TAKTAG: 통계와 규칙에 기반한 2단계 학습을 통한 품사 중의성 해결

신 상 현 이 근 배 이 종 혁
포항공과대학교 전자계산학과

TAKTAG: Two phase learning method for hybrid statistical/rule-based part-of-speech disambiguation

Sanghyun Shin Geunbae Lee Jong-Hyeok Lee
Dept. of Computer Science and Engineering, POSTECH

요 약

품사 태깅은 형태소 분석이후 발생한 모호성을 제거하는 것으로, 통계적 방법과 규칙에 기반한 방법이 널리 사용되고 있다. 하지만, 이들 방법론에는 각기 한계점을 지니고 있다. 통계적 방법인 은닉 마코프 모델(Hidden Markov Model)은 유연성(flexibility)을 지니지만, 교착어(agglutinative language)인 한국어에 있어서 제한된 윈도우로 인하여, 중의성해결의 실마리가 되는 어휘나 품사를 제대로 참조하지 못하는 경우가 있다. 반면, 규칙에 기반한 방법은 규칙 자체가 품사에 영향을 받으므로 인하여, 새로운 태그집합(tagset)이나 언어에 대하여 유연성이나 정확성을 제공해 주지 못한다. 이러한 각기 서로 다른 방법론의 한계를 극복하기 위하여, 본 논문에서는 통계와 규칙을 통합한 한국어 태깅 모델을 제안한다. 즉 통계적 학습을 통한 통계 모델이후에 2차적으로 규칙을 자동학습 하게 하여, 통계모델이 다루지 못하는 범위의 규칙을 생성하게 된다. 이처럼 2단계의 통계와 규칙의 자동 학습단계를 거치게 됨으로써, 두개 모델의 단점을 보강한 높은 정확도를 가지는 한국어 태거를 개발할 수 있게 하였다.

1. 서론

한 문장에서 단어는 문맥에 따라 다른 품사를 가진다. 그 문장에 가장 적당한 품사를 선택하는 과정을 품사 태깅이라 한다. 이러한 품사 태깅은 문자 인식 [14] ,음성 인식[11]등에 사용되어 자연어 처리의 초기 단계로서 중요한 역할을 하고 있다. 품사 태깅과정은 크게 각 단어에 대하여 가능한 모든 품사를 생성하는 모호성을 생성하는 부분과 여러 품사중에서 가장 적절한 품사를 선택하는 모호성을 해소하는 두 부분으로 구성된다. 한국어의 경우 교착어로 형태소 단위의 품사 태깅을 위하여 형태소 분석과정을 거쳐야 하는데, 형태소분석을 거치는 과정이 모호성을 생성하는 부분에 해당한다.

모호성을 해소하는 방법으로 통계적 방법이 널리 사용되고 있다[2, 8, 9, 11]. 이러한 통계적 방법의 성능은 부족한 자료로 인하여 발생하는 추정(estimation) 에러와 해결하려는 문제에 대한 완전한 지식의 부족으로 발생하는 모델링(modeling) 에러에 크게 영향을 받는다. 추정 에러를 줄이기 위하여 여러가지 평탄화(smoothing)방법을 사용한다[1]. 반면 모델링 에러를 줄이기 위하여 확률에 기반한 방법[15],결정 트리(decision tree)를 이용하는 방법[4]등이 사용되고 있다. 본 시스템 TAKTAG(Two phase learning Architecture for Korean part-of-speech TAGger)는 Eric Brill 스타일[6, 7]의 규칙학습을 통하여 교착어인 한국어에 맞게 통계적 방법의 추정에러와 모델링 에러를 해결하는 방법론을 채택한다.

2. 한국어에서의 품사 태깅

한국어는 형태론적으로 볼 때 일본어, 필란드어, 헝가리어, 그리고 터키어와 같이 교착어에 해당한다. 교착어는 실질 형태소에 형식 형태소가 결합하여 한 어절을 구성한다. 여기서 실질 형태소는 어휘적 의미를 나타낸다. 실질 형태소는 홀로 쓰일 수 있는 자립 형태소일 수도 있고 홀로 쓰일 수 없는 의존 형태소일 수도 있다. 형식 형태소는 문법적 관계를 지시하며 모두가 의존 형태소이다. 이러한 교착어는 체언에 여러가지 조사가 붙거나 , 용언의 어간에 어미가 결합되는 활용(conjugation)을 한다. 이처럼 교착어인 한국어에 있어서 한 어절이 여러개의 형태소로 구성될 수 있기 때문에, 어절을 형태소단위로 분리하는 과정이 필요하다. 이러한 과정을 형태소 분석기가 접속정보와 형태소 사전을 참조하여 수행하게 된다. 한국어에서는 형태소 분리로 인하여 형태소단위로 품사의 중의성이 발생하는 것 외에도 다양하게 중의성이 발생하게 된다. 예를 들면 어절 '감기는'은 다음에서 처럼 2개 나 3개의 형태소로, 즉 다른 갯수의 형태소로 분리된다.

[감기]:[보통명사] + [는]:[보조사]

[감기]:[동사] + [는]:[전성어말어미]
[감]:[동사] + [기]:[전성어말어미] + [는]:[보조사]
어절'나는'은 아래처럼 다른 형태소'나'와 '날'로 분리된다.
[나]:[동사] + [는]:[전성어말어미]
[나]:[대명사] + [는]:[보조사]
[날]:[동사] + [는]:[전성어말어미]
그러므로 영어처럼 동형 이품사만 가지는 언어와는 상대적으로 한 차원 더 높은 처리과정을 거쳐야 한다[2]. 이러한 한국어의 특성을 반영하여, TAKTAG에서는 통계이후의 규칙학습 단계에서 어절과 어절내의 형태소를 모두 고려하게 하여 규칙 template를 구성할 수 있게 하였다.

3. TAKTAG시스템 구조

TAKTAG시스템에서 사용하는 태그집합은 논문 [3]를 참조하기 바란다. TAKTAG시스템은 2단계의 학습과정을 거친다. 첫번째 통계 학습단계에서는 형태소 단위로 분할된 말뭉치를 이용하여 Baum-Welch의 알고리즘으로 HMM을 재추정하는 학습 단계를 거치게 한다. 하지만 HMM태거는 bigram model로 학습된 HMM을 참조하게 되는데, 현재의 형태소, 형태소 태그 그리고 이전의 형태소 태그만을 고려하므로, 교착어인 한국어을 제대로 반영하지 못하고 있다. 이러한 HMM의 한계를 극복하기 위하여 2차 학습인 규칙학습단계를 거치게 된다. 2차 규칙 학습기는 올바르게 태그된 말뭉치와 HMM 태거의 결과를 비교하여 규칙을 생성함으로써, HMM 자체를 견고하게 만들어 태깅의 정확률을 높인다.

전반적의 태깅단계는 형태소 분석기를 이용하여 형태소 단위의 품사중의성을 생성하고 1차로 HMM을 참조하여 HMM 태거가 초기의 태깅된 형태소 열을 출력하게 된다. 이 출력결과를 규칙 태거가 받아서 2차적으로 학습된 규칙을 참조하여 초기의 태그 열을 수정하여 최종의 태깅된 형태소 열을 출력하게 된다. 그림1은 이러한 TAKTAG시스템의 구조를 보여준다. 그림에서 실선은 태깅과정 , 그리고 점선은 학습과정을 보여주며 사각형은 프로세스, 라운드는 중간 결과물, 원은 프로세스에 필요한 자원(데이터)를 표시한다.

4. 통계를 이용한 품사 태깅

통계적 방법으로 HMM을 사용한다. HMM $\lambda = (A, B, \pi)$는 다음의 다섯가지 매개변수로 구성되어 있다.

1. N : 상태의 갯수

각 상태는 $1, 2, ..., N$ 으로 표기한다. 품사 태깅에서 상태는 하나의 품사에 해당하므로 상태의 갯수 N은 태그집합의 크기가 된다. 한 문장에서 t번째 형태소 품사 태그를 q_t로 표기한다.

2. M : 각 상태에서 발생할 수 있는 관측 심볼의 수

각 심볼은 $V = \{v_1, v_2, ..., v_M\}$으로 표기한다. 품사 태깅에서 관측가능한 심볼의 갯수는 사전에 등록된 형태소 수이다.

3. $A = \{a_{ij}\}$: 상태전이 확률 분포

$$a_{ij} = P[q_{t+1} = j | q_t = i], \quad 1 \le i, j \le N \tag{1}$$

태그 i에서 태그 j로 전이할 확률을 나타낸다. 즉 문장에서 태그 i 다음에 태그 j가 나올 확률을 나타낸다. 여기서

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{2}$$

4. $B = \{b_j(k)\}$: 상태 j에서의 관측 확률 분포

$$b_j(k) = P[q_t = v_k | q_t = j], \quad 1 \le k \le M \tag{3}$$

태그 j가 형태소 v_k일 확률을 나타낸다. 여기서

$$\sum_{k=1}^{M} b_j(k) = 1 \tag{4}$$

5. $\pi = \{\pi_i\}$: 초기 상태 분포

$$\pi_i = P[q_1 = i], 1 \le i \le N \tag{5}$$

태그 i가 문장의 맨 처음에 나타날 확률을 나타낸다. 여기서

$$\sum_{i=1}^{N} \pi_i = 1 \tag{6}$$

HMM을 태깅에 적용할 경우 두가지 문제를 해결해야 한다. 모델의 값인 A,B 와 π를 재추정하는 학습의 문제와 형태소 열로 이루어진 문장에 대한 최적의 품사 태그열을 구하는 문제이다.

모델의 값을 재추정하는 것은 Baum-Welch 알고리즘을 이용하고, 최적의 품사 태그일을 구하는 것은 Viterbi 알고리즘을 이용한다.

4.1 Baum-Welch 알고리즘을 이용한 HMM 학습

모든 모델 매개변수(model parameter)를 신뢰성있게 유지하기 위하여 많은 자료가 필요하다. 이를 위해 다중 관측열(multiple observation sequence)를 이용한다. 여기서 다중 관측열은 많은 문장(observation 혹은 morpheme sequence)을 포함한 학습 말뭉치가 된다.

K개의 문장을 $O = \{O^{(1)}, O^{(2)}, ..., O^{(K)}\}$로 표기한다. 여기서 $O^{(k)} = (O_1^{(k)} O_2^{(k)} ... O_{T_k}^{(k)})$ 은 k번째 문장이다. 각각의 문장은 다른 문장과 독립이라고 가정하면, Baum-Welch 알고리즘은 다음 확률을 최대화 하기위하여 모델 λ 의 매개변수를 조절하는 것이 된다.

$$P(O|\lambda) = \prod_{k=1}^{K} P(O^{(k)}|\lambda) \tag{7}$$

$$= \prod_{k=1}^{K} P_k \tag{8}$$

초기 상태 확률 π, 전이 확률 a , 그리고 관측 확률 b의 재추정(reestimation) 공식은 다음과 같다[13].

$$\overline{\pi_i} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \alpha_1^k(i) \beta_1^k(i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{j=1}^{N} \alpha_1^k(j) \beta_1^k(j)} \tag{9}$$

$$\overline{a_{ij}} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i) a_{ij} b_j(O_{t+1}^{(k)}) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i) \beta_t^k(i)} \tag{10}$$

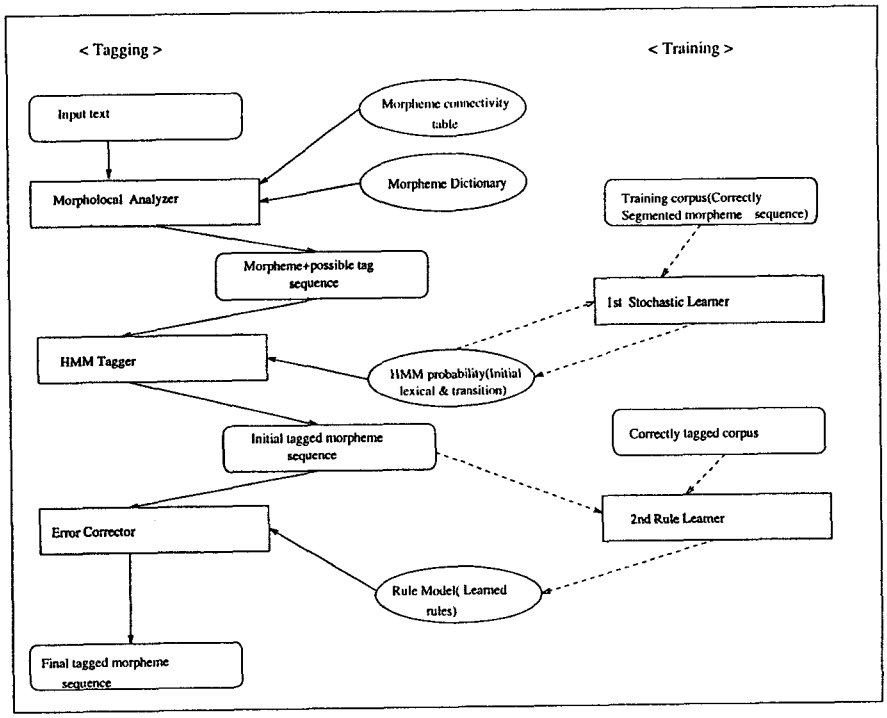

그림 1: TAKTAG시스템의 구조

$$\bar{b}_j(l) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1, s.t. O_t = v_l}^{T_k - 1} \alpha_t^k(i)\beta_t^k(i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i)\beta_t^k(i)} \quad (11)$$

전이확률a와 관측확률b는 0에서 1사이의 확률값을 가지므로 전방 전이 확률 α 와 후방 전이 확률 β 값은 계산 도중 쉽게 0으로 수렴된다. 이를 방지하기 위하여 scaling이 필요하다[5, 10, 13].

4.2 Viterbi 알고리즘을 이용한 품사 최적열 선택

입력 관측열이 주어 졌을때 이에 가장 적합한 태그열을 찾아 내는 알고리즘은 다음과 같다[13].

1. 초기화

$$\delta_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N \quad (12)$$

$$\psi_1(i) = 0. \quad (13)$$

2. 재귀적 계산

$$\delta_t(j) = max_{1 \le i \le N}[\delta_{t-1}(i)a_{ij}]b_j(o_t), \quad 2 \le t \le T, 1 \le j \le N \quad (14)$$

$$\psi_t(j) = arg\ max_{1 \le i \le N}[\delta_{t-1}(i)a_{ij}], \quad 2 \le t \le T, 1 \le j \le N \quad (15)$$

3. 재귀적 계산 종료

$$P^* = max_{1 \le i \le N}[\delta_t(i)] \quad (16)$$

$$q_T^* = arg\ max_{1 \le i \le N}[\delta_t(i)]. \quad (17)$$

4. backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T-1, T-2, ..., 1. \quad (18)$$

5. 규칙을 이용한 태깅 오류 수정

5.1 HMM의 한계 및 자동 규칙 학습

교착어인 한국어에서는 앞어절의 형태소들이 같은 품사를 가지게 되어도, 앞어절의 형태소 자체에 따라 현재 어절의 형태소는 다른 품사를 가지는 경우가 있다. 이 경우, 앞 어절의 형태소 품사만 보게 되는 통계 모델은 어쩔수 없이 모델링 에러를 지니게 된다. 이렇게 통계모델이 지닌 한계점을 극복하기 위하여, 이 모델이 해결할 수 없는 부분에 대하여 특별히 재학습하는 과정이 필요하다. TAKTAG에서는 이 부분에 대하여 Eric Brill 스타일의 재학습 과정을 두었다. 이 방법의 특별한 매력은 1)규칙을 자동으로 생성해주는 것,2)인접한 패턴이 상호 관련될 경우, 전체 말뭉치(corpus)에 대하여 인접 패턴이 개선된 상태에서 재학습을 거치므로 주위상황을 고려하여 가장 좋은 추정을 할 수 있다는 것, 그리고 3)독립된 규칙이 전체 말뭉치에 대하여 성능을 비교하게 되므로,결정트리와 비교해 볼때 상대적으로 과도학습(overtraining)을 막을 수 있다는 것이다[12].

5.2 규칙 학습 알고리즘

규칙 학습기(rule learner)는 HMM 태거의 결과와 손으로 태그한 말뭉치를 비교하여 규칙을 생성하게 되는데, 구체적 알고리즘은 다음과 같다.

1. 혼돈 행렬(confusion matrix)을 만든다. 여기서 혼돈 행렬은 HMM 태거의 잘못된 결과와 옳바르게 태그된 결과의 쌍으로 이루어진 테이블이다.

2. 혼돈 행렬의 각 원소에 대하여 규칙 template(5.3절 참

조)를 참조하여 규칙을 생성한다.
3. 생성된 규칙이 보정할 수 있는 에러의 갯수를 센다.
4. 가장 많은 에러를 보정하는 규칙을 선택한다.
5. 학습 text에 규칙을 적용한다.
6. 위의 1-5의 과정을 더 이상 에러수가 줄지 않을 때까지 한다.

5.3 규칙 template 형식(format)

$\{[N,P][1,2,3,...][F,L][MO,MT]\}^+$

$[N,P]$ 는 이전 어절을 볼 것인가, 아니면 다음 어절을 볼 것인가를 결정한다. $[1,2,3,...]$ 은 이전 혹은 이후 몇번째 어절을 볼 것인가를 결정한다. $[F,L]$ 는 한 어절내에서 처음 형태소를 볼 것인가 마지막 형태소를 볼 것인가를 결정한다. 한국어에서 주변의 품사에 영향을 미치는 형태소는 어절의 중간에 있는 형태소가 아니라 처음과 끝에 있는 형태소이다. $[MO,MT]$ 는 형태소를 볼것인가 아니면 형태소 태그를 볼 것인가를 결정한다. 여기서 $\{\}^+$ 는 하나 이상의 template를 의미한다.

위의 형식에 따라 TAKTAG에는 100여개 정도의 규칙 template가 있으나, 실제로 그중 일부만 규칙생성에 참여한다. 즉 한국어의 실정에 맞는 template만이 주로 사용되게 된다.

5.4 규칙 형식

위의 규칙 template를 참조하여 규칙 학습기(rule learner)는 아래 형태의 규칙을 생성한다.

[현재 형태소][현재 형태소 태그] ; [문맥] - > [변경될 형태소][변경될 형태소 태그]

[문맥] = { [규칙 template] [해당 형태소, 해당 형태소 태그] }$^+$
$\{\}^+$ 는 하나 이상을 의미한다.

5.5 규칙 적용 예

다음은 규칙 적용예를 보여 준다.

규칙 template:
P1LMO
학습된 규칙:

[있]:[D]_ ;[P1LMO]:[어] - > [있]:[H]

현재의 형태소가 '있'이고 '동사(D)'로 태그되어 있고 이전(P) 첫번째(1) 어절의 마지막(L) 형태소(MO)가 '어'일 경우 '있'의 태그를 '형용사(H)'로 바꾸어 준다.
통계 모델:

[우리의]=[우리]:[T]+[의]:[jC]
[몸은]=[몸]:[MC]+[은]:[jS]
[피부로]=[피부]:[MC]+[로]:[jC]
[둘러싸여]=[둘러싸이]:[D]+[어]:[mC]
[있다]=[있]:[D]+[다]:[mT]
규칙 적용 에러 교정:

[우리의]=[우리]:[T]+[의]:[jC]
[몸은]=[몸]:[MC]+[은]:[jS]
[피부로]=[피부]:[MC]+[로]:[jC]
[둘러싸여]=[둘러싸이]:[D]+[어]:[mC]
[있다]=[있]:[H]+[다]:[mT]

용언 '있다'는 대부분 '동사'의 역할을 하지만 특별히 보조연결어미 '아'나 '어'다음에 올 경우 현재의 상태를 나타내는 의미가 되어 '동사'가 아니라 '형용사'가 된다. HMM에서는 현재의 형태소와 앞 형태소의 품사를 보기 때문에 앞 어절의 형태소'아'나 '어'를 직접 보지 않을 경우 항상 '동사'로 태깅된다.

다음은 또다른 규칙 적용 예이다.

규칙 template:
N1FMO
학습된 규칙

-[와]:[jJ] ; [N1FMO]:[같이] - > [와]:[jC]

현재의 형태소가 '와'이고 '접속조사(jJ)'로 태그되어 있으며 다음(N) 첫번째(1) 어절의 처음(F) 형태소(MO)가 '같이'일 경우 형태소 '와'를 '격조사[jC]'로 바꾸어 준다.
통계 모델:

[보기와]=[보기]:[MC]+[와]:[jJ]
[같이]=[같이]:[B]
[다음]=[다음]:[MC]
[문장을]=[문장]:[MC]+[을]:[jC]
[바꾸어]=[바꾸]:[D]+[어]:[mC]
[써]=[쓰]:[D]+[어]:[mC]
[보자]=[보]:[D]+[자]:[mT]
규칙 적용 에러 교정:

[보기와]=[보기]:[MC]+[와]:[jC]
[같이]=[같이]:[B]
[다음]=[다음]:[MC]
[문장을]=[문장]:[MC]+[을]:[jC]
[바꾸어]=[바꾸]:[D]+[어]:[mC]
[써]=[쓰]:[D]+[어]:[mC]
[보자]=[보]:[D]+[자]:[mT]

조사 '와'는 주로 명사와 명사사이에 와서 '접속조사'의 역할을 한다. 하지만 '와'가 부사 '같이'와 공기관계(co-occurrence relation)를 가지게 되어 함께 쓰이게 될 경우에는 '격조사'의 역할을 한다.

6. 실험결과 및 문제점

손으로 태그한 말뭉치중 일부는 규칙학습용으로 그리고 나머지 일부는 테스트용으로 사용하였다. 약 1만 형태소를

실험말뭉치	전체 형태소수	모호한 형태소수	모호한 형태소에 대한 평균품사수	전체단어에 대한 정확률 HMM	전체단어에 대한 정확률 HMM + RM	형태소 분석기 성능
국민교육헌장	269	151	2.79	81.78	91	96.3
소설운명의힘	743	460	2.6	76.1	82.1	87.3
쓰기교과서	2227	710	2.8	79.7	92.5	95
자연교과서1	4660	1539	2.6	78.3	91.2	91.4
자연교과서2	3973	1329	2.5	78.1	93	94
총 계	11872	4189	2.6	78.4	91.5	92.8

표 1: 태그시스템 실험 결과

규칙학습용으로 사용하였다. 100여개 정도의 규칙 template를 이용한 결과 얻어진 규칙의 갯수는 450여개이다. 약 5만 형태소를 Baum-Welch에 의한 HMM 학습용으로 사용하였다. 테스트 용으로 약 1만 형태소를 사용하였다. 표1의 결과에서 알수 있듯이, Baum-Welch을 이용 1차학습만 거친 HMM을 단독으로 사용한 태거보다 2차적인 규칙학습을 거친 모델을 사용한 태거가 13%정도의 성능향상이 있어, 성능면에서 우수함을 알 수 있다.

표1에서 형태소 분석기의 성능은 형태소 분석기가 출력한 여러 결과중 하나라도 맞으면 옳다고 보았을 경우의 정확률이다. 결과를 보면 알 수 있듯이 형태소 분석기의 결과가 좋으면 태거의 성능도 좋다. 소설 운명의 힘의 경우 미등록어가 많아서 성능이 저조하다. 하지만 여전히 규칙의 에러 교정 능력은 확인할 수 있다. 태거의 역할은 형태소 분석기가 출력한 결과중 하나를 선택하기 때문에 형태소 분석기에 직접적으로 의존할 수 밖에 없다. 현재 약 93%정도의 성능을 형태소 분석기가 보여주는데 이를 개선해야 한다.

교착어인 한국어에 알맞는 규칙 template를 만드는 것도 역시 필요하다. 특히 한국어에서는 한 어절이 다른 갯수의 형태소로 분리될 수 있다. 틀리게 분리된 태그열을 올바르게 분리시키는 것과 관련된 template과 규칙이 필요하다.

7. 결론

이 논문은 통계 모델의 보강을 위하여, 2차적으로 규칙학습을 하게하는 2단계 학습기를 보유한 한국어 태깅 시스템 TAKTAG에 대하여 기술하였다. 이상적인 태거는 미등록어에 대하여 잘 처리할 수 있는 견고성(robustness), 거의 에러없이 처리하는 정확성(accuracy), 태그 집합의 변화가 있거나 언어가 바뀌어도 잘 처리할 수 있는 유연성(flexibility)을 보유해야 한다. 통계적 모델은 유연성을 보유하고 있지만, 모델 자체만으로는 고도의 정확성을 요구하는 응용에 사용하기에 무리가 있다. 통계적 모델의 단점을 보완하고자, 모델자체를 보강하는, 규칙에 근거한 모델을 제시하였다. 통계적 처리후에, 통계에서 다루지 못하는 범위에 대하여 2차적인 규칙 학습기를 두어 규칙을 학습함으로써, 유연성뿐만 아니라 정확성을 보유하게 하였다. 하지만 아직 모델에 많은 문제점을 보유한다. 앞으로 미등록어를 처리할 수 있는 부분이 필요하며, 한국어 태거는 형태소 분석기에 직접적인 영향을 받으므로 형태소 분석기의 수정 및 확장, 그리고 한국어의 실정에 맞으며 통계모델의 한계를 극복할 수 있게 규칙 template을 확장하는 것이 필요하다.

참고 문헌

[1] 김재훈,"자연언어 처리를 위한 한국어 품사 태그", 인공지능 연구센터, 기술문서*(CAIR-TR-94-55)*, 1993.

[2] 김 재훈, 임 철수, 서 정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사의 태깅", 한국정보과학회 논문지 제*22*권 제*1*호, 1995.

[3] 신 상현,이 근배,홍 남희,이 종혁,"확률과 규칙을 사용한 품사 태깅", 한글 및 한국어 정보 처리 학술 대회 학술 발표 논문집, 1994.

[4] Andre Kempe ,"Probabilistic tagging with feature structures", *The 15th international conference on computational linguistics,coling 94*, 1994.

[5] Dong Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun, "A practical part-of-speech tagger", *Proceedings of the 3rd conference on applied natural language processing*, 1992.

[6] Eric Brill, "A simple rule-based part of speech tagger", *Proceedings of the 3rd conference on applied natural language processing*, 1992.

[7] Eric Brill, "A report of recent progress in transformation-based error-driven learning", *Proceedings of the DARPA workshop on human language technology*, March, 1994.

[8] Eugene Charniak,Curtis Hendrickson,Neil Jacobson, and Mike Perkowitz, "Equations for part-of-speech tagging", *Proceedings of the eleventh national conference on artificial intelligence*, July, 1993.

[9] Eugene Charniak, "Statistical language learning", *A Bradford book. The MIT press* , 1993.

[10] Evangelos Dermatas and George Kokkinakis, "Automatic stochastic tagging of natural language texts". *Computational linguistics vol. 21, Number 2*,1995.

[11] Kenneth Ward Church, "A stochastic parts program and noun phrase parser for unrestricted text", *Proceedings of the Second conference on Applied Natural language processing*, 1988.

[12] Lance A. Ramshaw, Mitchell P. Marcus "Exploring the statistical Derivation of transformational rule sequences for part-of-speech tagging", *proceedings of the ACL balancing act workshop*,1994.

[13] Lawrence Rabiner, Biing-Hwang Juang, " Fundamentals of speech recognition ", *Prentice hall* 1993.

[14] Rohini K. Srihari and Charlotte M. Baltus, "Combining statistical and syntactic mothods in recognizing handwritten sentences", *Fall symposium series, probabilistic approaches to NL, AAAI*, 1992.

[15] Yi-Chung Lin, Tung-Hui Chiang and Keh-Yih Su, "Automatic model refinement" *Proceedings of the 15th international conference on computational linguistics,coling 94*, 1994.