발점으로 구별되는 특징이 있다.

그 사람은 학교에서|**장소** 자퇴했다.
수도꼭지에서|**출발점** 물이 잘 나온다.

예 2. '에서'의 의미역 결정의 예

<예 2>에서 알 수 있듯이 '나오다'의 경우 '자퇴하다'와 다르게 이동의 의미를 가지기 때문에 출발점으로 의미역이 결정된 것을 알 수 있다. 그러나 형태소 분석 및

구분 분석의 결과만으로는 동사의 이동이나 변화의 의미에 대해서는 파악할 수 없기 때문에 '에서'의 의미역을 정하는데 어려움이 있다. 추후 이러한 특징을 반영한다면 성능을 개선할 수 있으리라고 생각된다.

	정확률	실행시간
기존 방법	79.36%	11m 10.392s
수정된 방법	81.81%	5m 32.978s

표 3. Self-training 알고리즘 성능 비교

<표 3>[3]은 일반적인 self-training 알고리즘과 수정된 self-training 알고리즘의 성능과 실행시간을 나타낸다. 수정된 self-training 알고리즘이 약 2.5% 정도의 성능 향상을 보였으며 약 2배의 실행시간 단축을 보였다.

그 이유는 일반적인 self-training 알고리즘은 모델을 학습할 때 한번 반복이 끝나고 일괄적으로 확률모델을 수정하지만 제안된 알고리즘은 의미역이 결정되는 즉시 확률모델을 수정하여 이후 부분을 보다 나은 모델로 학습할 수 있기 때문이다. 이때 수정된 확률모델로 의미역을 결정하게 되면 일반적인 self-training 알고리즘과 달리 추가적으로 의미역이 결정되는 부분이 생겨 전체 반복횟수가 줄어들어 실행시간도 단축시키는 결과를 보였다.

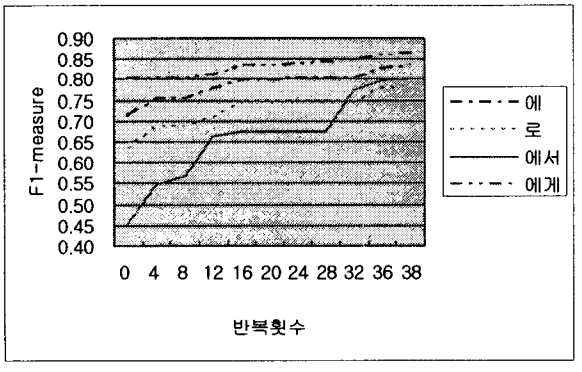


그림 6. 반복학습에 따른 학습 곡선

<그림 6>은 반복학습에 따른 학습곡선을 나타낸다. 확률모델을 반복적으로 학습함에 따라 초기에 격틀모델로 의미역이 결정되지 않은 대상들의 의미역이 결정된다. 격틀모델의 결과 적용률이 가장 낮았던 '에서'의 경우 성능 향상의 폭이 가장 컸으며, 적용률이 가장 높았던 '에게'의 경우 성능 향상의 폭이 가장 낮았다. '에게'를 제외한 나머지 부사격 조사들은 10% 이상의 성능 향상을 보였고 학습 곡선이 증가하는 방향으로 올바르게 학습 과정이 이루어지고 있음을 알 수 있었다.

[3] 이 실험은 Pentium4 2.4G, RAM 1G 컴퓨터 환경에서 수행하였다.

	에	로
격틀모델 1	71.75%	54.46%
격틀모델 2	73.62%	64.84%

표 4. 초기 학습 말뭉치 양에 따른 확률모델 결과

<표 4>는 학습 말뭉치의 양에 따른 확률모델의 성능을 나타낸다. 격틀모델 1은 격틀선택에 임계값을 두어 정확률이 높지만 재현률이 낮은 학습 말뭉치를 구축하는 모델(F1-measure = 0.2622)이고 격틀모델 2는 임계값을 두지 않고 정확률은 낮지만 재현률을 높은 학습 말뭉치를 구축하는 모델(F1-measure = 0.6852)이다.

위의 결과에서 알 수 있듯이 초기 학습 말뭉치가 부족한 경우 학습 예제 부족 문제가 발생할 수 있다. '에'와 '로' 모두 학습 예제가 충분한 격틀모델 2에서 좋은 성능을 나타내었다. 특히 '로'와 같이 애매성이 큰 경우 학습 말뭉치의 영향이 큰 것으로 보인다. 본 연구에서는 충분한 초기 학습 말뭉치 확보를 위해 격틀모델 2을 사용하였다.

8. 결론 및 향후 연구

본 논문은 격틀사전을 이용하여 자동으로 초기 학습 말뭉치를 구축하고 수정된 self-training 알고리즘에 따라 비지도 학습을 하는 의미역 결정 시스템을 제시하였다. 세종전자사전의 용언사전으로부터 격틀사전을 구축하고 이를 이용하여 확률모델 학습에 필요한 학습 말뭉치를 확보하였으며, 이를 점진적으로 학습하면서 의미역을 결정하였다.

실험 결과, 평균적으로 81.81%의 정확률을 보였다. 직접적인 비교는 어렵지만 같은 세종전자사전을 이용한 [13]의 결과와 비교했을 때 '에'의 경우 약 2%, '로'의 경우 6%의 성능 차이를 보였다. 그러나 [13]의 경우, 지도 학습(Supervised learning)을 기반으로 한 방법을 사용하였다는 점과 형태소 분석, 구문 분석, 단어의 의미 중의성 해소 등 이전 단계에서 발생하는 오류를 수정하여 제어하였다는 점을 생각해 볼 때 본 연구에서 제한된 방법의 성능을 예상해 볼 수 있다.

향후 연구로는 세 가지 측면에서 현재의 연구를 보완할 계획이다. 첫째, co-training 알고리즘을 의미역 결정에 적용하여 제안된 self-training 알고리즘과의 비교 실험을 하여 제안된 방법의 성능을 검증할 예정이다. 둘째, 확률모델에 구문 트리에서 얻을 수 있는 서술어와 논항의 정보(형태소, 파생접사, 보조사 등), 주격과 목적격의 정보들을 반영하여 보다 정확한 의미역 결정이 이루어지도록 확률모델을 보완할 계획이다. 셋째, 본 연구에서는 단어의 의미 중의성 해소 과정을 거치지 않았지만 정확한 격틀선택 및 확률 정보 확보를 위해서 최대 유사도를 가지는 논항의 의미를 선택하는 현재의 방법을 개선할 방법을 연구할 예정이다.

감사의 글

이 논문은 2006년도 두뇌한국21사업과 첨단정보기술연구센터(AITrc)를 통한 과학재단의 지원을 받았습니다.

참고 문헌

[1] 강신재, 박정혜, 대규모 말뭉치와 전산 언어 사전을 이용한 의미역 결정 규칙의 구축, 한국정보처리학회 논문지 B 제 10-B권 제 2호, 2003

[2] 양단희, 송만석, 기계학습에 의한 단어의 격 원형성 자동 획득, 정보과학회지, 25권 제7호, pp.1116-1127, 1998

[3] 이성헌, 전자사전에서의 기능동사 구문 처리문제-세종 체언사전의 경우, 한국사전학회 제 5회 학술대회 발표자료집, 2004

[4] 이희자, 이종희, 한국어 학습용 어미•조사사전, 2001

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

[6] Avrim Blum and Tom Mitchell, Combining labeled and unlabeled data with co-training, In COLT '98, 1998

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

[8] Daniel Gildea and Daniel Jurafsky, Automatic Labeling of Semantic Roles, In Proceedings of ACL 2000

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

[10] 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

[11] Kaml Nigam and Rayid Ghani, Analyzing the effectiveness and applicability of co-training, In CIKM, pp.86-93

[12] 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

[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] 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

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

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

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

[18] 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

[19] S. Clark, J.R Curran, and M.Osborne, Bootstrapping POS tagger using unlabelled data, Proceedings of CoNLL-2003, pp 49-55, 2003

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