어휘 정보의 자동 추출과 이를 이용한 한국어 품사 태깅†

강인호° 김도완 이신목 김길창
한국과학기술원 전산학과
{ihkang, dwkim, smlee}@csone.kaist.ac.kr, gckim@cs.kaist.ac.kr

Korean Part-of-Speech Tagging using Automatically Acquired Lexical Information

In-Ho Kang° DoWan Kim SinMok Lee Gil Chang Kim
Dept. of Computer Science
Korea Advanced Institute of Science and Technology

요 약

본 연구는 형태소 분석에 필요한 언어 지식과 품사 태깅에 필요한 확률 정보를 별도의 언어 지식 추가 없이 학습 말뭉치를 통해서 얻어내는 방법을 제안한다. 먼저 품사 부착된 학습 말뭉치로부터 형태소 사전과 결합 정보를 추출한다. 그리고 자주 발생하는 어절 및 해석상 모호성이 많은 어절에 대해서는 학습 말뭉치에서 발견된 형태소 분석 결과를 저장하여 형태소 분석에 소요되는 시간과 형태소 분석의 정확률을 높인다. 또한 미등록어의 많은 부분을 차지하는 인명, 지명, 조직명에 대해서는 정보 추출 분야에서 사용하는 고유 명사 분류법으로 해결한다. 품사 태깅을 위해서는 품사열 정보와 품사열 정보로는 해결할 수 없는 경우를 위한 어휘 정보를 학습 말뭉치에서 추출한다. 품사열 정보와 어휘 정보는 정형화 과정을 거쳐 최대 엔트로피 모델의 자질로 사용되어 품사 태깅 시스템을 위한 확률 분포를 구성한다. 본 연구에서 제안하는 방법은 학습 말뭉치를 기반으로 한다는 특성에 의해 다양한 영역에 사용하기 쉽다. 또한 어휘 정보로 품사 문맥 정보를 보완하기 때문에 품사 분류 체계와 형태소 해석 규칙에 영향을 적게 받는다는 장점을 가진다. MATEC '99 데이터 실험 결과 형태소 단위로 94%의 재현률과 93%의 정확률을 얻을 수 있었다.

1 서론

문장에서 각 단어의 품사를 결정짓는 것을 품사 태깅이라고 부르며 크게 규칙에 기반하는 방법과 확률 정보에 기반하는 방법 등이 연구 되었다(강인호, 1999). 근래에는 대규모의 말뭉치(corpus)가 만들어지면서, 고급의 언어 지식 없이도 자동으로 언어 현상을 추출하여 품사 태깅을 행하는 확률 정보 기반의 방법들이 많이 사용된다. 확률 정보를 기반으로 하는 대부분의 품사 태깅 모델들은 단어의 품사를 주변의 품사열과 같은 제한된 문맥정보를 통해서 결정할 수 있다는 가정을 사용한다(Charniak, 1993). 그러나 한국어에 있어서 올바른 품사 태깅을 위해서는 형태소의 의미, 기능, 형식 정보를 필요로 한다(남기심 고영근, 1994). 이러한 품사 분류 정보는 품사열과 같은 제한된 문맥 정보로는 설명할 수 없다. 더군다나 이러한 주변 문맥으로 사용하는 품사 정보가 세분화되어 있지 않을 경우 학습 말뭉치에서 추출할 수 있는 정보가 줄어들어 해석 가능한 형태소 분석 결과가 늘어나며 그에 따라 품사 태깅의 모호성이 증가한다.

본 연구에서 대상으로 하는 말뭉치의 품사 분류 체계(컴퓨터·소프트웨어기술 연구소, 1999)는 기존 품사 분류 체계(김재훈 외, 1996)에 비해 세분화되지 않았다. 품사의 종류가 세분화되지 않을 경우, 적은 수의 품사를 사용하므로 공간이나 속도면에서 효율적이다. 그렇지만 서로 다른 문맥에서 사용되는 형태소들이 하나의 품사로 취급되어 품사 사이의 변별력이 떨어져서 문맥 정보의 신뢰도가 저하된다. 본 연구에서는 품사 분류 단순화에 따라 품사열 외의 정보가 필요함을 지적하고, 학습 말뭉치에서 추출한 어휘 정보를 이용하여 해결하는 방법을 보인다. 또한 정보 추출(Information Extraction)에서 사용하는 고유 명사 분류법(Named Entity Classification)을 이용하여 미등록어 추정에 사용하는 방법도 보인다.

†본 연구는 제 1회 형태소 분석기 및 품사 태거 대회(MATEC '99) 일환으로 수행되었다.

2 관련 연구

2.1 확률 기반 품사 태깅

$W = \{w_1, w_2, \ldots, w_n\}$을 문장을 구성하는 단어열이라고 하고, $T = \{t_1, t_2, \ldots, t_n\}$을 그 단어열에 해당하는 품사열이라고 할 때, 품사 태깅의 문제는 식 1과 같이 품사열 $\Phi(W)$를 찾는 문제로 정의된다.

$$\Phi(W) = arg\ \max_T p(W|T)p(T) \quad (1)$$

식 1에서 전이 확률 $p(T)$와 어휘 확률 $p(W|T)$는 Markov Independence Assumption에 의해서 단순화된다(강인호, 1999). 품사의 발생 확률이 바로 앞 두 개의 품사로 결정된다고 볼 경우, 이를 tri-tag model이라고 하며 식 1은 식 2로 단순화된다.

$$p(T|W)p(W) = p(t_1)p(t_2|t_1) [\prod_{i=3}^{n} p(t_i|t_{i-1}, t_{i-2})] [\prod_{i=1}^{n} p(w_i|t_i)] \quad (2)$$

2.2 품사 태깅의 단위

영어에서는 단어 단위로 품사가 결정된다. 그러나 한국어 품사 태깅에서는 영어의 단어에 해당하는 것을 한국어의 어절로 볼 것이냐, 형태소로 볼 것이냐가 문제가 된다. 어절 단위의 품사 태깅 방법은 단어를 이루는 형태소들의 분석에 상관없이 품사 개수가 단어 개수와 같기 때문에 영어 품사 태깅에 사용하는 수식을 그대로 사용할 수 있다는 장점을 가진다. 반면 형태소 단위의 품사 태깅은 확률 정보 추출의 용이라는 장점을 가지나 어절에 대한 형태소 개수 정규화(Normalization)의 문제점(이운재, 1993)을 가진다. 본 연구에서는 형태소 단위로 품사 태깅이 이루어지며 형태소 개수에 대한 정규화 문제를 고려하지 않는다(이상호, 1995). 이는 정규화 문제를 고려하지 않고도 좋은 성능을 얻을 수 있으며 모델이 간단해지기 때문이다.

2.3 *MATEC* '99 품사 분류의 문제점

본 연구에서 대상으로 하는 말뭉치는 총 27개의 품사로 구성되었다. 이러한 품사 분류 체계(컴퓨터·소프트웨어기술 연구소, 1999)는 기존의 세분류 품사 체계(김재훈 외, 1996)에 비해 품사가 줄 수 있는 정보가 줄어들어, 형태소 해석 능력을 떨어뜨린다. 대표적인 예로 명사의 구분을 들 수 있다[1]. 서술성 명사와 비서술성 명사를 구분하지 않음으로 해서 형태소 해석에서 모호성이 더 증가한다. 예를 들어 "학생하고"라는 어절을 고려할 경우 '학생'을 단지 명사로만 본다면 서술성 명사일때의 해석과 비서술성 명사일때의 해석 두 가지가 다 가능하게 된다(그림 1). 즉 그림 1[2]에서 '학생/nc+하/xsv+고/ec'와 '학생/nc+하/xsm+고/ec' 두 후보를 제외시킬 수 없다. 서술성 명사로 해석되는 경우를 제외하기 위해서는 품사 이외의 별도의 언어 지식을 필요로 한다.

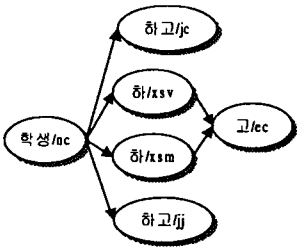

그림 1: 어절 "학생하고"에 대한 형태소 분석 결과

품사열 정보의 단순화에 따라 품사 태깅의 학습에도 문제가 생긴다. 즉 품사열에 의해서 구분할 수 있는 경우가 그 만큼 줄어들기 때문이다. 그림 2와 같이 어절 "침범합니다"에 대한 형태소 분석 결과를 고려할 때 '침범'이라는 어휘가 동작성 명사임을 알지 못할 경우, 주변 품사열이 똑같게 되어 '하'의 품사를 정해줄 수 없게 된다. 이럴 경우 학습 말뭉치의 확률적인 특성에 따라 거의 일률적인 정답만을 제시하는 형태가 되어 품사 태깅의 많은 오류 부분을 차지한다.

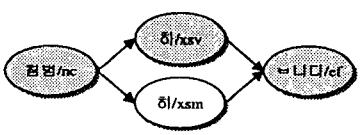

그림 2: 어절 '침범합니다'에 대한 형태소 분석 결과

이와 같은 예로 조사 '와/과, 하고, 랑'을 들 수 있다. 이들은 격조사와 접속격 조사 두 가지로 사용되는데 뒤 따르는 용언의 성질에 따라서 구분된다(강인호, 김재훈, 김길창, 1998), 이러한 경우들은 기존의 확률 모델에서 사용하는 품사열 정보만으로는 해결할 수 없다. 이에 따라 품사열 문맥 정보를 보완하는 어휘 정보의 추가가 필요하다. 본 연구에서는 품사열 정보와 어휘 정보를 함께 사용하기 위해 최대 엔트로피 모델을 사용한다.

3 최대 엔트로피 모델

본 연구에서 다양한 정보들을 결합하기 위해 사용하는 최대 엔트로피 모델(Maximum Entropy Model)에 대해서 알아본다. 최대 엔트로피 모델에서는 주어진 여러 정보를 자질 함수(feature function)라는 형태로 정의한다. 자질 함수는 trigger 형태로써, 정해놓은 제약조건을 만족하였는지 그렇지 않은

[1](김재훈 외, 1996)에서는 총 54개의 품사로 분류하며, 명사의 경우 서술성 명사, 비서술성 명사, 그리고 고유 명사로 구분하였다. 서술성 명사는 다시 동작성 명사와 상태성 명사로 구분한다.

[2]xsv는 동사 파생 접미사를 나타내며, xsm은 형용사 파생 접미사를 나타낸다.

지를 구분해주는 함수이다. 이러한 자질 함수를 통해 고려되고 있는 문맥에 사용하고자 하는 정보들이 적용가능한지를 결정한다.

$$f(h) = \begin{cases} 1 & \text{if } h \text{ meets some condition} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

최대 엔트로피 모델은 최대 엔트로피 원리(Maximum Entropy Principle)에 기반하여 만들어진다. 최대 엔트로피 원리란 랜덤 변수 $x_i(i = 1, 2, \ldots, n)$에 대한 확률 분포를 p_i라고 할 때, 자질 함수 f에 대해서 다음과 같은 제약 조건을 가한다(Berger, Pietra, and Pietra, 1996).

$$E[f_j] = \tilde{E}[f_j],\ 1 \le j \le k \quad (4)$$

$$E[f_j] = \sum_{h \in H, y \in Y} p(h,y) f_j(h,y) \quad (5)$$

$$\tilde{E}[f_j] = \sum_{i=1}^{n} \tilde{p}(h_i,y_i) f_j(h_i,y_i) \quad (6)$$

H와 Y는 각각 있을 수 있는 모든 문맥과 원하는 출력값의 집합이며, n은 학습 데이터에서 발견된 문맥 h와 y의 곱집합으로 얻을 수 있는 총 가지수로써 모델에서 고려하는 경우의 수를 의미한다. 그리고 k는 사용하는 자질 함수의 총 가지수를 나타낸다. 식 5에서 고려할 수 있는 H는 제대로 알기가 어려우며, 안다고 하더라도 너무 커서 평균값을 바로 구하는 것이 힘들다. 그래서 학습 데이터에서 발견한 경우만 고려하는 근사화 된 수식을 이용하여 계산한다. $\tilde{p}$와 $\tilde{E}$는 학습 말뭉치에서 얻어낸 값을 의미한다.

$$E[f_j] = \sum_{i=1}^{n} \tilde{p}(h_i) p(y_i|h_i) f_j(h_i,y_i) \quad (7)$$

그리고 p_i는 확률값이기 때문에 합쳐서 1이 된다. 이렇게 두 개의 제약조건을 만족하는 확률 분포는 근사화에 의해서 실제 가능한 모든 경우를 고려하지 않았기 때문에 고려하지 않은 영역에 대해서는 줄 수 있는 확률값이 여러개가 가능하기 때문에 확률 분포 또한 여러개가 가능하다. 가능한 확률 분포 중에서 엔트로피(식 8)가 최대인 모델을 선택하는 것이 최대 엔트로피 원리이다.

$$H(p) = -\sum_{h \in H, y \in Y} p(h,y) \log p(h,y) \quad (8)$$

엔트로피를 최대로 만든다는 것은 특정 부분에 치우치지 않는 분포를 구하겠다는 뜻이다. 즉 최대 엔트로피 원리는 제약조건(식 4, 5, 6)을 이용하여 알려진 또는 사용하고자 하는 정보에 대해서는 확실히 지켜주고, 고려하지 않은 경우나 모르는 경우에 대해서는 어느 하나가 낫거나 덜하다는 근거가 없기 때문에 동등하게 가중치를 줌으로 해서 특정 부분에 치우치지 않는 분포를 구한다는 뜻이다(강인호, 1999).

최대 엔트로피 원리에 의한 파라미터 추정법은 (Jaynes, 1957)에 의해 제시되었고, 이를 수치적으로 측정하는 GIS(Generalized Iterative Scaling)방법이 (Darroch and Ratcliff, 1972)에 의해 고안되었다. 요즘은 GIS 방법을 발전시킨 IIS(Improved Iterative Scaling) 방법이 많이 사용된다(식 9).

$$p(h,y) = \pi \prod_{i=1}^{k} \alpha_j^{f_j(h,y)} \quad (9)$$

식 9에서 π는 정규화를 위한 상수이고 α_j는 모델 파라미터[3]로써, 자질 f_j가 미치는 영향 정도를 나타낸다.

최대 엔트로피 모델에 기반하여 한국어 품사 태깅 모델을 정의하면 식 10과 같다. 주어진 문장 (S)에 대해서 가장 적당한 품사열을 찾는 것이다.

$$p(t_1, \ldots, t_k|S) = \prod_{i=1}^{k} p(t_i|h_i) \quad (10)$$

$p(t_i|h_i)$는 식 11과 같이 조건부 확률식으로 구할 수 있으며, $p(t_i, h_i)$는 식 9에서 얻을 수 있다.

$$p(t|h) = \frac{p(h,t)}{\sum_{t' \in T} p(h,t')} \quad (11)$$

4 정보 추출

본 절에서는 형태소 분석기와 품사 태거에 필요한 언어 지식 및 확률 정보를 학습 말뭉치에서 얻는 방법을 보인다.

4.1 형태소 분석기

4.1.1 형태소 사전과 결합 정보

형태소 분석기를 만들기 위해서 학습 말뭉치로부터 형태소 사전을 만든다. 형태소 사전은 형태소와 그 형태소가 가질 수 있는 품사 집합으로 구성된다. 형태소 사전을 이용하여 주어진 어절에 대해서 가능한 형태소들을 얻어 낸다. 이렇게 얻어낸 형태소들을 무조건 다 제시할 경우 어절에 대해 가능한 형태소 해석수가 너무 많아진다. 그래서 연속적으로 나타나는 형태소가 결합이 가능한지를 알려주는 결합 정보를 사용한다. 결합 정보는 학습 말뭉치에서 발견된 연속된 품사쌍으로 구성된다. 품사쌍으로 구성된다. 이때 가능한 형태소 후보 분석 결과수를 줄이기 위해 폐쇄류(Closed Class)에 해당하는 형태소에 대해서는 형태소와 품사를 같이 결합한 형태소 결합 정보를 얻어낸다.

[3] α_j는 e^{λ_j}의 형태이다.

폐쇄류는 특성상 그 개수가 한정되어 있으며 품사 분류로는 부족한 부분을 보완하기 위함이다. 추출해내는 결합 정보의 예를 보이면 표 1과 같다.

표 1: 형태소 결합 정보 예

좌 측	우 측	예
nc	pv	자리/*nc+잡/pv+고/ec*
nc	pa	연맹/*nc+같/pa+은/etm*
끼리/xsn	spa	그/*np+들/xsn+끼리/xsn*
pa	ㄹ지/ef	어떠하/*pa+ㄹ지/ef+?/s*
들/xsn	이/jc	짐승/*nc+들/xsn+이/jc*

형태소 결합 정보 외에 음절 결합 정보도 추출한다. 음절 결합 정보는 어절 안에서 연속적으로 나타나는 두 개 형태소의 품사 정보 외에 앞 형태소의 마지막 음절과 뒤 형태소의 첫 음절로 구성된다. 음절 결합 정보는 해당 어절이 미등록어를 포함하고 있을 경우 추정 형태소들이 연결 가능한지를 알아볼 때 사용한다. 예를 들어 미등록어를 포함하는 "쪼들리는데다가"라는 어절에 대해서 '쪼들리느/*pv+ㄴ데다가/ec*'의 연결같이 자연스럽지 않은 해석을 막아서 후보수를 줄인다.

4.1.2 입력 문장 전처리 및 패턴을 이용한 미등록어 추정

입력 문장에 나타난 주석이나 삽입문에 대해서 전처리를 하여 입력 문장의 길이를 줄이며 직접적으로 연관 있는 문맥을 사용한다. 예 (1)과 같은 문장의 경우 괄호 안의 삽입 구문을 제거한 문장과 삽입 문장 두개의 입력으로 나눌 수 있다. 이에 따라 형태소 '과'에 대한 문맥 정보를 '('가 아닌 '개발국'을 사용한다.

(1) 지구 전체로 보면 개발국*(인구의 20%*가 상품의 *80%*를 소비한다*)*과 저개발국간의 불균형은 점점 더 심화되고 있다.

인명, 지명, 단체의 경우 괄호를 이용하여 한자나 영어 표기를 덧붙이기도 한다. 이런 경우 원 단어에 대한 의미 설명 및 한글에 대한 원어를 알려주는 역할을 한다. 괄호안의 표기를 이용해서 괄호 밖의 형태소를 추정할 수 있다. 한자의 경우 그 독음이 괄호 밖의 글자와 같다면 형태소가 비록 사전에 들어있지 않는 미등록어인 경우에도 처리할 수 있다. 또한 영어의 경우에는 음차 표기 방식(transliteration)(강인호 김길창, 1999)을 이용해서 비교를 할 수 있다. 본 연구에서는 괄호 표기가 나올 경우 독음과 음차 표기를 이용해서 처리한다.

(2) 달콤한 향료 네롤리*(Neroli)* 유를 추출해서

(3) 한방의 약명으로는 상기생*(桑寄生)*, 우목*(寓木)*, 기동수*(奇童樹)*, 기생수*(奇生樹)*라고도 하며

또한 기자나 의장과 같은 직위에 관련된 형태소와 성과 이름의 형태로 사람 이름의 가능성이 있는 형태소가 연달아 나타날 경우 인명으로 추정한다. 이는 정보 추출(Information Extraction)에서 사용하는 고유 명사 분류(Named Entity Classification) 기술을 응용한 것이다.

4.1.3 형태소 분석의 과도한 후보 분석 제거

품사 태깅의 성능을 향상시키기 위해서는 형태소 분석에 걸리는 시간과 형태소 분석기에서 제시하는 후보수의 감소와 후보간의 모호성을 줄여야 한다. 학습 말뭉치에서 출현 빈도수를 기준으로 상위 10%에 속하는 어절에 대해서는 발생 가능한 모든 형태소 분석이 학습 말뭉치에 있다고 가정하고 형태소 분석을 수행 않고 학습 말뭉치의 분석을 출력하는 것으로 형태소 분석을 대신한다. 또한 학습 말뭉치에서 발견할 수 있는 형태소 분석 결과가 일정수(15)를 초과하는 어절에 대해서도 가능한 모든 형태소 분석을 얻어낸 것으로 가정하고 학습 말뭉치의 결과를 형태소 분석기의 결과물로 대체한다.

4.2 품사 태깅

4.2.1 품사 태깅을 위한 어휘 정보

품사열 정보로 해결할 수 없는 품사 구분을 위해서 어휘 정보를 추출한다. 어휘 정보는 폐쇄류 형태소에 대해서 식 12와 같은 형태로 추출한다.

$$f(h_i,t_i)=\begin{cases}1 & if\ m_{i-1}=\mathcal{M}_{i-1}\ \&\ t_{i-1}=\mathcal{T}_{i-1}\ \&\\ & t_i=\mathcal{T}_i\ \&\ m_{i+1}=\mathcal{M}_{i+1}\ \&\\ & t_{i+1}=\mathcal{T}_{i+1}\\ 0 & otherwise\end{cases} \quad (12)$$

식 12에서 $\mathcal{M}$과 $\mathcal{T}$는 특정 어휘와 품사를 나타내며, m_i는 i번째 형태소를 나타낸다. 이렇게 추출한 어휘 정보들은 학습 말뭉치의 오류를 제거하기 위해서 특정 빈도수 이상 발생하는 것만을 사용한다. 추출된 어휘 정보들은 최대 엔트로피 원리에 의해서 그 유용성에 따라 가중치가 결정된다.

4.2.2 미등록어 추정

영어 품사 태깅의 경우에는 미등록어를 해결하기 위해서 특정 접미사가 나타난 단어들의 확률 분포를 보거나, 대/소문자의 여부, 하이픈의 사용 여부 등의 정보를 이용한다(Weischedel et al., 1993). 그러나 한국어에는 해당하지 않기 때문에 주변 형태소와 품사를 이용해서 미등록어의 품사와 형태소를 추정한다. 본 연구에서는 어절의 첫번째 형태소가 미등록어라는 가정을 사용한다. 학습 말뭉치에서 미등록어 가능성이 있는 어절의 첫번째 형태소를 주변으로 앞, 뒤 형태소의 어휘와 품사 정보를 사용하는 형태로 정형화하였다. 여기서 $\mathcal{D}$는 사전을 의미한다.

$$f(h_i,t_i)=\begin{cases}1 & if\ m_i\notin\mathcal{D}\ \&\ t_{i-1}=\mathcal{T}_{i-1}\ \&\\ & t_i=\mathcal{T}_i\ \&\ m_{i-1}=\mathcal{M}_{i-1}\\ 0 & otherwise\end{cases} \quad (13)$$

$$f(h_i, t_i) = \begin{cases} 1 & if\ m_i \notin \mathcal{D}\ \&\ t_{i+1} = \mathcal{T}_{i+1}\ \& \\ & \quad t_i = \mathcal{T}_i\ \&\ m_{i+1} = \mathcal{M}_{i+1} \\ 0 & otherwise \end{cases} \quad (14)$$

본 연구에서 품사 태깅에 사용하는 정보는 품사열 정보인 bigram, trigram과 함께 형태소와 품사간의 확률 관계를 나타내는 어휘 확률 정보 그리고 폐쇄류 형태소를 중심으로 한 어휘 정보, 미등록어를 위한 어휘 정보이다. 이렇게 추출한 정보로 구성된 품사 태거는 학습 말뭉치에 실험 된다. 실험을 통해 발견한 오류는 폐쇄류 형태소와 함께 식 12 형태로 다시 어휘 정보를 추출하여 오류를 해결하기 위한 주변 어휘 문맥 정보 추출에 사용한다.

5 실험 및 결과 분석

학습 데이터로 사용한 문장은 2만여 문장으로써 소설 분야와 비소설 분야로 구성된다. 품사 부착된 말뭉치를 통해 총 22,000 엔트리의 형태소 사전을 만들었다. 그리고 품사 부착 가이드 책자(컴퓨터·소프트웨어기술 연구소, 1999)를 이용하여 축약 정보와 불규칙 정보를 얻었다.

표 2: 학습 말뭉치 구성

영역	어절수
비소설	96,365
소설	158,049
전체	254,414

시험용 데이터로 사용한 문장은 소설, 비소설 그리고 뉴스 영역으로 구성된다. 여기서 뉴스 영역은 학습 데이터에 사용되지 않은 영역으로 사람 이름이나 고유 명사의 발생이 빈번하다. 뉴스 영역의 경우 고유 명사 추정 기법을 사용하기 위해 인명 추정 사전을 사용하였다. 인명 추정 사전은 만 여명의 사람 이름으로 구성 되었다. 따라서 3만 형태소 사전을 이용해서 품사 태깅을 수행하였다.

품사 태깅을 위해서 먼저 폐쇄어를 대상으로 추출한 어휘 정보와 학습 말뭉치 실험 결과 오류 형태소를 대상으로 추출한 어휘 정보를 합치면 약 1만 여 개가 된다. 그리고 미등록어를 위해 추출한 어휘 정보는 약 3만 여 개가 된다. 이렇게 약 4만 여 개의 어휘 정보와 품사열 정보를 최대 엔트로피 원리에 기반하여 결합하였다. 이렇게 만들어진 품사 태깅 모델로 실험을 한 결과는 표 3과 같다.

6 결론 및 향후 연구

본 연구에서는 품사열 문맥 정보만으로는 품사 분류 기준을 제대로 설명할 수 없으므로 어휘 정보로 보완 해야 함을 보였다.

표 3: 어절 단위 평가 결과

분야	어절 단위		형태소 단위	
	어절수	정확률	재현률	정확률
뉴스	12,649	91%	94%	94%
비소설	12,194	92%	95%	94%
소설	9,012	88%	93%	92%
전체	33,855	90%	94%	93%

또한 형태소 분석 결과를 저장하는 방식으로 과도한 형태소 분석 후보 제시를 막았다. 그리고 전처리 기법과 고유 명사 분류법을 이용하는 미등록어 추정법도 보였다. 약 3만 어절에 대해서 실험을 한 결과, 형태소 단위로 94%의 재현률과 93%의 정확률을 얻었다. 비록 좋은 결과는 아니지만 특별한 언어 지식의 추가 없이 학습 말뭉치의 학습에 의해 얻은 결과이므로 다양한 분야에 대해서 쉽게 학습하여 적용할 수 있는 모델이라는 것을 알 수 있다.

품사 태깅은 그 사용 목적에 따라서 품사 분류가 결정되어야 한다. 정보 검색을 위한 품사 태깅일 경우에는 서술성 명사와 비서술성 명사의 구분이 그렇게 중요한 정보가 되지 못한다. 그러나 이것은 어디까지나 품사 태깅 결과물 사용자 입장에서의 관점이다. 실제 형태소 분석을 하고 품사 태깅을 하는 관점에서는 이러한 분류 체계가 내부적으로 되어 있거나 이를 구분할 수 있는 정보가 있어야 한다. 따라서 학습에 사용되는 데이터에는 이러한 정보를 쉽게 얻어 낼 수 있어야 할 것이고 품사의 단순화는 그 정보를 이용해서 나오는 결과에 대한 신뢰도와 목적에 따라서 이루어져야 된다.

앞으로 불규칙 및 축약 현상에 대한 자동 규칙 추출에 대한 연구가 이루어진다면 품사 태깅 모델을 위한 수작업 및 고급 지식이 상당히 줄어들 것이다. 본 연구에서 사용한 소설 분야 시험 데이터 경우에는 대화체 문장 및 사투리의 사용으로 인해 축약 현상이 빈번하였다. 축약에 대해서도 자동적으로 축약 정보 사전을 작성할 수 있는 방법에 대한 연구가 필요하다. 즉 학습 말뭉치에서 뽑아낼 수 있는 정보들에 대한 연구가 지속되으로 되어야 한다.

참고문헌

강인호, 김재훈, 김길창. 1998. 최대 엔트로피 모델을 이용한 한국어 품사 태깅. 제 *10*회 한글 및 한국어 정보처리 학술대회, pages 9–14.

강인호. 1999. 최대 엔트로피 모델을 이용한 한국어 품사 태깅. 한국과학기술원 전산학과, 석사학위논문.

강인호 김길창. 1999. 복수 음운 정보를 이용한 영한 음차 표기. 제 *11*회 한글 및 한국어 정보처리 학술대회.

김재훈 김덕봉 외. 1996. 통합국어정보 베이스 제 2차년도 최종 보고서.

김재훈. 1996a. 오류-보정 기법을 이용한 어휘 모호성 해소. 한국과학기술원 전산학과, 박사학위논문.

남기심 고영근. 1994. 표준 국어 문법론. 탑 출판사.

이운재. 1993. 한국어 문서 태깅 시스템의 설계 및 구현. 한국과학기술원 전산학과, 석사학위논문.

이상호. 1995. 미등록어를 고려한 한국어 품사 태깅 시스템 구현. 한국과학기술원 전산학과, 석사학위논문.

정성영. 1996b. 마코프 랜덤 필드를 이용한 영어 품사 태깅 시스템. 한국과학기술원 전산학과, 석사학위논문.

컴퓨터·소프트웨어 기술 연구소, 지식정보연구부. 1999. 품사 부착 말뭉치 구축 지침. 한국전자통신연구원.

Berger, A., V. Della Pietra, and S. Della Pietra. 1996. A maximum entropy approach to natural language processing. In *Computational Linguistics*, pages 39–71.

Charniak, Eugene. 1993. *Statistical Language Learning*. The MIT press.

Darroch, J.N. and D Ratcliff. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, volume 43, pages 1470–1480.

Jaynes, E.T. 1957. Information theory and statistical mechanics. *Physics Reviews106*, pages 620–630.

Pietra, Stephen Della, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. *IEEE Trans. Pattern Anal. Machine Intell.*

Ratnaparkhi, Adwait. 1996. A maximum entropy part-of-speech tagger. *In Proceedings of the Empirical Methods in Natural Language Processing Conference*, pages 17–18, May.

Ratnaparkhi, Adwait. 1997. A simple introduction to maximum entropy models for natural language processing. Technical Report IRCS 97-08.

Rosenfeld, Ronald. 1994. Adaptive statistical language modeling: A maximum entropy approach. Technical report, CMU-CS-94-138.

Weischedel, R., R. Schwartz, J. Palmucci, M. Meteer, and L. Ramshaw. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, pages 359–382.