# 국소 문맥을 이용한 형태적 중의성 해소

이충희 윤준태* 송만석
연세 대학교 컴퓨터과학과, *서울시 강남구 삼성동 154-10
{forever, mssong}@december.yonsei.ac.kr
*jtyoon@daumsoft.com

## Morphological disambiguation using Local Context

Chung-Hee Lee Jun-Tae Yoon* Man-Suk Song
Dept. of Computer Science, Yonsei University
*DaumSoft

**요 약**

본 논문은 국소문맥을 사용하여 만들어진 Decision List를 통해 단어의 형태적 중의성을 제거하는 방법을 기술한다. 최초 종자 연어(Seed Collocation)로 1차 Decision List를 만들어 실험 말뭉치에 적용하고 태깅된 결과를 자가 학습하는 반복과정에 의해 Decision List의 수행능력을 향상시킨다. 이 방법은 단어의 형태적 중의성 제거에 일정 거리의 연어가 가장 큰 영향을 끼친다는 직관에 바탕을 두며 사람의 추가적인 교정을 필요로 하지 않는 비교사 방식(대량의 원시 말뭉치에 기반한)에 의해 수행한다. 학습을 통해 얻어진 Decision List는 연세대 형태소 분석기인 MORANY의 형태소 분석 결과에 적용되어 태깅시 성능을 향상시킨다.

실험 말뭉치에 있는 중의성을 가진 12개의 단어들에 본 알고리즘을 적용하여 긍정적인 결과(90.61%)를 얻었다. 은닉 마르코프 모델의 바이그램(bigram) 모델과 비교하기 위하여 '들었다' 동사만을 가지고 실험하였는데 바이그램 모델의 태깅결과(72.61%)보다 뛰어난 결과(94.25%)를 얻어서 본 모델이 형태적 중의성 해소에 유용함을 확인하였다.

### 1. 서론

교착어인 한국어는 어절단위로 문장을 이루며 어절은 다시 형태소단위로 나뉘어진다. 형태소 분석은 각 어절에 대한 가능한 형태소와 품사의 쌍들을 분리해 내는 것으로 주어진 문장에 대한 형태소 분석 결과는 여러 가지가 나올 수 있다. 그러한 형태소 분석의 형태 품사적 중의성을 해소하기 위하여 품사 태깅 방법을 사용한다.

품사 태깅 방법은 일반적으로 통계 기반 방법(Statistical Method)[1][2][3][4]과 규칙 기반 방법(Rule-based Method)[5]으로 나뉘어지며 근래에는 두 가지 방법을 같이 사용하는 연구가 이루어지고 있다[6][7][8].

그 중 많이 사용되고 있는 통계적 방법에서는 기본적으로 은닉 마르코프 모델이 이용되는데 마르코프 가정과 독립가정을 사용하여 품사 전이 확률과 어휘 확률의 곱으로 (1)식과 같이 나타낸다.

$$P(T_{1,N}|W_{1,N}) = \Pi_{i=1,N} P(W_i|T_i) \times P(T_i|T_{i-1}) \quad (1)$$

위 식의 문제점은 독립가정을 적용함으로써 N개의 형태소에 대한 품사가 바이그램의 품사 전이 확률과 각 품사에 대한 어휘 확률에 의해 형태소와 품사가 결정된다는 점이다. 따라서 품사가 동일하게 전이되는 경우, 예컨대 동사-동사 중의성을 가지는 경우, 어휘 확률에만 의존하기 때문에 주어진 단어의 빈도수에 따라 빈도수가 높은 특정 형태 하나로만 일률적으로 태깅되게 된다. 또한 수정된 마르코프 모델로서 바로 앞의 어휘 정보를 문맥에 넣는 경우도 생각해 볼 수 있는데, 이 경우에도 바로 앞 어휘 정보에만 의존한다는 문제와 함께 현재 한국어 품사 부착 말뭉치의 양을 고려해 볼 때 충분한 어휘 정보를 얻을 수 없다는 근본적인 문제점을 가지고 있다. 결국 보다 넓은 범위의 어휘 정보를 이용할 필요가 있고 그러기 위해서는 충분히 큰 말뭉치가

필요하다. 즉, 품사를 결정하는 많은 분류자들이 실제로는 주위의 어휘에 의해 결정되는 일이 많은데 현재의 품사 태깅 모델들에서는 제한된 말뭉치의 양으로 인해 충분한 어휘 문맥을 고려하지 못해 특정 상황에 대해서는 정확한 형태소-품사 정보를 할당할 수 없다. 특히 한국어에서는 활용 시 불규칙 변화 등의 음운적 변이로 인해 하나의 어절에 대해 같은 품사로 된 두 개의 형태소가 나타나기도 한다. 예를 들어 '들어'라는 어절은 '듣(동사)+어(어미)'와 '들(동사)+어(어미)'의 두 가지 형태소 분석 결과를 가지며 이들 중의성 해소에는 주위의 문맥이 매우 중요하다.

본 논문에서는 원시 말뭉치로부터 품사 태깅을 위한 중의성 해소 정보를 학습하는 문제에 대해 기술한다. 일반적으로 원시 말뭉치를 이용하는 비교사 학습은 부착된 정보의 부재로 인한 정보원에 대한 불확실성으로 인해 일괄 이용하기 어려운 문제점이 있다. 본 논문에서 제시하는 방법은 전체 품사 중의성을 해소하는 품사 태깅 방법이라기 보다는 동사-동사 중의성과 같이 일반적인 통계적 방법으로 해결하기 힘든 특이 현상을 다룬다. 본 방법론은 품사 태깅의 전처리 과정으로 이용될 수 있으며 원시 말뭉치를 이용한 비교사 학습이 다양한 언어 현상을 다루기 위해 어떻게 이용될 수 있는지를 보여준다.

논문의 구성은 2장에서 관련연구에 대하여 알아보고, 3장에서 시스템의 구성에 대하여 알아본다. 4장에서 실제 시스템의 알고리즘을 각 과정별로 보이고, 5장에서 중의성을 가진 단어 하나에 대하여 여러 가지 방법에 의한 실험을 하고, 그 결과로 얻어진 최종 모델을 새로운 예제에 적용한 결과를 보인다. 6장에서 결론을 맺고 앞으로의 연구방향에 대하여 제시한다.

## 2. 관련 연구

최근의 품사 태깅 방법은 통계 모델과 통계와 규칙을 같이 사용하는 혼합 모델이 많이 사용된다.

통계 모델은 말뭉치에 기반한 방법으로 말뭉치로부터 빈도수에 관련된 정보를 자동으로 추출하여 사용한다. 기본 모델은 길이가 N인 주어진 문장 $W_{1,N}$에 대하여 말뭉치로부터 추출된 확률 정보를 이용하여 다음 식을 만족하는 최상의 태그열 $T_{1,N}$을 찾는다.

$$P(T_{1,N}|W_{1,N}) = \frac{P(W_{1,N}|T_{1,N}) \times P(T_{1,N})}{P(W_{1,N})} \quad (2)$$

$$= P(W_{1,N}|T_{1,N}) \times P(T_{1,N}) \quad (3)$$

$$= \Pi_{i=1,N} P(W_i|T_i) \times P(T_i|T_{i-1}) \quad (4)$$

(2)식은 베이지언 정리(Bayes' theorem)를 이용해 나타낸 식으로 분모의 값은 일정한 값을 가지므로 제거된 후 (3)식으로 나타낼 수 있다. 여기서 데이터 부족(data sparseness)을 완화시키기 위해 두 가지 가정 즉, 어휘간 독립가정과 품사간 전이에 관해 "현재 품사의 발생은 바로 이전 품사에만 의존한다"는 마르코프 가정을 도입한다. (3)식에 이와 같은 가정이 적용된 모델이 (4)에 주어진 식이다. 일반적으로 통계 모델은 (4)식을 기반으로 변형된 식을 사용한다.

(4)식은 동형 이품사 문제만을 가지는 영어에 적용되는 식으로, 이형 동품사 문제와 동형 이품사 문제, 그리고 서로 다른 위치에서 다른 개수의 형태소로 분리되는 비동기 문제를 가지는 한국어에 그대로 적용하기에는 문제가 있다. 이에 대해 [1]은 영어에 적용되어진 은닉 마르코프 모델을 한국어에 사용할 수 있도록 변형하고 공유 단어열과 가상 단어에 의하여 중복된 연산을 줄인 모델을 사용하여 89.13%의 정확률을 보였다. 이 모델은 비교사 학습 방식이지만 형태소 열로 만들어진 말뭉치를 사용하기 때문에 원시 말뭉치를 그대로 사용하지 못하고 형태소 열로 가공해야하는 문제가 있고 고려되는 문맥도 단순히 이전 품사만을 사용한다. 이러한 [1]의 개념을 일반화한 모델로 [3]은 가중치 망 모델을 제안하였는데 주어진 문장에 대한 형태소 해석 결과를 입력으로 받아 격자 구조로 표현하고, 각 연결선(edge)과 정점(node)에 품사 중의성을 해소하기 위한 적절한 가중치(weight)를 주었다. 품사 태깅은 최적의 경로를 찾으면 되고 실험 결과 95.85%의 높은 정확률을 보였다. 가중치 망을 사용할 경우 인접 어휘 문맥을 쉽게 사용할 수 있지만 확장 문맥을 고려하지는 못하였다.

위와 같은 영어 모델을 한국어에 적용하는 문제 이외에 통계 모델의 문제점은 첫째, 학습 말뭉치에 존재하는 많은 정보 중 일부만을 사용하여 많은 유용한 정보가 상실된다는 문제와 둘째, 학습 말뭉치에는 없고 실제 적용하는 말뭉치에는 있기 때문에 발생하는 데이터 부족 문제가 있다.

첫 번째 문제인 정보 상실 문제를 보완하기 위해 [2]는 어절 단위로 문장을 이루는 한국어의 특성을 고려하여 어절내 품사 전이 확률과 구분되는 어절간 품사 전이 확률을 (4)식에 추가하여 우수한 결과(94.4% 정확률)를 얻었다. 하지만 어휘 수준의 문맥 정보를 고려하기 어렵다

는 문제점을 가지고 있다. 다른 방법으로 [4]는 품사 분류에 유용한 어휘 정보, 연어 정보, 통사 정보를 말뭉치에서 뽑아 정형화하여 기존의 품사열 정보와 결합시킨 최대 엔트로피 모델을 구성하였다. 이 모델은 뛰어난 성능(97.43% 정확률)을 보이지만 바로 인접한 정보만 고려한다는 문제점을 아직 가지고 있다.

[1][2][3][4] 등 지금까지 연구된 방법들은 모두 약간의 차이는 있지만 근본적으로 인접 문맥만을 고려하였기 때문에 언어의 특성인 장거리 의존 관계를 제대로 반영하지 못하였다.

두 번째 문제는 기존의 연구들은 대부분 품사 부착 말뭉치를 사용해 학습하는 교사 학습 방식을 사용하였으므로 활용할 수 있는 데이터 양이 극히 제한되어 심한 데이터 부족 문제를 안고 있다. 현재 이들 연구에서 사용한 품사 부착 말뭉치의 양은 대부분 20만 어절 이하로서 품사의 할당에 중요한 역할을 할 수 있는 어휘 정보를 이용하는 데에는 한계가 있다.

지금까지 알아본 태깅과 관련된 연구 외에 본 논문과 관련된 연구로는 비교사 학습을 이용한 의미 중의성 해소 방법이 있다[9][12][13][14][15].

[12]는 비교사 학습 방식에 의해 단어 의미 중의성을 해소한 것으로 두 가지 가정을 기반으로 한다. 첫 번째 가정은 "One sense per collocation"으로 인접한 연어에 의하여 단어의 의미가 하나로 결정될 수 있다는 것이고, 두 번째 가정은 "One sense per discourse"으로 하나의 담화 내에 있는 동일한 단어들은 모두 같은 의미로 쓰인다는 것이다. 실험은 두 가지 중 첫 번째 가정을 주요 수단으로 사용하고 두 번째 가정은 선택적, 보조적 수단으로 사용하였다. Yarowsky는 실험에서 'plant'를 A(식물), B(공장) 두 가지 의미로 태깅하였다. 그 과정은 처음에 종자 연어(seed collocation)로 'life'와 'manufacturing'을 주고 Decision List(DL[1])를 만들어 예문들을 A, B로 태깅한다. 그리고 태깅된 결과들을 사용한 교사 학습으로 DL를 새로 갱신한 후 예문들을 다시 태깅하는 과정을 학습 파라미터들이 일정해질 때까지 반복한다. 실험 결과, 첫 번째 가정만을 사용할 경우 95.5%의 성능을 보여 교사 학습 방식의 성능(96.1%)보다는 못하지만 좋은 결과를 얻었다. 그리고 두 번째 가정을 추가하여 오류를 정정할 경우 96.5%의 성능을 보여 교사학습 방식보다 우수한 결과를 얻었다.

[9]는 [12]의 실험이 대상 단어의 의미 구분을 두 가지로 제한하여 문제의 난이도를 너무 낮췄기 때문에 가능하였고 방법론의 일반성 문제를 상실하였다고 지적하면서 지식원으로 국소 문맥과 국소 문맥을 보완하기 위하여 사전으로부터 얻은 공기 정보를 사용한 비교사 학습 방식의 명사 의미 중의성 해소 방법을 제안하였다.

[14]와 [15]는 비교사 방식의 Co-Training 방법을 제안하였는데 두 개의 독립적인 집합으로 Feature를 나눌 수 있는 데이터집합에 적용한다. 이때 두 개의 Feature들은 각각 독립적으로 사용되어도 데이터집합을 분류할 수 있어야 한다. 그 방법은 분류기1(Feature1에 대한)에 의해 분류된 결과로 분류기2(Feature2에 대한)를 학습하고 또한 그 결과를 분류기1에 주어서 학습하는 과정을 주어진 횟수만큼 반복한다. 이러한 Co-training에 의해 분류하는 것이 두개의 Feature를 동시에 사용하여 분류하는 것보다 우수한 성능을 보였다.

본 논문에서는 확장된 문맥을 고려하고 비교사 학습 방법으로 원시 말뭉치로부터 얻어진 지식을 이용하여 단어의 형태적 중의성을 해결하는 방법을 제안한다. 일반적으로 태깅에 있어서 원시 말뭉치의 이용은 품사 태깅에 대한 선험 지식 부족으로 정확성이 떨어지고 많은 양의 말뭉치에 대한 학습이 어렵다는 단점을 가지고 있다. 본 논문에서 제시하는 방법은 비교사 학습 품사태깅 모델이라기보다는 품사 부착 말뭉치로부터는 충분한 지식을 얻을 수 없어 해결하지 못하는 동사-동사 중의성과 같은 국소 문제를 해결하기 위한 것으로, 원시 말뭉치를 태깅에 어떻게 이용할 수 있는가에 대한 새로운 방법론을 제시한다. 즉, 본 연구는 원시 말뭉치를 이용함으로써 정보 부착 말뭉치에서 줄 수 없는 지식의 자동 학습 가능성과 품사 태깅에 교사 학습과 비교사 학습을 결합함으로써 보다 나은 결과를 가져올 수 있음을 보여준다.

## 3. 시스템 구성

### 3.1 국소 문맥

본 논문에서 고려하는 국소 문맥은 연어만을 사용하고 다른 부가적인 정보는 모두 제외한다. 다른 연구에서는 고려하지 못한 장거리 의존 관계를 이용하기 위해 연어는 바로 인접한 연어 외에 2이상의 거리를 가진 연어 또한 고려하여 확장 문맥을 반영한다. 연어의 대상단어와의 거리는 너무 원거리까지 고려하면 잡음 데이터가 많아져 도리어 성능을 저하시킬 우려가 있어서 이번 실험에서는 거리를 3으로 제한한다.

1) Decision List의 줄임말. 앞으로는 DL로 표기한다.

### 3.2 Decision List

본 시스템에서 중의성 제거에 사용하는 DL는 표1과 같은 형식을 가진다.

**표 1 Decision List의 예**

| | |
|---|---|
| 1 말을 듣고 5.345 | 3 소리/0 듣고 4.126 |
| 3 소리를 듣고 4.971 | 1 소리/0 듣고 6.238 |
| : | : |
| : | : |

표1의 왼쪽에 있는 것은 연어의 어절 형태에 해당하는 예이고 오른쪽은 형태소 형태에 해당하는 것인데 형태소 형태는 형태소와 품사를 같이 나타낸다. 각 예의 첫 번째 인자는 대상 단어와의 거리로서 바로 앞에 있는 연어인 경우 '1'을 가지고 2이상의 거리에 있는 연어는 '3'을 가진다. 두 번째와 세 번째 인자는 연어와 대상단어를 가리킨다. 네 번째는 연어의 함수값을 가리키며 클수록 대상단어와의 관련도가 높다.

### 3.3 Log-likelihood Ratio

DL를 만들 때 사용하는 함수값은 다음 식 (5)에 의해 구한다.

$$LogL = Log(\frac{P(Tag_A|Collocation_i)}{P(Tag_B|Collocation_i)}) \qquad (5)^{2)}$$

(5)식은 동일한 연어에 대해 대상단어를 A 또는 B로 태깅할 상대적인 확률값에 의하여 구하여진다. A로 태깅할 확률이 B로 태깅할 확률보다 높으면 결과값은 양의 값을 가지고 반대면 음의 값을 가진다. (5)식이 양이면 $Collocation_i$를 A와 관련된 연어로 결정하고 음이면 B와 관련된 연어로 결정한다. 해당 연어의 함수값은 $|LogL|$를 사용한다.

### 3.4 구성도

전체적인 시스템의 구성은 그림1과 같다.

그림1에서 $DL_1$부터 $DL_N$까지는 형태적 중의성을 가진 N개의 대상 단어들 각각에 대한 것이다. 또한 각 대상 단어에 대한 DL의 종류도 3가지인데 어절형태, 형태소형태 그리고 품사에 대한 것으로 분류된다.

그림2는 형태적 중의성 제거 모델을 세부적으로 나타낸다.

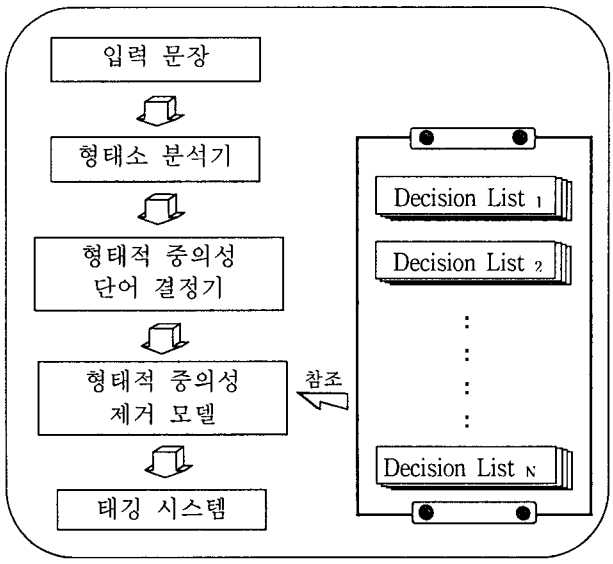


**그림1 시스템의 전체적인 구성**

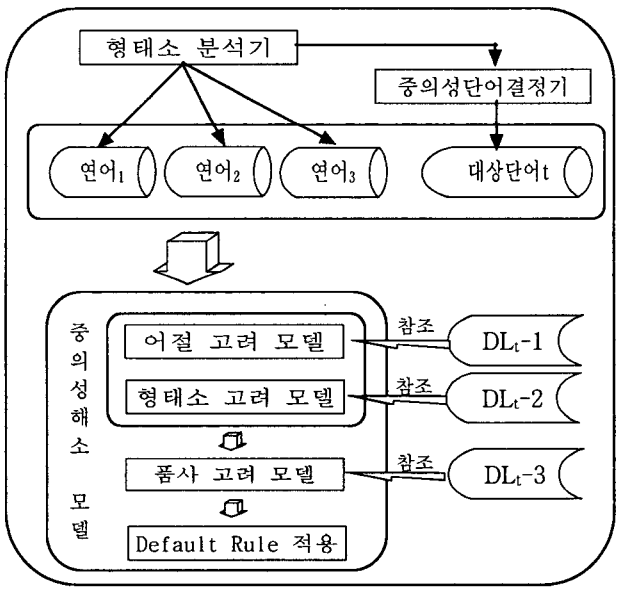


**그림2 형태적 중의성 제거 모델**

그림2에서 입력문장의 어절들 중 대상단어는 중의성 단어 결정기에 의해 선택되고 연어는 대상단어 앞 3이내의 거리에 있는 어절들이 고려된다. 연어들과 대상단어가 결정되면 대상단어와 관련된 DL들(DL-1,2,3)을 참조하여 세 단계의 과정을 거쳐 중의성을 제거한다.

첫 번째 단계는 어절과 형태소를 같이 고려하여 중의성을 제거하고 실패할 경우 두 번째 단계인 품사 고려 모델에서 대상 단어의 바로 앞 어절의 마지막 품사에 의해 중의성을 제거한다. 두 번째 단계에 의해서도 실패할 경

2) $Tag_A$ : '듣다'로 태깅('들었다'의 예에서)
$Tag_B$ : '들다'로 태깅('들었다'의 예에서)
$Collocation_i$ : 대상단어의 거리3 이내에 나오는 연어

우 Default Rule에 의해 무조건 하나의 형태로 태깅한다. 중의성이 제거된 결과는 태깅 시스템의 입력으로 보내진다.

다음으로 중의성 제거 모델의 알고리즘에 대하여 각 과정별로 설명한다.

## 4. 알고리즘

초기 DL를 만들어 학습 말뭉치에 적용하고 적용된 결과를 비교사 학습 방식에 의해 재적용하는 과정을 STEP별로 보인다. 각 STEP은 대상 단어 '들었다'에 실제 적용하는 예를 들어 설명한다.

STEP 1: 문장 추출
- 학습 말뭉치에서 '듣고'가 나오는 문장을 뽑아냄 (File 1) ← 종자 연어
- 학습 말뭉치에서 '들고'가 나오는 문장을 뽑아냄 (File 2) ← 종자 연어
- 학습 말뭉치에서 '들었다'가 나오는 문장을 뽑아냄 (File 3) ← 대상 단어

STEP 2: 형태소 분석
- STEP 1에서 뽑은 File 1과 File 2를 MORANY[3]를 사용하여 형태소 분석 (MOR 1, MOR 2)
- STEP 1에서 뽑은 File 3를 형태소 분석(MOR 3)

STEP 3: 빈도수 계산

3.1 어절 단위 빈도수 계산
- STEP 1의 File 1과 File 2를 사용하여 '듣고'와 '들고'의 거리3 이내에 있는 어절들의 빈도수를 계산

3.2 형태소 단위 빈도수 계산[4]
- STEP 2의 MOR 1과 MOR 2를 사용하여 거리3 이내에 있는 형태소들의 빈도수를 계산

STEP 4: Decision List 생성
- STEP 3.1에서 구해진 빈도수를 사용하여 $DL_t$-1을 생성(decision.eojeol)
- STEP 3.2에서 구해진 빈도수를 사용하여 $DL_t$-2를 생성(decision.morph)

STEP 5: 중의성 제거(태깅)
- STEP 4에서 만들어진 DL들을 STEP 2에서 구한 MOR 3에 있는 문장들의 대상단어에 적용

STEP 6: 학습(Learning)

6.1 빈도수 추가
- STEP 5에서 태깅된 결과들을 가지고 STEP 3의 빈도수를 재계산

6.2 반복(Iteration)
- STEP 4, STEP 5, STEP 6.1을 태깅되는 대상단어가 더 이상 없을 때까지 반복

STEP 7: 새로운 문장에 적용
- STEP 6에서 최종적으로 구해진 DL들을 새로운 실험 말뭉치에 적용

STEP 8: 품사 고려 모델 적용
- STEP 7에서 실패할 경우 대상 단어의 바로 앞 어절의 마지막 형태소 품사에 의하여 대상단어를 태깅 ($DL_t$-3을 사용)

STEP 9: Default Rule 적용
- STEP 8에서도 실패할 경우는 빈도수가 높은 형태로 무조건 태깅

## 5. 실험 및 평가

실험은 첫 번째, 원시 말뭉치로부터 얻을 수 있는 정보들의 유용성을 알아보기 위해 '들었다'의 형태적 중의성을 제거하는 실험을 다양한 방법으로 수행하였다. 두 번째는 첫 번째에서 실험한 방법들 중에서 성능이 좋은 방법들을 조합하여 최종 모델을 만들고 형태적 중의성을 가진 12개의 단어들에 적용하여 기존 모델의 결과와 비교하여 성능을 평가하였다. 사용되는 말뭉치는 국어 정보 베이스 II CD-ROM에 있는 1000만 어절의 원시 말뭉치 중 500만 어절을 학습 말뭉치 및 실험 말뭉치로 사용하였다. 실험 결과는 재현률, 정확률, F-measure 세 가지로 나타내며 F-measure에서는 재현률과 정확률의 중요도를 동등하게 보았다.

### 5.1 다양한 정보들에 대한 실험('들었다'에 적용)

이번 실험에서 비교 평가한 정보들은 다음과 같다.

첫째, 연어의 형태로 어절 형태, 형태소 형태, 그리고 둘 모두를 고려한 세 가지 방법에 대하여 실험

둘째, 형태소 형태의 연어 고려시 중요하다고 생각되는 품사 유형으로 3품사(명사,동사,형용사), 4품사(명사,대명사,동사,형용사), 5품사(명사,대명사,고유명사,동사,형용사)에 대하여 실험

셋째, DL를 적용할 때 전체를 한꺼번에 적용하는 방법과

---

3) 연세대 한글정보처리연구실에서 만든 형태소 분석기

4) 제한사항1:품사가 명사,동사,형용사 중 하나인 형태소만 고려
제한사항2:단일 형태소분석결과를 가지는 어절만 고려(N개의 분석결과를 가지는 어절은 무시)

함수값이 높은 N개만을 적용하는 방법을 비교

넷째, DL 생성 시 임계값을 적용한 방법과 적용하지 않은 방법을 비교하고 가장 우수한 결과를 내는 임계값을 실험을 통해 결정

다섯째, [14]에서 제안한 Co-training 방법을 적용하여 적용하지 않은 방법과 비교

여섯째, 기존의 통계적 방법과의 비교를 위해 Bayes Rule에 독립가정을 사용한 (8)식을 적용하여 실험

$$P(Tag_A|C_1, C_2, C_3)^{5)}$$

$$= \frac{P(C_1, C_2, C_3|Tag_A) \times P(Tag_A)}{P(C_1, C_2, C_3)} \quad (6)$$

$$= P(C_1, C_2, C_3|Tag_A) \times P(Tag_A) \quad (7)$$

$$= P(C_1|Tag_A)P(C_2|Tag_A)P(C_3|Tag_A) \times P(Tag_A) \quad (8)$$

일곱째, 은닉 마르코프 모델의 기본모델과 어휘 정보를 추가한 수정모델을 비교하여 어휘 정보의 중요성을 실험

여덟째, 인접한 연어일수록 높은 가중치를 주는 거리별 가중치 실험.

위의 여덟가지 정보들에 대하여 실험하고 비교한 결과는 표2에서 표9까지 나와있다.

그 외의 실험으로 '들었다'의 예에서 '듣다'와 '들다' 모두 타동사이므로 목적어가 매우 중요할 것으로 보고 목적어가 거리3 이내에 있을 경우 목적어를 최우선으로 고려한 모델을 만들어서 실험하였는데 성능이 1.7% 정도 떨어졌다. 그 이유는 '집어 들었다'에서 보듯이 목적어가 아닌 더욱 중요한 연어들을 무시하기 때문이다.

다른 실험으로 태깅할 때 DL 함수값이 가장 높은 연어 하나만을 보지 않고 거리3 이내에 있는 모든 연어들의 DL 함수값들의 합에 의해서 대상단어를 태깅하였는데 0.6% 정도의 성능 저하가 있었다. 잡음 데이터(Noise data)가 많이 들어가는 것이 성능저하의 원인이었다.

**표 2 연어 형태 실험 결과**

| 연어 형태 | 재현율(%) | 정확률(%) | F-measure(%) |
|---|---|---|---|
| 어절 | 87.25 | 92.41 | 89.76 |
| 형태소 | 84.03 | 91.56 | 87.63 |
| **어절+형태소** | **91.58** | **94.33** | **92.93** |

5) $Tag_A$: '들었다'의 경우, '듣다'로 태깅될 경우
$C_1, C_2, C_3$: 대상단어('들었다')의 앞에 나오는 연어들

**표 3 형태소 품사 실험 결과**

| 품 사 | 재현율(%) | 정확률(%) | F-measure(%) |
|---|---|---|---|
| **3 품사** | **84.75** | **93.05** | **88.71** |
| 4 품사 | 84.75 | 92.96 | 88.67 |
| 5 품사 | 84.03 | 91.56 | 87.63 |

**표 4 DL 적용 방법 실험 결과**

| 적용 방법 | 재현율(%) | 정확률(%) | F-measure(%) |
|---|---|---|---|
| **전체** | **91.58** | **94.33** | **92.93** |
| N best | 79.83 | 97.96 | 87.97 |

**표 5 임계값 적용**

| 임계값 | 재현율(%) | 정확률(%) | F-measure(%) |
|---|---|---|---|
| 0.3 | 84.58 | 92.27 | 88.26 |
| **0.5** | **84.50** | **92.69** | **88.40** |
| 1.0 | 83.83 | 92.72 | 88.05 |
| 2.0 | 79.83 | 93.55 | 86.15 |
| × | 84.03 | 91.56 | 87.63 |

**표 6 Co-training 적용**

| 적용 유무 | 재현율(%) | 정확률(%) | F-measure(%) |
|---|---|---|---|
| **No** | **91.58** | **94.33** | **92.93** |
| Yes | 89.17 | 94.19 | 91.61 |

**표 7 Bayes Rule을 적용**

| | 재현율(%) | 정확률(%) | F-measure(%) |
|---|---|---|---|
| **제안된 모델** | **94.25** | **94.25** | **94.25** |
| Bayes Rule | 86.33 | 92.58 | 89.35 |

**표 8 은닉 마르코프 모델 적용(bigram)**

| | 재현율(%) | 정확률(%) | F-measure(%) |
|---|---|---|---|
| 기본 모델 | 69.58 | 75.91 | 72.61 |
| **수정 모델** | **69.17** | **77.50** | **73.10** |

**표 9 거리별 가중치 적용**

| 가중치 | | | 재현율(%) | 정확률(%) | F-measure(%) |
|---|---|---|---|---|---|
| 거리1 | 거리2 | 거리3 | | | |
| 1.5 | 1.0 | 0.7 | 92.00 | 94.76 | 93.36 |
| 2.0 | 1.0 | 0.7 | 91.83 | 94.59 | 93.19 |
| 1.7 | 1.0 | 0.7 | 92.00 | 94.76 | 93.36 |
| **1.5** | **1.0** | **0.5** | **92.17** | **94.94** | **93.53** |
| 1.5 | 1.0 | 0.3 | 92.17 | 94.94 | 93.53 |
| 1.0 | 1.0 | 1.0 | 91.58 | 94.33 | 92.93 |

앞에서 실험한 결과들을 토대로 5.2절에서 여러 대상단어들에 적용할 최종 모델의 요소들은 다음과 같다.

- **연어 형태 : 어절형태 + 형태소형태**
- **DL 적용방법 : 전체 방법**
- **형태소 품사 : 3품사(명사, 동사, 형용사)**
- **DL 임계값 : 0.5**
- **거리별 가중치 : 1.5, 1.0, 0.5 (거리1,2,3)**

### 5.2 새로운 대상단어들(12개)에 대한 실험

5.1절에서 만들어진 최종 모델을 새로운 대상단어들에 적용한 결과는 표10과 같다.

실험에서 결과는 R(재현률), P(정확률), F(F-measure)로 나타낸다. 모델종류에서 Baseline은 학습을 하지 않고 처음에 주어지는 종자 연어들(Seed Collocation)만으로 태깅하는 모델로써 비교사 학습방식에 의한 성능 향상을 비교하기 위해 사용된다.

표10의 실험 결과에서 90%이하의 성능을 보이는 대상단어들 대부분은 대상단어를 가진 문장들의 수가 100개 이하로 학습이 제대로 이루어질 수 없었기 때문에 성능이 저조하다. 전체 평균은 약간 낮지만 몇 개의 성능이 저조한 것들을 제외하면 훌륭한 결과를 얻을 수 있었다.

### 6. 결론 및 향후 연구 방향

최근 연구 방향을 고려하고 태깅 방법에 대한 기존 연구들의 특정 문제점을 보완하기 위해, 본 논문에서는 확장된 문맥을 고려하여 형태적 중의성을 제거하는 모델을 제안하고 실험하였다. 또한 품사 부착 말뭉치로부터는 해결하지 못한 동사-동사 중의성과 같은 국소 문제를 해결하기 위해 원시 말뭉치를 이용하는 방법을 제시하였다.

실험에서는 원시 말뭉치와 형태소 분석된 결과로부터 얻을 수 있는 여러 가지 정보들에 대하여 다양한 실험을 거쳐 가장 우수한 결과를 얻을 수 있는 최종 모델을 만든 후 실제 대상단어들에 대하여 실험하였다. 12개의 대상단어들에 대한 실험에서 약간 저조한 결과(90.61%)를 얻었지만 학습이 제대로 되지 못한 몇 개의 대상 단어들을 제외한다면 우수한 결과를 얻을 수 있었다. 여기서 성능 저하의 원인이 된 학습 문제는 제안된 모델이 비교사 방식이므로 학습 말뭉치로 쓰인 원시 말뭉치의 양을 늘린다면 해결될 수 있기 때문에 그다지 큰 문제가 되지 않는다. 기존의 태깅 방법과 비교하기 위하여 '들었다' 단어 하나를 기존의 방법으로 실험하여 비교하였는데 기존 모델의 문제점을 보완할 수 있다는 결론을 얻었다. 또한 확장된 문맥으로부터 얻은 어휘 정보가 여러 언어 현상을 해결하는데 결정적 역할을 한다는 것을 실험을 통해 알 수 있었고 정보 부착 말뭉치에서 줄 수 없는 지식을 원시 말뭉치로부터 자동 학습할 수 있다는 가능성을 보여주었다.

본 논문에서 제안한 모델은 독립적으로 사용될 수도 있지만 그러기보다는 다른 시스템의 보조적인 수단으로 사용될 수 있는 모델이다. 이번 실험에서도 MORANY와 태깅 시스템 사이에 첨가하여 태깅 시스템의 전처리기로서 태깅 성능을 향상시켰다.

앞으로의 연구방향은 비교사 방식으로 인하여 발생하는 잡음 데이터를 보정할 수 있는 방법을 모색하고, 중의성이 제거된 대상단어를 통해 인접 단어의 형태소 분석 결과의 중의성을 제거할 수 있는 방법을 연구할 계획이다. 미등록어의 처리 문제도 형태소 분석 결과의 중의성 제거 방법과 관련해서 연구될 수 있다.

**표 10 대상단어별 실험 결과**

| 대상 단어 | 모델 종류 | R(%) | P(%) | F(%) |
|---|---|---|---|---|
| 들었다 (듣다, 들다) | New | 94.25 | | |
| | Baseline | 82.25 | 89.48 | 85.71 |
| 물었다 (묻다, 물다) | New | 92.23 | | |
| | Baseline | 66.10 | 91.15 | 76.63 |
| 걸었다 (걷다, 걸다) | New | 84.99 | | |
| | Baseline | 61.82 | 88.21 | 72.70 |
| 길었다 (긷다, 길다) | New | 96.43 | | |
| | Baseline | 76.79 | 97.73 | 86.00 |
| 끄는 (끌다, 끄다) | New | 96.64 | | |
| | Baseline | 82.55 | 98.40 | 89.78 |
| 까는 (깔다, 까다) | New | 80.95 | | |
| | Baseline | 42.86 | 90.00 | 58.06 |
| 쓰는 (쓸다, 쓰다) | New | 94.97 | | |
| | Baseline | 76.06 | 97.42 | 85.43 |
| 파는 (팔다, 파다) | New | 83.72 | | |
| | Baseline | 45.85 | 84.66 | 59.48 |
| 나는 (날다, 나다) | New | 72.73 | | |
| | Baseline | 45.45 | 75.00 | 56.60 |
| 사는 (살다, 사다) | New | 94.64 | | |
| | Baseline | 72.55 | 95.55 | 82.47 |
| 주는 (줄다, 주다) | New | 99.55 | | |
| | Baseline | 86.19 | 99.48 | 92.36 |
| 이어 (잇다, 이다) | New | 96.15 | | |
| | Baseline | 43.85 | 94.98 | 60.00 |
| **평 균** | **New** | **90.61** | | |

### 7. 참고 문헌


[1] 김재훈, 임철수, 서정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사의 태깅", 한국 정보과학회 논문지 제22권 제1호, pp.136-146, 1995

[2] 김진동, 임희석, 임해창, "Twoply HMM:한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델", 정보과학회논문지(B) 제24권 제12호, pp.1502-1512, 1997

[3] 김재훈, "가중치 망을 이용한 한국어 품사태깅", 정보과학회논문지(B) 제25권 제6호, pp.951-959, 1998

[4] 강인호, 김재훈, 김길창, "최대 엔트로피 모델을 이용한 한국어 품사 태깅", 제10회 한글 및 한국어 정보처리 학술대회 발표집, pp.9-14, 1998
[5] 임희석, 김진동, 임해창, "어절 태그 변형 규칙을 이용한 한국어 품사 태거", 정보과학회논문지(B) 제24권 제6호, pp.673-684, 1997
[6] 신상현, 이근배, 이종혁, "통계와 규칙에 기반한 2단계 한국어 품사 태깅 시스템", 정보과학회논문지(B) 제24권 제2호, pp.160-169, 1997
[7] 임희석, 권철중, 이재원, 오기은, "영한 기계 번역을 위한 혼합형 N-best 품사 태거", 제10회 한글 및 한국어 정보처리 학술대회 발표집, pp.15-19, 1998
[8] 이상주, 류원호, 김진동, 임해창, "품사태깅을 위한 어휘문맥 의존규칙의 말뭉치기반 중의성주도 학습", 정보과학회논문지(B) 제25권 제1호, pp.178-189, 1999
[9]이승우, 이근배, "국소 문맥과 공기 정보를 이용한 비교사 학습 방식의 명사 의미 중의성 해소", 정보과학회 논문지: 소프트웨어 및 응용 제27권 제7호, pp.769-783, 2000
[10]황이규, 이현영, 이용석, "형태소 및 구문 모호성 축소를 위한 구문단위 형태소의 이용", 정보과학회 논문지: 소프트웨어 및 응용 제27권 제7호, pp.784-793, 2000
[11]이호, 백대호, 임해창, "분류 정보를 이용한 단어 의미 중의성 해결", 정보과학회 논문지(B) 제24권 제7호, pp.779-789, 1997
[12]Yarowsky, David. "Unsupervised word sense disambiguation rivaling supervised methods", in Proceedings of the 33rd Annual Meeting, pp.189-196, Cambridge, MA, June. Association for Computational Lisguistics, 1995
[13]Collins, Michael and Singer, Yoram. "Unsupervised Models for Named Entity Classification", in Proceedings of EMNLP '99 , 1999
[14]Blum, Avrim and Mitchell, Tom. "Combining Labeled and Unlabeled data with Co-training", in Proceedings of the 11th Annual Conference on Computational Learning Theory(COLT'98), pp.209-214, 1998
[15]Nigam, Kamal and Ghani, Rayid. "Understanding the Behavior of Co-training", in KDD-2000 Workshop on Text Mining, 2000
[16]Yarowsky, David. "Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French", in Proceedings of the 32nd Annual Meeting, pp.88-95, Las Cruces, NM. Association for Computational Linguistics, 1994
[17]Mitchell, Tom. "The Role of Unlabeled Data in Supervised Learning", in Proceedings of the Sixth International Colloquium on Cognitive Science, San Sebastian, Spain, 1999 (invited paper).
[18]Jeongmi, Cho., Jungyun, Seo. and Gilchang, Kim. "Verb sense disambiguation based on dual distributional similarity", in Natural Language Engineering 5(2), pp.157-170, 1999