래티스상의 구조적 분류에 기반한 한국어 형태소 분석 및 품사 태깅

나승훈○, 김창현, 김영길
한국전자통신연구원
nash@etri.re.kr, chkim@etri.re.kr, kimyk@etri.re.kr

Lattice-based discriminative approach for Korean morphological analysis and POS tagging

Seung-Hoon Na○, Chang-Hyun Kim, Young-Kil Kim
Natural Language Processing Laboratory
Electronics and Telecommunication Research Institute

요 약

본 논문에서는 래티스상의 구조적 분류에 기반한 한국어 형태소 분석 및 품사 태깅을 수행하는 방법을 제안한다. 제안하는 방법은 입력문이 주어질 때 어휘 사전을 참조하여, 형태소를 노드로 취하고 인접형태소간의 에지를 갖도록 래티스를 구성하며, 구성된 래티스상 가장 점수가 높은 경로상에 있는 형태소들을 분석 결과로 제시하는 방법이다. 실험 결과, ETRI 품사 부착 코퍼스에서 기존의 1차 linear-chain CRF에 기반한 방법보다 높은 어절 정확률 그리고 문장 정확률을 얻었다.

주제어: 래티스, 구조적 분류, 형태소 분석, 품사 태깅

1. 서론

한국어 형태소 분석을 위한 *규칙기반 방법* [6,7,8,9,18,20,22,23,26]은 분석에 필요한 규칙을 수작업으로 구축하기 때문에, 개발 비용이 높고, 새로운 도메인에 대해 적응력이 떨어지는 단점을 지닌다. *통계기반 방법*은 대규모 품사 부착 말뭉치로부터 형태소 분할 및 품사 태깅에 필요한 규칙 및 확률 모델을 자동 또는 반자동으로 학습하는 방식으로 [3,4,11,14,15,16,24,25,27], 수작업이 거의 필요 없고, 성능이 우수하며, 타 도메인으로의 적용성이 높고, 기존의 어휘 사전과의 하이브리드가 가능하다는 점 등의 장점으로 인해, 현대의 대부분의 품사 태깅 연구가 이에 기반을 두고 있다.

그러나, 자동 번역 등과 같은 실제 응용을 목표로 하는 *응용 지향 형태소 분석기*에서는 어휘 사전은 여전히 필수적인 자원이다. 특히, 자동 번역에서는 목적언어의 대역어가 원시언어의 어휘로부터 얻어지기 때문에, 어휘 사전은 기본적인 리소스가 된다. 응용 지향 형태소 분석기에서는 최종 응용에서의 성능을 높이기 위해, 오랜 기간 동안 튜닝과정을 거쳐 사전의 규모화가 이루어진 경우가 많다. 이러한 사전의 규모화는 최근 웹의 크기가 빅데이터화되고, 개체명 (named entity)을 자동으로 추출하는 웹 마이닝 기법이 지속적으로 발전되고 있는 추세 속에서 더욱 용이해지고 있다.

반면, 응용 지향 형태소 분석기에서는 일반적으로 통용되는 형태소 단위 (Sejong코퍼스에서 정의되는 단위)를 사용하지 않고 자체적으로 형태소 단위를 재정의하여 사용하기도 한다 [19]. 그런데, 이러한 경우 해당 단위의 품사 부착 말뭉치는 규모화가 이루어져 있지 않고 소규모에 그치고 있기도 하다. 예를 들면, ETRI의 자동번역을 위해 구축된 품사 태깅 말뭉치는 10만 문장 정도로, 이는 2011년에 배포된 세종 말뭉치의 80만 문장보다 현저히 적은 규모이다.

본 논문에서는 어휘 사전은 규모화되어있으나 학습 데이터가 소규모인 경우, 형태소 분석의 정확도를 높이기 위해 학습 기반 *구조적 분류 모델 (structured classification)*을 적용하는 방법을 제안한다. 제안하는 방법은 래티스 기반의 구조화 분류 방법으로, 먼저 입력문으로부터 어휘 사전을 참조하여, 사전에 나타난 개별 형태소를 노드로 취하고 인접 형태소간에 에지를 구성하여, 입력문의 래티스 (*lattice*)를 구성한다. 다음으로, 이렇게 얻어진 래티스상에서 가장 점수가 높은 최적의 경로를 찾아, 이 최적 경로 상에 있는 형태소열을 분석 결과로 제시한다. 경로의 점수는 구성하는 에지의 점수의 합으로 이루어지며, 각 에지의 점수는 에지의 자질 벡터와 자질 가중치 벡터의 내적으로 정의된다. 이때, 자질 가중치 벡터는 학습 데이터로부터 구조적 분류 알고리즘을 통해 학습된다. 이러한 래티스 기반 한국어 형태소 분석은 기존의 전통적인 규칙 기반 방법에서 활용되었던 방식이나, 기존의 연구는 모두 HMM의 생성 모델에 기반을 둔 반면, 본 논문은 래티스상의 구조적 분류 모델을 적용한다는 점에서 기존 연구와 차별점이 있다.

ETRI 한국어 품사 부착 코퍼스상에서 실험 결과 제안 방법은 1차 linear-chain CRF기반 방법에 비해 높은 성능을 보여주었다.

2. 관련 연구

제안 방법과 가장 유사한 기존 연구는 일본어 형태소 분석을 위해 CRF를 적용한 연구인 [2]이다. [2]에서는, 본

연구와 마찬가지로, 입력문에 대해 어휘 사전으로부터 래티스를 구성하고, 래티스상의 임의의 경로에 대한 확률을 입력문에 대한 조건부 확률 (conditional probability)로 모델링하여, 가장 높은 확률을 갖는 경로를 찾는 문제로 형태소 분석 문제를 형식화했다. 그러나, 제안 방법과 기존 연구[2]와는 다음의 차이가 있다. 첫째, [2]에서는 가질 가중치 학습 방법이 CRF로, 분류 모델(discriminative model)에 속하나, 제안 방법은 별도의 조건부 확률을 정의하지 않는 SVM과 같은 분류 함수(discrimiantive function)방법에 속한다. 둘째, [2]에서는 사용되는 자질이 에지에 참여하는 두 노드로 국한된 first-order 방식만을 사용하고 있으나, 제안 방법은 참고하는 자질이 에지의 갯수가 2개인 second-order의 자질의 사용까지 포함한다. 셋째, [2]는 미등록어에 대한 문제를 남겨두었으나, 본 연구에서는 linear-chain기반 CRF를 이용하여 미등록어를 추가하는 방법도 함께 제시한다.

본 연구과 밀접하게 관련된 다른 연구로 [5]의 그래프 기반 의존 파싱을 들 수 있다. 제안 방법과 유사하게, [5]에서는 의존 파싱 문제를, 입력문에 대해 의존성을 에지로 하여 그래프를 구성하고, 해당 그래프에서 최적의 트리 (tree)를 찾는 문제로 파싱문제를 형식화했다. 본 논문의 방법은 [5]의 그래프 기반 의존 파싱에서 그래프가 래티스로, 트리가 경로로 제한된 특수한 경우라고 볼 수 있다. 그러나, 본 논문은 그래프 기반 분류 방법을 품사 태깅 문제에 적용하여, [5]의 의존 파싱 목적과 근본적인 차이가 있다. 저자의 지식에 따르면, 현재까지 래티스 기반 구조적 분류 방식을 한국어에 적용하여 성능 평가를 수행한 연구는 없었다.

3. 제안 방법

그림 1은 제안하는 래티스상 구조적 분류에 기반한 한국어 형태소 분석의 흐름도를 보여준다. 다음 절에서 제안 방법을 보다 형식적으로 기술하기로 한다 (이해를 위해, [5]의 표기를 주로 채택한다).

3.1 문제 정의

입력 문 $\boldsymbol{x}$에 대해서, 래티스 $G=(V,E)$는 방향성 그래프로, V은 입력문 x에서 부분문자열 중 사전에 있는 모든 형태소 [표층 문자열, 품사 태그]를, E는 인접하는 모든 형태소간의 가중치를 갖는 에지 (weighted edge)집합을 의미한다. 모든 래티스는 {0,1}이라는 2개 특수 노드를 갖는다. 여기서, 0은 문장의 시작을 가리키는 시작 형태소를 1은 문장의 끝을 가리키는 마지막 형태소이다. 에지 (i,j)는 노드 i에서 노드 j로의 전이 (transition)를 의미한다.

입력문의 래티스상의 경로 (path) $\boldsymbol{y}$는 시작 노드 0과 마지막 노드 1를 연결하는 경로로, 구성하는 에지 순차열 (edge sequence)이라고 볼 수 있다.

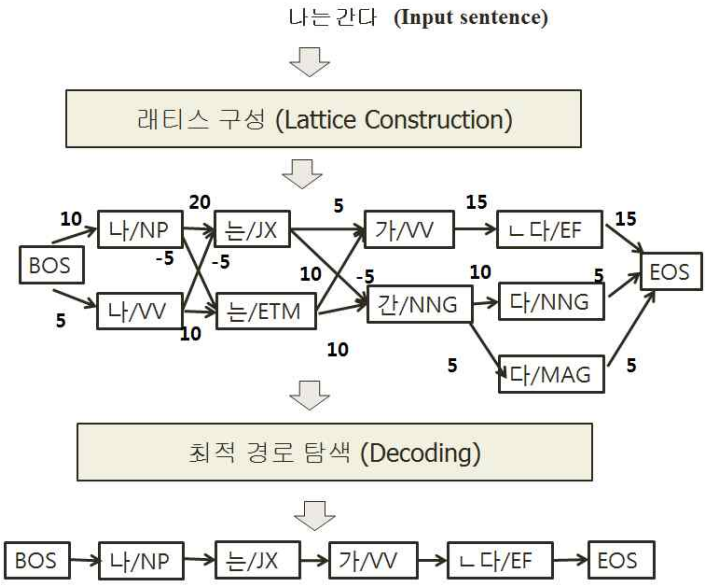

그림 1. 제안 방법의 흐름도

경로 $\boldsymbol{y}$의 점수는 $s(\boldsymbol{y})$로 표기하며, 다음과 같이 구성 에지의 점수의 합으로 정의된다.

$$s(\boldsymbol{y}) = \sum_{(i,j)\in \boldsymbol{y}} s(i,j) = \sum_{(i,j)\in \boldsymbol{y}} \boldsymbol{w} \cdot \boldsymbol{f}(i,j) \qquad (1)$$

식 (1)에서, $\boldsymbol{f}(i,j)$는 래티스 G의 노드 i,j에 대한 자질 벡터 (feature vector)를, $\boldsymbol{w}$는 자질 가중치 벡터 (feature weight vector)이다. 입력문 $\boldsymbol{x}$에 대해 시작노드 0과 마지막 노드 1을 연결하는 모든 가능한 경로 집합을 $P(\boldsymbol{x})$라고 하자. 형태소 분석 문제는 $P(\boldsymbol{x})$중에서 점수가 가장 높은 경로를 찾는 것으로, 다음의 최적화 문제를 푸는 것으로 귀결된다.

$$\hat{\boldsymbol{y}} = \arg max_{\boldsymbol{y}\in P(\boldsymbol{x})} (s(\boldsymbol{y})) \qquad (2)$$

그림 1은 입력문 "나는 간다"에 대한 래티스의 예를 보여준다. 그림에서, BOS와 EOS는 각각 시작 노드 0와 끝 노드 1를 가리킨다.

2차 자질 사용

기존의 그래프 기반 의존 파싱 연구와 유사하게, 앞의 형식화는 2차 (second-order)자질을 고려하여 확장될 수 있다. 1차 (first-oder) 자질은 자질 벡터를 구성할 때 참조하는 에지가 1개인 경우를 말하며, 2차 자질은 참조하는 에지들이 연속된 두 개의 인접한 에지인 경우를 일컫는다. 결국, 2차 자질은 연속된 세 개의 형태소 노드로부터 추출된다. 2차 자질을 사용하여 식 (1)을 확장하면 다음과 같다.

$$s(\boldsymbol{y}) = \sum_{(i,j,k)\in \boldsymbol{y}} s(i,j,k) \qquad (3)$$

여기서, $s(i,j,k)$로 다음과 같이 형식화 된다

$$s(i,j,k) = \boldsymbol{w} \cdot \boldsymbol{f}(i,j,k) \quad (4)$$

$\boldsymbol{f}(i,j,k)$는 노드 i,j,k에서 추출된 2차 자질 벡터이다. [5]의 그래프 의존 파싱에서와 같이, 최종적으로 $s(i,j,k)$는 1차 자질의 점수와 통합되어 사용된다.

$$s'(i,j,k) = s(j,k) + s(i,j,k) \quad (5)$$

3.2 디코딩 알고리즘

식 (2)의 최적화 문제는 에지의 점수 $s(i,j)$를 $-s(i,j)$로 간주할 때, 그래프 상의 최단 경로를 찾는 문제로 Viterbi 알고리즘을 이용하여 풀 수 있다. 그림 2은 2차 자질을 사용할 때의 최적 경로를 찾는 Viterbi 알고리즘의 pseudo code을 보여준다.

```
T = { 0 }
while( T ≠ ∅ ){
k = T.pop()
visited[k] = 1
for each j∈prevs(k)
    S[j][k] = max_{i∈prevs(j)}(S[i][j] + s(i,j,k) + s(j,k))
  a[i][k] = argmax_{i∈prevs(j)}(S[i][j] + s(i,j,k) + s(j,k))
}

for each j∈nexts(k)
    if prevs(j)의 모든 노드가 이전에 방문되었으면,
        T.add(j)
}
```

그림 2. 래티스상 최적 경로를 찾는 Viterbi 알고리즘 (2차 자질을 사용하도록 확장된 버전)

3.3 자질 가중치 벡터 학습

자질 가중치 벡터 $\boldsymbol{w}$를 학습하기 위해서, 본 논문은 [1]의 *averaged perceptron*을 사용한다. 그림 3은 자질 가중치 벡터를 학습하기 위한 averaged perceptron알고리즘의 pseudo code을 보여준다.

```
Training data: T = {(x_t, y_t)}_{t=1}^{M}
w_0 = 0; v = 0; i = 0
for n : 1 ··· N
    for t : 1 ··· M
        ŷ_t = argmax_{y∈P(x_t)}(s(y))
        w_{i+1} = F(x_t, y_t) − F(x_t, ŷ_t)
                (where F(x,y) = Σ_{(i,j)∈y} f(i,j))
        v = v + w^{(i+1)}
        i = i + 1
w = v/(N • M)
```

그림 3. Averaged perceptron에 기반한 자질 가중치 벡터 학습 알고리즘

3.4 래티스 구성 (Lattice Construction)

래티스를 구성할 때, 일본어나 중국어와 달리, 한국어에서는 활용/변이로 인해 표층형과 사전형이 달라지는 문제를 고려해야 한다. 이른바 *표층형과 사전형 간의 불일치* 문제를 해결하기 위해, 본 논문에서는 별도로 음절 매핑 테이블 (syllable mapping table)을 참조하여 래티스를 구성한다[1]. 아래, 음절 매핑 테이블의 예들을 보여준다.

표층형	사전형
했	하았
게	거이, 것이
내	나의, 나어
왔	오았
와	오아

한국어에서 래티스 구성은 다중 입력문 (multiple input sentences)으로부터 각 입력문의 부분음절열에 대해 사전 참조하여 래티스를 구성하는 방식과 동일하다. 다시 말해, 주어진 음절열에 대해, 입력문내 각 음절마다 음절 매핑 테이블을 참조하여 사전형으로 치환하여 새로운 입력문을 만들어 낸다. 예를 들어, "처리했다"의 경우, "했"을 사전형 "하았"으로 치환하여 "처리하았다"라는 새로운 입력문을 파생시키는 것이다. 결국 최종 래티스는 가능한 모든 변이 입력문을 모아서 N개의 다중 입력문을 구성한 후에 각각의 입력문에 대해서 래티스를 생성 한 후, 이렇게 얻어진 N개의 래티스를 병합하여 얻어진다. 다중 입력문의 수 N는 매핑 테이블을 참조하여 음절 변이가 일어나는 음절의 갯수에 따라 지수적으로 증가하는데 (exponentially increasing), 이를 효율적으로 처리하기 위해 dynamic programming기법을 사용한다.

1) 음절 매핑 테이블은 학습코퍼스로부터 대부분 자동으로 획득된다.

CRF기반 미등록어 자동 추출

래티스 생성 단계는 어휘 사전에 등재된 형태소만을 대상으로 하기 때문에, 미등록어 문제를 해결하지 못한다. 미등록어 처리를 위해, 본 논문에서는 [11]의 linear-chain CRF를 입력문에 적용하여 얻어진 1-best 형태소 분석 결과를 어휘 사전에 추가하였다.

3.5 자질 벡터

본 논문에서 사용하는 자질은 1차 자질과 2차 자질으로 나뉜다. 사용되는 정보는 형태소 노드의 표층형 (surface form), 품사 (POS tag), 해당 표층형의 시작 음절 앞 또는 끝 음절 다음에 띄어쓰기가 있었는지 여부 (spacing information), 형태소 노드의 시작과 끝 음절이 변이를 통해 얻어졌는지 여부 등이다.

사용하는 자질 집합을 보다 구체적으로 기술하기 위해, 먼저, 노드 i에 대한 기본 자질 (basic features)을 다음과 같이 정의하도록 한다.

기본자질이름	정의
lspace(i)	현재 형태소 시작전에 띄어쓰기가 있는지 여부
rspace(i)	현재 형태소 다음에 바로 띄어쓰기가 있는지 여부
lopen(i)	현재 형태소의 시작 음절이, 음절 매핑테이블을 참조하여 변이되었는지 여부
ropen(i)	현재 형태소 마지막 음절이 음절매핑테이블을 참조하여 변이되었는지 여부
surface(i)	현재 형태소의 표층 문자열
tag(i)	형태소 i의 품사 태그
length(i)	형태소 i의 표층형 길이
surfaceext(i)	lopen(i), ropen(i)을 고려한 표층 문자열: surface(i), lopen(i), ropen(i)의 세 정보의 조합으로 구성된다.

여기서, lopen, ropen의 자질 값의 예를 들기 위해, “갔”으로부터 “갔=>가았”의 음절 매핑테이블을 참조하여, “가/VV”, “았/EP”의 두 가지 형태소 노드가 만들어졌다고 가정하자. 이렇게 얻어진 두 형태소 노드에 대한 lopen, ropen정보는 다음과 같다.

	lopen	ropen
가/VV	false	true
았/EP	true	false

반면, 주어진 형태소의 시작과 끝 음절이 모두 변이가 없이 얻어진 경우에는 lopen, ropen값이 모두 false가 된다.

1차 자질과 2차 자질은 위에서 정의된 기본 자질로부터 파생되는데, 간편한 기술을 위해, 다음과 같은 표기를 정의하도록 하자.

표기	내용
Pt	tag(i)
Ct	tag(j)
Nt	tag(k)
Plen	length(i)+lopen(i)+ropen(i)
Clen	length(j)+lopen(j)+ropen(j)
Pr	ropen(i)
Cr	ropen(j)
Pls	lspace(i)
Cls	lspace(j)
Pw	surface(i)
Cw	surface(j)
Nw	surface(k)
Pw2	surfaceext(i)
Cw2	surfaceext(j)
Nw2	surfaceext(k)

여기서 “+”는 여러개의 자질들의 조합 연산자이다.

최종적으로, 1차 자질과 2차 자질은 다음과 같다.

1) 1차 자질 $\boldsymbol{f}(i,j)$:

- CtClen, CtPlenClen, PtCtPlenClen, PtCtPr, CtPr, CtCr, CtCl, CtCr, CtClsCl, CtCrsCr, CtCw, CtCw2, PwCw, Pw2Cw, PtCtCw, PtCtCw2, PtCtCw, Pw2CtCw2, PtPwCw, Pw2CtCw2, PtPwCw, PtPw2Cw2, PtPwCt, PtPw2Ct, PtPwCtCw, PtPw2CtCtw2
- window기반 지역 문맥 정보 (local context): 각 형태소의 시작과 끝 음절을 기준으로 앞뒤 3음절까지의 부분음절열을 지역 정보로 사용한다.
- penality: 경로상 하나의 에지를 취할 때 드는 패널티로, 최장일치휴리스틱처럼, 경로의 노드의 수를 가능한한 적게 (또는 많게)하기 위해서 사용한다.

2) 2차 자질 $\boldsymbol{f}(i,j,k)$

- PtCwNt, PwCtNt, PtCtNw, PtCtNtNw, PtCtNtCwNw, PtCtNtPwCwNw, PtCtNtCw, PtCtNtPw, PtCtNtPwCw

1차 및 2차 자질 모두 자질명 (feature name)과 자질값 (feature value)으로 구성되며, penalty만 제외하고 나머지는 모두 이진 값(binary value)를 자질값으로 취한다.

4. 실험 결과

실험을 위해 ETRI품사 부착 대화체 말뭉치 약 10만 문장을 사용한다. 이중 90%를 학습용으로 나머지 10%는 테스트용으로 사용하였다. 참조하는 어휘 사전의 크기는 총 약 380만 엔트리로, 대부분 복합명사나 전문용어로 구성되었다. 최종 어휘 사전은 원래의 380만 어휘 사전 외에 학습데이터에 나타나는 어휘까지 추가하여 구성하였다 (대부분은 사전에 포함된다). 제안 방법을 포함 본 실험에서 평가 대상으로 하는 형태소 분석 방법들은 다음과 같다.

- **베이스라인 시스템 (CRF)**: [11, 25]의 linear-chain CRF에 기반한 방법을 확장한 것으로, [25]과 유사한 음절 태깅에 기반하여 형태소 분할과 태깅을 동시에 수행하는 결합 모델을 사용하였다. 복합형태소를 단위 형태소로 분해하기 위해 [12]의 방법을 사용한다. [12]에서는 복합형태소에 대한 기분석과 기능형태소에 대한 기분석의 두 가지 기분석 패턴을 사용하는데. 먼저, 복합형태소 전체에 대한 기분석을 적용하여 분해를 시도하고, 여기에 나타나지 않는 복합형태소에 대해서는 기능형태소에 대한 기분석에 참조하여 분해를 시도한다.
- **래티스 기반 구조화 분류 (Lattice)**: 제안하는 래티스 기반 구조화 분류를 적용한 방법으로, 별도의 미등록어 추출없이 원래의 어휘 사전만을 이용한 결과이다. 사용 자질 유형에 따라 두 가지 모델이 얻어지는데, **Lattice**는 first-order자질만 사용한 것을, **Lattice(2nd order)**는 second-order자질까지 함께 사용한 모델을 가리킨다.
- **래티스 기반 구조화 분류 + 미등록어 자동 추출 (Lattice+unkProc)**: 앞에서 설명한 제안 래티스 기반 구조화 분류 방법에 더하여, 3.4절에서 CRF기반 미등록어 자동 추출까지 적용한 방식을 가리킨다.

표 1. 제안 lattice기반 방법의 성능

방법	F-measure	EA	SA
CRF	0.9086	0.8649	0.5543
Lattice	0.9262	0.8886	0.6322
Lattice(2nd order)	0.9393	0.9103	0.6920
Lattice(2nd order)+ unkProc	**0.9407**	**0.9127**	**0.6938**

표 1은 네 가지 방법을 비교한 결과이다. F-measure는 형태소 단위의 F-measure를, EA (eojeol accuracy)는 어절 정확률을, SA는 문장 정확률 (sentence accuracy)을 가리킨다. 표 1에 보다시피, 제안 래티스 기반 방법은 CRF기반 방법대비 세가지 지표에서 모두 높은 성능을 보여주었다. 또한, 1차 자질에 더해 2차 자질까지 확장할 때, 성능이 더욱 향상됨을 알 수 있다. 미등록어 자동 추출을 적용하여 어휘 사전을 확장한 경우, 최종 성능에서 증가가 있었으나 그 차이는 크지 않았다. 이는 어휘 사전이 이미 충분히 있어, 미등록어 현상이 거의 발생하지 않았기 때문이다.

기존 연구 [11]과 비교하여, 어절 정확률이 낮은 이유는 상당부분 대부분 복합명사 분해 애매성으로부터 기인하는 것으로 보인다. 다시 말해, 복합 명사의 경우 형태소 단위가 코퍼스 구축자의 판단에 의존하는데, 분석 결과는 다른 관점에서 보면 올바르나, 정답은 한가지 경우만 기술되어 있어 나머지들이 모두 오류로 잡혀 정확률을 감소시킨 것이다. 향후, 복합명사 분해에 대해 보다 완화된 평가 방식을 고안하여, 더욱 정교하게 성능을 비교할 필요가 있다.

5. 결론

본 논문은 래티스 기반의 구조적 분류에 기반한 한국어 형태소 분석 및 품사 태깅 방법을 제안하였다. 실험 결과, ETRI 품사 부착 말뭉치에 대해, 제안 래티스 기반 방법은 기존의 linear-chain CRF기반 방법에 비해 우수한 성능을 보여주었다. 물론, 본 실험에서 사용된 linear-chain CRF은 [11]의 특정 1차 자질 셋에 기반을 둔 것이므로, 래티스 기반 방법이 모델적으로 (일반적으로) linear-chain CRF보다 우수하다고 결론을 이끌어낼 수는 없다. 더욱 객관적인 결론 도출을 위해서는 보다 다양한 자질 집합에서 엄밀한 실험을 수행하여야 할 것이다.

향후, Sejong코퍼스에서 본 논문에서 제안 방법의 실험을 확장하여, 코퍼스가 대규모인 경우에 어떠한 차이를 보이는지 비교해 볼 계획이다. 또한, 래티스 구성 시, 형태/음운론적 제약과 접속 정보 등을 사용하여 제안 방법에 대한 상세한 실험을 수행할 것이다. 그리고, 미등록어 자동 추출시 CRF결과의 1-best결과 외에 n-best결과를 함께 이용하는 방법도 향후 연구 주제로서 흥미로울 것이다. 마지막으로, 본 논문에서는 자질 가중치 벡터를 학습하기 위해 averaged perceptron을 이용하였으나, SVMStruct등의 다른 학습 방법도 활용하여 비교하는 것도 흥미로운 연구 주제가 될 것이다.

참고문헌

[1] Collins, M. (2002). Discriminative Traning Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. EMNLP, 1-8.

[2] Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. EMNLP, 230-237.

[3] Lee, D., & Rim, H. (2005). Probabilistic Models for Korean Morphological Analysis. IJCNLP (pp. 197-202).

[4] Lee, G. G., Cha, J., & Lee, J. (2002). Syllable-Pattern-Based Unknown- Morpheme Segmentation and Estimation for Hybrid Part-of-Speech Tagging of Korean. Computational Linguistics, 28(1).

[5] Mcdonald, R., Grammer, K., & Pereira, F. (2005). Online Large-Margin Training of Dependency Parsers. ACL, 91-98.

[6] Shim, K., & Yang, J. (2002). MACH - A Supersonic Korean Morphological Analyzer. COLING. 939-945.

[7] 강승식. (1994). 다층 형태론과 한국어 형태소 분석 모델. 한글 및 한국어정보처리 학술대회 (HCLT) 140-145.

[8] 강승식, & 장병탁. (1996). 음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기. 정보과학회

논문지, 23(5).
[9] 강승식. (2002). “한국어 형태소 분석과 정보 검색“, 홍릉과학출판사.
[10] 권오욱, 정유진, 김미영, 류동원, 이문기, & 이종혁. (1999). 음절단위 CYK알고리즘에 기반한 형태소 분석기 및 품사태거. MATEC ‘99 대회.
[11] 나승훈, 양성일, 김창현, 권오욱, 김영길. (2012). CRF에 기반한 한국어 형태소 분할 및 품사 태깅. 한글 및 한국어 정보처리 학술대회.
[12] 나승훈, 김창현, 김영길. (2013). CRF기반 한국어 형태소 분할 및 품사 태깅에서 두 단계 복합형태소 분해 방법. 한글 및 한국어 정보처리 학술대회.
[13] 박철제, 이종혁, & 이근배. (1997). 확률 접속표를 이용한 일본어 형태소 분석 결과의 후처리 기법. 정보과학회논문지, 24(1).
[14] 신준철, & 옥철영. (2012). 기분석 부분 어절 사전을 활용한 한국어 형태소 분석기. 한국정보과학회 논문지, 39(5), 415-424.
[15] 심광섭. (2013). 음절단위의 한국어 품사 태깅에서 원형 복원. 정보과학회논문지: 소프트웨어 및 응용, 40(3), 182-189.
[16] 심광섭. (2011). 형태소 분석기 사용을 배제한 음절 단위의 한국어 품사 태깅. 인지과학, 22(3), 327-345.
[17] 심광섭. (2011). CRF 를 이용한 한국어 자동 띄어쓰기. 인지과학, 22(2), 217-233.
[18] 심광섭, & 양재형. (2004). 인접 조건 검사에 의한 초고속 한국어 형태소 분석. 정보과학회논문지, 31(1), 89-99.
[19] 양성일, 홍문표, 김영길, 최승권. (2003). 띄어쓰기 정보를 이용한 한국어 복합 형태소 분석. 한국정보과학회 학술발표논문집, 612-616.
[20] 여상화, 김용호, 이학주, & 이정현. (1991). 다단계 필터링 능력을 갖는 형태소 분석기의 설계 및 구현. 한국정보과학회 가을 학술발표 논문집.
[21] 오진영, & 차정원. (2009). 엔트로피 지도 CRF를 이용한 한국어 어절 구문태그 예측. 한국정보과학회 논문지, 15(5), 395-399.
[22] 이성진, 김덕봉, 서정연, 최기선, & 김길창. (1992). Two-level 모델을 이용한 한국어 용언의 형태소 해석. 한국정보과학회 가을 학술발표논문집.
[23] 이은철, & 이종혁. (1992). 계층적 기호 접속정보를 이용한 한국어 형태소 분석기의 구현. 한글 및 한국어 정보처리 학술발표 논문집 (HCLT).
[24] 이재성. (2011). 한국어 형태소 분석을 위한 3단계 확률 모델. 정보과학회논문지, 38(5).
[25] 이창기. (2013). Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅. 한국컴퓨터종합학술대회, 604-606.
[26] 최재혁, & 이상조. (1993). 양방향 최장일치법에 의한 한국어 형태소 분석기에서의 사전 검색 횟수 감소 방안. 한국정보과학회 논문지, 20(10).
[27] 홍진표, & 차정원. (2008). 어절패턴 사전을 이용한 새로운 한국어 형태소 분석기. 한국컴퓨터종합학술대회 논문집 (Vol. 35).