최대 엔트로피 모델 기반 품사 태거의 성능 향상 기법

조민희[0] 김명선 박재한 박의규 나동열
연세대학교 전산학과
lplabmini@empal.com

Techniques for improving performance of POS tagger based on Maximum Entropy Model

Min-Hee Cho, Myoung-Sun Kim, Jae-Han Park, Eui-Kyu Park, Dong-Yul Ra
Dept. of Computer Science ,Yonsei University

요 약

한국어에서의 품사 결정 문제는 형태론적 중의성 문제도 있지만, 영어에는 발생하지 않는 동품사 중의성 문제로 더 까다롭다. 이러한 문제들은 어휘 문맥을 고려하지 않고서는 해결하기 어렵다. 통계 자료 부족 문제에 쉽게 대처하는 모델이 필요하며 문맥에 따른 품사를 결정하고자 할 때 서로 다른 형태의 여러가지 어휘 문맥 정보를 반영할 수 있는 모델이 필요하다. 본 논문에서는 이런 점에 가장 적합한 최대 엔트로피(maximum entropy : ME) 모델을 품사태깅 작업에 이용하는 문제에 대해 다룬다. 어휘 문맥 정보를 이용하기 위한 자질함수가 매우 많아지는 문제에 대처하기 위해 필요에 따라 어휘 문맥 정보를 사전화한다. 본 시스템의 특징으로는 어절 단위 품사 태깅을 위한 처리 기법, 어절의 형태소 분석열에 대한 어절 내부 확률 계산, ME 모델의 정규화 과정 생략에 의한 성능 향상, 디코딩 경로의 확장과 같은 점들이있다. 실험을 통하여 본 연구의 기법이 높은 성능의 시스템을 달성할 수 있음을 알게 되었다.

1. 서 론

품사 태거의 주요한 역할은 형태소 분석 결과를 통해 나온 다양한 품사들 가운데 올바른 품사 하나를 결정하는 것이다. 이 때 품사 태깅의 단위에 따라 형태소 단위 품사 태깅 모델과 어절 단위 품사 태깅 모델로 구분할 수 있다.

어절 단위의 품사 태깅은 한국어의 어절 구성이 다양하여 어절마다 품사를 정의하는 품사 수가 많아지는 단점이 있고, 형태소 단위의 품사 태깅은 한국어의 중요한 정보인 어절 정보를 이용하지 못하는 단점이 있다.

본 논문의 태깅 단위 모델은 이 두 단점을 보완한 어절 단위 품사 태깅 모델로 어절에 대하여 내용어 형태소 및 품사, 기능어 형태소 및 품사를 정의하여 사용하고 있다[4,8,10]. 따라서 한 어절의 다양한 분석열을 내용어, 기능어로 정의함으로써 복잡도를 줄이고, ME 모델에서 히스토리 범위가 고정된 윈도우(좌우 2개)인 경우 어절간의 주요 정보인 어말, 어두 정보를 쉽게 볼 수 있다.

기존의 품사 문맥 정보만을 이용한 품사 태거는 특히 어휘 중의성 및 동품사 중의성[1]을 해결하는데 변별력이 떨어지는 것으로 보인다[5,9]. 좌우의 어휘 문맥 정보가 고려되지 않는 한 중의성 문제는 해결하기 어렵다.

[그림 1]의 예를 보면 같은 "짝"이란 형태소가 각 어절마다 다른 품사로 해석되고 있다. (1), (2), (3)은 어절 "짝"의 문맥으로 "MM (짝) JK VV"의 동일한 품사열을 가지고 있다. 이 경우 품사 문맥만을 고려하면 3개 예제 모두 같은 품사를 내어 줄 가능성이 크다. 사람의 경우 (1)~(5) 모두 고려하는 지식이 다른 것 같다. 좌측 단어와의 연어 정보를 볼 것인지 우측 단어와의 연어 정보를 볼 것인지, 품사 문맥만 볼 것인지. 이러한 다양한 정보를 조화시켜 이용하는 모델이 좋을 것으로 생각되는데 ME 모델이 이에 적합한 것으로 알려져 있다. 위에서 얘기한 다양한 문맥 정보를 ME 모델의 자질들로 표현한다. 이 때 어휘 문맥 정보를 반영하는 자질함수가 매우 많아지는 문제에 대처하기 위해 필요에 따라 어휘 문맥 정보를 사전화함으로써 자질의 수를 줄인다.

---

1) "나는"에 대하여 날(fly)/VV+는/ETM, 나(sprout)/VV+는/ETM과 같이 어휘가 다른 동일한 품사열의 여러 후보가 가능한 중의성을 말함.

[명사:의존명사:부사:접미사] 중의성
(1)새/MM **짝/N**+을/JK 소개하/VV+다.
(2)갈비 두/MM **짝/NNB**+을/JK /배달하/VV+다.
(3)아무/MM **짝/NNB**+에/JK 못쓰/VV+겠다.

[그림 1] "짝"의 다양한 분석 예

전체적인 시스템 구성은 다음과 같다. 형태소 분석기는 문장의 각 어절에 대해 이를 구성하는 형태소 열을 파악한다. 이를 (형태소열) 분석후보라 하는데 보통 한 어절에 대하여 수많은 분석후보가 생성된다. 일반적으로 이 분석후보들이 태거에게 전달되나 본 시스템에서는 복합명사분석기를 거치도록 한다. 복합명사 분석기는 입력으로 받은 분석후보들 중에서 복합명사로 분석되는 후보(명사, 접두사, 접미사, 미등록어로 구성된 리스트)들을 모아서 조사가 있는 것과 없는 것 두 가지를 선택하고 나머지는 제거한다. 이 결과를 다른 분석후보(즉 용언이나 부사로 해석되는 후보들)들과 같이 태거에게 전달한다. 품사 태깅 작업은 이들 중에서 가장 적합한 후보를 선택하는 작업이라고 할 수 있다.

품사 태깅 시스템에서 어려운 작업의 하나는 미등록어를 인식하는 것이다[6]. 특히 미등록어가 다른 형태소와 결합하여 어절을 이루는 경우 어렵다. 이를 위하여 우리는 미등록어를 포함한 복합명사 분석을 위해 특별히 개발된 모듈을 이용한다[7].

본 논문은 ME 기반 품사 태거의 성능 향상을 위하여 어절 내부 확률 도입, ME 모델의 정규화 생략, 디코딩 경로의 확장 등의 방법을 3장, 4장에서 제안하고 실험을 통하여 높은 성능을 얻을 수 있음을 보인다.

2. 관련연구

품사 태거에 대한 연구는 현재까지 많이 진행되어 왔다. HMM[5], 복합확률 모형[10], ME 모델[3], 어휘규칙[11], 후처리 오류정정[4] 등의 다양한 연구가 있어 왔다.

[5]는 형태소 단위의 태깅이므로 어절간의 정보를 보기 위하여 어절간 전이와 어절 내 전이를 구별한다. 하지만 어절간의 품사 바이그램 확률과 (현재 형태소의) 어휘 확률만을 참조하므로 중의성 문제를 해결하는 것에 어려움이 있고, [10]은 어절기반의 단점을 보완하고 있지만 어절 간의 어말-어두 공기정보를 모으는데 40만 통계정보의 사용은 자료 부족 문제를 야기하므로 어휘 중의성 해소에 정확성이 많이 떨어진다. 또한 자료부족 문제로 어휘 발생확률이 존재하지 않을 경우에 대한 해결책이 제시되어 있지 않다. [11]은 어휘 규칙을 수동으로 획득하는 데 많은 시간과 비용을 요구한다. [4]는 후처리를 통한 오류 정정과정에서 어휘문맥을 사용한다. 이 때 어휘문맥에 대한 적당한 가중치를 부여하는 것이 어렵다.

[그림 1]을 보면 분석 어절에 가장 큰 영향을 미치는 문맥은 좌 문맥의 내용어 어휘, 기능어 어휘, 우문맥의 내용어 어휘임을 알 수 있다. 명사와 동사 사이에 부사와 같은 수식어가 삽입될 경우 좌우 각 2 토큰을 보는 고정된 윈도우에서는 형태소 단위 보다는 어절 단위로 처리하는 시스템이 더욱 많은 정보를 활용할 수 있다.

이러한 이유로 기존의 연구된 형태소 기반 시스템들은 어휘 문맥 반영을 위하여 후처리 작업[4], 모델 변형[6,10]등의 접근을 했다.

본 논문에서는 어절 단위 기반 시스템의 단점을 극복하고 어절을 기반으로 하여 생기는 장점을 취하는 시스템을 소개한다.

본 시스템은 또 다른 주요 특징은 ME 모델을 기반으로 한 것이다. ME 모델 기반 한국어 품사 태깅의 다른 연구로는 [3]이 있다. 이 시스템의 문제점은 형태소 기반 시스템이어서 이용하는 문맥의 범위가 제한적이라는 것이다. 우리는 기본적인 ME 모델 위에 몇가지 새로운 기법을 추가함으로써 성능 향상을 이루도록 하였다.

3. ME 모델 기반 태거의 성능 향상 기법

3.1 품사 태깅에서의 ME 모델

본 논문에서는 ME 모델에 기반하여 문장에 대한 가장 좋은 품사열을 결정하고자 한다. 현재 어절(w_i)의 좌우 문맥정보를 히스토리(history)라 정의한다[1].

$$h = \{w_{i-2}/t_{i-2}, w_{i-1}/t_{i-1}, w_i, w_{i+1}/t_{i+1}\}$$ 2)

$$w_i = \{m_{h,i}, m_{f,i}, t_{h,i}, t_{f,i}\}$$ 3)

좌우 문맥은 가능한 많이 보면 좋지만 제한된 정보를 이용하므로 현재 어절의 왼쪽 2개 오른쪽 1개 어절로 제한하였다. 문맥정보에 이용되는 것들은 어절의 내용어

2) w_i = 현재 어절

3) $m_{h,i}$ = 내용어 형태소, $m_{f,i}$ = 내용어 품사
$t_{h,i}$ = 기능어 형태소, $t_{f,i}$ = 기능어 품

형태소, 기능어 형태소, 내용어 품사, 기능어 품사이다.

주어진 문맥 h_i에서 현재 어절의 품사를 t_i로 결정할 확률은 식(1)과 같다. (f : 피쳐(즉 자질함수), t : 태그, h : 히스토리, n : 어절 수, k : 피쳐 수, T : 태그 수, i : 어절 인덱스, j : 피쳐 인덱스, l : 태그 인덱스).

$$p(t_i \mid h_i) = \frac{\prod_{j=1}^{k} \alpha_j^{f_j(h_i,t_i)}}{Z(h_i)} \qquad (1)$$

따라서 길이가 n인 어절열이 주어진 상황에서 문장에 대한 최적의 품사열을 찾기 위한 확률값을 계산하고자 하면 다음과 같다.

$$p(t_1,t_2 \ldots t_n \mid S) = \prod_{i=1}^{n} p(t_i \mid h_i) = \prod_{i=1}^{n} \frac{\prod_{j=1}^{k} \alpha_j^{f_j(h_i,t_i)}}{Z(h_i)}$$

$Z(h_i) = \sum_{l=1}^{T} (\prod_{j=1}^{k} \alpha_j^{f_j(h_i,t_l)})$ 는 정규화를 위한 상수이다.

(f : 피쳐(즉 자질함수), t : 태그, h : 히스토리, n : 어절 수, k : 피쳐 수, T : 태그 수, i : 어절 인덱스, j : 피쳐 인덱스, l : 태그 인덱스).

f_j는 자질함수로 정해진 조건에 만족을 하면 1, 만족하지 않으면 0의 값을 갖는다. α_j는 자질함수 훈련을 통하여 구해진 f_j에 대한 가중치 값이다.

말뭉치를 통하여 구한 자질 함수는 아래의 제약조건을 만족하면서 모델의 엔트로피가 최대화 되도록 한다.

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

$$E[f_j] = \sum_{h,t} p(h,t) f_j(h,t) \qquad \ldots\ldots(2)$$

$$\tilde{E}[f_j] = \sum_{i=1}^{n} \tilde{p}(h,t) f_j(h_i,t_i) \approx \sum_{i=1}^{n} \tilde{p}(h_i), p(t_i \mid h_i) f_j(h_i,t_i) \ldots\ldots(3)$$

(2)는 모델에서의 기대값이고 (3)은 훈련 말뭉치로부터의 기대값이다. 이 두 기대값이 같다는 제약조건을 만족하는 확률 분포 가운데 엔트로피를 최대화 하는 p^*를 찾는 것이 최대 엔트로피(ME) 모델의 핵심 이론이다.

모델 p^*를 계산하기 위해서는 α_j들로 표현되게 되는데 α_j (자질에 대한 문맥 가중치)값 추정에는 GIS, IIS 알고리즘 등이 이용되고 있으며 본 논문에서는 [2]를 이용하여 α_j 값을 구한다.

3.2 어절 기반에 의한 문맥의 확장

기존의 어절 단위 품사 태깅 모델은 어절 안의 분리된 형태소열을 한 덩어리로 보아서 품사를 부착하려 한다. 형태소의 가능한 조합 수만큼 품사도 늘어나 품사의 수가 커지고 어절 단위의 바이그램(bigram) 수도 커지는 단점을 가지고 있다.

형태소 단위의 품사 태깅은 어절 안의 분리된 형태소를 단위로 품사를 부착한다. 따라서 하나의 어절이 길이가 다른 여러 개의 형태소열로 분석이 가능하므로 정규화의 문제를 따르는 단점이 있다. 형태소 단위 모델은 어절 단위 모델의 단점은 없지만 어절 단위 모델의 장점인 어절간의 정보를 이용하지 못하는 문제가 있다.

한국어 품사 부착에서 나타나는 동품사 중의성 문제를 해결하기 위해서는 좌우 문맥의 의미정보를 잘 살펴야 한다. [3]의 논문은 앞 형태소의 최대 2개만 보고 있기 때문에 한 어절이 3개 이상의 형태소로 갈라 질 수 있을 때 그 앞의 어휘 정보를 보는데 어려움이 있다. 따라서 우리는 문맥 확장이 용이한 어절 단위 태깅을 선택한다.

[표 1] 품사 집합

N	명사	MA	부사	XSV	동접
NNB	의존명사	IC	감탄사	XSA	형접
NP	대명사	EE	어미	XSN	접미
NR	수사	ETM	관형형어미	XP	접두
VV	동사	JK	조사	EP	선어말어미
VA	형용사	SN	숫자	ETN	명전성어미
VX	보조동사	FN	외국어	VCP	지정사
MM	관형사	SY	심볼	UW	미등록어

어절 단위 품사 태깅의 단점을 해결하기 위하여 어절의 내용어 형태소, 내용어 품사, 기능어 형태소, 기능어 품사만을 이용하는 방식을 취해 어절 기반의 단점을 보완한다. 한 어절의 내용어는 용언의 어간, 접두사 및 접미사를 제외한 명사, 본용언과 보조용언의 결합경우 본용언의 어간, 복합명사의 경우 마지막 명사로 한다. 어절의 기능어는 그 어절 안에 있는 조사+조사, 어미+조사의 경우 앞에 기능어 형태소로 한다[10].

이렇게 되면 어절의 태그로 이용되는 태그 집합은 우리의 형태소 분석기가 가진 [표 1]의 24개의 품사 가운데 어절 품사가 될 수 없는 것들[4]을 제거한 16개([표 1]의 색칠한 품사)이다. 이 표에서 어절 품사에 포함되지 않은 품사들은 은닉된 상태로 존재하며 태깅 결정 후 최종 출력에서는 다시 노출되도록 한다. [1]

[표 2]를 보면 어절 태깅 예를 볼 수 있다. 예를 들어 형태소열 "대유/N+적/XSN"은 어절 태그열이 "대유/N+^/-1"인 후보로 변환되면서 (어절 기반) 품사 태깅 모듈의 입력으로 다른 후보와 함께 들어간다. 태깅 작업 결과 이 후보가 선택되면 원래의 형태소열인 "대유/N+적/XSN"이 결과로 출력된다. 따라서, 어절 태깅에 참여하지 않았던 형태소 태그들도 출력에 나타나게 된다.

이런 기법을 사용함으로써 어절 기반 태깅에서 어절의 품사의 수를 작은 범위로 제한할 수 있게 된다.

[표 2] 어절 태그 부착 예

형태소 분석후의 후보	어절 태그
이(MM)	이/MM__^/-1[5]
점/N__에서/JK__는/JK	점/N__에서/JK
관계/N__기관/N__의/JK	기관/N__의/JK
대유/N__적/XSN	대유/N__^/-1
표현/N__이/VCP__다/EE	표현이/VV__다/EE
가|VV__아|EE__버리|VV__어|EE	가|VV__어|EE
공부|N__하|XSV__다|EE	공부하|VV__다|EE

3.3 정규화의 생략

식 (1)은 현재 어절에 대한 문맥 h_i 상태에서의 특정 품사로의 확률이다. 식(1)을 따르면 영어의 경우 현재 어절을 결정할 때 히스토리(history)가 같기 때문에 $Z(h_i)$는 모든 가능한 태그에 대하여 변하지 않으며 $\sum_{l=1}^{T} p(t_l \mid h_i) = 1$ 임을 보장하는 정규화 상수이다.

그러나 한국어의 경우는 문제가 다르다. [그림 2]의 어절 "나는"을 같은 품사 VV로 보는 히스토리는 다음과 같이 여러 개가 나온다. (명사나 대명사로 보는 것들도 있으나 이 그림에서는 생략됨.) 히스토리가 고정되는 영어와 달리, 한국어에서는 여기의 "나는 화가"의 예처럼 내용어 어휘가 여러 가지로 나올 수 있으며 히스토리도 여러 가지로 바뀌는 현상이 있다. 즉 "나는"을 현재 토큰으로 볼 때 한 히스토리만 가능한 것이 아니라 아래와 같이 여러 히스토리가 가능하다.

{"BK"/BK,"BK"/BK,"나"/VV,"화"/N}[6]
"BK"=시작 어휘, BK=시작 태그를 나타낸다
{"BK"/BK,"BK"/BK,"날"/VV,"화"/N}
{"BK"/BK,"BK"/BK,"나"/VV,"화가"/N}
{"BK"/BK,"BK"/BK,"날"/VV,"화가"/N}

즉 여러 히스토리가 동시에 활성화된 상태에서 현재 어절 "나는"의 품사를 결정하는 상황이 된다. 이런 상황에서는 식 (1)에서 정규화 상수 $Z(h_i)$는 개입시키지 않는 것이 좋다.

따라서 우리는 앞의 식(1) 대신에 정규화를 생략한 다음 식 (4)를 이용한다.

$$p(t_i \mid h_i) = \prod_{j=1}^{k} \alpha_j^{f_j(h_i, t_i)} \quad .....(4)$$

h_i가 변화하는 한국어에는 적합하지 않으므로 확률의 정규화를 위한 $Z(h_i)$ 분모를 생략해 실험을 한 결과 3%의 성능 상승을 얻었다.

3.4 어절 확률의 이용

품사 태거는 형태소 분석기의 분석된 결과를 가지고 그 가운데 가장 좋은 분석 한 가지를 고른다. 우리가 제안하는 어절 기반 기법은 어절을 구성하는 모든 형태소를 고려하는 것이 아니고 내용어 어휘와 기능어 어휘만을 고려한다. 어절의 분해로는 좋지 않은 형태소열임에도 그 일부만 봄으로써 좋은 것처럼 보일 수 있다. 즉 결합할 가능성이 없는 분석열에 대한 내용어 어휘가 모델에서 좋은 자질로 평가될 수 있는 단점이 있다. 이를 보완하기 위해서 형태소 바이그램 확률을 이용하여 어절 내부 확률을 전체 확률에 반영하도록 한다. 세종 말뭉치 1,000만 어절을 이용하여 12가지 형태소 바이그램 통계정보를 추출하고 백오프(back_off) 평탄화 기법을 이용하여 어절 확률을 계산해낸다.[7]

"~조건으로 내세운 것은~" 라는 절이 있을 경우 "내세운"은 세 가지 분석이 나온다.

4) 접두사, 전성어미, 접미사류, 지정사 등아 하나의 어절을 이루는 경우는 없다고 가정한다.

5) ^는 조사가 없을 경우를 표시하는 어휘이고 이에 대한 품사는 1로 나타낸다.

6) "BK"=시작 어휘, BK=시작 태그를 나타낸

① 내세(N)+운(N)
② 내세우(VV)+ㄴ(ETM)
③ 내세(N)+우(N)+ㄴ(JK)

①, ③ 분석의 형태소열이 어절을 구성할 확률은 극히 작다. 그런데 어절의 내용어 어휘를 결정하려 할 때 복합 명사이기 때문에 뒤의 "운", "우"가 내용어 어휘로 된다. 실제 "운","우"란 명사는 ME 모델에서 좋은 자질로 평가되기 때문에 정답으로 선택될 오류가 생길 수 있다.

이러한 어절 기반 기법에서 어절을 구성하는 여러 형태소 중에서 내용어 및 기능어 어휘만을 보는 단점을 보완하기 위하여 백오프 기법을 통하여 어절 확률 $p(w_i)$을 계산해 내어 이를 식 (5)에서와 같이 현재 어절에서의 $p(t_i \mid h_i)$를 계산하는 데 이용한다.

$$p(t_i \mid h_i) = \prod_{j=1}^{k} \alpha_j^{f_j(h_i,t_i)} * p(w_i) \quad .....(5)$$

어절이 L개의 형태소로 구성되어 있을 때 각 형태소를 $(m_1,t_1)....(m_L,t_L)$이라 정의하고, $p(w_i)$를 구하여 보면 식(6)과 같다.

(6)

$$P(w_i) = P(m_1/t_1, m_2/t_2, ..., m_{L-1}/t_{L-1}, m_L/t_L) =$$

$$\left[P(m_1/t_1 \mid \#) * \{ \prod_{i=2}^{L} P(m_i/t_i \mid m_{i-1}/t_{i-1}) \} * P(\# \mid m_L/t_L) \right]^{\frac{1}{\lambda}}$$

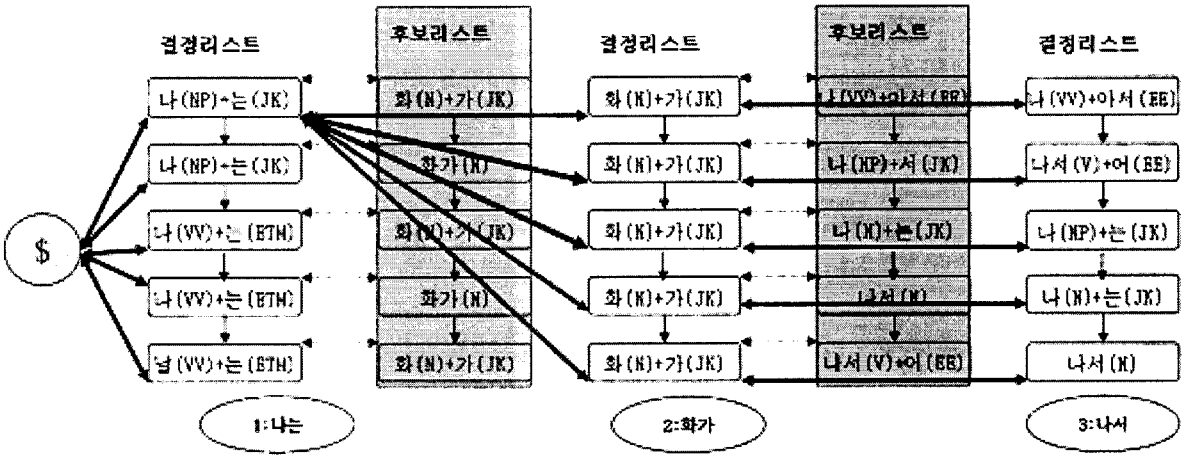
식 (6)에 대한 자세한 설명은 [7]에있다.

3.5 디코딩 경로의 확장을 통한 다음 어절 품사의 이용

문장에 태깅을 하기 전에 각 어절은 형태소 분석 결과를 얻어 어절의 내용어 어휘 및 품사, 기능어 어휘 및 품사의 여러 가능한 후보를 가지고 있다. 품사 태거는 좌우 문맥을 고려하여 형태소 분석 결과들 가운데 어절마다 하나를 정답으로 선택해야 한다. 문장의 어절마다 여러 가능한 후보가 존재하는 상황은 [그림 2]와 같이 래티스(lattice)와 같이 생각할 수 있다. ME 모델은 래티스의 가장 좋은 경로를 찾아 그것을 태깅 결과로 결정한다[1].

기존의ME 모델이 디코딩 기법에서는 래티스 내의 경로들이 현재 어절까지만 뻗은 상태이다. 그러나 우리 모델에서는 현재 어절의 결정 시에 래티스 경로는 다음 어절까지 뻗은 것들을 이용한다. 그 이유는 $p(t_i \mid h_i)$를 구할 때 다음 어절의 분석 후보들도 고려하기 때문이다.

[그림 2]에서 보면 1번째 어절의 품사를 결정할 때 다음 어절의 분석 후보들(2개)을 검토하면서 각 경로의 확률을 구한다. 이러한 경로들 중에서 상위 n 개를 결정리스트라 부르자. 결정리스트의 각 경로는 다음 어절의 특정 분석 후보로 연결되어 있다. 경로 확장이란 이



[그림 2] 문맥을 반영한 품사 탐색 경로

런 상황을 말하는 것이다. 결정리스트의각 경로가 연결된 다음 어절의 분석 후보들을 모은 것을 후보리스트라 한다. (결정리스트와 후보리스트는 갯수가 동일하다.) 결정리스트가 결정되면 현재 어절에 대한 처리가 끝나고 다음 어절을 새로운 현재어절로 하여 처리하는 단계로 넘어 간다. 새 어절에서는 후보리스트의 노드들을 이용하여 다시 경로 확장을 추구하여 새로운 결정리스트를 구하게 된다.

위의 그림은 리스트의 경로수를 n = 5로 제한하였기 때문에 5번째 경로의 확률 값 보다 작은 경로들은 버려진다. 이러한 방법을 보통 빔탐색(beam search)이라 부른다. 실험에 사용된 실제 노드 사이즈(n)는 15이다.

마지막 어절의 결정리스트까지 구해지면 가장 상위 경로를 쫓아가 최적의 품사 경로를 구할 수 있다.

4. 다양한 어휘 문맥 정보의 이용을 위한 자질함수

한국어는 어순이 자유롭기 때문에 어절 단위 태깅 과정에서 품사 문맥 정보는 품사 결정에 큰 영향을 미치지 못한다. 따라서 우리의 ME 시스템의 모든 자질은 어휘를 고려하는 특징이 있다. 여러 가지 템플릿(자질함수 생성 모형)을 품사태깅 말뭉치에 적용하여 자질함수를 추출한다. 템플릿에 사용되는 X, Y, Z, L, M, N 등은내용어 어휘및 품사, 기능어 어휘 및 품사 등을 나타내는 변수이다. 아래에서 인덱스 i 는 현재 어절, i-1, i +1 는 그 앞, 뒤 어절을 지칭한다.

▣ 기본적인 자질

[표 3]의 (1번과 2번)기본 자질들은 어절 간의 주요 정보인 기능어 어휘 및 품사를 이용하여 현재 어절의 품사를 결정하고자 한다. 3번은 내용어 어휘, 기능어 어휘 두 가지를 고려하는데 용언의 결정에 이용되는 자질함수이다. 4번은 현재 어절의 내용어의 어휘 및 품사를 보고 이 품사임을 주장하는 자질이다.

[표 3] 기본 자질 함수(Template : 템플릿)

#	문맥 정보	
1	$t_{h,i-1}$=X & $m_{f,i-1}=Y$ & $t_{f,i}$=Z	$t_{h,i}=L$
2	$m_{f,i}=X$ & $t_{f,i}=Y$ & $t_{h,i+1}$=Z	
3	$m_{f,i-1}=M$ & $m_{h,i}$=N & $t_{h,i}$=V [7]	
4	$m_{h,i}$=N & $t_{h,i}$=etc [8]	

▣ 공기 사전 이용을 위한 자질들

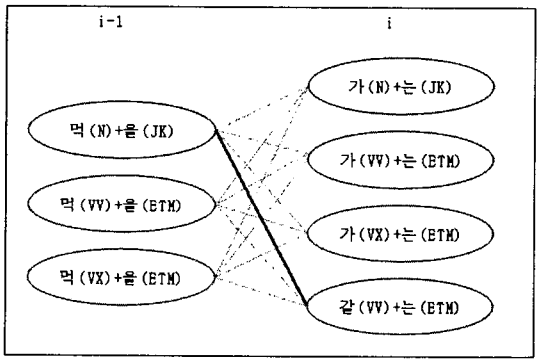
품사 태깅의 중의성 해소에는 어휘간의 공기정보가 매우 유용하게 이용되는 경우가 많다. 그러나 어휘 공기쌍 마다 따로 자질을 준비하면 자질의 수가 너무 많아진다. 그래서 우리는 모든 유용한 공기쌍을 사전[9]에 넣고 이 사전을 참조하여 특정 품사를 주장하는 자질함수를 품사마다 하나씩 두고자 한다.

- 명사, 대명사, 동사, 형용사를 위한 공기사전 참조 자질함수[10]

$$f_j(m_i,t_i,h_i)=$$
$$\begin{cases} 1, (m_{h,i\pm j}=L \;\&\; m_{f,i\pm j}=M \;\&\; m_{h,i}\text{=N}) \in \text{NV\_DIC} \\ \quad \&\; t_{h,i}\text{=Y} \\ 0, \text{otherwise} \end{cases}$$

이 자질함수에서 Y 에 들어가게 되는 명사, 대명사, 동사, 형용사마다 각각 자질함수를 만든다(4개의 자질함수). 여기서 N은 현재 어절의 내용어 어휘, M은 체언에 붙는 기능어 어휘, L은 주변의 한 어절의 내용어 어휘로 하여 (L,M,N)의 공기 여부를 사전 NV_DIC에서 참조한다. 있으면 1, 없으면 0을 출력한다. Y는 현재 어절 i의 내용어 품사로 채워진다. L이 용언이고 N을 수식하는 경우 기능어 어휘인 M은 참조하지 않는다. $i\pm j$ 의 의미는 현재 어절 위치 i에서 앞뒤로 임의의 j 만큼의 거리에 위치한 어절을 나타낸다.

[그림 3]을 보면 "가는"의 품사를 결정할 때 품사와 어휘를 같이 결정해야 한다. 이 경우 '먹/N+을/JK'이 '갈/VV'과 함께 NV_DIC사전에 나타나 있으면 (즉 훈련말뭉치에서 공기하였으면) 현재 태그를 동사로 결정하는데 위 자질이 활성화되어 이 결정을 지원한다(j = -1). 이 예에서 '먹'이 현재 어절일 때 '먹'의 품사(명사, 동사)를 결정하는 데에도 위와 같은 방법으로 우문맥의 '갈'이 i+j 위치의 어절로 대응되어(j = 1) 공기사전 참조가 이루어지고 '먹'이 명사로 결정되는데 위 자질이 이용된다.



[그림 3] "~먹을 가는~"의 형태소 분석 예

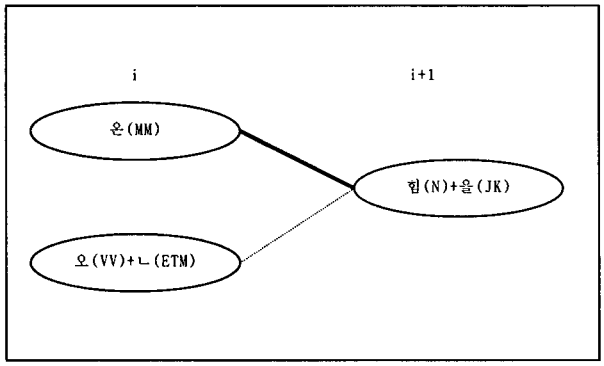
- 관형사

7) V=동사 형용사
8) etc= 동사,형용사,보조용언을 제외한 모든 품사

9) 체언과 용언간의 공기 사전 형태 :
NV_DIC : 체언어휘_체언품사_기능어어휘_용언어휘_용언품사,
MM_DIC : 관형사 사전은 관형사어휘_내용어어휘_내용어품사.
10) f 는 자질 함수로서 템플릿이 아님에 유의할 것.

$$f_j(m_i,t_i,h_i) = \begin{cases} 1, \ (\mathrm{m}_{h,i}=\mathrm{M} \ \& \ \mathrm{m}_{h,i+1}=\mathrm{N}) \in MM\_\mathrm{DIC} \\ \quad \& \ \mathrm{t}_{h,i}=\mathrm{MM} \\ 0, \text{ otherwise} \end{cases}$$

관형사의 경우 수사, 대명사, 명사, 용언으로 해석될 수 있는 중의성이 발생한다. 관형사는 우문맥에 나타나는 어휘를 수식하는 특징이 있다. 이 성질을 이용하여 자질을 설계하여 관형사를 결정하는데 도움을 준다.



[그림 4] "~온 힘을~"의 형태소 분석 예

[그림 4]를 보면 자질함수는 "온"의 중의성을 해결하기 위하여 현재 어절의 내용어 어휘 "온"과 다음 어절 내용어 어휘 "힘"이 관형사 관련 공기 사전(MM_DIC)에 나타나 있는지를 본다. 있으면 위 자질이 활성화되어 관형사 결정을 지원한다.

• 보조용언

$$f_j(m_i,t_i,h_i) = \begin{cases} 1, \ (\mathrm{m}_{f,i-1}=\mathrm{M} \ \& \ \mathrm{t}_{f,i-1}=\mathrm{EE} \ \& \ \mathrm{m}_{h,i}=\mathrm{N}) \in PX\_\mathrm{DIC} \\ \quad \& \ \mathrm{t}_{h,i}=\mathrm{PX} \\ 0, \text{ otherwise} \end{cases}$$

보조용언의 경우 동사, 형용사 등과 가장 많은 품사 모호성을 가지고 있다. 보조용언은 이전 단어의 특정 기능어 어휘에 의해 결정되는 특징이 있다. 이를 반영한 자질 함수를 설계하여 보조용언을 결정하는데 도움을 준다.

[표 4] PX_DIC 사전 예

$\mathrm{m}_{f,i-1}$=M	$\mathrm{m}_{h,i}$=N
아야	하
는가	싶
게	되
지	않

[표4]에 나타난 특정 기능어 어휘, 보조용언 사전을 이용하여 중의성이 있는 보조용언의 경우 도움을 받는다.

▣ 특정 품사 인식을 위한 자질들

한국어는 품사마다 보는 문맥 정보가 틀리므로 자질들도 품사마다 그 인식을 위하여 다르게 설계한다.

• 의존명사

#	문맥정보
1	$\mathrm{t}_{h,i-1}=V \ \& \ m_{f,i-1}=X \ \& \ \mathrm{m}_{h,i}=M$
2	$\mathrm{t}_{h,i-1}=etc. \ \& \ t_{f,i-1}=X \ \& \ \mathrm{m}_{h,i}=M$
3	$\mathrm{t}_{h,i-1}$=NR & $\mathrm{m}_{h,i-1}=M \ \& \ \mathrm{m}_{h,i}=M$

의존명사는 일반적으로 (단 음절) 일반 명사와 중의성이 가장 많이 나타나고, 의존명사의 앞 어절에 나타나는 수사, 관형사들도 일반 명사와 중의성이 많이 나타난다. 이 때 어휘 문맥을 고려하지 않으면 정확한 품사결정이 어렵다.

[예]

"먹을 수 있다."의 경우 앞 어절의'ㄹ'기능어 형태소 어휘는 현재 어절의 내용어 어휘 '수'를 의존명사로 결정하는데 도움을 준다.

"종류가 오만 가지나 된다."의 경우 2,3번 자질에 의해 의존명사로 결정되는데 도움을 받는다.

5. 실 험

5.1 실험 데이터

학습 데이터는 "세종 말뭉치 1999,2000,2001,2002"를 사용하였으며 680만여 어절로 구성되어 있다[표 5].

[표 5] 학습 말뭉치

	어절 수
1999	1,553,278
2000	1,945,341
2001	1,750,050
2002	1,584,542
총 어절 수	6,833,211

실험 데이터는 세종 말뭉치 2003에서 추출한 신문사설, 세계사, 수필 3가지로 29,453어절로 구성되어 있다. [표 6].

[표 6] 테스트 말뭉치

	어절 수
신문사설	9,865
소설	9,538
세계사	9,360

5.2 실험 결과 및 검토

3가지 경우에 대하여 실험을 하였는데, 실험 C는 최종 모델로서 3.4절의 수식 (5)을 이용하여 구성된 모델이다. 실험A는 실험 C에서 어절 내부 확률을 곱하지 않은 모델, 실험 B는 실험 C에서 정규화 $Z(h_i)$를 생략하지 않고 구성된 모델이다. A, B, C 모델 모두 미등록어 인식을 수행하는 모델이다.

[표 7] 품사 결정 모델의 성능

모델 종류	어절 정확도 (%)
A (어절확률 생략)	91.45
B (어절확률 이용, 정규화)	94.98
C (어절확률 이용, 정규화 생략)	98.03

[표 7]을 통하여 실험 B와 C의 정확도를 보면 p(tlh)를 정규화하는 분모인자$Z(h_i)$가 생략되어있을 때 3%의 정확률 상승을 보인다. 그리고 실험 A와 실험 C를 비교하여 볼 경우 6.5%의 정확률 상승이 보이는데, 이것은 어절 기반의 품사 부착시 생기는 문제점을 어절 확률이 충분히 해결해준다고 할 수 있다. 위 실험의 결과로 성능 향상에 사용된 두 기법이 유용함이 보여진다.

[표 8] 모델 C의 정확도 분석(정확도 %)

형태소열과 품사열 모두 일치	97.73
품사열만 일치	98.03

[표 8]의 실험을 보면 실험 말뭉치에 대하여 형태소열, 품사열 모두 일치한 경우 97.73%의 성능을 보였다. 이것은 자질들이 좌우 어휘 문맥을 고려하기 때문에 품사 결정시 어휘까지 충분히 고려하여 어휘 중의성 문제를 해결할 수 있었음을 보여준다.

오류 유형을 살펴보면 학습 문서의 오태깅으로 인한 잘못된 자질 추출, 시스템의 중간 과정인 복합 명사 분해의 오분석, 유용한 자질함수 부족 등이 있다.

말뭉치 오태깅으로 인한 문제가 가장 큰데 그 예로 본용언과 보조용언, 의존명사와 수사, 의존명사와 명사, 부사와 명사 등의 잘못된 태깅이다. 복합명사 분해기, 품사 태거 모두 통계치를 기반으로 하기 때문에 정확도가 높은 코퍼스를 사용하는 것이 매우 중요하다는 것을 실험을 통해 확인할 수가 있었다.

[표 9]는 실험환경(형태소 분석기, 학습 말뭉치, 테스트 말뭉치, 품사 집합 등)이 동일하지 않기 때문에 객관적인 비교는 될 수 없지만 기존의 보고된 시스템과 비교평가를 해 보았다.

[표 9] 품사 태깅 모델의 비교 평가

모델	모델 종류	어절 정확도(%)
어절 단위 기반	본 논문 (미등록어 포함)	98.03
	이하규 97	98.80
	김영길 03	97.08
형태소 단위 기반	강인호 00	93.40
	김진동 98	95.80

우리 모델과 같은 모델인 ME을 사용한 시스템으로 강인호 99[3]가 있는데 미등록어를 위한 특수 자질을 사용하였다. 하지만 본 연구의 시스템이 많은 성능 향상을 이룸을 볼 수 있다. 이하규 97[10]은 보고된 시스템들 가운데 가장 좋은 성능을 보인다. 하지만 내용어 어휘 통계정보가 없어 어휘 발생 확률을 구하지 못할 경우에 대한 해결방안이 제시되지 않아 미등록어를 포함한 어절이 나타난 문장의 경우 올바른 품사열을 결정하기 어렵다. 김진동 98[5]는 어절 정보의 반영을 위하여 띄어쓰기를 고려한 태거 시스템을 제안하였지만 어절 간 어휘 문맥 정보가 반영이 되지 않아 높은 성능을 달성하기 어렵다. 김영길 03[4]은 품사 결정시 주변 어절 간의 공기 정보를 반영하는 후처리 단계를 두어 성능을 향상시키고 있지만 우리 시스템은 이것을 품사 태깅 단계에서 하고 있다.

6.결론 및 향후 과제

본 논문에서는 ME 기반 한국어 품사 태깅 시스템의 성능 향상 방법들에 대해 살펴 보았다. 제안된 시스템의 특징으로는 어절 기반 태깅 작업, 어절 내부 확률의 이용, ME 모델의 정규화 작업 생략에 의한 성능향상 모색, 디코딩 시에 다음 어절까지 디코딩 경로의 확장 기법 사용, 어휘 공기 정보를 이용한 어휘 중의성 해소 및 자질 함수의 개수 감소 등을 들 수 있다. 본 시스템은 동품사 중의성과 같은 어휘 중의성의 해소 및 미등록어의 원활한 인식을 위하여 대량의 말뭉치에서 추출한 많은 통계 정보를 사용하였다. 어절 내부 확률 계산 과정에서 자료 부족 문제를 해결하기 위해 백오프 평탄화 기법을 사용하였다. 실험 결과 높은 성능의 시스템

개발이 가능함을 알게 되었다. 향후 연구로는 우리가 사용한 훈련말뭉치의 오태깅에 의한 시스템 성능 저하 정도의 파악, 부족한 자질 함수의 보완에 의한 성능 향상 모색 등이 있다.

참고 문헌

[1] Adwait Ratnaparkhi, "A maximum entropy part of speech tagger." In processing of the Empirical Methods in Natural Language Processing Conference, 1996

[2] Eric Sven Ristad,"Maximum entropy Modeling Toolkit", 1998

[3] 강인호, "최대 엔트로피 모델을 이용한 한국어 품사 태깅", KAIST 석사 학위 논문, 1999

[4] 김영길, 양성일, 홍문표, 박상규, "형태소 어휘 문맥에 기반한 태깅 오류 정정", 제15회 한글 및 한국어 정보처리 학술대회 논문집, 2003

[5] 김진동, 이상주, 임해창, "어절 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델", 제10회 한글 및 한국어 정보처리 학술대회 논문집, 1998

[6] 김진동, 임희석, 임해창, "Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델", 정보과학회 논문지(B), 제 24권, 제12호, 1997

[7] 박재한, 김명선, 노대욱, 나동열, "백오프 통계정보를 이용한 미등록어 포함 복합명사 분해", 2004

[8] 이상주, 임희석, 임해창, "은닉 마르코프 모델을 이용한 두단계 한국어 품사 태깅", 제6회 한글 및 한 국어 정보처리 학술대회 논문집, 1994

[9] 이상호, 미등록어를 고려한 한국어 품사 태깅 시스템 구현, KAIST 석사 학위 논문 ,1995

[10] 이하규, "어말-어두 공기 정보를 이용한 한국어 어휘 중의성 해소", 정보과학회 논문지(B), 제 24권, 제1호, 1997

[11] 임희석, "언어지식과 통계정보를 이용한 한국어 품사 태깅 모델", 고려대학교 박사 학위 논문, 1997