은닉 마르코프 모델을 이용한 두단계 한국어 품사 태깅

이 상 주, 임 희 석, 임 해 창
고려대학교 전산과학과

Two-Level Part-of-Speech Tagging for Korean Text Using Hidden Markov Model

Sang-zoo Lee, Heui-suk Lim, Hae-chang Rim
Dept. of Computer Science, Korea University

요 약

품사 태깅은 코퍼스에 정확한 품사 정보를 첨가하는 작업이다. 많은 단어는 하나 이상의 품사를 갖는 중의성이 있으며, 품사 태깅은 지역적 문맥을 이용하여 품사 중의성을 해결한다. 한국어에서 품사 중의성은 다양한 원인에 의해서 발생한다. 일반적으로 동형 이품사 형태소에 의해 발생되는 품사 중의성은 문맥 확률과 어휘 확률에 의해 해결될 수 있지만, 이형 동품사 형태소에 의해 발생되는 품사 중의성은 상호 정보나 의미 정보가 있어야만 해결될 수 있다. 그러나, 기존의 한국어 품사 태깅 방법은 문맥 확률과 어휘 확률만을 이용하여 모든 품사 중의성을 해결하려 하였다. 본 논문은 어절 태깅 단계에서는 중의성을 최소화하고, 형태소 태깅 단계에서는 최소화된 중의성 중에서 하나를 결정하는 두단계 태깅 방법을 제시한다. 제안된 어절 태깅 방법은 단순화된 어절 태그를 이용하므로 품사 집합에 독립적이며, 대량의 어절을 소량의 의사 부류에 사상하므로 통계 정보의 양이 적다. 또한, 은닉 마르코프 모델을 이용하므로 태깅되지 않은 원시 코퍼스로부터 학습이 가능하며, 적은 수의 파라메터와 Viterbi 알고리즘을 이용하므로 태깅 속도가 효율적이다.

I. 서 론

품사(part-of-speech)는 문장 내에서 단어의 구문적, 의미적 역할을 반영하므로, 문장의 정확한 해석은 품사에 의존적이다 [13]. 많은 단어들은 하나 이상의 품사를 가지며, 이러한 동형 이품사 단어[1]는 문장의 해석 과정에서 품사 중의성(categorial ambiguity)을 발생시킨다. 그러므로, 문장의 정확한 해석을 위해서는 단어의 올바른 품사를 결정해야 하며, 이러한 과정을 품사 중의성 해결(categorial ambiguity resolution) 또는 품사 태깅(part-of-speech tagging)이라 한다.

동형 이품사어는 다른 단어와 함께 사용될 때 단지 하나의 품사로만 사용되므로, 그 단어가 사용된 문맥을 고려하면 올바른 품사를 결정할 수 있다. 예를 들면, 단어 'tag'는 명사와 동사가 될 수 있지만, 문장 "A tag is a part-of-speech label."에서는 부정관사 'A'와 3인칭 단수 be 동사 'is' 사이에 나타나므로 명사로만 사용된 것을 알 수 있다.

영어에서는 단어의 굴절형도 하나의 단어로 취급된다. 단어의 품사는 사전 탐색에 의해 결정되고, 굴절형의 품사는 간단한 패턴 매칭에 의해 결정된다. 일반적인 영어의 품사 태깅은 단어가 가질 수 있는 모든 품사를 알아내고, 그 단어가 사용된 지역적 문맥(local context)을 고려하여 중의성을 해소한다.

한편, 첨가이인 한국어의 품사 태깅은 굴절어인 영어와는 다른 특성을 갖는다. 영어에서 품사 중의성은 동형 이품사어에 의해서만 발생하지만, 한국어에서 품사 중의성은 다양한 원인에 의해서 발생한다. 일반적으로 동형 이품사 형태소에 의해 발생되는 품사 중의성은 문맥 확률(contextual probability)[2]과 어휘 확률(lexical probability)[3]에 의해 해결될 수 있지만, 이형 동품사[4] 형태소에 의해 발생되는 품사 중의성은 상호 정보(mutual information)[5]나 의미 정보(semantic information)가 있어야만 해결될 수 있다. 그러나, 기존의 한국어 품사 태깅 방법은 문맥 확률과 어휘 확률만을 이용하여 모든 품사 중의성을 해결하려 하였다.

본 논문은 한국어 품사 태깅을 품사 중의성을 최소화하는 어절 태깅 단계와 최소화된 품사 중의성 중에서 하나를 결정하는 형태소 태깅 단계로 분리하는 두단계 품사 태깅 방법을 제시하고, 형태소 품사 집합과 학습 코퍼스에 대해 독립성을 유지하면서 품사 중의성을 최소화하기 위해 어절 태그와 은닉 마르코프 모델을 이용한 어절 태깅 방법을 제안한다.

1. 'tag(명사,동사)'와 같이 다른 품사를 갖는 같은 형태의 단어를 뜻한다. 'mail(우편,갑옷)'과 같이 다른 의미를 갖는 같은 형태의 단어를 뜻하는 동형 이의어(homograph)와는 다르다.
2. 인접된 품사간의 의존 관계를 나타낸다.
3. 단어가 특정 품사로 사용되는 확률을 나타낸다.
4. 같은 품사를 갖는 다른 형태의 형태소를 의미한다.
5. 단어가 특정 단어와 함께 발생하는 확률을 나타낸다.

II. 관련 연구

2.1 한국어의 품사 중의성

한국어의 어절은 의미를 나타내는 실질 형태소와 문법적 관계를 나타내는 형식 형태소의 결합으로 구성된다. 한국어에서는 형태소에 대한 품사만 정의되어 있으므로, 어절이 어떤 품사의 형태소로 구성되었는지를 알아내기 위해서는 형태소 분석이 필수적이다. 예를 들면, 어절 '나는'은

- '나:명사∨[6]대명사 + 는:조사'
- '나:자동사∨타동사∨보조동사 + 는:어미'
- '날:자동사∨타동사 + 는:어미'

와 같이 7가지의 형태소 결합으로 분석된다.

영어에서 띄어쓰기 단위는 단어이지만 한국어에서 띄어쓰기 단위는 어절이다. 그러므로, 영어의 단어와 한국어의 어절이 대응되며, 영어에서 단어의 품사는 어절의 구문 범주(syntactic category)[7]와 대응된다. 그러나 어절의 구문 범주는 이형 동품사 형태소의 결합에 의한 품사 중의성[8]을 분별할 수 없다. 예를 들면, 어절 '나는'의 구문 범주는 '명사+조사', '대명사+조사', '자동사+어미', '타동사+어미', '보조동사+어미'가 된다. 그러나, '나:자동사+는:어미'와 '날:자동사+는:어미'는 구성 형태소가 다르므로, 구문 범주는 같더라도 품사 중의성을 여전히 갖게 된다.

한국어의 품사 중의성은 형태소 분석 결과에 따라 다양하게 발생된다[22]. 한국어의 품사 중의성을 원인별로 유형을 분류하면 다음과 같으며, 일반적으로 몇 개의 유형이 복합적으로 품사 중의성을 발생시킨다.

[유형1] 동형 이품사 형태소

유형1의 품사 중의성은 어절의 구성하는 형태소가 여러 개의 품사를 가짐으로써 발생된다. 예를 들면, 형태소 '나'가 대명사, 자동사, 타동사의 품사를 갖는 동형 이품사 형태소이고, 형태소 '고'가 조사, 어미의 품사를 갖는 동형 이품사 형태소이기 때문에, 어절 '나고'는

- '나:대명사 + 고:조사'
- '나:자동사∨타동사 + 고:어미'

와 같은 형태소 결합으로 분석된다. 유형1에 속하는 어절에는 '말은', '불은' 등이 있다.

[유형2] 다른 형태소 분리 위치

유형2의 품사 중의성은 어절의 형태소 분리가 다른 위치에서 발생함으로써 발생된다. 예를 들면, 어절 '먹일랑'은

- '먹:명사 + 일랑:조사'
- '먹이:명사 + ㄹ랑:조사'

와 같은 형태소 결합으로 분석된다. 유형2에 속하는 어절로서 '소라도', '나래도' 등이 있다.

[유형3] 다른 형태소 갯수

유형3의 품사 중의성은 분리된 형태소의 갯수가 달라서 발생된다. 예를 들면, 어절 '감기는'은

- '감기:명사 + 는:조사'
- '감기:자동사 + 는:어미'
- '감:타동사 + 기:명사형어미 + 는:조사'

와 같은 형태소 결합으로 분석된다. 유형3에는 어절 '잘', '가면' 등이 속한다.

[유형4] 원형 왜곡

유형4의 품사 중의성은 불규칙이나 축약, 탈락 등의 현상에 의해 형태소의 원형이 왜곡되어 발생된다. 예를 들면, 어절 '걸을'은

- '걸:타동사 + 을:어미'
- '걷:자동사∨타동사 + 을:어미'

와 같은 형태소 결합으로 분석된다. 유형4에는 어절 '나는', '가는' 등이 속한다.

2.2 기존 연구 고찰

2.2.1 영어권의 품사 태깅

영어 품사 태깅 방법은 크게 규칙 기반 접근 방법과 통계 기반 접근 방법으로 분류할 수 있다[4]. 규칙 기반 접근 방법(rule-based approach)은 문맥을 고려하기 위한 규칙을 기술하고 그 규칙을 기반으로 중의성을 해소한다. 규칙의 기술 방법으로는 문맥틀(context frame)[8][11], 품사 trigram[9], 제약 문법(Constraint grammar)[10], 유한 상태 기계(finite-state machine)[12], 수정틀(patch template)[2] 등이 제안되었다. 규칙 기반 접근 방법은 규칙에 나타난 현상은 잘 처리할 수 있는 반면, 일관된 규칙을 찾기가 어렵고 많은 규칙들을 제어하기가 쉽지 않아 일반적으로 견고하지 못하다.

통계 기반 접근 방법(statistical approach)은 대량의 코퍼스로부터 통계 정보를 추출하고, 그 통계 정보를 기반으로 주어진 문장에 대한 발생 확률이 가장 높은 품사 태그열을 선택함으로써 중의성을 해소한다. 통계 기반 모델으로는 문맥 확률(contextual probability)과 어휘 확률(lexical probability)을 적용한 연관 확률(collocational probability) 모델[3][5][7], 은닉 마르코프 모델(Hidden Markov Model, HMM)[4][6][13], 신경망 모델[1][14] 등이 제안되었다. 통계적인 방법은 태깅된 코퍼스(tagged corpus)로부터 확률을 추출하는 지도 학습(supervised learning)과 태깅되지 않은 원시 코퍼스(raw corpus)로부터 확률을 추정하는 자율 학습(self learning)으로 분류되며, 일반적으로 규칙에 의한 태깅 방법보다 좋은 결과를 보이고 있다.

2.2.2 한국어 품사 태깅

한국어 품사 태깅에 대한 연구는 영어권에 비해 비교적 최근에 시작되었으며, 주로 통계 기반 접근 방법이 주류를 이루고 있다.

한국어의 품사 태깅은 태깅 단위가 어절인지 형태소인지에 따라서 어절 태깅(word-phrase tagging) 또는 형태소 태깅(morpheme tagging)으로 구분된다. 어절 태깅 방법은 어절을 구성하는 형태소의 품사를 나타내는 어절 태그(word-phrase tag)를 이용하여 어절의 구문 범주를 결정하는 것이고, 형태소 태깅 방법은 형태소 품사를 나타내는 형태소 태그(morpheme tag)를 이용하여 어절의 구성 형태소와 각 형태소의 품사를 결정하는 것이다.

현재까지 연구된 어절 태깅 방법에는 변형된 은닉 마르코

---

6. ∨는 '또는(or)'를 의미한다.

7. 어절이 어떤 품사의 형태소로 구성되었는지를 나타내며, 어절의 구문적 역할을 반영한다.

8. 이하 '이형 동품사 중의성'이라 표현한다.

프 모델을 이용한 방법[20], 신경망을 이용한 방법[19], Bigram 확률 모델과 묶인말 사전을 이용한 방법[18] 등이 있고, 형태소 태깅 방법에는 퍼지망을 이용한 방법[17], 다중 관측열에 기반한 은닉 마르코프 모델을 이용한 방법[21] 등이 있다.

2.3 한국어 품사 태깅의 고려 사항

일반적으로 한국어 품사 태깅에서 고려해야 할 사항에는 태깅 단위, 품사 세분화, 학습 방법 등이 있다.

2.3.1 태깅 단위

한국어 태깅 모델은 다음과 같은 측면을 고려하여 태깅 단위를 선택해야 한다.

첫째, 중의성의 표현 능력을 고려해야 한다. 어절 태그는 이형 동품사 중의성을 표현하지 못하지만, 형태소 태그는 표현 가능하다.

둘째, 하나의 어절이 여러 가지의 형태소 열에 대응되는 형태소 분리 문제를 고려해야 한다. 어절 태깅은 형태소를 분리할 필요가 없으므로 이러한 문제가 발생하지 않지만, 형태소 태그는 [21]에서와 같이 적절한 처리 방법이 요구된다.

셋째, 한국어는 어절간의 의존 관계가 중요하고 어절 사이에 부분적으로 자유로운 어순을 가지므로, 이러한 특성을 처리하기 위해 모델은 넓은 범위의 문맥을 고려할 수 있어야 한다. 어절 태그는 n-gram을 이용하여 문맥을 쉽게 확장할 수 있지만, 형태소 태그는 그렇지 못하다.

넷째, 문맥 확률의 크기를 고려해야 한다. 어절 태그를 이용하는 경우에는 새로운 어절 유형이 발생하므로 모든 어절 태그를 결정하기가 어렵고, 태그 갯수가 많아지며, 이로 인해 n-gram 문맥 확률의 양도 커지게 된다. 그러나, 형태소 태그를 이용하는 경우에는 모든 형태소 태그를 결정할 수 있고, 태그 갯수도 적으므로, 문맥 확률의 유지가 쉽고 저장 공간도 적게 필요하다.

다섯째, 어휘 확률의 크기를 고려해야 한다. 어절 태깅은 신뢰도 있는 어휘 확률의 추출과 저장 공간 문제로 인하여 어절의 어휘 확률을 직접 이용하는 것은 현실적으로 불가능하다. 이러한 문제를 해결하기 위해서 일반적으로 어절의 어휘 확률을 형태소 빈도로부터 계산하는 방법[20]이 이용되고 있다. 그러나, 형태소 태그를 이용하는 경우에는 어절 갯수보다 형태소 갯수가 비교적 적으므로 코퍼스로부터 추출된 어휘 확률을 직접 이용할 수 있다.

2.3.2 품사 세분화

품사가 세분되지 않으면 서로 다른 문맥에서 사용되는 형태소들이 하나의 품사로 취급되어 품사 사이의 분별력이 떨어지므로 중의성 해결을 위한 문맥 정보의 신뢰도가 저하되지만, 적은 수의 품사를 사용하므로 저장 공간이나 태깅 속도 면에서 효율적이다. 반면에 품사가 세분되면 각 품사가 사용되는 문맥이 쉽게 구별되므로 중의성 해결 정보의 신뢰도가 향상되지만, 유사하게 사용되는 형태소들이 서로 다른 품사를 갖게 될 수도 있고 더 많은 형태소들이 동형 이품사가 되어 오히려 태깅 성능이 저하되기도 한다. 그러므로, 품사 세분화는 가능한 한 품사 사이의 분별력이 높아지고, 단어의 중의성을 증가시키지 않는 범위 안에서 이루어져야 한다.

2.3.3 학습 방법

태깅 모델의 학습 방법에는 지도 학습과 자율 학습이 있다. 지도 학습(supervised learning)은 태깅된 코퍼스로부터 문맥 확률과 어휘 확률을 위한 빈도수를 추출하는 것이고, 자율 학습(self learning)은 태깅되지 않은 코퍼스로부터 확률 정보를 추정하는 것이다. 지도 학습은 자율 학습보다 비교적 높은 정확도를 보이지만, 신뢰도 있는 확률 정보를 추출하기 위해 대량의 태깅된 코퍼스가 필요하고 품사 집합이 변하면 코퍼스의 태깅 정보도 수정되어야 하는 단점이 있다.

현재 한국어에 대해서는 품사 집합에 대한 체계적인 연구가 미흡해서 품사 집합이 매우 불안정하며, 대량의 태깅된 코퍼스도 구축되어 있지 않은 상태이다. 그러므로, 품사 집합에 독립적이고 태깅된 코퍼스를 필요로 하지 않는 자율 학습이 이러한 상황에 잘 적응할 수 있을 것이다.

III. 두단계 한국어 품사 태깅

3.1 기본 개념

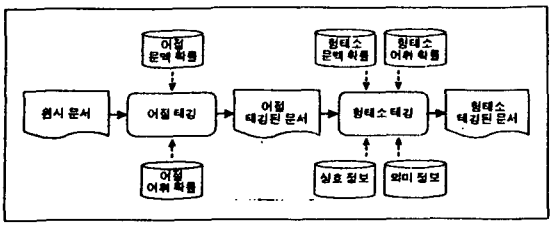
일반적으로 구문 범주가 표현할 수 있는 품사 중의성은 2장에서 분류한 품사 중의성 유형1과 유형3에 의해서 주로 발생되는 것으로, 문맥 확률과 어휘 확률을 적용한 확률 모델로 해결할 수 있지만, 구문 범주가 표현할 수 없는 이형 동품사 중의성은 품사 중의성 유형2와 유형3에 의해서 주로 발생되는 것으로, 문맥 확률이나 어휘 확률에 의해서 해결할 수 없다.

예를 들면, 어절 '나는'은 문장 '하늘을 나는 비행기는...'에서는 '날:타동사+는:어미'로 태깅되어야 하고, 문장 '시골에서 여름을 나는 사람들은...'에서는 '나:타동사+는:어미'로 태깅되어야 한다. 두 경우 모두 같은 문맥을 가지므로, 어휘 확률에 의해 둘 중 하나를 선택할 것이다. 만약 태깅된 코퍼스로부터 추출된 '날:타동사'의 어휘 확률이 '나:타동사'의 어휘 확률보다 높다면, 두 문장에서 모두 '날:타동사'가 선택되어 '날:타동사+는:어미'로만 태깅될 것이다. 즉, 이형 동품사 중의성을 갖는 어절은 태깅된 코퍼스로부터 추출된 어휘 확률에 종속적으로 태깅되므로 어휘 확률만으로는 중의성 해결이 불가능하다.

그러나, 이형 동품사 중의성은 그 형태소의 의미 정보를 이용하거나 주변 형태소와의 상호 정보를 이용할 수 있으면 해결할 수 있다. 형태소 '하늘' 또는 '비행기'와 형태소 '날:타동사'는 서로 자주 함께 발생하므로 그들간의 상호 정보가 형태소 '나:타동사'와의 상호 정보보다 높을 것이다. 또는, '하늘'과 '비행기', '날:타동사' 등이 같은 의미 범주(semantic category)를 갖도록 함으로써 형태소 '나:타동사'와 구별할 수 있을 것이다.

두단계 품사 태깅의 기본 개념은 "형태소 분석 정보와 문맥 및 어휘 확률에 의해 해결할 수 있는 수준까지 품사 중의성을 최소화하고, 나머지 품사 중의성은 상호 정보와 의미 정보를 이용할 수 있을 때까지 보류한다."는 것이다. 예를 들면, 위의 예에서 어절 '나는'의 품사 중의성을 구문 범주 '타동사+어미'까지 최소화한다는 것이다.

두단계 품사 태깅은 어절 태깅(중의성 최소화) 단계에서는 품사 중의성을 구문 범주 수준으로 최소화하고, 형태소 태깅(중의성 해결) 단계에서는 어절 태깅에서 최소화된 중의성을 해결하는 전략이다. [그림 1]은 두단계 품사 태깅의 개념도이다.



[그림 1] 두단계 품사 태깅의 개념도

어절 태깅과 형태소 태깅을 분리함으로써 얻을 수 있는 잇점은 어절 태깅을 품사 집합과 코퍼스의 변화에 독립적으로 적용할 수 있다는 것이다. 본 논문에서는 어절 태깅 모델로서 자율 학습이 가능한 은닉 마르코프 모델을 채택한다.

3.2 어절 태깅 단계

은닉 마르코프 모델을 어절 태깅 문제에 적용하는 간단한 방법은 은닉 마르코프 모델에서 상태(state)를 어절 태그에 대응시키고, 은닉 마르코프 모델에서 관측 심볼(observation symbol)을 어절에 대응시키는 것이다. 그러나, 이 방법은 다음과 같은 문제들을 가진다.

첫째, 기존의 어절 태그를 이용하면 새로운 어절 태그가 발생할 때마다 모델을 변경해야 하며, 대량의 확률 저장 공간이 필요하다.

둘째, 어절을 관측 심볼에 대응시키는 것은 모든 어절을 관측 심볼에 대응시키는 것도 불가능할 뿐만 아니라 저장 공간이 엄청나게 필요하므로 불가능하다.

셋째, 초기 모델 설계 시에 모델의 파라메터에 동일한 확률 분포(uniform probability distribution)를 부여하면 수렴 속도가 느리다.

넷째, 은닉 마르코프 모델에서 관측열에 문장이 대응되므로 모델의 학습은 문장 단위로 이루어진다. 그러나, 보통 하나의 문장은 열 개 내외의 단어로 구성되므로, 불충분한 관측열 문제를 일으킨다.

본 논문에서 제안하는 어절 태깅 모델은 이러한 문제를 해결하기 위해 기존의 어절 태그 대신에 단순화된 어절 태그(simplified word-phrase tag)를 이용하고, 어절 대신에 의사 부류(pseudo class)[9]를 이용한다. 또한 초기 모델 설계 방법으로 문법적 제약(grammatical constraint)을 적용하고 동일한 확률 분포를 부여하는 방법을 이용하고, 보간법(interpolation)을 이용하여 불충분한 관측열 문제를 해결한다.

3.2.1 어절 태그 단순화

일반적인 어절 태그는 구성 형태소 품사의 결합으로 만들어지므로 새로운 어절 유형이 발생하면 새로운 어절 태그가 생성되어 문맥 확률의 수정이 불가피해 진다. 또한 형태소 품사 집합이 변경되면 어절 태그를 재구성해야 한다는 단점도 있다. 이러한 문제는 어절 태그를 단순화하여 형태소 품사 집합과 어절 유형에 독립적으로 만듦으로써 해결할 수 있다.

어절 태그의 단순화 방법은 다음과 같다.

첫째, 한글 이외의 문자[10]를 어절에서 분리하고 하나의 어절로 간주하여 각각 품사를 부여한다.

둘째, 한글로만 구성된 어절[11]을 형태소 분석하여 구성 형태소의 결합을 찾아낸다. 형태소 분석기는 비한글 어절 다음에 사용된 조사나 서술격조사 이절을 분석할 수 있도록 중의성을 고려한 형태소 분석기[22]를 수정하여 사용하였다. 예를 들면, 어절 'computer이다.'로부터 분리된 한글 어절 '이다'는 '이:서술격조사+다:종결어미'로 분석된다.

셋째, 형태소 품사를 유사한 구문 특성에 따라 단순화한다. 일반적으로 유사한 구문 특성을 갖는 품사들은 중의성 해결 정보가 코퍼스에 의존적이거나, 중의성 해결을 위해 상호 정보나 의미 정보를 필요로 하므로 이러한 품사들을 묶어 하나의 품사에 대응시킨다. [표 1]은 형태소 분석기가 사용한 형태소 품사와 단순화된 품사를 보여준다.

단순화된 품사	형태소 품사
체언	UNI - 명사,의존명사,대명사,수사
관형사	DET (이, 새, 한, 역사적, ...)
부사	ADV (빨리, 아니, 그리고, ...)
감탄사	INT (허허, 저, 얼씨구, ...)
본용언	MAJ - 자동사,타동사,형용사
보조용언	AUX - 보조동사,보조형용사
의존용언	DEP (체하다, 듯하다, 법하다, ...)
조사	jos (이, 를, 의, 와, 도, ...)
서술격조사	pjo (이)
종결어미	emf (다, 구나, ㄹ까, ...)
연결어미	emc (고, 면, 어, 게, 지, ...)
관형형어미	emd (을, 는, 라는, ㄴ다는, ...)
명사형어미	emn (음, 기, ㅁ)
조사어미	emj (는가를, 는지가, ...)
	/선어말어미 (시, 었, 겠, ...)
	/접미사 (들, 님, 쯤, 간, ...)
	/용언화접미사 (하, 되, 답, ...)
	/부사화접미사 (없이, 껏, 스레, ...)

[표 1] 형태소 품사와 단순화된 품사

넷째, 어절 사이의 구문적 의존 관계에 무관한 품사를 제거한다. 이러한 품사에는 선어말어미, 접미사 등이 있다. 예를 들면, 어절 '사람들은'은 '사람:명사+들:접미사+은:조사'로 분석되는데, 접미사 '들'은 실제로 복수를 의미할 뿐 어절의 구문 범주에 영향을 주지 못한다.

다섯째, 파생 접미사가 있는 어절은 파생된 품사로 대체한다. 부사 파생 접미사의 경우는 부사로, 용언화 접미사의 경우는 본용언으로 대체한다.

어절 태그 간단화 방법으로 10만 어절의 코퍼스로부터 50개의 어절 태그를 추출하였다. 추출된 어절 태그 집합은 [표 2]와 같다. [표 2]에서 번호는 어절 태그 번호를 의미한다. 새로운 어절 유형이 발생하거나 형태소 품사 집합이 변화하는 경우에도 기존의 50개 어절 태그에 대응시킬 수 있으므로 어절 태그가 비교적 고정적이다.

3.2.2 의사 부류

9. 중의성 유형(ambiguity pattern)에 대응되는 것으로 [13]에서 이용한 word equivalence class, [4]에서 사용한 ambiguity class와 유사한 개념이다.

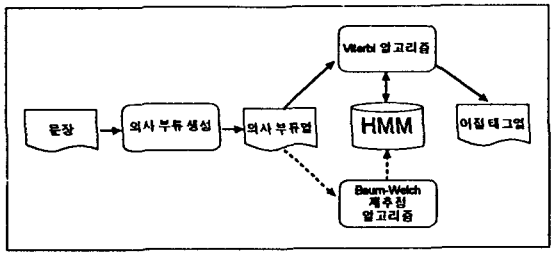
10. 이하 편의상 '비한글 어절'이라고 표현하며, 영문자(Eng), 숫자(Dgt), 좌괄호(Lft), 우괄호(Rht), 종결부호(Prd), 쉼표(Com), 인용부호(Qte), 기타문자(One) 등이 있다.

11. 이하 편의상 '한글 어절'이라 표현한다.

3]과 같다.

$$\pi_i = \Pr(q_1 = Q_i) \quad (1 \le i \le N)$$
$$1 = \sum_{i=1}^{N} \pi_i \qquad \text{[식 3]}$$

초기 어절 태그 확률 분포는 $\Pi = \{\pi_i \mid 1 \le i \le N\}$이다.



[그림 2] 어절 태깅 모델의 개념도

어절 태깅 모델의 학습과 태깅은 Baum-Welch 재추정 알고리즘과 Viterbi 알고리즘을 직접적으로 이용하여 해결할 수 있다. [그림 2]는 어절 태깅 모델의 개념도이다.

3.2.4 초기 모델 설계

초기 모델을 구성하는 일반적인 방법은 모델의 파라메터 A, B, Π에 동일한 확률 분포를 배정하는 것이다. 이것은 초기 모델 설계가 쉽다는 장점이 있는 반면, 파라메터 수렴 속도가 느린 단점이 있다. 다른 방법은 태깅된 코퍼스로부터 추출한 확률 분포를 배정하는 것이다. 이것은 대량의 태깅된 코퍼스가 있다면 효율적인 방법이지만 코퍼스에 대해 의존적인 단점이 있다.

코퍼스에 독립적이면서 수렴 속도를 향상시킬 수있는 초기 모델 구성 방법은 각 확률 분포에서 제약 정보에 의해 불가능하다고 결정된 확률들에는 0 또는 최소값을 배정하고, 나머지에 대해서는 동일한 확률값을 배정하는 것이다.

문법적 제약을 받는 확률은 0 또는 최소값으로 배정하고 나머지에 대해서는 동일한 확률값을 배정하는 것이다.

본 어절 태깅 모델에서 사용한 어절 태그 전이 확률 분포의 문법적 제약은 다음과 같다. 첫째, 관형형어미로 끝나는 어절 태그는 체언이나 의존 용언[13]으로 시작하는 어절 태그로만 전이한다. 둘째, 관형사 어절 태그는 체언이나 관형사로 시작되는 어절 태그로만 전이한다. 셋째, 보조 용언으로 시작하는 어절 태그는 연결어미로 끝나는 어절에서만 전이된다. 넷째, 조사 및 서술격 조사로 시작하는 어절 태그는 비한글 어절 태그에서만 전이된다. 다섯째, 종결어미로 끝나는 어절 태그는 종결 부호 등으로만 전이한다.

초기 어절 태그 확률 분포의 문법적 제약은 조사 및 서술격조사로 시작하는 어절 태그, 쉼표, 종결 부호, 우괄호 등의 어절 태그, 보조 용언으로 시작되는 어절 태그, 의존 용언으로 시작되는 어절 태그는 문장 처음에 올 수 없다는 것이다.

의사 부류 발생 확률 분포의 제약은 다음과 같다. 의사 부류는 어절 태그의 조합이므로, 각 어절 태그에서 관측될 수 있는 의사 부류는 그 어절 태그를 포함하는 의사 부류이다. 미등록어 의사 부류는 모든 어절 태그에서 관측될 수 있다.

3.2.5 학습

학습 문제(training problem)는 모델에서 주어진 의사 부류열이 발생될 확률을 최대로 하기 위해 모델의 파라메터인 A, B, Π를 재추정하는 것이다.

모델 파라메터를 재추정하기 위해서는 먼저 전향 변수와 후향 변수를 계산해야 한다[14]. 전향 변수는 다음 [식 4]와 [식 5]로 계산된다.

$$\alpha_1(i) = \pi_i b_i(o_1) \quad (1 \le i \le N) \qquad \text{[식 4]}$$

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \right] b_j(o_t) \quad (2 \le t \le T, 1 \le j \le N) \qquad \text{[식 5]}$$

후향 변수는 다음 [식 6]과 [식 7]로 계산된다.

$$\beta_T(i) = 1 \quad (1 \le i \le N) \qquad \text{[식 6]}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \quad (1 \le i \le T-1, 1 \le j \le N) \qquad \text{[식 7]}$$

모델의 학습은 다음의 반복적인 재추정에 의해 이루어진다[15]. 먼저, 초기 모델을 이용하여 주어진 단어열에 대해서 위의 전향 변수와 후향 변수를 계산한다. 계산된 전향 변수와 후향 변수를 다음 [식 8], [식 9], [식 10]에 적용하여 모델의 파라메터를 재추정한다.

$$\overline{\pi_i} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{N} \alpha_1(j)\beta_1(j)} \quad (1 \le i \le N) \qquad \text{[식 8]}$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \quad (1 \le i, j \le N) \qquad \text{[식 9]}$$

$$\overline{b_j(k)} = \frac{\sum_{t=1, o_t = v_k}^{T} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} \quad (1 \le j \le N, 1 \le k \le M) \qquad \text{[식 10]}$$

다음으로 재추정된 모델과 초기 모델에서 주어진 단어열 W의 발생 확률을 다음 [식 11]에 의해서 계산한다.

$$\Pr(W|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad \text{[식 11]}$$

만약 재추정된 모델의 확률과 초기 모델의 확률의 차이가 어떤 임계값(limiting point)에 도달하면 재추정을 멈추고, 그렇지 않으면 재추정된 모델을 초기 모델로 대체하고 위의 재추정 과정을 반복한다.

어절 태깅에서는 대개 열 개 내외의 어절을 포함하는 문장을 학습열로 이용하므로, 모든 파라메터를 재추정하기가 어렵다. 이 문제를 해결하기 위해 보간법을 이용한다. 보간법에 의해 학습된 모델 λ_t의 파라메터는 초기 모델 λ_i의 파라메터에서 재추정된 모델 λ_r의 파라메터를 다음 [식 12]와 같이 보정하여 구한다. 식에서 w는 재추정된 모델의 가중치값이다.

$$\lambda_t = (1-w) \times \lambda_i + w \times \lambda_r \qquad \text{[식 12]}$$

3.2.6 태깅

태깅 문제[16]은 [그림 5]에서처럼 문서 내의 문장들을 의사 부류열로 변환하고, 변환된 의사 부류열이 학습된 모델에서 발생될 확률이 최대가 되는 태그열을 결정하는 과정으로,

13. '체하다', '듯하다' 등으로, 보조 동사에서 분리하였다.

14. 이 과정을 forward-backward procedure라고 한다.

15. 이 과정을 Baum-Welch reestimation이라고 한다.

16. 일반적인 은닉 마르코프 모델에서는 해석 문제(decoding problem)에 해당한다.

[1] ADV	[26] MAJ+emn+pjo+emc
[2] AUX+emc	[27] MAJ+emn+pjo+emd
[3] AUX+emd	[28] MAJ+emn+pjo+emf
[4] AUX+emf	[29] MAJ+emn+pjo+emj
[5] AUX+emj	[30] MAJ+emn+pjo+emn
[6] AUX+emn	[31] MAJ+emn+pjo+emn+jos
[7] AUX+emn+jos	[32] One
[8] Com	[33] Prd
[9] DEP+emc	[34] Qte
[10] DEP+emd	[35] Rht
[11] DEP+emf	[36] UNI
[12] DEP+emj	[37] UNI+jos
[13] DEP+emn	[38] UNI+pjo+emc
[14] DEP+emn+jos	[39] UNI+pjo+emd
[15] DET	[40] UNI+pjo+emf
[16] Dgt	[41] UNI+pjo+emj
[17] Eng	[42] UNI+pjo+emn
[18] INT	[43] UNI+pjo+emn+jos
[19] Lft	[44] jos
[20] MAJ+emc	[45] pjo+emc
[21] MAJ+emd	[46] pjo+emd
[22] MAJ+emf	[47] pjo+emf
[23] MAJ+emj	[48] pjo+emj
[24] MAJ+emn	[49] pjo+emn
[25] MAJ+emn+jos	[50] pjo+emn+jos

[표 2] 어절 태그 집합

어절의 어휘 확률은 어절의 어휘적 특성을 반영한다. 예를 들면, 어절 '나는'의 어휘 확률은 어절 '나는'이 어절 태그 '체언+조사', '본용언+관형형어미', '보조 용언+관형형어미' 중에서 어떠한 어절 태그로 잘 사용되는가를 나타낸다.

의사 부류는 '동일한 품사 중의성을 갖는 어절들은 비슷한 어휘적 특성을 가진다.'고 가정한다. 예를 들면, 어절 '가는'과 어절 '사는'은 같은 품사 중의성을 가지며, 이 두 어절은 보통 '본용언+관형형어미'로 잘 사용된다.

의사 부류는 어절의 가능한 어절 태그 유형, 즉 품사 중의성을 뜻한다. 의사 부류의 구성 방법은 다음과 같다.

첫째, 해당 어절의 가능한 모든 어절 태그를 생성한다.

둘째, 중복된 어절 태그를 제거한다.

셋째, 어절 태그를 연결한다.

예를 들어, 어절 '나는'은 어절 태그로 '보조용언+관형형어미', '본용언+관형형어미', '체언+조사'를 가지므로, 의사 부류는 'AUX+emd/ MAJ+emd/ UNI+jos'이다.

위와 같은 방법으로 10만 어절 코퍼스로부터 291개의 의사 부류가 추출되었다. [표 3]은 추출된 의사 부류의 일부를 보여준다. 괄호 안의 숫자는 의사 부류 번호를 나타낸다. 의사 부류는 새로운 중의성 유형이 발생하는 경우에 증가되지만 어절에 비해 비교적 고정적이다.

3.2.3 은닉 마르코프 모델

일반적인 은닉 마르코프 모델[12]은 상태 전이를 결정하는 은닉 처리(hidden process)와 관측 심볼을 출력하는 관측 처리(observation process)를 갖는 이중 통계 처리(doubly stochastic process)이다[15]. 단순화된 어절 태그와 의사 부류를 이용한 은닉 마르코프 모델의 파라메터는 다음과 같이 정의할 수 있다.

12. 일반적인 은닉 마르코프 모델에 대한 자세한 설명은 [15] [16]에 있다.

[1] ADV
[2] ADV/AUX+emc/AUX+emf/MAJ+emc/MAJ+emf/UNI
[3] ADV/AUX+emc/AUX+emf/MAJ+emc/MAJ+emf/UNI+jos
...
[97] AUX+emd/MAJ+emd
[98] AUX+emd/MAJ+emd/UNI
[99] AUX+emd/MAJ+emd/UNI+jos
[100] AUX+emd/MAJ+emd/UNI+jos/jos
...
[146] DET
[147] DET/INT/UNI
[148] DET/INT/UNI/jos
...
[217] MAJ+emd/UNI+jos
[218] MAJ+emd/UNI+jos/UNI+pjo+emd
[219] MAJ+emd/UNI+jos/jos
...
[286] UNI/UNI+pjo+emn
[287] UNI/jos
[288] UNI/pjo+emd
[289] Unk // 형태소 분석 불가 어절의 의사 부류
[290] jos

[표 3] 의사 부류 집합

다.

(1) N : 어절 태그 갯수

어절 태그 집합 Q = {Q_1, Q_2, ..., Q_N}에서 i번째 어절 태그는 Q_i로 표기하며, 문장에서 t번째 어절의 어절 태그는 q_t로 표기한다.

(2) M : 의사 부류 갯수

의사 부류 집합 V = {V_1, V_2, ..., V_M}에서 j번째 의사 부류는 V_j로 표기하며, 문장에서 t번째 의사 부류는 o_t로 표기한다.

(3) A : 어절 태그 전이 확률 분포

어절 태그 Q_i에서 어절 태그 Q_j로 전이할 확률 a_{ij}은 연관 확률 모델에서 문맥 확률에 대응되며, 다음 [식 1]과 같다.

$$a_{ij} = \Pr(q_t = Q_j | q_{t-1} = Q_i) \quad (1 \le i,j \le N)$$
$$1 = \sum_{j=1}^{N} a_{ij} \qquad \text{[식 1]}$$

어절 태그 전이 확률 분포는 A = {a_{ij} | $1 \le i,j \le N$}이다.

(4) B : 의사 부류 발생 확률 분포

은닉 마르코프 모델의 관측 심볼 확률(observation symbol probability) 분포에 해당한다. 어절 태그 Q_j에서 의사 부류 V_k가 발생할 확률인 $b_j(k)$는 연관 확률 모델에서 어휘 확률에 대응되며, 다음 [식 2]와 같다.

$$b_j(k) = \Pr(o_t = V_k | q_t = Q_j) \quad (1 \le j \le N, 1 \le k \le M)$$
$$1 = \sum_{k=1}^{M} b_j(k) \qquad \text{[식 2]}$$

의사 부류 발생 확률 분포는 B = {$b_j(k)$ | $1 \le j \le N, 1 \le k \le N$}이다.

(5) Π : 초기 어절 태그 확률 분포

은닉 마르코프 모델의 초기 상태 확률(initial state probability) 분포에 해당한다. 문장의 처음 위치에 어절 태그 Q_i가 나타날 확률인 π_i는 연관 확률 모델에서 문장 시작 표시와 어절 태그 사이의 문맥 확률에 대응되며, 다음 [식

Viterbi 알고리즘으로 해결할 수 있다.

먼저 알고리즘에 사용되는 변수 $\delta_t(i)$와 $\psi_t(i)$를 [식 13]와 [식 14]에 의해 초기화한다.

$$\delta_1(i) = \pi_i b_i(o_1) \quad (1 \le i \le N) \qquad [식\ 13]$$

$$\Psi_1(i) = 0 \quad (1 \le i \le N) \qquad [식\ 14]$$

계속해서 두 변수의 값을 [식 15]와 [식 16]를 이용하여 재귀적으로 계산한다.

$$\delta_t(j) = \max_{1 \le i \le N} [\ \delta_{t-1}(i) a_{ij}\ ]\ b_j(o_t) \quad (2 \le t \le T, 1 \le j \le N) \qquad [식\ 15]$$

$$\phi_t(j) = \arg\max_{1 \le i \le N} [\ \delta_{t-1}(j) a_{ij}\ ] \quad (2 \le t \le T, 1 \le j \le N) \qquad [식\ 16]$$

마지막으로 다음 [식 17]과 [식 18]을 이용하여 태그열을 역추적(backtracking)하면, 최선의 태그열 I를 구할 수 있다.

$$I_T = \arg\max_{1 \le i \le N} [\ \delta_T(i)\ ] \qquad [식\ 17]$$

$$I_t = \phi_{t+1}(I_{t+1}), \quad t = T-1, T-2, \ldots, 1 \qquad [식\ 18]$$

IV. 실험 및 평가

본 논문에서는 실험 코퍼스로서 국민학교 교과서의 1813개의 문장과 국민 교육 헌장을 선정했다. 어절 태깅 모델의 학습에 사용하기 위해 실험 코퍼스 중에서 국민학교 교과서의 일부(300 문장)를 추출하였다. 실험 코퍼스에 대한 통계는 [표 4][17]와 같다. [표 4]에서 평균 중의성이 비교적 낮은 이유는 단순한 어절 태그를 이용하기 때문이다.

통계 \ 코퍼스		국민 교육 헌장	국민학교 교과서	학습에 포함된 부분	학습에 포함되지 않은 부분
문장 수		8	1813	300	1513
의사 부류	분석 불가	0	82	5	77
		0.0%	0.59%	0.22%	0.66%
	비한글	22	2483	346	2137
		14.29%	17.89%	15.07%	18.45%
	중의성 무	105	7983	1425	6558
		68.18%	57.53%	62.06%	56.63%
	중의성 유	27	3328	520	2808
		17.53%	23.89%	22.65%	24.25%
	총 수	154	13876	2296	11580
어절 태그 후보 수		188	18647	3002	15645
평균 중의성		1.22	1.34	1.31	1.35

[표 4] 실험 코퍼스의 통계 분석

다른 가중치로 반복하여 학습한 뒤, 선정된 국민학교 교과서의 전체를 태깅한 결과 [표 5]와 같았다. [표 5]에서 각 항의 숫자는 태깅 오류 갯수와 오류율[18]이다. 가중치가 높으면 학습열에 대한 수렴 속도는 빠르지만, 파라메터가 특정 학습열에 쉽게 적응되므로 학습이 불안정하게 된다. 실험 결과, 재추정된 모델의 가중치를 0.005로 정하고 학습열을 5회 반복하여 학습한 모델의 오류율이 가장 낮았다.

0.005의 가중치로 5회 반복 학습한 모델을 이용하여 국민

17. 어절 태그 후보는 문서 내의 의사 부류가 가질 수 있는 어절 태그의 총 수이며, 평균 중의성은 하나의 의사 부류가 가질 수 있는 평균 어절 태그의 수를 의미한다.

가중치 \ 반복횟수	1	2	3	4	5
0.05	854	893	939	939	987
	6.15%	6.44%	6.77%	6.77%	7.11%
0.01	905	868	878	922	921
	6.52%	6.26%	6.33%	6.64%	6.64%
0.005	940	892	851	839	833
	6.77%	6.43%	6.13%	6.05%	6.00%

[표 5] 반복 학습된 모델의 태깅 성능

교육 헌장과 국민학교 교과서를 태깅한 결과를 학습된 부분과 학습되지 않은 부분을 분리하여 평가한 결과는 [표 6]과 같다. 학습에 포함된 부분의 오류율이 비교적 낮았다. 국민 교육 헌장의 경우는 학습에 사용되지 않았음에도 비교적 적은 오류율을 보이는데, 이것은 평균 중의성이 낮고, 중의성이 있는 어절의 수가 적기 때문이다.

코퍼스 \ 성능	오류 갯수	오류율
국민 교육 헌장	8	5.19%
학습에 포함된 부분	103	4.49%
학습에 포함되지 않은 부분	730	6.30%

[표 6] 학습 여부와 태깅 성능

국민학교 교과서에서 발생한 오류 833개 중에서 빈도가 높은 것들에 대해서 혼동 행렬(confusion matrix)을 만들면 [표 4]와 같다. [표 4]에서 세로축은 태깅된 코퍼스에 있는 올바른 어절 태그 번호를 나타내고, 가로축은 태깅 결과로 생성된 잘못된 어절 태그 번호를 나타낸다.

오류를 분석한 결과, 보조용언과 본용언을 혼동하는 오류[19]가 가장 많은 400개로 전체 오류의 48%를 차지했다. 일반적으로 보조용언과 본용언의 구별은 어려운 문제여서 이러한 오류는 수동 태깅 시에도 자주 발생된다.

두번째로 많은 오류는 어절 태그 '체언+조사'를 '본용언+관형형어미'나 '본용언+명사형어미+조사'로 태깅하는 오류[20]로 71개가 발생했다.

다음으로는 '본용언+연결어미'를 '부사'로 태깅한 것[21]이다. 수동 태깅된 문서와 비교한 결과, 어절 '어떻게'는 '부사'로 태깅되어야 함에도 '형용사+연결어미'로 잘못 태깅되어 있었다. 일반적으로 수동 태깅에서 일관성이 부족하거나 과다 분석이 문제가 되는데, 이 예는 수동 태깅이 형태소 분석기의 분석 수준에서 이루어져야 함을 시사한다.

마지막으로 대략 40개의 오류가 종결어미를 연결어미로 잘못 태깅한 것[22]이었다.

연결어미의 품사를 세분화하여 보조적 연결어미를 추가함으로써 보조용언과 본용언의 분별 능력을 높이고, 고빈도 어절에 대해서는 독립적인 의사 부류를 할당함으로써 고빈도 어절의 어휘적 특성을 같은 중의성 유형을 갖는 저빈도 어절과 분리시키고, 신뢰도 있는 파라메터 추정을 위해 모델의 학습 단위를 여러 개의 문장으로 확장함으로써 태깅 성능이 향

18. 태깅 오류 갯수 / 의사 부류 총 갯수.
19. [표 7]에서 4를 22로, 2를 20으로, 21을 3으로 잘못 태깅한 오류이다.
20. [표 7]에서 37을 21이나 25로 혼동하는 오류이다.
21. [표 7]에서 20을 1로 태깅한 오류이다.
22. [표 7]에서 4를 2로, 22를 20으로 혼동하는 오류이다.

수동 태깅 \ 자동 태깅	1	2	3	4	15	20	21	22	25	36	37
1						5					
2				0		80					
3							6				
4		13						288			
15										26	
20	34	2						1			
21			32								5
22				5		27					
25											1
36					3						
37							39		32		

[표 7] 고빈도 오류의 혼동 행렬

상되리라 생각된다.

V. 결 론

본 논문에서는 품사 태깅 문제를 형태소 분석 수준에서 얻을 수 있는 정보를 이용한 중의성 최소화 단계와 보다 복잡하고 대량의 확률 정보인 상호 정보나 의미 정보를 이용하여 최소화된 중의성을 해결하는 단계로 분리하는 두단계 품사 태깅 방법을 제안했다.

중의성 최소화를 위한 어절 태깅 모델은, 첫째로 코퍼스에 독립성을 가질 수 있도록 자율 학습이 가능한 은닉 마르코프 모델을 채택했고, 둘째로 품사 집합의 변화에 잘 적응할 수 있도록 어절 태그를 단순화시켰으며, 세째로 코퍼스에 의존적인 어절의 어휘 정보를 이용하는 대신에 중의성 유형을 나타내는 의사 부류를 이용했다. 제안된 어절 태깅 모델은 태깅된 코퍼스를 이용하지 않으므로 확률 추출을 위한 수동 태깅의 노력이 필요없으며, 적은 수의 어절 태그와 의사 부류를 이용하므로 관리해야 할 확률 정보의 양을 줄일 수 있다.

제안된 어절 태깅 모델은 국민학교 교과서 중에서 1813개의 문장과 국민 교육 헌장으로 이루어진 코퍼스를 구성하고, 태깅 정확도 검사를 위해 반자동 수동 태깅을 했다. 코퍼스의 일부 300개 문장을 학습열로 구성해서 은닉 마르코프 모델을 반복적으로 학습시키고, 어절 태깅한 결과, 대략 94%의 태깅 정확도를 보였다. 이 수치는 모델이 자율 학습 방법임을 고려할 때 상당히 높은 정확도이다.

전체의 태깅 오류 중 48%의 오류가 사람도 구분하기 어려운 본용언과 보조 용언을 구별하는 오류였고, 상당수의 오류가 체언과 본용언의 구별, 본용언과 부사의 구별, 종결어미와 연결어미의 구별에서 발생했다. 앞으로 품사 집합에 대한 더 정밀한 연구를 거쳐 어절 태그를 결정하고, 고빈도 어절에 독립적인 의사 부류를 대응시켜 의사 부류가 단어의 특성을 잘 표현할 수 있도록 한다면 더 좋은 성능을 보일 수 있을 것이다.

향후에는 상호 정보와 의미 정보를 획득하고 관리하는 방법과 형태소 태깅 모델에 적용 방법에 대한 연구를 통해 궁극적으로 중의성을 해결해야 할 것이다.

참고 문헌

[1] J. Benello et al., "Syntactic category disambiguation with neural networks", Computer Speech and language, 3, pp.203-217, 1989.

[2] E. Brill, "A Simple Rule-Based Part of Speech Tagger", Proc. 3rd Conference on Applied NLP, Trento, Italy, pp.153-155, 1992.

[3] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text", proc. 2nd Conference on Applied NLP, Austic, TX, pp.136-143, 1988.

[4] D. Cutting et al., "A Practical Part-of-Speech Tagger", proc. the 3rd Conference on Applied NLP, Trento, Italy, pp.133-140, 1992.

[5] S. J. DeRose, "Grammatical category disambiguation by statistical optimization", Computer Linguistics, vol.14, no.1, pp.31-39, 1988.

[6] A. M. Derouault and B. Merialdo, "Natural language modelling for phoneme-to-txt transcription", IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8, pp.742-249, 1986.

[7] R. Garside, G. Leech, and G. Sampson, The Computational Analysis of English: A Corpus-based Approach, London : New York : Longman Group UK limited, 1987.

[8] B. B. Greene and G. M. Rubin, Automatic grammatical tagging of English, Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, 1971.

[9] D. Hindle, "Acguiring Disambiguation Rules from Text", Proc. ACL-89, pp.118-125, 1989.

[10] F. Karlsson, "Constraint grammar as a framework for parsing running text", proc. COLING-90, vol.3, pp.168-173, 1990.

[11] S. Klein and R. F. Simmons, "A Computational Approach to Grammatical Coding of English Words", JACM, vol.10, pp.334-347, 1963.

[12] K. Koskenniemi, "Finite-state parsing and disambiguation", proc. COLING-90, vol.2, pp.229-232, 1990.

[13] J. Kupiec, "Robust part-of-speech tagging using a hedden Markov model", Computer Speech and Language, vol.6, 225-242, 1992.

[14] M. Nakamura and K. Shikano, "A study of English word category prediction based on newral networks", proc. ICASSP-89, pp.731-734, 1989.

[15] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models", IEEE ASSP Mag., vol.3, no.1, pp.4-16, Jan. 1986.

[16] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", IEEE Proc., vol.77, no.2, pp.257-286, Feb. 1989.

[17] 김재훈 외, "퍼지망을 이용한 한국어 품사 태깅", 제 5 회 한글 및 한국어 정보처리 학술발표 논문집, pp.593-603, 1993.

[18] 박혜준 외, "말뭉치 품사꼬리달기 시스템 구현", 한국정보과학회 봄 학술발표 논문집, vol.21, no.1, pp.829-832, 1994.

[19] 이선정, 신경망을 이용한 한국어 단어빈주 예측 및 애매성해소, 서울대학교 박사학위논문, 1994.

[20] 이운재 외, "한국어 문서 태깅 시스템", 한국정보과학회 봄 학술발표 논문집, vol.20, no.1, pp.805-808, April 1993.

[21] 임철수, HMM을 이용한 한국어 품사 태깅 시스템 구현, 한국과학기술원 석사학위논문, 1994.

[22] 임희석, 중의성 유형 분류에 근거한 한국어 형태소 분석기, 고려대학교 석사학위논문, 1994.