# 일반화된 미등록어 처리와 오류 수정규칙을 이용한 혼합형 품사태깅

차정원 이원일 이근배 이종혁
포항공과대학교 전자계산학과

# Hybrid POS Tagging with generalized unknown word handling and post error-correction rules

Jeongwon Cha Wonil Lee Geunbae Lee Jong-Hyeok Lee
Dept. of Computer Science and Engineering
Pohang University of Science and Technology

## 요 약

본 논문에서는 품사 태깅을 위해 여러 통계 모델을 실험을 통하여 비교하였으며 이를 토대로 통계적 모델을 구성하였다. 형태소 패턴 사전을 이용하여 미등록어의 위치와 개수에 관계없는 일반적인 방법의 미등록어 처리 방법을 개발하고 통계모델이 가지는 단점을 보완할 수 있는 오류 수정 규칙을 함께 이용하여 혼합형 품사 태깅 시스템인 POSTAG[i]를 개발하였다. 미등록어를 추정하는 형태소 패턴 사전은 한국어 음절 정보와 용언의 불규칙 정보를 이용하여 구성하고 다어절어 사전을 이용하여 여러 어절에 걸쳐 나타나는 연어를 효과적으로 처리하면서 전체적인 태깅 정확도를 개선할 수 있다. 또 오류 수정 규칙은 Brill이 제안한 학습을 통하여 자동으로 얻어진다. 오류 수정 규칙의 자동 추출시에 몇 가지의 휴리스틱을 사용하여 보다 우수하고 일반적인 규칙을 추출할 수 있게 하였다. 10만의 형태소 품사 말뭉치로 학습하고 학습에 참여하지 않은 2만 5천여 형태소로 실험하여 97.28%의 정확도를 보였다.

## 1. 서 론

품사 태깅은 문자 기반의 정보추출, 음성인식 그리고 음성합성 등을 포함한 자연어 처리의 기본이 되는 작업이다. 품사 태깅은 자료 부족 문제, 미등록어 문제, 정거리 의존 등 많은 문제점들이 있다.

그 중에서 미등록어 문제는 품사 태깅과 태깅 응용에서 가장 심각한 문제가 된다. 현재까지의 미등록어 추정 방법들은 형태소 분석이 실패한 어절들에 대해서 조사, 어미 사전들을 이용하여 기능 형태소를 분리해내고 이들과 접속 가능한 품사를 미등록어 품사로 예측하는 방법을 취해왔다. 그러나 이러한 방법은 두 가지의 가정으로부터 시작되는데 하나는 미등록어는 어절의 앞부분에 존재한다는 것이고 또 하나는 어절에서 미등록어는 하나라는 것이다. 그러나 접두사가 붙는 경우는 미등록어의 위치는 어절의 앞부분이 아닐 수 있고 복합어의 경우는 미등록어가 하나가 아닐 수 있다. 예를 들어 "정기세미나에서"에서 "세미나"가 미등록어인 경우 어절의 앞부분이 아니다. 또한 "황장엽비서가"라는 어절에서 '가'라는 조사를 분리하고 이와 접속 가능한 미등록어로써 추정을 한다면 "황장엽비서" 모두를 미등록어로 추정하는 오류를 범하게 된다.

이러한 문제를 해결하게 위해 본 논문은 미등록어용 형태소 패턴 사전을 이용하여 어절내의 위치와 개수에 관계없이 추정이 가능한 방법을 사용하고 한국어 음절정보를 확장하여 미등록어용 패턴 사전 구성과 형태적 모호성 축소에 이용하며 [1]에서 제안한 다어절어(多語節語, multi-word) 사전을 이용하여 정확도를 높였다.

이러한 방법은 형태소 분석기의 구조를 간단하게 만들 수 있고 여러 어절에 걸친 한국어 연어의 일관성 있는 처리가 가능하다. 또한 위치와 개수에 관계없는 미등록어 추정이 가능해짐으로써 인터넷 문서와 같이 다양한 형태의 문서에 대한 분석에 훌륭히 적용될 수 있다. 일반화된 미등록어 처리의 자세한 내용은 [1]을 참조하기 바란다.

앞으로 우리 연구실에서 개발중인 세션 기반 인터넷 IR (Information Retrieval) 시스템인 air-web[ii]시스템의 웹문서 인덱싱에 사용하게 될 것이다.

또한 일반적으로 품사 태깅은 은익 마코프 모델(HMM)을 기본으로 하는 통계적인 방법을 이용한다[2,3,4]. 하지만 통계적 모델은 제한된 윈도우 안의 형태소들만을 참조하기 때문에 윈도우를 벗어난 문맥을 참조할 수 없고 그 모델의 특성상 뒤에 오는 형태소를 참조할 수 있는 방법이 없다. 또한 동일한 문맥을 갖는 다른 형태소를 구분할 수 있는 방법이 없다.

이러한 통계적 모델의 단점을 극복하기 위해서 말뭉치로부터 자동으로 규칙을 구하는 연구들이 최근에 많이 연구되고 있다[5]. 그러나 규칙에 의한 방법은 일반적으로 새로운 언어나 새로운 품사에 적용할 때 강건함이 부족하며 성능에서도 통계적 방법보다 우수하지 못하다.

따라서 통계적 방법의 이식성과 강건함을 이용하면서 통계적 방법의 단점을 보완해주는 규칙에 의한 방법을 결합한 hybrid method 가 필요하다.

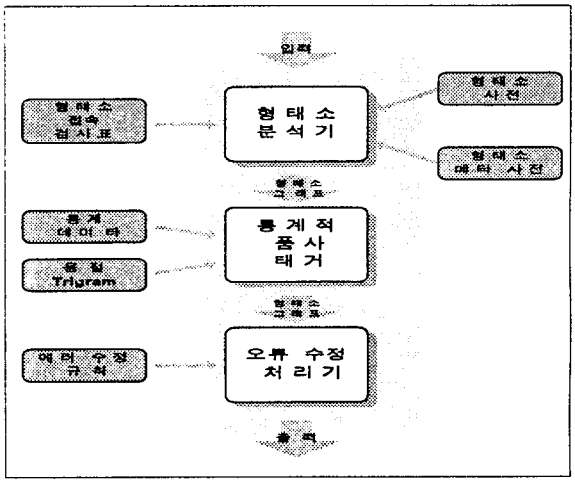
본 논문의 구성은 다음과 같다. 2 장에서는 혼합형 태거 관련 연구를 설명하고 3 장에서는 기본 개념을 4 장에서는 시스템의 세부 설명한다. 그리고 5 장에서는 실험 결과를 설명하고 마지막으로 6 장에서는 결론을 내린다.

## 2. 관련 연구

통계적 방법과 오류 수정 규칙을 이용한 방법은 [6]에서 최초로 시도되었다. 이 논문에서는 POSTAG 와 동일하게 통계 태거와 오류 수정 태거를 직렬로 연결한 모델을 사용하였다. 오류 수정 규칙은 형태소 단위로 문맥을 고려하여 수정할 수 있게 하였으며 여러 개가 AND 나 OR 로 연결되어 적용될 수 있다. [7]에서는 어절 단위로 확장한 규칙을 사용하였다. 본 논문에서는 형태소에 기반한 자료구조 위에서 형태소와 어절이 결합된 오류 수정 규칙을 새로히 개발하여 사용한다.

## 3. 기본 개념

그림 1 은 본 논문에서 제안한 hybrid 품사 태깅 시스템을 보여준다. 시스템은 형태소분석기, 통계 적 품사 태거, 에러 수정 규칙 처리기로 구성되어 있으며 이들은 각각 단순히 직렬 연결되어 있는 것이 아니라 상호 영향을 미친다. 형태소 분석과 통계적 품사 태거는 문장을 분석하는 과정에서 서로의 결과를 참조하면서 진행하고 에러 수정 규칙 처리기는 통계 태거의 오류로부터 학습되었으므로 통계 태거의 오류 특성과 밀접히 관련되어 있다.




[그림 1] 시스템 구조

형태소 분석기는 형태소 사전과 다어절어 사전, 그리고 형태소 패턴 사전을 이용하여 등록어와 미등록어의 형태소를 동일한 방법으로 분리하고 원형을 복원하며 접속 검사표를 이용하여 접속을 검사한다. 접속이 이루어진 형태소들은 부분적인 형태소 그래프를 형성하고 이 그래프에 대하여 통계정보를 이용하여 Viterbi 탐색을 이용하여 태깅을 한다. 이 때 미등록어의 통계 정보는 음절 Trigram 을 이용하여 구해진다. 한 문장에 대하여 최적의 품사열이 정해지면 에러 수정 처리기가 통계 정보 태거의 오류들을 올바르게 수정하여 최종적으로 가장 올바른 결과를 출력한다.

## 4. 시스템의 세부 설명

### 4.1 통계 태깅 모델의 비교 실험

통계적 처리에 의한 품사 태깅 방법은 근본적으로 주어진 문장 $W$ 에 대하여 (1)을 만족하는 품사열 $T$ 를 찾는 것이다.

$$\phi(W) \equiv \arg\max_{T_i} P(T_i|W) = \arg\max_{T_i} \frac{P(T_i,W)}{P(W)} = \arg\max_{T_i} P(T_i,W) \quad (1)$$

여기서 현재 단어의 품사는 $k$ 개의 이전의 품사

에만 영향을 받고 현재 단어는 현재 품사에 관한 정보에만 영향을 받는다는 Markov 가정을 적용한 은익 마코프 모델(HMM)을 얻을 수 있다.

$$\phi(W) = \arg\max_{T} \prod_{i=1}^{n} P(t_i|t_{i-k},\ldots,t_{i-1})P(w_i|t_i) \quad (2)$$

현재 사용되고 있는 많은 모델은 $P(t_i|t_{i-k},\ldots,t_{i-1})$에서 $k$를 2,3으로 사용하여 품사의 문맥정보를 제한하고 있다. 또한 식(2)를 변형한 모델로 식(3)과 같은 모델을 사용하는 시스템도 있다[8,9]

$$\phi(W) = \arg\max_{T} \prod_{i=1}^{n} P(t_i|t_{i-k},\ldots,t_{i-1}) \frac{P(w_i)P(t_i|w_i)}{P(t_i)} \quad (3.1)$$

$$\equiv \arg\max_{T} \prod_{i=1}^{n} P(t_i|t_{i-k},\ldots,t_{i-1})P(t_i|w_i) \quad (3.2)$$

본 연구에서는 위에서 열거한 각 모델들을 다음과 같이 비교 실험해서 가장 우수한 모델을 선택하여 사용했다.

실험환경은 "계몽사 백과사전"중 357 문장을 대상으로 실험했으며 전체 어절수는 4537 (한 문장당 12.71 어절)개이며 전체는 10204 형태소(한 문장당 28.58 형태소)로 굉장히 길고 복잡한 문장들이다.

여기서 $\alpha$와 $\beta$는 가중치를 의미하며 $\alpha = 0.4, \beta = 0.6$을 사용하여 실험했다.

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

모 델 2: $\arg\max_{T} \prod_{i=1}^{n} \alpha P(t_i|t_{i-1})\beta P(w_i|t_i)$

모 델 3: $\arg\max_{T} \prod_{i=1}^{n} P(t_i|t_{i-1})P(t_i|w_i)$

모 델 4: $\arg\max_{T} \prod_{i=1}^{n} \alpha P(t_i|t_{i-1})\beta P(t_i|w_i)$

모 델 5: $\arg\max_{T} \prod_{i=1}^{n} P(t_i|t_{i-1}) \frac{P(t_i|w_i)}{P(t_i)}$

모 델 6: $\arg\max_{T} \prod_{i=1}^{n} \alpha P(t_i|t_{i-1})\beta \frac{P(t_i|w_i)}{P(t_i)}$

실험결과 각 모델별 정확도는 다음과 같다.

(단위: %)

| | 모델 1 | 모델 2 | 모델 3 | 모델 4 | 모델 5 | 모델 6 |
|---|---|---|---|---|---|---|
| 어절 | 86.80 | 90.48 | 89.40 | 89.62 | 91.73 | 92.48 |
| 형태소 | 91.32 | 94.93 | 94.40 | 94.48 | 95.77 | 96.12 |

실험결과 모델 6 이 형태소 정확도 96.12%로 가장 우수했다.

### 4.2 통계 태깅 모델

4.1 의 실험결과를 바탕으로 본 시스템의 통계 모델을 결정하였다

**어휘확률**

어휘확률은 4.1 의 실험 결과를 바탕으로 모델 6 의 어휘확률($\frac{P(t_i|w_i)}{P(t_i)}$)을 사용했다. 미등록어의 경우는 [1]에서 제안한 음절 Trigram 을 사용한다. 그러나 음절 trigram 또한 학습을 위해 많은 량의 말뭉치가 필요하므로 식(4)와 같이 평탄화(smoothing)하여 사용한다.

$$W = m_1 m_2 \ldots m_n$$

$$P(W|t) \equiv P_t(m_1|\#,\#)P_t(m_2|\#,m_1)\prod_{i=3}^{n} P(m_i|m_{i-2},m_{i-1})P(\#|m_{n-1},m_n)$$

$$P_t(m_i|m_{i-2},m_{i-1}) \equiv \frac{f_t(m_{i-2},m_{i-1},m_i)}{f_t(m_{i-2},m_{i-1})} + \frac{f_t(m_{i-1},m_i)}{f_t(m_{i-1})} + \alpha(t) \quad (4)$$

여기서

$W$: 형태소

$m$: 음절

$t$: 품사

#: 형태소경계

여기서 $f_t(m_1,m_2,m_3)$은 음절 $m_1$, $m_2$, $m_3$가 동시에 나오는 횟수를 나타낸다.

**문맥확률**

문맥확률은 식(5)와 같은 bigram 을 이용한다. 확률의 정확도를 위하여 평탄화 과정을 거친다.

$$\text{문맥확률} = P(t_i|t_{i-1}) + P(t_i) + \beta$$

통계 태거는 10 만 형태소 크기의 학습 말뭉치에서 획득한 통계 데이타를 이용였으며, 2 만 5 천 형태소의 훈련되지 않은 실험 말뭉치에서 96.04%의 정확도를 보였다.

### 4.3 에러 수정 규칙

오류 수정 규칙에서 사용한 형식은 그림 2 와 같다.

(현재 분석, 틀, 참조대상) -> 옳바른 분석
{[AND|OR] (현재 분석, 틀, 참조대상) -> 옳바른 분석}*
틀 : [P|N]n[F|L][M|T|B]

[그림 2] 규칙 틀의 형식

[그림 2]의 Pn 은 현재 어절 전 n 번째 어절 위치를 나타내고 Nn 은 현재 어절 후 n 번째 어절 위치를 나타낸다. F 는 참조대상의 어휘 형태소를 나타내고 L 은 기능 형태소를 의미한다. M 은 참조대상에서 형태소를 참조함을 나타내고 T 는 품사를 참조함을 나타낸다. B 는 형태소와 품사를 모두 참조함을 나타낸다. 예를 들어, P1FM 는 "현재 어절 이전의 첫번째 어절의 어휘 형태소의 형태소를 참조하라"를 나타낸다.

이러한 규칙은 오류와 옳바른 분석을 비교하여 만들어지는 혼동행렬에 따라 형태소 단위, 어절단위 모두를 수정하는 규칙을 만들 수 있다. 예를 들어, "나는 학교에 간다."라는 문장이 있을 때, "나는"은 "나/대명사+는/조사", "나/규칙동사+는/어미", "날/불규칙동사+는/어미"로 분석될 수 있다. 주어진 문장에서 혼동행렬이 "나/대명사+는/조사 : 날/불규칙동사+는/어미"로 만들어지면 어절 단위로 수정 가능한 규칙이 만들어지고 "여름을 나는 사람은 ..."이라는 문장에서 혼동 행렬이 "날/불규칙동사 : 나/규칙동사"로 만들어 지면 형태소 단위의 규칙을 만들 수 있다.

또한 문맥에서도 형태소 단위와 어절 단위를 모두 고려할 수 있다.

### 오류 수정 규칙의 학습

오류 수정 규칙의 학습에는 [5]에서와 같은 방법을 사용하였다. 통계 태거의 결과와 올바르게 태깅된 말뭉치를 입력으로 받아서 통계 태거의 오류 특성을 수정할 수 있는 규칙을 만들어낸다.

학습에서 고려할 사항은 오류를 가장 많이 수정하면서 잘못된 수정으로 인해 올바르게 태깅된 결과를 오히려 오류로 만들지 말아야 한다. 이런 규칙을 얻기위해서는 scoring 을 하여 적정 기준값(threshold) 이상의 규칙만 실재 규칙으로 취해야 한다. scoring 에서 또 한 가지 고려해야 할 것은 규칙을 적용할 때에도 score 가 큰 규칙부터 적용하여야 한다. 따라서 2 번 적용하여 2 번 모두 성공하는 규칙과 20 번 적용하여 20 번 성공하는 규칙은 그 score 가 달라야 한다. 이러한 score 를 계산하기 위해서 [10]에서 사용한 다음과 같은 계산식을 사용한다.

$$p = \frac{\text{올바르게 수정한 수} + 0.5}{\text{규칙을 적용한 수}(n) + 1}$$

$$score = p - 1.65 * \sqrt{\frac{p(1-p)}{n}}$$

그러므로 2 번 적용하여 2 번 성공한 규칙은 0.3985 이며 20 번 적용하여 20 번 성공한 규칙은 0.9199 이 된다. 따라서 오류를 많이 수정하는 규칙을 선택할 수 있다.

오류 수정 규칙은 10 만 형태소에 대해서 100 여 개를 획득하였다.

### 보다 일반적인 규칙을 위한 휴리스틱

학습 말뭉치에서 획득한 규칙을 전혀 다른 실험 말뭉치에 적용할 때 일반화 과정이 필요하다. 이러한 규칙을 일반 규칙이라 부르기로 하고 이 과정을 거치지 않은 규칙을 세부 규칙이라 하자. 예를 들어 "이름(name)"을 통계 태거가 항상 "이르/DI+ㅁ/eCNMM"으로 분석하여 "(이르/DI+ㅁ/eCNMM, P1FT, S) -> 이름/MC" 이라는 규칙을 만들었다고 해서 이 규칙이 바로 테스트 말뭉치에 적용되지는 않는다.

그러나 일반화 과정을 통하여 만들어진 규칙의 과도한 적용으로 인해 오히려 오류가 증가할 수도 있다.

따라서 일반적인 경우에 적용될 수 있는 양질의 규칙을 만들기 위해서는 일반화 과정에 적당한 제한을 가하야 한다. 아래의 규칙은 학습과정에서 보다 일반적인 규칙을 얻기위해 사용한 일반화 규칙이다.

일반화 규칙 1:

용언(D, H)+명사형 어미(eCNMM)는 앞에 주격이나 목적격이 있을 경우에 해당하고 수식을 받으면 전체 어절이 명사(MC)

일반화 규칙 2:

나열을 나타내는 이나(접속조사), 그리고, 혹은, 및, ','의 앞뒤는 같은 품사다.

일반화 규칙 3:

'다른' 과 같이 "다르/형용사+ㄴ/관형형 어미"과 "다른/관형사"로 될 수 있는 형태소는 앞 문맥의 용언으로 사용되면 "형용사+관형어미"이고 아니면 "관형사"이다.

일반화 규칙 4:

체언은 N 으로 용언은 V 로 그룹화한다

일밤화 규칙 5:

체언을 수식하는 품사는 mn, 용언을 수식하는 품사는 mv 로 한다.

일반화 규칙 6:

고유명사에 속하는 품사는 caching 역할[iii]을 하게 규칙을 만든다

즉, 고유명사류는 사전에 등록할 수 없으므로 score에 영향을 받지 않고 규칙을 만든다.

예를 들어 위의 "이름"은 일반화 규칙 1 에 의해서 다음과 같이 일반화 될 수 있다.

(이르/DI+ㅁ/eCNMM, P1FT, S) -> 이름/MC
(*/V+*/eCNMM, P1FT, mn) -> */MC

여기서 '*'는 메타문자를 나타낸다

이러한 과정으로 15 개의 일반화 규칙 패턴을 얻었다.

**규칙의 적용**

오류 수정 규칙은 오류를 가장 많이 수정할 수 있고 잘못된 수정을 가장 작게 할 수 있어야 한다. 따라서 score 가 큰 규칙부터 적용한다. 그리고 세부 규칙을 먼저 적용하고 적용되지 않으면 일반 규칙을 적용한다.

## 5. 실험 결과

실험은 인터넷 기사, 대화체 문장, 교과서, 소설, 백과사전 등 다양한 말뭉치를 대상으로 하였다. 이렇게 실험한 이유는 우리의 POSTAG 는 다양한 형태의 문장이 존재하는 인터넷 문서를 위한 인덱싱과 같은 응용에 앞으로 이용될 수 있어야 하기 때문이다.

학습은 100,000 형태소의 말뭉치로 이루어졌으며 평가는 백과사전(계몽사), 기사(인터넷 문서), 대화체 문장 등 25730 형태소를 대상으로 이루어졌다.

(단위 : %)

| | A | B | C | D |
|---|---|---|---|---|
| 말뭉치 | 87.12 | 89.53 | 96.04 | 97.28 |

(형태소 단위 비교)

A : 미등록어를 모두 틀린 것으로 처리할 경우
B : 미등록어를 모두 "보통명사"로 추정할 경우
C : 일반화된 미등록어 처리와 통계 태거만을 사용할 경우
D : C 에 오류 수정 규칙을 더하여 실험한 경우

실험 말뭉치는 문장당 형태소가 19.22 개의 긴 문장이고 미등록어의 수가 2531 개로 전체 형태소에 9.84%에 해당할 정도로 많은 문장들이다.

## 6. 결론

본 논문에서는 인터넷 문서 등 미등록어가 많고 길이가 긴 문서를 효율적으로 분석하고 품사 태깅하는 혼합형 품사 태거인 POSTAG 을 제안했다.

제안한 POSTAG 에서는 형태소 패턴 사전을 이용하여 미등록어 위치와 개수에 제한이 없는 미등록어 추정이 가능하고 형태소 분석 과정내에 미등록어 추정 과정을 포함하기 때문에 모델이 간단해지는 장점이 있다.

또한 다어절어 사전을 도입하여 한국어에서 발생하는 다어절 연어, 축약 등의 분석에서의 효율과 분석의 정확도를 모두 얻을 수 있다.

또한 비교 실험을 통하여 한극어에 적합한 통계 모델을 새로 만들어 사용하였으며 오류 수정 규칙을 통하여 통계 태깅 오류를 자동으로 수정할 수 있게 하는 등, 품사 태깅의 많은 새로운 idea 들을 새로 개발하였다.

앞으로 이 태거를 이용하여 실제 무제한 인터넷 문서의 태깅 및 indexing 에 도전해 볼 예정이다.

## 8. 참고문헌

1. 차정원, 이원일, 이근배, 이종혁, "형태소 패턴 사

전을 이용한 일반화된 미등록어 처리", 인공지능 연구회 학술발표 논문집, pp37-42, 1997.
2. D. Cutting, J. Kupiec, J.Pedesen, and P.Sibun. "A Paraticlal part-of-speech tagger." In Proceedings of the conference on applied natural language processing, 1992.
3. J. Kupiec. Robust part-of-speech tagging using hidden markov model." Computer speech and language, 6:225-242, 1992.
4. R. Weischedel, R. Scewartz, J.Ralmucci, M. Meteer, and L. Rawshaw. "Coping with ambiguity and unknown words through probabilistic model." Computational linguitics, 19(2):359-382, 1993.
5. E. Brill, "A simple rule-based part-of-speech tagger." In Proceedings of the conference on applied natural language processing, 1992.
6. 신상현, TAKTAG : 통계와 규칙에 기반한 혼합형 한국어 품사 태깅 시스템, 포항공대 전자계산학과 석사학위 논문, 1996
7. 임희석, 김진동, 임해창, "한국어 특성에 적합한 변형 규칙 기반 한국어 품사 태깅." 인공 지능 연구회 춘계 학술 발표 논문집, pp3-10, 1996.
8. Kenneth Ward Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", Proceedings of Applied Natural Language Processing, Austin, Texas, 1988.
9. S. J. DeRose, "Grammatical Category Disambiguation by Statistical Optimization," Amer. J. of Computational Linguistics, vol. 14, no. 1, pp. 31-39, 1988.
10. Andrei Mikheev, "Unsupervised Learning of Word-Category Guessing Rules", cmp-lg/9604022

---

[i] POStech TAGger or Part-Of-Speech TAGger.

[ii] Agent-based natural language Interaction Retrieval on the Web.

[iii] 분석 결과를 미리 저장하였다가 다음 분석에 재사용하는 것