접사 정보를 이용한 영어 미등록어의 품사부착 성능개선

김형철○ 김재훈, 최윤수
한국해양대학교 / 한국과학기술정보연구원
yhdosu@nate.com, jhoon@hhu.ac.kr, armian@kisti.re.kr

Performance Improvement of POS tagging for English Unknown words Using Affixes

Hyung-Chul Kim○, Jae-Hoon Kim, Yun-Soo Choi
Korea Maritime University / Korea Institute of Science and Technology Information

요 약

품사 부착은 각종 자연어처리의 기본적인 요소이며, 크게 규칙 기반 방법과, 통계 기반 방법으로 나눌 수 있다. 대부분은 통계 기반의 기계학습을 이용하고 있으며, 대개 95% 이상의 성능을 보여주고 있다. 그러나 미등록어에 대해서는 성능이 그다지 높지 않다. 이 논문에서는 단어의 접사 정보를 이용해서 미등록어에 대한 품사 부착의 성능을 높이는 방법을 제안한다. 제안된 시스템은 CRF(Conditional Random Fields)를 이용하며, 그 자질의 일부로 접사 정보를 이용한다. 그 결과 미등록어에 대해서 약 40%의 성능이 개선되었다. 앞으로 미등록어에 적합한 자질을 연구하고 개발할 필요가 있을 것으로 생각된다.

주제어: 품사부착, CRF, 접사, 미등록어

1. 서 론

품사 부착은 자연언어처리뿐 아니라 정보검색, 음성인식, 문자인식 등 여러 분야에서 사용되고 있다. 물론 분야에 따라 품사의 종류도 다르고 그 기능도 조금씩 다를 수 있으나 근본적으로 여러 개의 품사를 가진 단어에 대해서 하나의 품사를 결정한다는 점에서는 모두 같다. 품사 부착은 일반적으로 다른 자연언어처리 시스템들의 전처리 과정으로 사용되는 경우가 많다. 이러한 경우 품사 부착에서 생긴 오류는 응용 시스템에 그대로 전파되므로 품사 부착시스템의 성능이 매우 중요하다. 더구나 자연언어에서 미등록어 문제는 피할 수 없는 문제 중 하나이다.

이 논문은 영어 품사 부착에서 미등록어 문제를 다룬다. 영어에서 미등록어에 대한 품사 추정 방법으로 가장 간단한 방법은 학습 말뭉치 내에서 가장 높은 빈도를 보이는 품사를 부착하거나, 혹은 말뭉치가 충분하다고 가정 한다면 미등록어는 고유명사를 제외하고는 나타날 수 없으므로 미등록어에 대해서는 고유명사를 품사로 결정하는 방법이다 이 방법들은 실용적인 품사 품사 부착 시스템으로 사용하기에는 부적합하다고 생각된다 일반적으로 흔히 사용되는 방법으로는 미등록의 품사를 개방어(open-class words)에 속하는 모든 품사를 할당하고 품사 부착 과정에서 이를 해결하도록 하는 방법이다[1]. 이 방법은 통계 기반 품사 부착에서 흔히 사용되는 방법 중 하나이기도 하다. 또 다른 방법으로는 규칙 기반 방법으로 미등록어의 품사를 예측하는 방법이다[2]. 이 방법은 통계 기반 혹은 규칙 기반의 품사 부착 방법 모두에 그대로 사용할 수 있다는 장점이 있다. 또한 접미어의 여러 가지 통계적인 특성을 이용해서 통계 기반 품사 부착 방법이 있다[1-3]. 이 방법들은 복잡한 방법에 의해서 미등록에 대한 단어 확률을 계산하거나 미등록어에 대한 사전 정보를 분석하여 이들 정보를 품사 부착 모델에서 사용한다. 또한 [4]에서는 접사(접미어와 접두어)의 정보를 이용하는데 그 길이가 1~4까지인 모든 접사를 사용한다. 따라서 미등록어의 품사를 추정하기 위해서 추가적으로 8개 자질을 사용하며 이들에 대한 매개변수를 추정해야 한다. 본 논문에서는 비교적 간단한 접사 정보만으로도 충분히 실용적으로 사용할 수 있는 영어 품사 부착 시스템을 구현하였다. 본 논문에서는 단어의 접사 정보를 품사 부착 시스템의 자질로 사용하여 기존의 방법과 그 성능을 비교해 본다.

2장에서는 관련연구에 대하여 살펴본다. 3장에서는 미등록어 문제를 해결할 수 있는 자질들을 제시하며, 4장에서는 3장에서 제시된 자질들을 사용한 품사 부착 모델을 실험을 통하여 그 유용성을 살펴본다. 5장에서는 기존 연구들과의 비교를, 끝으로 6장에서는 향후의 연구 과제를 살펴보고 결론을 맺고자 한다.

2. 관련 연구

2.1 CRF(Conditional Random Filed)

1장에서 언급했던 것처럼 이 논문에서는 품사 부착 시스템에서 미등록어 문제를 해결하기 위하여 입력 단어의 접사들을 자질로 사용한다. 하지만 기존에 사용되던 HMM(hidden Markov Model)을 이용한 품사 부착 시스템은 이러한 모델을 적용하기 힘든 점이 있다.

HMM은 관찰열의 확률과 정답 태그열의 순서를 이용한 확률의 결합 확률을 이용한 생성 모델이다품사 부착의 경우 입력단어에 대한 확률과 품사열 순서. 하지만 이러한 모델의 특성상 두 가지의 큰 문제점이 발생한다. 첫 번째는 다중의 서로 다른 자질들을 사용하기가 매우 어렵다는 것이다. 다중의 자질을 사용하고자 하면 그 자질들을 조합하여 하나의 자질로 통합하여야 한다. 하지만 이렇게 하더라도 자질들의 종류가 늘어날 때마다 통합된 자질의 클래스가 크게 늘어나게 되고 계산 속도에 큰 문제를 가져오게 된다. 두 번째 문제는 품사열의 순

서를 이용할 때 범위에 대한 제약이다.

이러한 문제를 해결하기 위해 Conditional Model이 적용된 MEMM(maximum entropy Markov model)이 개발되었다. MEMM는 기존의 두 가지 문제를 해결하면서 대부분의 경우에 HMM보다 높은 성능을 보장하였지만, 유한 상태 모델을 사용함으로써 Label Bias 문제를 가지게 되었다. 최근에는 label bais 문제를 해결하면서 MEMM 이상의 성능을 보장하는 CRF 모델이 개발되어 널리 사용되고 있다[5].

CRF는 품사 부착과 같은 연속적인 자료의 라벨(label)을 결정하는데 매우 유용하게 사용되는 분별 확률 모델(discriminative probability model)이며(식 (1)), 즉 주어진 입력 벡터 $\mathbf{x}$에 대해서 조건부 확률 $p(\mathbf{y}|\mathbf{x})$를 최대로 하는 라벨 $\mathbf{y}^*$를 선택하는 비방향성 그래프 모델이다[5].

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \quad (1)$$

여기서 $p(\mathbf{y}|\mathbf{x})$는 CRF의 종류에 따라 다양하게 정의될 수 있다. 품사 부착의 경우에는 선형연쇄(linear chain) 모델이 적합하며 식 (2)과 같이 구한다.

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right) \quad (2)$$

여기서 $f_k(\cdot)$는 자질 함수(feature function)이며 자질 k에 따른 특성 함수(characteristic function)이다. 즉 주어진 입력 $y_{t-1}, y_t, \mathbf{x}$에 자질 k가 포함되어 있으면 1을 반환하고 그렇지 않으면 0을 반환한다. λ_k는 매개변수이며 자질 k의 가중치가 된다. λ_k의 학습 방법은 일반적으로 기울기 하강 알고리즘(gradient descent algorithm: Generalized Iterative Scaling(GIS), Improved Iterative Scaling(IIS))과 준뉴턴 방법(quasi-Newton method: limited memory BFGS(L-BFGS))를 주로 사용한다. $Z(\mathbf{x})$는 정규화 요소이다.

2.2 미등록어 추정을 위한 자질

이 논문은 품사 부착을 위해서 앞 절에서 언급한 기계학습 방법 중 하나인 CRF를 이용한다. 일반적으로 기계학습 방법에서는 주이진 입력을 기계학습에 적합한 입력 즉 자질 집합으로 변환하여 사용한다. 즉 기계학습 방법의 입력은 자질 집합이다. 품사 부착에서도 예외는 아니다. 이 절에서는 품사 부착을 위한 자질 집합을 살펴보자. 가장 널리 사용되는 품사 부착 모델로서 은닉 마르코프 모델(hidden Markov model, HMM)[6]과 최대 엔트로피 모델(maximum entropy model, MEM)[7]과 지지 벡터 기계(support-vector machine, SVM)[4, 8]와 에 대해서 자질 집합은 살펴보면 단어w_i를 중심으로 이전 두 단어(w_{i-2}, w_{i-1})와 이후 두 단어(w_{i+1}, w_{i+2})에 대한 단어 자신들의 정보를 사용하고 이전 단어에 대해서는 품사 정보도 함께 사용한다. 이들 모델들이 사용하는 미등록어를 위한 자질로는 1~4개 문자의 접미사와 접두사[1, 4, 7, 8], 단어 속성(대소문자, 숫자, 하이픈 등)[1, 8], 단어길이[8] 등을 이용한다. 대부분의 모델들은 미등록어를 해결하기 위해서 매우 다양한 모델의 자질 집합을 사용한다. 이 논문에서는 성능이 크게 떨어지지 않는 범위에서 매우 단순한 자질 집합을 사용한다. 즉 길이 2인 접두사와 길이 3인 접미사만을 미등록어를 추정하기 위한 자질 집합으로 사용한다.

3. 품사 부착을 위한 자질 선택

2장에서 언급한 것처럼 영어에서 접미사는 그 단어의 품사를 추정하는 중요한 정보가 된다. (표 1)은 접미사에 의해서 품사를 추정할 수 있는 몇 가지 예를 보이고 있다.

<표 1> 접사를 통한 품사 추정의 예

접사	품사	예
~tion	명사	introduction, application, description
~ly	부사	actually, lonely, lately, slowly
~ble	형용사	disable, possible, lovable

이 논문에서는 CRF를 이용하여 미등록어에 대한 품사 부착 정확도를 개선하기 위해 단어의 접사 정보를 자질들을 사용하여서 모델을 구성하였다. 선택된 자질들은 다음과 같다.

- **품사 자질(part-of-speech features)** : 이 논문에서는 선형연쇄(linear chain) CRF 모델[9]을 사용하기 때문에 이전 품사(t_{i-1})만을 자질로 사용한다.
- **어휘 자질(lexical features)** : 현재 단어(w_i)와 그 주변 단어(w_{i-1}, w_{i+1})의 어휘를 그대로 자질로 사용한다.
- **접사 자질(affix features)** : 현재 단어의 접두사(s_i) 및 접미사(p_i)를 자질로 사용한다. 이전 연구들은 1~4까지의 모든 접사를 사용하는 반면에 이 논문에서는 다양한 길이에 대한 접두사와 접미사의 결합에 대한 성능을 분석하고 가장 성능이 우수한 하나의 결합 형태를 미등록어 추정을 위한 자질 집합으로 제안한다

이 논문에서는 기본 자질은 <표 2>같으며 <표 2>의 예는 문장 "... she/PRP returned/VBD to/TO Greenville/(Unknown Word) two/CD days/NNS before/IN ... "에서 추출한 것이다.

<표 2> CRF 모델을 위한 기본 자질 집합

자질	예
w_{i-1}	to
w_i	Greenville
w_{i+1}	two
t_{i-1}	TO

이 논문에서는 접사 자질이 미등록어 품사 부착에 어떤 영향을 끼치며, 접사의 길이가 얼마일 때 최적의 성능을 보이는지를 살펴보고자 한다. 자세한 실험은 4장에서 기술할 것이다.

4. 실험 및 평가

4.1 학습 및 실험 말뭉치

이 논문에서 학습 말뭉치는 Penn Treebank 3[10]의 Brown Corpus를 사용한다. 총 말뭉치의 크기는 49,587 문장이며 1,100,028 단어로 구성되어 있다. 이 말뭉치를 학습 말뭉치와 실험 말뭉치로 분리하였으며 실험 말뭉치에 가능하면 많은 미등록어가 나타날 수 있도록 학습 말뭉치와 실험 말뭉치의 크기를 동등하게 나누었다. 그 결과 학습 말뭉치의 크기는 22,416 문장(552,422 단어)이고 실험 말뭉치는 27,171 문장(547,606 단어)이다.

4.2 성능 척도

흔히 사용하는 품사 부착의 성능은 아래와 같은 정확률 A를 사용한다. 여기서 C는 정확하게 품사를 부착한 단어 수이고 N은 전체 단어 수이다.

$$A = \frac{C}{N} \times 100(\%)$$

이 논문에서는 3장에서 소개한 자질에 접사의 자질이 추가되었을 때 기본 자질 집합에 비해 얼마의 성능이 개선되었는지를 나타내는 개선율 I을 사용한다. 여기서 C^K는 현재 모델 K이 정확하게 품사를 부착한 단어 수이고 C^B는 기본 모델이 정확하게 품사를 부착한 단어 수이다.

$$I = \frac{C^K - C^B}{N - C^B} \times 100\ (\%)$$

4.3 접사 자질

이전 연구[[1, 4, 7, 8]에서 주로 접사 자질로 1~4개의 문자를 그대로 사용했다. 이 논문에서는 어떤 접사가 가장 좋은 성능을 보이는지를 평가하여 자질 수를 줄였다. 자질 수는 실행 속도를 개성하는데 많은 도움을 준다. 이 논문에서 사용한 접사 자질은 <표 3>과 같다.

<표 3> 실험에 사용한 접사 자질의 조합

모델명	접사 자질	예: eatable
P2	길이가 2인 접두사	^ea
P3	길이가 3인 접두사	^eat
S2	길이가 2인 접미사	le$
S3	길이가 3인 접미사	ble$
S4	길이가 4인 접미사	able$
P2S3	P2와 S3의 결합	^ea, ble$
P3S3	P3와 S3의 결합	^eat, ble$
P3S4	P3와 S4의 결합	^eat, able$

어떤 자질 조합이 가장 높은 성능을 나타내는 지를 알아보기 위하여 품사 자질과 어휘 자질(기본 자질 집합)은 고정시키고 접사 자질만 다르게 적용하여 실험한다. 이러한 자질들이 조합된 모델들을 이용하여 전체 품사 부착의 정확도와 미등록어의 품사 부착 정확도를 살펴볼 것이다.

4.4 성능 평가

<표 4>는 실험 말뭉치를 대상으로 기본 모델에 대한 성능이다. <표 4>에서 보는 바와 같이 실험 말뭉치에 속한 547,729 단어에 대해서 기본 모델의 정확률은 약 92.3%이다. 29,354 미등록어(전체 단어 약 5.3%)에 대한 정확률은 약 49.4%이며, 미등록어를 제외한 단어의 정확률은 약 94.8%(전체 단어의 정확률 92.32%)이다.

<표 4> 기본 모델의 성능

	단어수	정답	정확률
전체 단어	547,729	505,665	92.32
미등록어	29,354	14,500	49.40
미등록어 제외	518,375	491,165	94.75

<표 5>는 접사 자질에 따른 정확률과 개선율을 보이고 있다. 개선율은 <표 5>의 기본 모델의 성능을 기준으로 얼마나 성능이 개선되었는지를 말하며, 모든 접사 자질에 대해서 성능이 개선되었음을 알 수 있다. 일반적으로는 접두사 혹은 접미사 정보를 개별적으로 사용하는 것보다는 두 접사 정보를 결합하여 사용하는 것이 더 좋았다. 특히 두 문자의 접두사 자질(P2)과 세 문자의 접미사 자질(S3)이 결합될 경우에 가장 좋은 결과인 95.4%의 정확률과 40.0%의 개선율을 보였다.

<표 5> 접사 자질 집합에 따른 품사 부착의 성능 변화

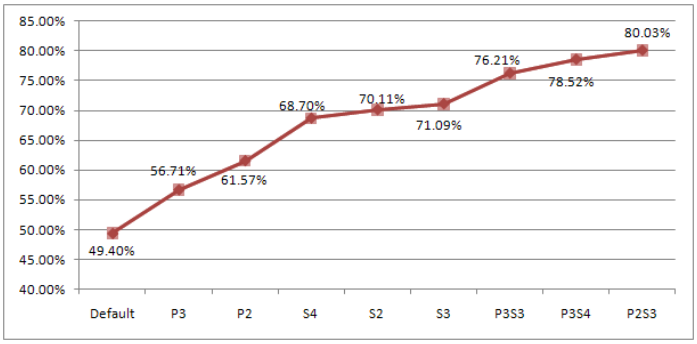
	정답 (Words)	정확률 (%)	개선율 (%)
기본 모델	505,665	92.32	-
P2	511,435	93.37	13.72
P3	509,977	93.11	10.25
S2	516,880	94.37	26.66
S3	518,101	94.59	29.56
S4	517,695	94.52	28.60
P2S3	**522,479**	**95.39**	**39.97**
P3S3	521,310	95.18	37.19
P3S4	522,243	95.35	39.41

<표 6>는 접사 자질에 따른 미등록어에 대한 정확률과 개선율을 보이고 있다. 미등록어에 대해서도 모든 접사 자질에 대해서 성능이 개선되었음을 알 수 있으며 접사 자질의 종류에 따라 조금씩 성능의 차이는 보이지만 전체 단어를 사용하는 경우와 비슷한 결과를 보였다. 두 문자의 접두사 자질(P2)과 세 문자의 접미사 자질(S3)이 결합될 경우에 가장 좋은 결과인 80.0%의 정확률과 60.5%의 개선율를 보였다.

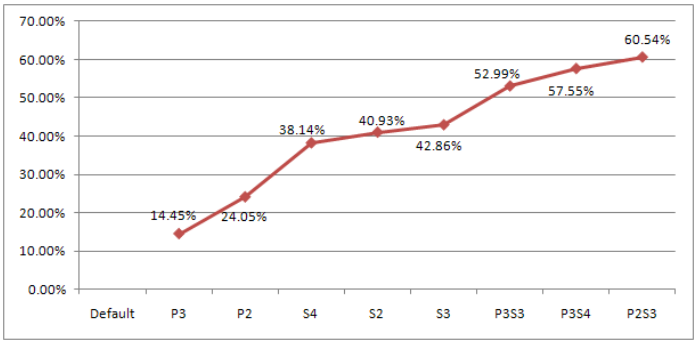
<표 6> 미등록어에 대한 품사 부착의 성능 변화

	정답 (Words)	정확률 (%)	개선율 (%)
기본 모델	14,500	49.40	-
P2	18,073	61.57	24.05
P3	16,646	56.71	14.45
S2	20,580	70.11	40.93
S3	20,867	71.09	46.86
S4	20,166	68.70	38.14
P2S3	**23,493**	**80.03**	**60.54**
P3S3	22,371	76.21	52.99
P3S4	23,049	78.52	57.55

(그림 1)은 미등록어를 대상으로 각 접사 자질에 따른 성능 변화를 그래프로 표현한 것이다. (그림 1)의 X축은 정확률이 증가하는 순서로 재배치하였다. 전체적으로 살펴보면 접두사 보다는 접미사 자질을 사용할 경우가 더 좋은 성능을 보였다. 접두사의 경우는 세 문자의 자질보다 오히려 두 문자의 자질이 더 좋은 성능을 보였다. 접미사의 경우에는 접두사의 경우와 달리 길이에 따라서 큰 변화는 보이지 않으나 세 문자의 자질에 대해서 가장 좋은 결과를 보였다. 결국 이 두 자질을 결합했을 때 미등록어에 대해서 약 80%의 정확률을 보였으나 기본 모델에 비해 약 61%의 개선율을 보였다(그림 2).



(그림 1) 미등록어만을 고려했을 경우, 접사 자질에 따른 정확률 변화



(그림 2) 미등록어만을 고려했을 경우, 접사 자질에 따른 개선율 변화

5. 기존 연구와의 비교

[4]는 SVM을 이용한 품사 부착 시스템이며 전체 정확률은 97%이고 미등록어에 대한 정확률은 86.3%이다. 이 시스템은 학습말뭉치의 크기는 1,000,000이고 실험말뭉치의 미등록어 비율은 2.2%이다. 이 시스템이 사용하는 자질 집합은 <표 7>과 같다.

<표 7> [4]의 자질 집합

품사 자질	$t_{i-1}, t_{i-2}, t_{i+1}, t_{i+2}$
어휘 자질	$w_{-1}, w_{-2}, w_{+1}, w_{+2}$
접사 자질	$\|\text{prefix}(w_i)\| \le 4, \quad \|\text{suffix}(w_i)\| \le 4,$

이 시스템의 자질 집합으로 현재 단어를 중심으로 다음에 부착될 정확한 품사 정보(t_{+1}, t_{+2})를 사용하고 있다. 실험에서는 가능하지만 현실적으로 불가능한 자질이다. 따라서 품사 정보를 제외한 자질을 사용할 경우가 이 논문과 비교할 수 있는 자질 집합이다. 이 경우 미등록어에 대해서 약 80%의 정확률을 보이고 있다. 이 결과는 학습 방법(SVM)과 환경(말뭉치의 크기 약 2배, 미등록어 비율 약 0.5배, 자질 집합)은 전적으로 차이가 있지만 이 논문의 결과와 거의 같다. 이 논문에서 제안한 방법은 [4]의 자질 집합보다 훨씬 작은 규모의 자질 집합을 사용할 뿐 아니라 학습 말뭉치의 크기도 약 1/2 정도의 크기임을 감안하면 좋은 결과라고 말할 수 있다.

[7]은 최대 엔트로피 모델을 사용한 품사 부착 시스템이며 이 시스템의 성능은 96.3%이고 미등록어에 대한 정확률은 86.3%이다. 이 시스템은 학습말뭉치의 크기는 962,687이고 실험말뭉치의 미등록어 비율은 0.3%%이다. 이 시스템이 사용하는 자질 집합은 <표 8>과 같다.

<표 8> [7]의 자질 집합

품사 자질	$t_{i-1}, \ t_{i-2}t_{i-1}$
어휘 자질	$w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}$
접사 자질	$\|\text{prefix}(w_i)\| \le 4, \quad \|\text{suffix}(w_i)\| \le 4$
기타 자질	숫자 포함, 하이픔 포함, 대문자 포함

이 시스템은 미등록어 추정한 자질로 1~4개의 문자로 구성된 접두사와 접미사를 모두 사용하고, 학습말뭉치의 크기도 이 논문에 비해서 약 2배이며 미등록어의 비율이 매우 낮다.

6. 결론 및 향후 연구

이 논문에서는 영어 단어에 대한 품사 부착 시스템에서 미등록어 문제에 대한 해결 모델을 제안하였다. 제안한 모델은 HMM모델의 단점들을 해결하고, 접사 정보를 새로운 자질로 사용하여 품사 부착에서의 미등록어 문제를 효과적으로 해결할 수 있다. 접사 정보를 사용함으로써 미등록어의 품사 부착 성능이 향상된다는 것은 물론 두 문자의 접두사와 세 문자의 접미사를 사용하는 것이 가장 좋은 성능을 보이며 접사 정보를 이용하지 않을 때 보다 약 61%의 개선율을 보였다. 더구나 접사 정보를 사용한 품사 부착 시스템의 경우 미등록어 뿐 아니라 학습말뭉치에 존재하지만 잘못된 품사를 부착했던 단어에 대해서도 높은 개선율을 보인다는 사실을 알 수 있었고, 이를 이용하면 접사 정보를 이용하지 않고도 높은 성능을 보이는 기존의 품사 부착 시스템의 성능을 한 단계 더 향상시킬 수 있는 가능성을 내

포하고 있다.

향후 연구로는 본 논문에서는 접사 정보를 이용하기 위하여 단순히 부분문자열을 그대로 사용하고 있다. 이 경우 접사 정보가 품사와 상관성이 없는 단어에 대해서도 품사 결정에 영향을 줄 수 있다는 문제점이 있다. 이 문제점을 해결하기 위하여 접사 정보를 몇 개의 범주로 분류하여 품사 부착 시스템의 자질로 사용하는 연구가 필요하다.

감사의 글

본 연구는 한국과학기술정보연구원에서 수행하는 교육과학기술부 차세대 정보유통 핵심기술 연구・개발 사업의 위탁연구로 수행되었습니다.

참고 문헌

[1] R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer and L. Ramshaw, Coping with ambiguity and unknown words through probabilistic models, Computational Linguistics, Vol. 19, No. 2, pp. 359-382, 1993.

[2] A. Mikheev, Automatic rule induction for unknown-word guessing, Computational Linguistics, Vol. 23, No. 3, pp. 405-423, 1997.

[3] S. Thede, Predicting part-of-speech information about unknown words using statistical methods, Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, pages 1505-1507, 1998.

[4] T. Nakagawa, T. Kudoh and Y. Matsumoto, Unknown word guessing and part-of-speech using supporting vector machines, Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, pp. 325-331, 2001.

[5] J. Lafferty, A. McCallum, and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Procdings of the 18th International Confenence on Machine Learning, pp. 282-289, 2001.

[6] C. Sutton and A. McCallum, An introduction to conditional random fields for relational learning, Introduction to Statistical Relational Learning (eds), L. Getoor and B. Taskar, pp. 93-127, 2007.

[6] J. Kupiec, Robust part-of-speech tagging using a hidden Markov model, Computer Speech and Language, pp. 225-242, 1992.

[7] A. Ratnaparkhi, A maximum entropy model for part-of-speech tagging, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 133-142, 1996.

[8] J. Gimenez and L. Marquez, Fast and accurate part-of-speech tagging : The SVM approach revisited, Proceedings of the International Conference RANLP, pp. 153-163, 2003.

[10] The Penn Treebank Project, http://www.cis.upenn.edu/~treebank/.