(제 10회 한글 및 한국어 정보처리 학술대회)

최대 엔트로피 모델을 이용한 한국어 품사 태깅

강인호* 김재훈** 김길창*
*한국과학기술원 전산학과
{ihkang, gckim}@csone.kaist.ac.kr
**한국해양대학교 컴퓨터공학과
jhoon@hanara.kmaritime.ac.kr

Korean Part-Of-Speech Tagging based on Maximum Entropy Model

In-Ho Kang* Jae-Hoon Kim** Gil Chang Kim*
*Department of Computer Science, KAIST
**Department of Computer Engineering, Korea Maritime University

요약

주어진 문자열에 품사를 정해주는 방법으로 현재 많이 사용되고 있는 것 중의 하나로 통계적 방법을 들 수 있다. 대부분의 통계적 방법은 품사 태깅을 위해 주변 품사열만으로 이뤄진 단순한 정보를 사용하고 있는데, 품사 태깅 문제는 본래 품사열 정보 뿐 아니라 단어에 대한 어휘 정보, 통사 정보, 연어 정보 등 다양한 정보들이 종합되어야 하는 문제이다.

이에 본 논문에서는 품사 태깅에 유용한 정보를 정형화하여 성능 향상을 얻어내는 방법을 제안한다. 제안된 방법은 먼저 품사열 정보만을 이용한 품사 태깅의 주된 오류인 조사, 용언, 연결어미의 구분 문제와 복합어의 형태소 분석 문제를 해결하기 위한 정보를 품사 분류 기준으로부터 얻어낸다. 얻어낸 정보들은 정형화 과정을 거쳐 최대 엔트로피 모델의 자질로 사용된다. 이렇게 얻어낸 모델을 가지고 수행된 실험 결과, 품사열 정보만을 이용한 품사태깅보다 좋은 성능을 얻을 수 있었다.

1 서론

품사는 문법적으로 성질이 공통된 단어끼리 모아 놓은 단어의 갈래를 말하는 것으로서, 형태소 해석이나 구문 해석을 하는데 있어서 중요한 정보가 된다. 주어진 단어의 품사를 결정하기 위해서는 전체 문서의 내용과 언어지식 그리고 인간의 기본 지식이 함께 사용되어야 한다. 그렇지만 이를 구현하기란 쉬운 일은 아니다. 따라서 현재 사용되고 있는 품사 태깅 모델의 대부분이 단어들의 품사는 주변의 품사열정도의 제한된 정보 만으로 결정될 수 있다는 가정을 사용하고 있다[9][10][11].

그러나 본래 품사 분류의 기준으로는 일반적으로 의미, 기능, 형식의 셋을 들 수 있다. 여기서 의미란 개별 단어의 어휘적 의미가 아닌 형식적인 의미로서 사물의 이름을 나타내느냐 그렇지 않으면 움직임이나 성질, 상태를 나타내느냐 하는 것이다[14]. 그리고 기능은 한 단어가 문장 가운데서 다른 단어와 가지는 관계를 가리키며 형식은 단어의 형태적 특징을 의미한다. 본 논문에서는 이러한 품사 분류의 기준이 되는 의미, 기능, 형식을 함축하는 정보를 품사열 정보와 결합시켜 좋은 성능을 얻어내는 방법을 제안한다. 본 논문의 구성을 보면 다음과 같다. 2절에서는 본 논문에서 해결하고자 하는 문제에 유용한 정보를 보이고, 3절에서는 품사 태깅 모델을 위한 최대 엔트로피 모델 설명과 정보의 정형화 방법을 보인다.

2 관련 연구

2.1 기존의 확률 모델

W를 단어열이라고 하고 T를 그 단어열에 해당하는 품사열이라고 할 때 품사 태깅의 문제는 다음과 같이 품사열 $\Phi(W)$를 찾는 문제로 정의 된다.

$$\Phi(W) = \arg\max_{T} P(W \mid T)P(T)$$

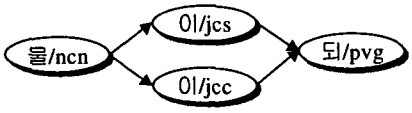
단, $W = w_1, w_2, ..., w_n \quad T = t_1, t_2, ..., t_n$

w_i는 i번째 위치한 단어이고, t_i는 그 단어에 해당하는 품사이다. 현재 나와 있는 대부분의 품사 태깅 모델이 위의 수식을 기반으로 다양하게 근사화 시켜 사용하고 있다. 전체에 대한 확률을 부분의 확률값의 곱으로 나타내는 등 대부분 **n-gram**의 결합형태로 나타난다**[6][9][10][11][12].**

2.2 문제점

주변 품사열 정보만을 이용해서 품사를 결정할 수 있다는 가정을 가지고 품사 태깅을 할 경우, 품사의 애매성을 해결하지 못하는 경우가 있다. 앞 단어의 품사 정보인 바이그램**(bigram)**과 어휘의 품사 확률 정보인 유니그램**(unigram)**을 사용하는 품사 태깅 모델을 예로 들어보자. "*얼음이 물이 되었다.*" 라는 문장에 대해서 '이'가 보격 조사**(jcc)**인지 주격 조사**(jcs)**인지를 결정하기 위해서, 바로 앞 품사인 일반 명사**(ncn)** 정보와 형태소 '이' 의 유니그램 정보를 사용한다**(**그림 **1).** 그리고 다음 형태소인 '되'의 품사를 결정하면서 앞의 품사를 고려하지만 '되**/pvg**'라는 형태소가 보격 조사를 필요로 한다는 특성은 이러한 품사열에는 나타나지 않는다. 즉 보격 조사에 대한 확률값을 올려줄 정보가 없다.

이와 같이 품사열 정보만으로는 품사를 결정하지 못하는 경우가 있음을 알 수 있다. 이와 유사한 경우들을 들면 주격 조사와 보격 조사의 구분, 접속격 조사와 공동격 조사의 구분, 대등적 연결어미와 종속적 연결어미, 본용언과 보조용언, 서술 명사와 비서술 명사의 구분, 복합명사에서의 형태소 결정 문제 등이 있다. 이들은 품사열을 기반으로 하는 태깅 시스템에 빈번한 오류들로써**[12]** 품사열만으로는 쉽게 해결이 되지 않는다. 이러한 품사들의 실제 분류 기준과 말뭉치를 만들 때 사용할 수 있는 정보들을 보면 다음과 같다**[6][7][14].**



ncn jcs + 이/jcs　　jcs pvg + 되/pvg
ncn jcc + 이/jcc　　jcc pvg + 되/pvg

그림 1　품사 태깅의 예

■ **주격/보격 조사**

보격조사는 형태적으로 주격조사 '이/가'와 같기 때문에 혼동이 될 수 있다. 그러나 서술격 조사의 부정어인 '아니다'와 동사 '되다'의 지배를 받는 '이/가'라는 점에서 주격조사와 구분 할 수 있게 된다.

■ **접속격/공동격 조사**

조사 '와/과' 그리고 '하고'는 공동격 조사로써, 형태적으로 접속격 조사와 같기 때문에 혼동이 된다. 그러나 공동격조사 '와/과'나 '하고'는 그 뒤에 '만나다, 부딪치다, 헤어지다, 사귀다, 싸우다, 친하다, 닮다, 비슷하다, 다르다, 어긋나다' 등과 같은 이른바 교호성이 있는 대칭동사들이 오는 특징을 가진다**[6].**

■ **대등적/종속적 연결어미**

대등적 연결어미는 나열, 선택, 반복 등의 의미로 두 용언을 연결하고, 종속적 연결어미는 두 문장을 주종관계로 연결시키는 말로서 구속, 가정, 이유, 필연, 방임, 양보, 설명, 도급, 추정, 의도, 비교 등의 의미로 서로 연결한다. 이는 어휘적으로는 구분하기가 힘들며, 다만 앞 절이 뒷 절 속으로 자리 옮김이 가능하면 종속적인 접속으로 볼 수 있다. 이에 연결어미를 중심으로 한 주위의 의미적 관계를 알아야 한다.

■ **본 용언/보조 용언**

자립성을 가지고 실질적인 의미를 나타내며 단독으로 서술능력을 가지는 동사와 형용사를 본용언이라 이르며, 보조 용언은 겉 모습은 일반적인 동사나 형용사와 다름이 없지만 자립성이 없거나 약하여 본 용언에 기대어 그 말의 뜻을 도와 주는 동사와 형용사로서 '-게 되다, -게 만들다, -게 지다' 등이 있다. 보조 용언은 같이 나타나는 본 용언을 이용하여 보조 용언과 구분할 수 있다. 그러나 많은 경우에 있어 의미 정보가 필요하다.

■ **서술성 명사/ 비서술성 명사**

서술성 명사는 '-하다'나 '-되다'가 붙어서 동사나 형용사가 된다. 예를 들어 '말하다'에서 '말'은 애매성을 가지는 형태소로서 비서술성 명사인 말(馬)인지, 서술성 명사인 말(言)인지 구분이 힘들다. 그러나 '하다'를 이용한다면 서술성 명사 '말(言)'인 것을 알 수 있듯이 의미정보를 이용하면 애매성을 많이 줄일 수 있다.

■ **복합 명사**

독립적으로 사용될 수 있는 두 개 이상의 어근이 결합하여 하나의 단어를 형성하는 복합명사의 경우, 형성 방식이 매우 다양하며, 복잡한 형태의 의미적 구조를 가지고 있다**[13].** 예를 들어 '*농산물생산자수*' 라는 어절에 대해서 발생할 수 있는 형태소 분석 결과인 '*농산물+생산자+수*' 와 '*농산물+생산+자수*'를 구분하는 것은 품사열만으로는 할 수가 없다. 그러나 인접된 각 형태소의 상호연관성을 이용하면 '*생산*'과 '*자수*'가 결합되는 경우를 제거할 수 있다.

이렇게 위의 문제들은 품사열만으로 해결될 수 있는 문제가 아니다. 따라서 품사열 정보가 아닌 품사의 분류기준이 함축된 정보가 품사 태깅에 사용되어야 한다. 그러나 이러한 정보들이 서로 어떠한 관계를 가지며 품사 결정에 어떻게 관여하는지는 알 수 없다. 최대 엔트로피 모델(**Maximum Entropy Model**)은 주어진 정보들이 영향을 주는 것은 알지만 어떤 관계가 있는지 정확하게 알지 못할 때 적합한 모델로서 동형정보(**Homogeneous Information**)가 아니더라도 쉽게 결합하여 확장할 수 있다는 장점을 가지고 있다[2][3][4][11]. 이에 본 논문에서는 앞에서 제시한 문제들을 해결하기 위해 추가되는 정보를 최대 엔트로피 모델을 이용해서 결합한다.

3 최대 엔트로피 모델

3.1 품사 태깅을 위한 최대 엔트로피 모델

본 논문에서는 품사 태깅을 주어진 어절열에 대해 가장 적당한 형태소 분석을 구하는 것으로 최대 엔트로피 모델에 기반하여 정의한다. 즉 주어진 문장(S)에 대해서 다음의 확률값을 가장 좋게 하는 품사열 $\{t_1, ..., t_n\}$을 찾는 것이다.

$$p(t_1, \ldots, t_n \mid S) = \prod_{i=1}^{n} p(t_i \mid h_i)$$

여기서 h_i 는 품사 t_i를 계산할 때 사용할 수 있는 문맥 환경을 말한다.

단, $$p(t \mid h) = \frac{p(h,t)}{\sum_{t' \in T} p(h,t')} \qquad (3.1.1)$$

T는 발생 가능한 모든 품사를 말한다.여기에서 $p(h,t)$는 최대 엔트로피 모델에서 얻어질 수 있다[2][3][4].

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$$

π 는 정규화를 위한 상수이고 $\{\mu, \alpha_1, \ldots, \alpha_k\}$ 는 모델 파라미터(**parameter**)이며, $\{f_1, \ldots, f_k\}$ 는 자질(**feature**)들로서 $f_j(h,t) \in \{0,1\}$ 이다. 각각의 자질 f_j 에는 해당하는 α_j 가 존재하며, 주어진 문맥에서 이 자질이 미치는 영향정도를 나타낸다.

최대 엔트로피 모델은 최대 엔트로피 원리(**Maximum Entropy Principle**)에 기반하여 만들어진다. 최대 엔트로피 원리란 랜덤 변수 x_i (**i=1, 2, ..., n**)에 대한 확률 분포를 p_i라고 할 때, 자질 f에 대해서 아래와 같은 제약조건을 가한다[2].

$$E[f_j] = \tilde{E}[f_j], \quad 1 \le j \le k$$

$$E[f_j] = \sum_{h \in H, t \in T} p(h,t) f_j(h,t)$$

$$\tilde{E}[f_j] = \sum_{i=1}^{n} \tilde{p}(h_i, t_i) f_j(h_i, t_i)$$

여기에서 H와 T는 있을 수 있는 모든 문맥과 품사의 집합이다. 위의 수식은 모델에서의 자질의 평균값 $E[f_j]$와 학습 데이터에서 얻어낸 경험치에 의한 자질의 평균값 $\tilde{E}[f_j]$가 일치해야 된다는 제약 조건을 뜻한다. 여기에서 $\tilde{p}(h_i, t_i)$는 학습 데이터에서 얻어낸 것을 말한다. 그러나 계산하는 과정에 있어서 H가 너무 크기 때문에 평균값을 바로 구하는 것이 힘들다. 그래서 아래와 같이 근사화 시킨 수식을 이용하여 계산한다.

$$E[f_j] = \sum_{i=1}^{n} \tilde{p}(h_i) p(t_i \mid h_i) f_j(h_i, t_i)$$

여기서 $\tilde{p}(h_i)$는 학습데이터에서 뽑아낸 값을 말한다. 그리고 p_i는 확률값이기 때문에 합쳐서 **1**이 된다는 제약 조건이 주어진다.

이렇게 두 개의 제약조건을 만족하는 확률 분포 p_i값들 중에서 확률 분포 p_i의 엔트로피가 최대가 되도록 모델을 구성하는 것이 최대 엔트로피 원리이다.

$$H(p) = -\sum_{h \in H, t \in T} p(h,t) \log p(h,t)$$

엔트로피를 최대로 만든다는 것은 알려진 모든 정보를 이용하면서 특정 부분에 치우치지 않는 분포를 구하겠다는 뜻이다. 최대 엔트로피 원리에 의한 파라미터 추정법은 **Jaynes[1]**에 의해 제시되었고, 이를 수치적으로 측정하는 **GIS(Generalized Iterative Scaling)**방법이 **Darroch**와 **Ratcliff[2][4]**에 의해 고안되었다.

3.2 미등록어를 고려한 모델

처리 중인 문장에 사전에 있지 않은 단어가 사용될 수 있는데 이를 미등록어라고 한다. 미등록어 문제의 자연스러운 접근 방법은 미등록어가 갖고 있는 특징으로 그 품사를 예측하는 것이다. 영어 품사 태깅의 경우에는 특정 접미사(**suffix**)가 나타난 단어들의 확률 분포를 보거나, 대/소문자의 여부, 하이픈의 사용 여부 등의 정보를 이용한다[3][4][11]. 그러나 한국어에는 해당하지 않기 때문에, 제일 낮은 확률값을 일률적으로 주거나[9], 미등록어 주위의 형

태소 및 품사를 이용하여 확률값을 추정하기도 했다[10].

최대 엔트로피 모델은 주어진 환경에 대해 어떤 사건의 발생 가능성을 계산하는 방식이다. 즉 실제 세계에서 발생할 수 있는 다양한 환경에 대해서 학습을 시킬수록 좋은 결과를 보인다. 이에 학습 말뭉치의 일부분(90%)에 속하지 않는 어휘들을 미등록어로 표시하여, 미등록어가 발생하는 실제 세계에 대해서 학습을 시킨다. 이러한 인위적인 미등록어로 인해 본 논문에서 사용할 여러 자질들에 대해서, 미등록어가 발생하여 적용이 되지 않는 경우와 형태소 정보를 볼 수 없는 경우 등의 다양한 환경을 만들어 미등록어일 경우에도 자연스럽게 학습이 된다.

3.3 문제 해결을 위한 특수 자질들

앞에서 논의된 품사열 정보만을 이용하는 품사 태깅에서 많은 문제가 되고 있는 품사들에 대해서 품사의 분류 기준이 함축된 정보를 정형화하여 최대 엔트로피 모델의 자질로 사용한다.

■ **어휘 정보**

이는 일반적인 **n-gram** 형태에 형태소까지 같이 보는 형태가 된다. 보격조사를 필요로 하는 '되다'나 '아니다'와 같이 어휘의 고유한 정보를 넣을 수 있게 된다. 또한 공동격 조사와 같이 나타나는 용언의 정보도 포함할 수 있게 되어 접속격 조사와 구분할 때 도움이 된다.

$$f(t_i, h) = \begin{cases} 1, & t_{i-1}=X \ \& \ t_i=Y \ \& \ m_i=M \\ 0, & otherwise \end{cases}$$

■ **통사 정보**

조사를 중심으로 한 품사열의 정보 형태가 된다. 예를 들어 접속격 조사와 공동격 조사는 조사를 중심으로 한 품사들의 대칭성을 가지고 어느 정도 유추할 수 있다. 즉 명사와 명사 또는 동사와 동사가 같이 나타나면 접속격 조사, 그렇지 않고 앞에서 제시된 용언들이 따라오거나 하면 공동격 조사가 될 확률이 높다. 이렇게 조사를 중심으로 품사들을 조사함으로써 접속격 조사와 공동격 조사의 구분외에 어절간의 연결관계 또한 얻어 낼 수 있다.

$$f(t_i, h) = \begin{cases} 1, & t_{i-1}=X \ \& \ t_{i+1}=Z \ \& \ t_i=Y \ \& \ m_i=M \\ 0, & otherwise \end{cases}$$

■ **연어 정보**

대등적/종속적 연결어미나 서술성/비서술성 명사는 그 의미가 고려 되어지지 않는 한 해결하기 어렵다. 하지만 주위 연어정보를 가지고 판단을 하면, 많은 도움을 줄 수 있을 것이라는 가정하에 주변의 어휘를 같이 보는 정보를 추가한다. 또한 복합어에 있어 형태소끼리의 연관성에 대해서도 정보를 줄 수 있게 된다.

$$f(t_i, h) = \begin{cases} 1, & m_{i-1}=M \ \& \ m_i=N \ \& \ t_i=Y \\ 0, & otherwise \end{cases}$$

이렇게 정형화된 자질들을 말뭉치에서 추출하는 것을 예를 들면 아래와 같다.

중국/nq+의/jcm 대기/ncn+오염/ncn+이/jcs

'오염/ncn'에서 뽑아낼 수 있는 품사열 자질[4]과 특수 자질들은 아래와 같다. 여기서 '오염'은 조사가 아니기 때문에 통사 정보는 만들어지지 않는다.

정보유형	뽑혀진 자질
Bigram	t_{i-1} = ncn & t_i = ncn
Trigram	$t_{i-2}t_{i-1}$ = jcm ncn & t_i = ncn
Unigram	m_i = '오염' & t_i=ncn
어휘정보	t_{i-1}= ncn & t_i=ncn & m_i= '오염'
연어정보	m_{i-1}='대기'& m_i= '오염' & t_i=ncn

표 1 학습 말뭉치에서 추출되는 자질의 예

4 실험 및 결과 분석

4.1 학습 말뭉치

본 논문에서 사용하는 품사 태그는 통합 국어정보베이스[7]에서 정의한 것을 기반으로 한다. 총 56개의 품사 태그로 구성되어 있다. 실험에 사용하는 **KAIST** 말뭉치의 특성을 간략히 살펴보면 아래와 같다.

#	출처	문장 수	어절 수	형태소 수
1	동아일보 사설	3,120	36,169	81,324
2	한겨레신문 생활면	283	4,259	9,989
3	농민이야기 주머니(조성우)	3,239	41,666	87,018
4	국민학교 6학년 교과서	5,485	50,208	108,175
5	㈜한글과컴퓨터의 홍보물	115	2,729	5,690
6	겨울이야기 (소설)	4,881	40,493	89,356
	전체	17,123	175,524	381,552

표 2 KAIST 말뭉치 요약

본 논문에서는 말뭉치를 세 부분으로 나누어서 사용한다. 전체의 **10%**는 시험용 말뭉치로 남기고, 또 다른 **10%**는 시스템 성능 개발을 위한 개발용 말뭉치로 가지며 나머지는 학습용으로 사용된다.

4.2 실험

본 논문에서는 세가지의 모델을 실험한다. 본 논문에서 제시하고 있는 모델과의 비교 평가를 위해 현재 사용되고 있는 확률 태깅 모델 중 좋은 성능을 보이고 있는 가중치 망 기반의 품사 태깅 모델(모델A)[12]을 같이 실험한다. 모델A는 트라이그램 정보까지 사용한다. 그리고 추가된 자질과 최대 엔트로피 모델의 유용성을 보이기 위해 트라이그램, 바이그램, 그리고 유니그램을 사용했을 때의 최대 엔트로피 모델(모델B)과 앞에서 논의 되어진 특수 자질을 추가한 모델(모델 C)을 실험한다. 본 논문에서 사용하는 형태소 분석기는 김재훈[12]의 초기 모델을 사용한다.

표[3]은 가중치 망 모델과 본 논문에서 얘기되어진 자질을 사용한 최대 엔트로피 모델의 품사 태깅 결과이다. 여기서 정확률은 어절 단위로 계산되었다. 아래 표[4]는 각 모델에서 발생한 오류 중의 일부를 나타내는 것으로서 본 논문에서 사용한 특수 자질들의 유용성을 보기 위함이다.

통계치의 종류		모델A	모델B	모델C
어절 (18,536)	오류수	1629	2544	1600
	정확률	91.21%	86.28%	91.37%
형태소 (39,817)	오류수	2243	3119	2150
	정확률	94.37%	92.17%	94.60%

표 3 실험 결과1

오류 유형	모델 A	모델 B	모델 C
보격/주격 조사	66	63	21
접속격/공동격 조사	32	48	26
대등적/종속적/보조적 연결어미	165	432	291
서술성/비서술성 명사	58	172	55
본용언/보조용언	83	387	122
복합 명사	32	37	31

표 4 오류 유형 개수1

위와 같이 다소 낮은 정확도를 보이는 이유는 학습 말뭉치가 다영역으로 구성되어 있어 미등록어가 많았기 때문이다. 형태소 분석기의 성능을 보이면 아래와 같다.

	후보 개수/어절	정확률
확장 전	6.01	96.1%
확장 후	6.08	99.2%

표 5 형태소 분석 정확도

표[5]에서와 같이 형태소 사전을 확장하였을 경우에 같은 실험을 한 결과는 표[6]과 같다.

통계치의 종류		모델A	모델B	모델C
어절 (18,536)	오류수	917	1174	872
	정확률	95.05%	93.67%	95.29%
형태소 (39,817)	오류수	1194	1351	1025
	정확률	97.00%	96.60%	97.43%

표 6 실험 결과2

오류 유형	모델 A	모델 B	모델 C
보격/주격 조사	70	70	8
접속격/공동격 조사	39	36	19
대등적/종속적/보조적 연결어미	164	189	179
서술성/비서술성 명사	10	15	8
본용언/보조용언	106	179	95
복합 명사	9	20	11

표 7 오류 유형 개수2

위 실험의 결과를 통해 새로 추가된 자질들이 품사 태깅에 유용한 정보임을 알 수 있다. 또한 제시된 모델이 현재 확률 정보를 사용하는 모델 중 좋은 성능을 보이고 있는 가중치망 모델보다 나은 결과를 나타내고 있다.

5 결론 및 앞으로의 연구

본 논문에서는 품사 태깅에 유용한 정보들을 정형화하고 이를 최대 엔트로피를 이용해서 결합시키는 방법을 제안했다.

품사열만을 사용하는 품사 태깅 시스템은 주격/보격 조사의 구분, 접속격/공동격 조사의 구분, 보조용언/본용언의 구분, 대등적/종속적 연결어미의 구분, 그리고 복합명사의 형태소 해석등과 같은 문제를 잘 해결하지 못한다. 이에 품사 분류 기준을 함축한 어휘 정보, 연어 정보, 통사 정보를 정형화하여 말뭉치에서 뽑아내었다. 이렇게 뽑혀진 자질들은 최대 엔트로피 원리를 이용해서 기존의 품사열 정보와 결합되어져서 최대 엔트로피 모델을 구성한다. 약 19,000어절에 대해서 실험을 한 결과, 품사 태깅의 정확률이 24.4% 향상된 97.43%를 나타내었고, 이는 가중치망 모델보다 나은 결과를 보인다.

본 논문의 실험을 통해 품사 태깅에 유용한 자질이 주어질 경우 이를 최대 엔트로피 모델을 이용하여 성능 향상을 가져올 수 있음을 알았다. 본 논문에서 제안된 방법을 통하여 품사 태깅에 유용한 자질들을 합칠 수 있기 때문에, 이제는 정보들을 어떻게 결합하는가 보다는 품사 태깅에 도움이 되는 유용한 정보들을 찾는 작업이 중요할 것이다. 그러나 무조건적인 정보의 추가는 오히려 모델 전체의 성능 저하를 가지고 올 수 있다. 기존의 자질들과 될 수 있는 한 영향을 주지 않으면서 상호 보완하여 전체 성능을 향상 시킬 수 있는 자질을 선택할 필요(feature selection)가 있다. 즉 가능성이 있는 자

질들에 대해서 도움을 주는 정도를 순서화하여 뽑아내는 선택 과정이 필요하다.

또한 본 논문에서는 바로 주변 단어의 연어 정보를 이용해서 의미적 정보를 추출해내려고 하였다. 그러나 보다 나은 정확률을 위해서, 바로 주변의 문맥이 아닌 좀 더 큰 범위의 문맥에서도 연관된 내용을 가져와서 사용할 수 있게 모델링하는 방법도 연구되어져야 한다.

마지막으로 한국어 품사 태깅은 영어 품사 태깅처럼 단어가 고정되어 있고 그 단어의 품사를 결정하는 것이 아니라 형태소까지 결정하면서 품사도 결정한다. 따라서 최대 엔트로피 모델을 구성하고 품사를 구하는 과정 또한 이러한 형태소 정보도 포함하는 형태로 나타나야할 것이다.

참고문헌

[1] **E.T. Jaynes. 1957. Information Theory and Statistical Mechanics, *Physics Reviews106, 620-630***

[2] **J.N. Darroch and D Ratcliff. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43:1470-1480**

[3] **Ronald Rosenfeld. 1994. Adaptive Statistical Language Modeling: A Maximum Entropy Approach. *CMU-CS-94-138***

[4] **Adam L.Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra 1996. A Maximum Entropy Approach to Natural Language Processing. *In Computational Linguistics vol. 22.***

[5] **Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. *In proceedings of the Empirical Methods in Natural Language Processing Conference***

[6] 김재훈, 김덕봉 등, **1996** 통합국어정보 베이스를 위한 한국어 형태·통사 태그 설정. *한국과학기술원 전산학과 컴퓨터 시스템 실험실 내부메모*

[7] 과학기술처 **1996.** 통합 국어 정보 베이스 제**2**차년도 최종보고서

[8] 이공주, 김재훈 등 **1996.** 한국어 구문 트리 태깅코퍼스 작성을 위한 구문태그. *한국과학기술원 전산학과 기술문서(CS-TR-96-102)*

[9] 이운재 **1993.** 한국어 문서 태깅 시스템의 설계 및 구현. *석사학위논문, 한국과학기술원 전산학과*

[10] 이상호 **1995.** 미등록어를 고려한 한국어 품사 태깅 시스템 구현. *석사학위논문, 한국과학기술원 전산학과.*

[11] 정성영 **1996.** 마코프 랜덤 필드를 이용한 영어 품사 태깅 시스템. *석사학위논문, 한국과학기술원 전산학과*

[12] 김재훈 **1996.** 오류-보정 기법을 이용한 어휘 모호성 해소. *박사학위논문, 한국과학기술원 전산학과*

[13] 정동환 **1993.** 국어 복합어의 의미 연구, *서광학술 자료사*

[14] 남기심, 고영근 **1994.** 표준 국어 문법론, *탑출판사*