어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사태깅

신중호°, 한영석, 박영찬, 최기선
한국과학기술원 전산학과

An HMM Part-of-Speech Tagger for Korean Based on Wordphrase

Shin Jung-ho°, Han Young-seok, Park Young-chan, Choi Key-Sun
Department of Computer Science, KAIST

요 약

말뭉치에 품사를 부여하는 일은 언어연구의 중요한 기초가 된다. 형태소 해석의 모호한 결과로부터 한 가지 품사를 선정하는 작업을 태깅이라고 한다. 한국어에서 은닉 마르코프 모델 (Hidden Markov Model)을 이용한 태깅은 형태소 관계만 혹은 어절관계만을 이용한 방법이 있어 왔다. 본 논문에서는 어절관계와 형태소관계를 동시에 은닉 마르코프 모델에 반영하여 태깅의 정확도를 높인 모델을 제시한다. 제안된 방법은 품사의 변별력은 떠어나지만 은닉 마르코프 모델의 노드의 수가 커짐으로써 형태소만을 고려한 방법보다 더 많은 학습데이타를 필요로 한다. 실험적으로 본 논문의 방법이 기존의 방법보다 높은 정확성을 가지고 있음이 검증되었다.

1. 서 론

하나의 단어가 여러가지 가능한 품사를 갖는 모호성을 가질 수 있으며 이러한 품사의 모호성을 해소하는 과정을 품사태깅(part-of-speech tagging)이라고 하는데 본 논문에서는 간단히 태깅이라고 칭한다. 형태소 분석을 거쳐 태깅된 말뭉치는 자연언어처리나 정보검색등에서 중요한 기초 데이타로 쓰인다. 지금까지 이용된 태깅방법은 크게 규칙을 이용하는 방법과 말뭉치로부터 추출된 통계정보를 이용하는 방법으로 나눌 수 있다. 규칙 접근방법은 규칙을 기술하기가 어렵고 다른 영역으로의 적응성이 떨어지므로 일반적으로 통계정보를 이용하는 방법이 많이 사용되고 있다[이하규94]. 본 논문에서는 통계적인 처리의 일환으로 은닉 마르코프 모델(Hidden Markov Model)을 이용한다.

한국어는 영어의 경우와 같이 단어의 열이라기 보다는 어절의 열로서 문장이 만들어진다. 어절 간의 순서가 자유로운 편이지만, 통계적으로는 뚜렷한 순서를 발견할 수 있다. 이러한 어절 간의 의존성은 태깅과정에서 손쉽게 쓰일 수 있는 일종의 구문정보이다. 품사 간의 의존성은 영어와 같은 언어에서 많이 쓰이고 있으며, 한국어 태깅의 경우에도 유용하다.

본 논문에서는 어절간의 의존성과 형태소간의 의존성을 모두 반영하는 은닉 마르코프 모델의 구성 방법을 제안한다. 어절을 구성하는 형태소 패턴은 어절에 따라 다른 형태를 가진다. 그리고 같은 형태소라도 그것이 속해있는 어절의 형태에 따라 주변 형태소의 분포가 다르게 나타날 수 있으며, 의존확률(전이확률)도 달라질 수 있다. 이러한 상황을 무시한 형태소단위의 은닉 마르코프 모델에서는 여러가지 상황이 섞이게 됨으로써 변별력을 잃게 된다. 형태소 단위의 방법 외에 어절간의 관계에 촛점을 둔 방법도 제시되었다[이운재94]. 이 방법은 형태소 간의 관계를 직접적으로 모델링하지 않았으며, 가능한 모든 형태소 조합을 어절품사로 봄으로써 모델의 노드수가 많아져 정확도면에서 문제가 있었다. 이외에, 의미정보와 같은 언어학적 정보를 이용해서 태깅의 성능을 높이려는 시도가 있었으나[김충원94] 본 논문은 형태소정보와 어절정보에 한정하여 태깅하는 것을 목표로 하기 때문에 직접적인 비교는 무의미하다.

제안된 방법은 우선 어절을 32개로 분류하고 학습자료로부터 어절 간의 관계를 찾아내어서 1차적인 망을 구

성한 후, 각 어절을 이루는 형태소 관계를 찾아내어 각 어절 망에 채움으로써 하나의 커다란 은닉 마르코프 모델을 구성한다. 일단 망이 완성이 되고 나면, 어절 간의 관계는 간접적으로 표현되고 전체 망구조는 형태소 단위의 망으로 보이게 된다. 수동으로 태깅된 5만5천 어절을 이용하여 네가지 방법으로 은닉 마르코프 모델을 구축하고, 5천 어절을 자동태깅한 결과에 의하면 본 논문에서 제안된 모델이 가장 높은 정확률을 보이고 있다.

2. 문제정의

단어들의 기능, 형태, 의미 등에서 공통적인 특징을 보이는 것끼리 분류한 것을 품사라고 한다[남기심85]. 하나의 단어에 대해 여러 개의 품사가 존재하는 경우 품사의 모호성이 있다고 한다. 품사 모호성의 해소가 품사태깅의 목표이다. 한 단어에 대한 품사 모호성은 주변단어 그 자체나 주변단어의 품사정보를 이용하여 어느 정도 해결될 수 있다. 태깅은 주어진 단어열 $w_{1,n}$에 대해 가장 적절한 품사열 $t_{1,n}$을 부여하는 것으로 정의한다. 이때, $w_{1,n}$은 n개의 단어열이고, $t_{1,n}$은 $w_{1,n}$의 각각의 단어에 대응할 수 있는 품사열 중의 하나이다. 이 $t_{1,n}$중 $w_{1,n}$에 가장 적합한 품사열을 $T(w_{1,n})$이라고 하자. 즉, 다음과 같다[Charniak93a].

$$
\begin{aligned}
T(w_{1,n}) &= \arg\max_{t_{1,n}} P(t_{1,n}|w_{1,n}) \\
&= \arg\max_{t_{1,n}} \frac{P(t_{1,n}, w_{1,n})}{P(w_{1,n})} \qquad (1) \\
&= \arg\max_{t_{1,n}} P(t_{1,n}, w_{1,n}).
\end{aligned}
$$

단, $\arg\max_{t_{1,n}} P(x)$는 확률값 $P(x)$를 최대로 하는 $t_{1,n}$을 구하는 것을 의미한다.

이용할 수 있는 정보를 품사로 한정하고 단어의 품사 모호성의 판별을 주변 단어에 의존하여 결정한다는 모델은 단어 문맥의 구조적 정보가 없다는 조건하에서는 보편화된 착안이다[Marcus93, Charniak93b].

3. 은닉 마르코프 모델을 이용한 품사태깅

품사의 모호성을 주위 단어와 품사정보를 이용하여 해결하려고 할 때 어떠한 정보를 얼마나 이용할 것인가가 중요한 문제가 된다. 가능한 모든 정보를 이용하는 것이 모호성을 최대로 해결해 줄 수 있겠지만 기술적, 물리적 한계 때문에 이용할 수 있는 정보를 제약해야만 한다. 이러한 제약조건하에서는 현재의 모호성을 갖는 단어가 어떤 정보에 가장 의존하는가를 파악해야만 한다. 일반적으로 단어 열에서 가까운 곳에 위치한 단어일수록 더 많은 연관성이 있다고 할 수 있다. 이러한 제반의 조건들을 고려할때 마르코프 모델은 태깅에 적합한 모델로 평가된다. 태깅문제를 은닉 마르코프 모델의 최적의 상태열을 찾는 문제로 변환하는 과정은 다음과 같이 전개될 수 있다. (1)에서 $P(t_{1,n}, w_{1,n})$를 사슬규칙(Chain Rule)을 이용하여 전개하면 (2)와 같은 식으로 변환된다[Charniak93a].

$$
\begin{aligned}
& P(t_{1,n}, w_{1,n}) \\
&= P(t_1)P(w_1|t_1)P(t_2|t_1, w_1)P(w_2|t_{1,2}, w_1)\cdots \\
&\quad P(w_n|t_{1,n-1}, w_{1,n-1})P(t_n|t_{1,n-1}, w_{1,n}) \qquad (2) \\
&= \prod_{i=1}^{n} P(t_i|t_{1,i-1}, w_{1,i-1})P(w_i|t_{1,i}, w_{1,i-1})
\end{aligned}
$$

품사태깅을 마르코프 과정으로 간주하면 마르코프 가정을 이용하여 (3)과 같이 간략화된 확률식을 사용할 수 있다[Charniak93a].

$$
\begin{aligned}
& P(t_i|t_{1,i-1}, w_{1,i-1}) = P(t_i|t_{i-1}) \qquad (3) \\
& P(t_i|t_{1,i}, w_{1,i-1}) = P(w_i|t_i)
\end{aligned}
$$

(3)은 현재의 품사(t_i)는 이전의 품사(t_{i-1}) 하나에만 의존하여 결정되고 현재의 단어(w_i)는 자신의 품사(t_i)에만 의존하여 결정된다는 가정을 반영한 것이다. 이와 같은 가정을 토대로 태깅은 (4)로 표현될 수 있다[Charniak93a].

$$
T(w_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i) \qquad (4)
$$

(4)를 구현하는 은닉 마르코프 모델은 다음과 같이 정의될 수 있다[임철수94, Rainber89].

< 태깅을 위한 은닉 마르코프 모델 정의 >

1. N: 상태의 갯수 = 품사의 총갯수
2. M: 사건의 갯수 = 단어의 총갯수
3. A: 상태전이 확률 = 품사 t_i가 발생했을때 품사 t_j가 발생될 확률

$$a_{ij} = P(t_j|t_i), \quad 1 \le i, j \le N$$

4. B: 관측 심볼 확률 = 품사 t_i가 발생했을때 단어 w_k가 발생될 확률

$$b_i(w_k) = P(w_k|t_i), \quad 1 \le j \le N, 1 \le k \le M$$

5. Π: 초기 상태 전이 확률 = 품사 i가 문장의 맨처음에 나타날 확률

$$\pi_i = P(t_i), \quad 1 \le i \le N$$

위와 같이 태깅을 위한 은닉 마르코프 모델을 정의하게 되면 상태전이 확률은 문장내에서 문맥확률을 나타내고 관측심볼 확률은 어휘확률을 나타낸다. 그리고 (4)는 단어열로 이루어진 하나의 문장에 대해 가장 적합한 품사열을 찾는 수식이 되고 은닉 마르코르 모델에서 최적의 상태열을 찾는 Viterbi 알고리즘을 적용하여 구할 수 있다[임철수94].

4. 태깅관점에서 본 한국어 특징

영어, 프랑스어 등의 구미어는 문장이 단어들의 조합으로 구성된다. 그러나 한국어는 형태소들이 어절을 구성하고 어절들이 문장을 구성하게 된다. 형태소들이 어절을 형성할 때 변형이나 생략이 일어나게 된다. 하나의 어절은 다양한 해석이 가능하며 이로 인해 품사 뿐만 아니라 형태소의 단위 또는 형태소 갯수도 다르게 분석될 수 있다. 감기는은 감기+는 또는 감+기+는으로 분리될 수 있는데 이는 품사, 단위, 갯수가 모두 다른 예이다[임철수94]. 이러한 특징은 영어에서 발견되지 않는 것이기 때문에 영어 태깅모델을 그대로 한국어에 적용하는 것은 문제가 있다.

한국어의 어절은 크게 실질 형태소와 형식 형태소로 나뉘는데 형식 형태소가 문법적 기능을 나타내준다[남기심85]. 이처럼 문법적 기능이 명시적으로 드러난 점은 태깅문제를 푸는데 크게 도움이 될 수 있다. 한국어는 문장이 어절단위로 이루어지고 또한 어절은 띄어쓰기 단위로 쉽게 구분되므로 태깅모델을 어절단위로 구성할 수 있다. 한국어의 어절과 비슷한 예로 영어에서 구(Phrase)를 들 수 있지만 띄어쓰기 단위로 쉽게 구분되는 어절과는 달리 구는 문장에서 따로 분리되기가 어렵다. 본 논문에서는 한국어가 가지는 유용한 정보인 어절구조를 직접적으로 반영함으로써 태깅의 정확도를 높일 수 있는 모델을 제안한다.

5. 한국어 어절구조를 반영한 계층 은닉 마르코프 모델

은닉 마르코프 모델을 이용하여 구현된 한국어 태깅 시스템은 형태소를 영어의 단어와 같은 단위로 간주하는 방식[임철수94]과 문맥확률에서는 어절사이의 관계를 이용하고 어휘확률은 형태소 단위의 정보를 이용하는 방법이 있었다[이운재93]. 첫번째 방식은 위의 4절에서 설명했던 한국어의 중요한 특징인 어절을 고려하지 않았다는 문제점이 있다. 그리고 두번째 방법은 어절단위의 문맥확률만을 이용하였으므로 형태소 사이의 관계정보는 이용하지 못하며 형태소들의 조합을 어절품사로 정했으므로 어절품사의 갯수가 너무 많다는 문제점도 있다. 그러므로 본 논문에서는 원형복원이 쉬운 형태소 단위의 태깅을 하되 어절구조를 은닉 마르코프 모델의 망구조에 반영하여 어절정보를 얻는 방식을 취했다. 그리고 어절품사를 정함에 있어서 품사의 갯수는 작게 가지고 문법적 기능은 충분히 고려되는 분류방식을 채택하고자 했다. 전체 과정은 우선 어절단위의 분류체계를 정한 후 어절구조를 골격으로 형태소 단위의 망을 구축하고 어휘확률을 확장시켜 본 논문에서 제안하는 알고리즘을 완성한다.

5.1 어절단위의 분류

어절은 보는 관점에 따라 여러 형태로 분류될 수 있는데 본 논문에서는 각각의 어절이 실질 형태소와 형식 형태소로 나뉘는 성질을 이용하여 실질 형태소와 형식 형태소의 대표값의 조합으로 모든 어절을 표현하였다. 대표값의 지정방법은 어절 앞부분의 형태소 2개를 분석하여 실질 형태소 대표값을 정하고 어절 뒷부분의 형태소 2개를 분석하여 형식 형태소의 대표값을 지정하였다.

은닉 마르코프 모델에서 어떤 노드로부터 나가는 출력링크는 그 노드에서 나갈 수 있는 모호성의 가짓수를 표현하기 때문에 출력링크의 수가 적을 수록 바람직한 구조라고 할 수 있다. 어절을 대표하는 실질 형태소와 형식 형태소를 분류할 때 어절간의 충분한 변별력을 유지하면서 어절 간 링크의 수가 최소화 되도록 하였다.

본 논문에서는 어절품사를 나타내는 두 대표값으로 실질 형태소는

S(기호), A(부사성), I(감탄사),
M(관형사), N(명사류), P(용언류)

로 정의하고 형식형태소는

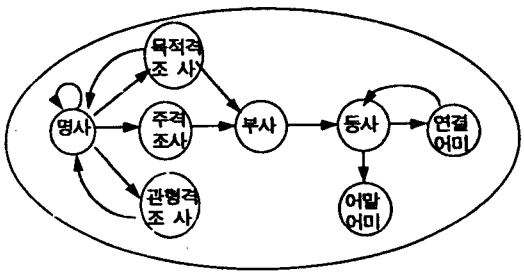
A(부사격조사), C(연결어미), F(어말어미),
S(주격조사), O(목적격조사), M(관형격조사),
X(보조사), Y(접속조사)

로 정의하였다. 이 대표값끼리의 모든 조합이 가능한 것은 아니기 때문에 32개의 조합형태로 어절을 분류할 수 있다. 예를 들어 '자신에게 벌을 준 학생입니다'는

'자신에게 벌을 준 학생입니다'로 어절단위의 분류를
NA NO PM NF 할 수 있다.

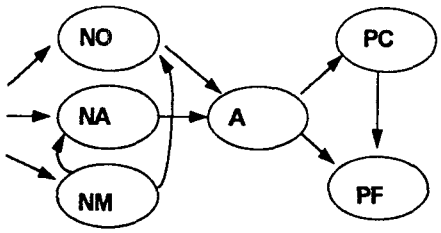
5.2 망구조 설계및 구현

기존의 연구[임철수94]에서는 형태소간의 이진그램(bigram)을 이용하여서 <그림2a>와 같이 직접 망구조를 설계하였다. 본 논문에서 제시하는 망구조 설계는 어절단위의 전체적인 골격 구성단계와 형태소 단위의 세부구조 구성단계로 나뉘어 진다.



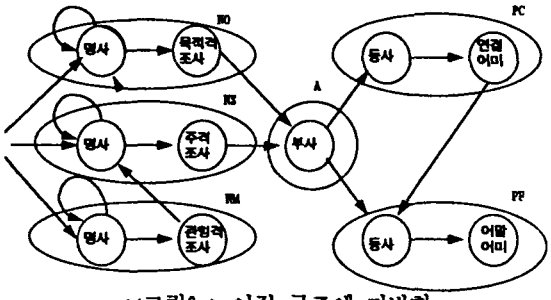
<그림2a> 형태소단위의 망구조의 예

우선 5.1에서 제시한 어절 분류를 이용해서 <그림2b>에서와 같이 골격 구조를 형성한다.



<그림2b> 어절단위의 망구조의 예

그렇게 형성된 구조내부에서는 기존의 연구들[임철수 94]에서와 같이 형태소간의 이진그램으로 세부망을 구성한다. <그림2c>는 이렇게 하여 최종적으로 형성된 망의 예이다.



<<그림2c> 어절 구조에 기반한
형태소 단위의 망구조의 예

최종적으로 구축된 망구조는 거쳐가는 경로를 더욱 세밀하게 분리하게 되므로 원래의 망구조보다는 더 큰 변별력을 가지고 있다. 여기에서의 각각의 노드는 어절정보가 고려된 태그 즉 [어절태그, 형태소태그]를 부여받게 된다.

예를 들어 <그림2a>에서의 어떤 노드에 해당하는 태그 t_i가 '명사'라면 그에 상응해서 어절정보가 고려된 태그 t'_i는 그림 d에서처럼 [NS, 명사] [NA, 명사] [NM, 명사]로 나뉘어 표현된다.

이에 대응되는 알고리즘은 다음과 같은 수식으로 표현되는데 수식에서 t'_i가 주어지면 t_i는 유일하게 결정될 수 있다.

$$(5)\quad T(w_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t'_i|t'_{i-1})P(w_i|t'_i)$$

여기에서 어휘확률 $P(w_i|t'_i)$를 $P(w_i|t'_p t'_{i-1})$로 확장시키면 단어와 이전 품사의 연관관계를 구할 수 있게 된다. 그러므로 다음과 같은 수식을 본 논문에서 제안하는 태깅 알고리즘으로 정했다.

$$(6)\quad T(w_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t'_i|t'_{i-1})P(w_i|t'_i,t'_{i-1})$$

$$P(t'_i|t'_{i-1}) \cong \frac{freq(t'_i,t'_{i-1})}{freq(t'_i)}, P(w_i|t'_i,t'_{i-1}) \cong \frac{freq(w_i,t'_i,t'_{i-1})}{freq(t'_i,t'_{i-1})}$$

$P(w_i|t'_p t'_{i-1})$을 어휘확률로 이용할 때는 $P(w_i|t'_i)$을 어휘확률로 이용할 때보다 더 많은 학습 데이타를 요구하므로 자료희귀 문제가 발생할 수 있다. 그러므로 다음과 같이 값을 보정하여 자료희귀 문제에 대처하였다 [Charniak93a].

$$(7)\quad P(w_i|t'_i,t'_{i-1}) \cong \lambda P(w_i|t'_i,t'_{i-1}) + (1-\lambda)P(w_i|t'_i) \qquad 0 \le \lambda \le 1$$

5.3 제안하는 알고리즘에 의한 변별력 향상

본 논문에서 제안된 방법에 의해 기대되는 성능 향상의 경우를 분류하면 다음과 같다.

5.3.1 어절 구조의 반영을 통한 성능향상

(가) 형태소와 어절 관계의 호응을 보아야 하는 경우

자신은 보통명사(자신감)로도 쓰이고 인칭대명사(자기 자신)로도 쓰이는 모호성을 갖는 단어이다. 그러나 주위의 형태소만 고려할 때 보통명사와 인칭대명사는 인접하여 사용되는 품사들이 비슷하므로 주위의 품사들만을 고려해서는 해결될 수가 없다. 이에 반해 주위의 어절들을 고려할 경우 자신(보통명사)은 목적격으로 주로 쓰이고(예: 자신을 가지고) 자신(인칭대명사)의 경우는 부사격(예: 자신에게 말했다)이나 관형격(예: 자신의 책) 혹은 주격(예: 자신이 입던 옷)으로 많이 쓰인다.

즉 P(자신 | 보통명사)와 P(자신 | 인칭대명사) 값만을 이용해야 하는 경우를 어절구조를 망에 반영함으로써 P(자신 | [주절,보통명사])와 P(자신 | [주절,인칭대명사]) 또는 P(자신 | [관형절,보통명사]) 와 P(자신 | [관형절,인칭대명사])으로 혹은 P(자신 | [목적절,보통명사])와 P(자신 | [목적절,인칭대명사])으로 세분하여 구별할 수 있으므로 더 큰 변별력을 가질 수 있다.

(나) 어절내에서 형태소 조합 정보를 이용하는 경우

그대로가 좋다에서 그대로를 형태소 분석하면 그대로(부사어)+가(주격조사)와 *그대로(부사어)+가(동사)+아(보조적 연결어미) 등의 중의성을 가진다. 이때 [실험 1]과 같은 환경에서는 P(그대로 | 부사어)*P(주격조사 | 부사어)*P(가 | 주격조사)*P(동사 | 주격조사)와 P(그대로 | 부사어)*P(동사 | 부사어)*P(가 | 동사)*P(보조적 연결어미 | 동사)를 비교하게 되고 두번째의 경우가 더 많이 사용되는 형태이므로 그대로(부사어)+가(동사)+아(보조적 연결어미)를 적합한 예로 잘못 추정하게 된다. 그러나 [실험 2]와 같은 환경에서는 P(그대로 | [PC,부사어])*P([PC,동사] | [PC,부사어])*P(가 | [PC,동사])*P([PC,보조적연결어미] | [PC,동사])를 고려할 때 PC어절내에서는 부사어+동사+보조적 연결어미는 자주 발생하지 않는 형태이므로 낮은 값을 가지게 된다. 그

러므로 그대로(부사어)+가(주격조사)로 정확하게 추정을 할 수 있게 된다.

5.3.2 어휘확률의 확장을 통한 성능향상

준을 형태소 분석하면 주다(동사)+ㄴ(관형형 어미)와 줄다(동사)+ㄴ(관형형어미)의 모호성을 가지게 된다. 주다와 줄다는 모두 동사이므로 문맥확률이 같게 된다.

이 경우에 모호성 해소가 P(주다 | 동사)와 P(줄다 | 동사)에만 의존하게 되므로 오류가 생길 수 있다. 이에 반해 주위 어절을 고려할 경우 주다는 목적절(예: 돈을 주다)이나 부사절(예: 그에게 주다)과 많이 쓰이고 줄다는 주절(예: 식사량이 줄다)과 많이 쓰인다는 정보를 이용해 더 큰 변별력을 가지게 된다.

즉 P(주 | 동사)와 P(줄 | 동사)값만을 이용하여 구별해야 되는 경우를 어휘확률을 확장시키면 P(주 | 목적격조사, 동사)와 P(줄 | 목적격조사, 동사), P(주 | 부사격조사, 동사)와 P(줄 | 부사격조사,동사), 그리고 P(주 | 주격조사, 동사)와 P(줄 | 주격조사,동사)로 세분하여 구분할 수 있으므로 더 큰 변별력을 갖게 된다.

6. 실험 및 결과분석

6.1 실험의 목적및 환경

실험의 목적은 제안하는 모델이 충분히 훈련된 같은 환경 하에서 기존의 모델보다 더 정확하게 태깅할 수 있음을 보이는데 있다. 한국어의 전체 현상을 반영하는 대량의 균형된(balanced) 코퍼스를 이용하여 실험하는 것이 이상적이겠지만 현실적인 제약으로 인해서 이번 실험에서는 5만5천 어절을 이용하여 은닉 마르코프 모델을 학습시키고 이 중에서 5천 어절을 임의로 추출하여 실험 데이타로 이용하였다. 5만 어절로 학습하였을 때 어절구조가 반영된 은닉 마르코프 모델은 413개의 노드와 10,421개의 링크로 구성되어 있다.

태깅에 사용되는 품사의 갯수는 총 52개였고[김재훈], 사용한 형태소 해석기는 형태소들 간에 결합될 수 있는 모든 가능성을 유효한 결과로 내주기 때문에 의미적으로 결합이 불가능한 결과도 내줄 수 있다. 어절당 평균 모호성의 수는 약 3.5개이다. [이상호]

태깅의 정확도는 사용하는 품사집합과 학습 데이타에 크게 달라질 수 있으므로 수치적으로 주어지는 정확률만으로는 객관적인 평가가 어렵다. 그러므로 실험은 상대적인 평가를 위해서 다음의 네 가지 모델을 같은 환경하에서 실험하여 비교하였다. [실험1]은 기존의 시스템에서 주로 이용되어 왔던 방법으로 형태소 단위의 은닉 마르코프 모델을 이용한 실험이고 [실험2]는 4-(2)에서 제시했던 방법 중에서 어절구조만을 망구조에 반영한 은닉 마르코프 모델을 이용하여 실험한 것이다. [실험3]은 형태소 단위의 은닉 마르코프 모델에서 어휘확률을 확장시킨 은닉 마르코프 모델을 이용한 실험이고 [실험4]는 어절구조를 반영한 은닉 마르코프 모델에서 어휘확률을 확장시킨 모델을 이용하여 실험한 것이다. [실험1]과 [실험2] 그리고 [실험3]과 [실험4]를 비교함으로써 어절구조를 망구조에 반영하였을 때의 성능향상을 측정할 수 있다.

[실험1] 형태소 단위의 모델

$$T(t_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i)$$

[실험2] 어절구조가 반영된 모델

$$T(t_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t'_i|t'_{i-1})P(w_i|t'_i)$$

[실험3] 어휘확률이 확장된 모델

$$T(t_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i, t_{i-1})$$

[실험4] 어절구조, 어휘확률이 반영된 모델

$$T(t_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t'_i|t'_{i-1})P(w_i|t'_i, t'_{i-1})$$

6.2 실험결과

위 각각의 실험에 대하여 사람의 태깅과 비교하면 다음과 같은 정확률을 보인다.

실험 모델	어절단위의 정확률
실험 1	97.26 %
실험 2	97.34 %
실험 3	97.46 %
실험 4	98.26 %

<표1> 각 모델별 성능비교

위의 결과에서 어절구조를 반영한 은닉 마르코프 모델이 일반적으로 성능이 향상을 볼 수 있다. 특히 어절구조만 반영하거나[실험2] 어휘확률만을 확장하여[실험3] 모델을 구성했을 때보다 두 가지를 조합한 모델의 경우 [실험4]가 급격한 성능향상을 보였다. 이는 어절구조를 반영함으로써 얻어지는 정보와 어휘확률을 확장함으로써 얻어지는 정보를 종합하여 이용할 때 변별력의 뚜렷한 증가를 가져온다는 결론을 내릴 수 있다.

7. 결 론

본 논문에서는 한국어의 특성을 반영한 태깅을 위한 은닉 마르코프 모델구성 방법에 대해서 소개하였다. 한국어가 형태소와 어절단위로 구성되는 점에 입각한 제안된 방법은 논리적 동기가 충분할 뿐만 아니라 실험적으로도 입증이 되었다. 은닉 마르코프 모델의 노드수가 많아짐으로써 더 많은 학습 데이타가 필요하다는 점이 단점이 될 수 있으나, 가까운 미래에 많은 학습데이타를 이용할 수 있을 것이라고 본다면 커다란 문제점은 아니라고 할 수 있다.

한국어의 특성으로 인해 영어태깅보다 복잡한 은닉 마르코프 모델구축 방법을 제시하였지만, 이와 비슷한 작업이 영어에서는 할 수 없는데 반해 한국어에서는 어절 정보라는 태깅하기에 손쉬운 정보를 찾을 수 있다는 긍정적인 결론을 내릴 수 있다.

참고문헌

[Charniak93a] E. Charniak, C. Hendrickson, N. Jacobson, M. Perkowits, "Equations for Part-of-Speech Tagging", *AAAI-93*, 1993.

[Charniak93b] E. Charniak, "Statistical Language Learning" , *The MIT Press*, 1993.

[Kupiec92] J. Kupiec, "Robust Part-of-Speech Tagging Using a Hidden Markov Model", *Computer Speech and Language*, vol. 6, pp. 225-242, 1992.

[Marcus93] M. P. Marcus, B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linquistics*, Vol. 19, No.2, 1993.

[Rainber89] L. R. Rainber, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", *Proc. of the IEEE*, Vol 77, No.2, pp.257-286, Feb. 1989.

[김재훈94] 김재훈, 서정연, '자연언어 처리를 위한 한국어 품사 태그', 한국과학기술원, 인공지능 연구센타, CAIR-TR-94-55, 1994.

[김충원94] 김충원, 임권묵, 송만석, '의미 정보를 이용한 형태소 중의성 해결', 한국정보과학회 가을 학술발표논문집, Vol. 21, No. 2, pp.649-652, 1994.

[남기심85] 남기심, '표준국어 문법론', 탑 출판사, 1985.

[이운재93] 이운재, 최기선, 김길창, '한국어 문서 태깅 시스템', 정보과학회 봄 학술발표논문집, Vol. 20 No. 1 pp. 805-808, 1993.

[이하규94] 이하규, 김영택, "통계정보에 기반을 둔 한국어 어휘중의성해소", 한국통신학회논문지 '94-2 Vol. 19 No. 2, 1994.

[임철수94] 임철수, '은닉 마르코프 모델을 이용한 한국어 품사태깅 시스템 구현', 한국과학기술원 석사학위 논문, 1994.

[이상호94] 이상호, 김재훈, 조정미, 서정연, '부분 분석 결과를 공유하는 한국어 형태소 분석', 제 11회 통신 및 신호처리 워크샵 논문집, 1994.