# LHMM기반 영어 형태소 품사 태거의 도메인 적응 방법

# (Domain Adaptation Method for LHMM-based English Part-of-Speech Tagger)

권 오 욱 † 김 영 길 ††
(Oh-Woog Kwon) (Young-Gil Kim)

**요 약** 형태소 품사 태거는 언어처리 시스템의 전처리 기로 많이 활용되고 있다. 형태소 품사 태거의 성능 향상은 언어처리 시스템의 전체 성능 향상에 크게 기여할 수 있다. 자동번역과 같이 복잡도가 높은 언어처리 시스템은 최근 특정 도메인에서 좋은 성능을 나타내는 시스템을 개발하고자 한다. 본 논문에서는 기존 일반도메인에서 학습된 LHMM 이나 HMM 기반의 영어 형태소 품사 태거를 특정 도메인에 적응하여 높은 성능을 나타내는 방법을 제안한다. 제안하는 방법은 특정도메인에 대한 원시코퍼스를 이용하여 HMM이나 LHMM의 기학습된 전이확률과 출력확률을 도메인에 적합하게 반자동으로 변경하는 도메인 적응 방법이다. 특허도메인에 적응하는 실험을 통하여 단어단위 태깅 정확률 98.87%와 문장단위 태깅 정확률 78.5%의 성능을 보였으며, 도메인 적응하지 않은 형태소 태거보다 단어단위 태깅 정확률 2.24% 향상(ERR: 66.4%)과 문장단위 태깅 정확률 41.0% 향상(ERR: 65.6%)을 보였다.

키워드 : 형태소 품사 태거, 도메인 적응 방법, LHMM, HMM

*Abstract* A large number of current language processing systems use a part-of-speech tagger for pre-processing. Most language processing systems required a tagger with the highest possible accuracy. Specially, the use of domain-specific advantages has become a hot issue in machine translation community to improve the translation quality. This paper addresses a method for customizing an HMM or LHMM based English tagger from general domain to specific domain. The proposed method is to semi-automatically customize the output and transition probabilities of HMM or LHMM using domain-specific raw corpus. Through the experiments customizing to Patent domain, our LHMM tagger adapted by the proposed method shows the word tagging accuracy of 98.87% and the sentence tagging accuracy of 78.5%. Also, compared with the general tagger, our tagger improved the word tagging accuracy of 2.24% (ERR: 66.4%) and the sentence tagging accuracy of 41.0% (ERR: 65.6%).

Key words : part-of-speech tagger, domain adaptation method, LHMM, HMM

## 1. 서 론

형태소 품사 태거는 입력 문장의 각 단어(형태소)에 최적의 형태소 품사를 설정하는 작업을 한다. 많은 언어처리 시스템은 전처리를 위하여 형태소 품사 태거를 이용한다. 그러므로 형태소 품사 태깅 성능은 이를 이용하는 언어처리 시스템에서 매우 중요한 역할을 한다. 이러한 태깅 성능 향상을 위하여, Hidden Markov Model(HMM)[1,2], Lexicalized Hidden Markov Model (LHMM)[3], Conditional Random Field(CRF) 모델[4], Transformation-based Learning[5], Memory-based Learning[6], 결정 트리[7], Maximum Entropy Model [4,8] 등의 통계적 학습 방법론들이 연구되었다. 또한, 성능 향상을 위하여, 다른 태깅 모델들을 결합하여 그 결과를 "투표"하여 결정하는 방법론[9]도 연구되었다.

기존 대부분의 형태소 품사 태깅 방법들은 주로 형태소 품사가 태깅된 코퍼스를 활용한 통계적 학습 방법을 이용한다. 형태소 품사 태깅된 코퍼스를 기반한 형태소 품사 태깅 방법들은 영어의 경우에 일반적으로 96%~97% 정도의 단어단위 품사 태깅 정확률을 보인다[3]. 단어 단위로 처리하는 언어처리 시스템의 전처리기로는 좋은 성능이지만, 자동번역(Machine Translation)과 같이 문장 단위로 처리하는 언어처리 시스템에서는 결코 좋은 성능이 아니다. 예를 들어, 평균 20단어로 구성된 100문장을 자동번역하고자 할 때, 형태소 품사 태깅 단어 정확률이 97%이라면 60단어(= 0.03 × 2,000 단어)가 오류가 있다. 60단어가 각기 다른 문장에 오류가 발생한 것이라고 하면 전체 번역할 100문장 중에서 60문장이 각기 하나의 품사 태깅 오류를 가지고 구문분석 이후의 과정을 수행하게 된다. 그러므로, 아직도 형태소 품사

· 이 논문은 2010 한국컴퓨터종합학술대회에서 'LHMM기반 영어 형태소 품사 태거의 도메인 적응 방법'의 제목으로 발표된 논문을 확장한 것임

† 정 회 원 : 한국전자통신연구원 언어처리연구팀 선임연구원
ohwoog@etri.re.kr
†† 비 회 원 : 한국전자통신연구원 언어처리연구팀 팀장
kimyk@etri.re.kr

논문접수 : 2010년 8월 6일
심사완료 : 2010년 9월 16일

태깅 모델의 성능 향상에 대한 여지와 그 기대는 크다.

본 논문에서는 특정 도메인에 적합한 품사 태깅된 코퍼스가 없이, 일반 도메인의 품사 태깅된 코퍼스로부터 학습한 형태소 품사 태깅 모델을 도메인에 적응하는 방법을 제안한다. 본 논문에서 제안하는 품사 태깅 모델의 적응 방법은 영어권에서 널리 사용되는 PennTree Bank 코퍼스로 학습한 LHMM 기반 영어 형태소 품사 태거를 이용하였다.

## 2. LHMM 기반한 영어 형태소 품사 태거

본 논문의 영어 형태소 품사 태거에서 사용되는 LHMM 방법은 널리 알려진 HMM 방법과 거의 동일하다. 단지, 품사간의 전이만이 아니라, 어휘도 포함하는 것이 다른 점이다[3]. 기존 HMM 방법은 형태소 품사간만의 전이를 고려하여서 어휘들 간에 달라지는 전이에 대해서는 고려하지 않는 단점을 가진다. 이러한 단점으로 인하여 어휘 정보를 활용할 수 있는 Maximum Entropy Model과 CRF 등의 방법론보다 성능이 낮은 경향이 있었다[4,8]. 이러한 어휘 자질을 HMM 방법에 이용하는 방법이 LHMM 방법이다[3].

형태소 품사 태깅을 위한 LHMM은 식 (1)과 같은 일반적인 HMM에 의해서 형태소 품사 태그들의 나열이 최대가 되는 확률 값을 찾는 문제로 정의할 수 있다.

$$\bar{T} = \arg\max_{t_1 \cdots t_n} \left( \prod_{1..n} P(t_i \mid t_{i-1}, t_{i-2}) \cdot P(w_i \mid t_i) \right) \quad (1)$$

위의 식 (1)은 길이가 n인 입력 단어 나열 $w_1$, …, $w_n$이 주어졌을 때, 각 단어에 상응하는 품사 태그 나열이 최대가 되는 확률을 찾는 문제이다. 식 (1)은 기존의 HMM 수식과 동일하다. 단지 LHMM과 HMM의 차이는 품사 집합에 어휘와 결합한 품사들을 추가하는 차이만 있을 뿐이다. 식 (1)에서 $P(w_i|t_i)$은 출력확률이라고 하고, $P(t_i|t_{i-1},t_{i-2})$는 전이확률이라고 한다. 출력확률과 전이확률은 품사 태깅된 코퍼스로부터 각 확률에 대한 통계값을 예측치로 사용한다[1,3].

본 논문에서 사용하는 영어 형태소 품사 태거는 [3]에서 제시하는 어휘 자질을 이용하는 방법을 도입하여 100개 어휘들을 선택하였다. 선택한 100개 어휘들과 그 품사들로 구성된 253개의 어휘-품사를 추가하여 기존 PennTree Bank 품사 집합을 확장하였다. 선택한 어휘는 문법적 역할이 큰 단어(be, have 동사 등)들과 전치사, 접속사, 관사, 조동사 등과 같이 주위의 단어들의 품사를 구분하는 단어들을 포함하였다. 이러한 어휘들은 학습코퍼스에서 그 어휘 빈도가 큰 어휘들(85개)을 대상으로 하여 선택하였다. 또한 학습코퍼스를 대상으로 실험하여 오류율이 높은 고빈도 단어(15개)들을 추가 선택하여 품사 태깅이 자주 틀리는 단어들도 포함하였다. 표 1은 본 논문에서 사용하는 어휘 자질을 이용하여 추가된 어휘-품사 집합의 예를 보이고 있다.

표 1 어휘 자질을 이용하여 추가된 어휘-품사

| 추가된 어휘품사 (어휘/품사로 표현) |
|---|
| 's/VBZ, 's/POS, a/DT, a/NN, an/DT, about/IN, about/RB, according/VBG, and/CC, after/CONJ, after/IN, after/RB, all/PRP, all/RB, all/DT, all/PDT, along/RB, along/IN, another/PRP, another/DT, any/DT, any/PRP, any/RB, are/VBP, as/CONJ, as/IN, as/RB, at/IN, back/RB, back/NN, back/VB, back/VBP, back/JJ, be/VB, been/VBN |

LHMM으로의 확장은 품사 집합을 크게 하여 식 (1)의 전이확률 $P(t_i|t_{i-1},t_{i-2})$의 희소데이터 문제(sparse-data problem)를 가중시킬 수 있다. 본 논문의 영어 형태소 품사 태거는 [1]에서 제시한 smoothing 방법을 이용하였다. [1]과의 차이점은 어휘화된 품사의 경우에 어휘화되지 않은 품사들까지 같이 *smoothing*에 활용하였다. 즉, "is/VBZ NN of/IN"인 trigram의 전이확률 P(of/IN|is/VBZ,NN)을 얻기 위하여, 학습 코퍼스에서 예측한 P~(of/IN), P~(of/IN|NN), P~(of/IN|is/VBZ,NN) 뿐만 아니라, P(of/IN|VBZ,NN)와 P(IN|VBZ,NN)도 예측하여 smoothing하였다. 이와 같은 smoothing 방법에 의해서 어휘화된 품사를 포함한 전이확률이 학습 코퍼스에 나타나지 않는 최악의 경우에도 단순 품사들에 의하여 smoothing된 낮은 전이확률을 계산할 수 있어서, 희소성 문제로 발생하는 오류를 줄였다.

또한, 영어에서 단어 첫 글자의 대소문자 구분이 품사 태깅에 중요한 역할을 하므로, [1]에서 제시한 것처럼 품사 집합을 단어 첫 글자의 대소문자 구분에 따라 구분하였다. 첫 글자의 대문소자 구분 자질은 전체 품사 집합의 품사 수를 두 배로 증가시킨다. 표 2는 본 영어 형태소 품사 태깅 시스템에서 사용하는 품사 집합이 어휘 자질과 대소문자 자질을 사용함으로써 품사 개수의 증가를 보여 준다.

미등록어에 대한 품사 추정 문제를 해결하기 위해, [1]에서 제시한 단어의 접미사를 이용하는 영어 품사 추정 방법을 도입하여 LHMM 방법과 통합하였다.

본 논문에서 사용하는 LHMM 기반의 영어 형태소

표 2 자질 추가에 따른 영어 품사 집합 수의 증가

| 종류 | 증가 품사 수 | 총 품사 수 |
|---|---|---|
| PennTree Bank 품사 집합 | 36 개 | 36 개 |
| 기호 품사 추가 | 8 개 | 44 개 |
| 어휘화에 따른 어휘-품사 추가 | 253 개 | 297 개 |
| 첫글자 대소문자 구분 | 297 개 | 594 개 |

품사 태거는 CRF나 Maximum Entropy Model 방법에 비하여 어휘나 접미사, 대소문자 등의 자질을 기존 HMM 방법에서 사용하기 어려운 문제을 해결하였다. 그러므로, LHMM 기반의 영어 형태소 품사 태거는 HMM의 장점인 모델적 간략성과 더불어 다양한 자질을 이용할 수 있는 장점을 가진다.

## 3. LHMM 기반 영어 태거의 도메인 적응 방법

코퍼스 학습 기반의 형태소 품사 태깅 모델을 도메인에 적합하도록 적응하는 방법은 간단하게 두 가지 방법으로 나누어 볼 수 있다. 첫 번째 방법은 특정 도메인에 적합한 형태소 품사 태깅된 코퍼스를 이용하여 새로이 학습하는 방법이다. 이 방법은 대용량의 코퍼스를 수작업으로 형태소 품사를 부착해야 하므로, 시간과 비용이 많이 드는 단점을 가진다. 두 번째 방법은 기존 학습된 태깅 모델을 도메인의 원시코퍼스를 이용하여 태깅 모델에서 이미 학습한 값을 적용하고자 하는 도메인에 적합하게 변경하는 방법이다. 이 방법에 의하면 새로운 도메인 학습코퍼스를 구축하지 않는 장점을 가지지만, 도메인 원시코퍼스를 이용하여 어떻게 기존 학습된 모델을 변경할 수 있는가에 대한 문제가 남는다. 본 논문에서는 두 번째 방법의 도메인 적응 문제를 해결하는 방법을 제안하고자 한다.

LHMM 기반의 형태소 품사 태깅 방법에서 학습코퍼스로부터 학습된 값은 전이확률 $P(t_i|t_{i-1},t_{i-2})$과 출력확률 $P(w_i|t_i)$이다. 그러므로, 본 제안하는 방법에서는 전이확률과 출력확률을 적용하고자 하는 도메인에 적합하도록 재구성하고자 한다. 제안하는 도메인 적응 방법은 먼저 특정도메인의 원시코퍼스를 대량으로 기학습된 품사 태거로 자동태깅하여 자동태깅된 코퍼스를 구축한다. 구축된 자동태깅된 도메인 코퍼스로부터 전이확률 $P'(t_i|t_{i-1}, t_{i-2})$과 출력확률 $P'(w_i|t_i)$을 추출한다. 기존 학습코퍼스로부터 추출한 확률값들과 자동태깅된 도메인 코퍼스로부터 추출한 확률값들 간의 차이가 큰 어휘와 trigram을 추출한다. 적응 도메인과 기학습된 도메인 간의 차이가 큰 추출된 어휘와 trigram에 대한 전이확률과 출력확률을 도메인에 맞도록 전문가가 언어적 직관으로 휴리스틱하게 조정한다. 일반적으로 학습된 통계값을 사람에 의해 수정하기가 매우 어렵다고 알려져 있다. 다음 절에서는 언어적 직관에 의하여 확률 값들을 조정할 수 있는 방법을 자세히 기술하고 있다.

### 3.1 LHMM의 출력확률에 대한 도메인 적응 방법

LHMM의 출력확률을 도메인에 맞도록 조정할 어휘들을 먼저 추출하기 위해서, $abs(P(w_i|t_i)- P'(w_i|t_i))$ 가 임의의 임계치 $W_t$를 넘는 어휘 $w_i$ 들을 추출한다.

출력확률 $P(w_i|t_i)$은 임의의 품사 $t_i$에서 임의의 단어 $w_i$가 출력 또는 발생할 확률이다. 예를 들어, 출력확률 $P(w_i="write"|t_i=VB)$는 VB(동사원형) 품사일 경우에 "write"가 나타날 확률이다. 이 확률값은 언어 직관력으로는 그 확률값의 범위대를 예상하기가 매우 어렵다. 하지만, 반대로 write 단어가 NN(명사)와 VB(동사원형), VBP(현재 복수형 동사)를 품사로 가질 때, "write"이 특정 도메인에서 어떤 품사 분포로 나타날 지에 대해서는 자신만의 언어직관력으로 표현할 수 있다. 즉, "write"이 동사로 어느 정도 나타날 것이며 명사로는 어느 정도로 나타날 것인가를 예측할 수 있다. 그러므로, 사람의 언어적 직관에 의해서 출력확률 $P(w_i|t_i)$은 표현하기는 어렵지만, 임의의 단어가 특정 품사가 될 확률 $P(t_i|w_i)$은 언어적 직관력으로 보다 더 정확하게 표현할 수 있다. 이러한 언어적 직관으로 영어 전문가가 출력확률을 수정하기 위하여, 본 논문에서는 출력확률 $P(w_i|t_i)$을 어휘품사확률 $P(t_i|w_i)$을 이용하여 식 (2)와 같이 표현한다.

$$P(w_i|t_i) = P(t_i|w_i) \times P(w_i) / P(t_i) \\ = P(t_i|w_i) \times f(w_i) / f(t_i) \quad (2)$$

식 (2)에서 $f(w_i)$는 단어 $w_i$가 전체 코퍼스에서 나타나는 빈도수를 의미하고 $f(t_i)$는 품사 $t_i$의 전체 빈도수를 의미한다. 본 시스템에서는 각 단어와 각 품사에 대한 빈도수 $f(w_i)$와 $f(t_i)$를 저장하고, 또한 어휘품사확률 $P(t_i|w_i)$을 저장하였다가, LHMM에서 출력확률 $P(w_i|t_i)$이 필요한 경우에 식 (2)에 의해서 계산한다.

출력확률을 도메인에 맞게 조정할 전문가에게는 추출된 어휘 $w_i$와 어휘 $w_i$에 대한 기존 학습코퍼스에서의 빈도수와, 그 어휘가 가질 수 있는 모든 품사들에 대한 기존 출력확률 $P(t_i|w_i)$를 제공하고, 더불어 자동태깅된 도메인 코퍼스에서 추출한 어휘 $w_i$의 빈도수와 $P'(t_i|w_i)$를 같이 제공한다. 또한, 어휘 $w_i$에 대한 각 품사로 태깅된 예문들을 도메인 코퍼스에서 추출하여 최대 10문장씩 제공한다. 전문가는 예문으로 제공된 문장에서 그 단어가 어떤 품사로 태깅되어야 하는가를 표시한다. 임의 단어 $w_i$가 m개의 품사 $t_1, \cdots, t_m$을 가진다고 할 때, 전문가가 품사 당 10문장씩 올바르게 태깅한 품사에 의하여 식 (3)과 같이 조정 가능한 어휘품사확률 $P"(t_p|w_i)$을 전문가에 제시한다.

$$P"(t_p|w_i) = \sum_{j=1}^{m} (P(t_j|w_i) \times f(t_p|t_j) / N_j) \quad (3)$$

식 (3)에서 $f(t_p|t_j)$는 품사 $t_j$으로 태깅된 예문에서 실제 정답이 품사 $t_p$인 문장 수이며, $N_j$는 품사 $t_j$의 전체 예문의 수이다. 식 (3)에 의해서, 도메인 코퍼스에서 자동 태깅된 예문이 모두 맞으면 기존 학습된 어휘품사확률을 조정하지 않아도 된다. 많이 틀리는 품사에 대해서

는 어휘품사확률 조정이 크게 발생한다.

전문가는 식 (3)에 의해서 제시되는 어휘품사확률 $P''(t_p|w_i)$을 근거로 하여 기존 어휘품사확률 $P(t_p|w_i)$과 도메인 코퍼스에서 추출한 어휘품사확률 $P'(t_p|w_i)$을 총체적으로 판단하여 최종 도메인에 적합한 어휘품사확률로 변경한다.

### 3.2 LHMM의 전이확률에 대한 도메인 적응 방법

LHMM의 전이확률은 대상언어의 문법적 특성을 나타낸다. 본 논문에서는 도메인이 달라지더라도 대상언어가 가지는 기본적 문법적 특성은 그대로 많이 유지되지만, 도메인에 따라서 추가적으로 필요한 문법적 특성들이 나타날 것으로 생각한다. 그러므로, 기존 학습코퍼스가 도메인 밸런스가 이루어진 코퍼스라면 높은 전이확률을 가지는 trigram은 모든 도메인에서도 같을 것이라고 가정한다. 본 논문에서 도메인에 따라 달라지는 전이확률들을 반영하는 방법은 도메인의 문법적 특성을 나타내어 새로 출현한 trigram이나 대상언어 전반적으로 작게 출현하였지만 도메인에서는 매우 높게 출현하는 trigram을 추출하여 그 trigram에 대한 전이확률을 조정한다.

본 논문에서는 LHMM의 전이확률을 도메인 적응하기 위한 trigram을 추출하기 위해서, 도메인 코퍼스에서의 전이확률 $P'(t_i|t_{i-1},t_{i-2})$가 높은 순서로 trigram을 정렬한 후에 기존 학습코퍼스에서 임의의 임계치 $T_t$이하의 확률값을 가진 $P(t_i|t_{i-1},t_{i-2})$(확률값 0도 포함) 만을 남긴다. 이렇게 정렬된 상위의 trigram들을 이용하여 기존 학습된 전이확률에 추가하여 사용한다. 선택된 trigram은 전문가에 의해 문법적으로 정확한 것인가를 예문을 통하여 확인하여 사용한다.

선택된 trigram $t_{i-2},t_{i-1},t_i$에 대하여 도메인에 적합하게 전이확률을 조정하는 방법은, 먼저 선택되지 않은 기존 학습된 trigram들 중에서 전이확률 $P'(t_i|t_{i-1},t_{i-2})$와 가장 가까운 전이확률 $P'(t_q|t_{q-1},t_{q-2})$값을 가진 trigram $t_{q-2}$, $t_{q-1},t_q$을 찾는다. 그리고, 도메인 적응된 전이확률 전이확률 $P(t_i|t_{i-1},t_{i-2})$은 식 (4)와 같이 계산하여 도메인에 적합한 전이확률로 조정한다.

$$P(t_i|t_{i-1},t_{i-2}) = P'(t_i|t_{i-1},t_{i-2}) \times \frac{P(t_q|t_{q-1},t_{q-2})}{P'(t_q|t_{q-1},t_{q-2})} \quad (4)$$

## 4. 실 험

제안한 LHMM 기반의 영어 형태소 품사 태거에 대한 도메인 적응 방법을 실험하기 위하여, 영어 특허문서 도메인에 LHMM 기반 영어 형태소 품사 태거를 적응하는 실험을 하였다.

다음 2가지 시스템에 대하여 비교 실험하였다.

- LHMM 태거: PennTree Bank에서 학습한 LHMM기반의 영어 형태소 품사 태거
- LHMM 도메인 태거: 약 30만 특허문서 원시코퍼스를 이용하여 6,000 어휘에 대한 어휘품사확률을 조정하고 1,500 trigram에 대한 전이확률을 조정한 도메인 적응 LHMM 기반 영어 형태소 품사 태거(전문가 3인에 의해서 1개월 작업 분량)

LHMM 도메인 태거는 6,000 어휘에 대하여 어휘품사확률을 조정한 것 이외에, 영어 특허문서에서 -ed 형태의 거의 모든 단어들이 VBD(과거형 동사)로 사용되지 않고 VBN(과거분사)로 사용되는 특성을 이용하여 일괄적으로 VBD의 어휘품사확률 값을 30%로 감소하고 감소된 VBD의 확률값을 기존 VBN의 어휘품사확률 값에 더하도록 하였다.

실험 집합은 100만 특허문서에서 랜덤하게 추출한 200문장(평균 28.76단어)으로 구성하였다. 표 3은 비교 실험한 두 시스템에 대한 단어단위 태깅 정확률과 문장단위 태깅 정확률을 보이고 있다. 표 3은 본 논문에서 제안한 방법이 문장단위 태깅 정확률 측면에서 크게 성능이 향상한 것을 보이고 있다. 도메인 적응을 통하여 단어단위 태깅 정확률을 2.24%를 올려 문장단위 태깅 정확률은 41.0%나 상승시켰다. 또한, 이것은 각각 오류감소율 ERR(Error Reduction Rate)이 66.49%와 65.5%로 오류 비율을 크게 줄임을 보였다.

표 3 도메인 적응에 대한 비교 실험

| 비교 대상 | 단어단위 태깅 정확률 | 문장단위 태깅 정확률 |
|---|---|---|
| LHMM 태거 | 96.63% | 37.5% |
| LHMM 도메인 태거 | 98.87% (ERR: 66.49%) | 78.5% (ERR: 65.5%) |

본 LHMM 도메인 태거는 특허 도메인에 적응하기 위하여, 수식과 과학기호와 관련된 토큰분리 등의 처리와 LHMM의 통계기반에서 자주 발생하는 오류에 대한 후처리 및 주위문맥으로 해결되지 않는 태깅 모호성에 대한 처리를 추가 개발하였다. 이러한 개발을 포함한 특허 도메인용 영어 형태소 품사 태거는 같은 실험집합에서 단어단위 태깅 정확률 99.62%와 문장단위 태깅 정확률 90.5%를 보였다. 이러한 높은 문장단위 태깅 정확률을 가진 형태소 품사 태거는 자동번역과 같이 문장단위의 처리를 요구하는 언어처리 시스템의 성능을 크게 향상할 수 있다.

## 5. 결 론

본 논문에서는 LHMM이나 HMM 기반의 형태소 품사 태거를 품사 태깅된 도메인 코퍼스가 없이 원시코퍼스만을 이용하여 특정 도메인에 적응하는 방법을 제안

하였다. 기학습된 LHMM 기반의 영어 형태소 품사 태거를 이용하여 도메인 원시코퍼스를 자동 태깅하여, 자동 태깅된 도메인 코퍼스에서 추출한 LHMM의 전이확률과 출력확률과 기존 학습된 전이확률과 출력확률의 비교를 통하여 반자동으로 도메인 적응 가능한 확률값으로 조정하였다. 제안한 방법으로 도메인 적응한 태깅 시스템은 특허 도메인에서 단어단위 태깅 정확률 2.24% 향상과 문장단위 태깅 정확률 41.0% 향상을 보였다. 이러한 문장단위 태깅 성능 향상은 문장단위로 처리하는 자동번역과 같은 언어처리시스템의 성능 향상에 크게 기여할 것이다.

## 참 고 문 헌


[1] Brants, T., "TnT - a statistical part-of-speech tagger," *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, pp.224-231, 2000.

[2] Merialdo, B., "Tagging English text with a probabilistic model," *Computational Linguistics*, vol.20, no.2, pp.155-171, 1994.

[3] Ferran Pla and Antonio Molina, "Improving Part-of-speech Tagging Using Lexicalized HMMs," Natural Language Engineering, vol.10, no.2, pp.167-189, 2004.

[4] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proceedings of the Eighteenth International Conference on Machine Learning 2001*, pp.282-289, 2001.

[5] Brill, E., "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," Computational Linguistics, vol.21, no.4, pp.543-565, 1995.

[6] Daelemans, W., Zavrel, J., Berck, P. and Gillis, S. "MBT: A memory-based part-of-speech tagger generator," *Proceedings 4th Workshop on Very Large Corpora*, pp.14-27, 1996.

[7] Ma'rquez, L., Padro', L. and Rodr'ıguez, H, "A machine learning approach to POS tagging," Machine Learning, vol.39, no.1, pp.59-91, 2000.

[8] Ratnaparkhi, A., "A maximum entropy part-of-speech tagger," *Proceedings 1st Conference on Empirical Methods in Natural Language Processing*, E.

[9] Brill, E. and Wu, J., "Classifier Combination for Improved Lexical Disambiguation," *Proceedings Joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL*, pp.191-195. Montr'eal, Canada, 1998.