Semi-CRF or Linear-Chain CRF? 한국어 형태소 분할 및 품사 태깅을 위한 결합 모델 비교

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

Semi-CRF or Linear-chain CRF? A Comparative Study of Joint Models 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 한국어 형태소 분할 및 품사 태깅 방법을 위한 결합 모델로 Semi-CRF와 Linear-chain CRF에 대한 초기 비교 실험을 수행한다. Linear-chain방법은 출력 레이블을 형태소 분할 정보와 품사 태그를 조합함으로써 결합을 시도하는 방식이고, Semi-CRF는 출력의 구조가 분할과 태깅 정보를 동시에 포함하도록 표현함으로써, 디코딩 과정에서 분할과 태깅을 동시에 수행하는 방법이다. Sejong품사 부착 말뭉치에서 비교결과 Linear-chain방법이 Semi-CRF방법보다 우수한 성능을 보여주었다.

주제어: 성능 비교, Semi-CRF, Linear-chain CRF, 형태소 분석, 품사 태깅

1. 서론

한국어 형태소 분석을 위한 통계기반 방법은 대규모 품사 부착 말뭉치로부터 형태소 분할 및 품사 태깅에 필요한 규칙 및 확률 모델을 자동 또는 반자동으로 학습하는 방식으로 [3,4,10,12,13,14,15,16,17], 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.

통계 기반 방법 중, 최근 들어 입력문 전체에 대해 CRF나 structured SVM등과 같은 기계 학습 방법을 한국어 형태소 분석에 적용하는 연구들이 제안되었다 [10, 14,15,16]. 이들 연구 중, [10]에서는 한국어 분석을 중국어의 단어 분할 및 품사 태깅 방법을 그대로 적용할 수 있는 방법을 제안했다. 이 방법은 실험 결과 SEJONG코퍼스에서 높은 효과성을 보였으며, [16]에서는 음절 기반 결합 모델로 성능이 더욱 향상되었다.

본 연구에서는 형태소분할과 품사 태깅을 동시 수행하는 결합 모델로 linear-chain CRF [2]와 Semi-CRF [5]을 성능 평가를 수행하여 경험적으로 비교한다. 먼저, linear-chain CRF방식은 [16]의 음절 기반 결합 모델로, 음절의 태깅 출력 레이블을 형태소 분할 정보와 품사태그를 조합하여 구성함으로써, 결합을 수행하는 방식이다. Semi-CRF는 디코딩 과정에서 분할과 태깅을 동시에 수행할 수 있는 방법이다. Sejong품사 부착 말뭉치에서 비교 결과, 기본 자질 집합에서, linear-chain CRF가 Semi-CRF보다 더욱 높은 성능을 보여주었다.

2. 관련 연구

최근의 한국어 형태소 분석의 연구는 입력문 전체에 언어학적 형태소 분석이 없이 통계기반 모델을 수행하는 방식으로 연구가 진행되고 있다. 이들 연구는 표층형과 사전형 간의 불일치 문제를 해소하는 방법에 따라 다음의 두 가지 접근법으로 구분할 수 있다.

먼저, 표층형과 사전형 불일치를 *복합형태소* (compound morpheme)의 도입을 통해 해결하는 방식이다 [10,16]. [10,16]에서는 순차 태깅을 통해, 형태소 분할과 태깅을 수행, 입력문을 단위형태소 또는 복합형태소로 분해한다. 이후, 복합형태소는 후처리 단계에서 기분석 패턴을 참조하여 단위형태소로 상세 분해한다 [10,15]. 기분석 패턴에 나타나지 않은 미등록 복합형태소를 위해 [10]은 이에 더 나아가 래티스 (lattice)기반 HMM방식을 제안하였다.

다른 접근법은 [13,14]의 음절 단위의 접근법이다. 이 방법은 [12]과 기본적으로 동일하나, 표층형과 사전형간의 불일치를 해소하기 위해 복합형태소를 도입하는 대신, 복합 음절을 도입한다. 복합 음절은, 두 가지 서로 다른 태그의 부분음절들이 표층화 단계에서 한음절로 표현된 것으로, [13,14]은 복합 음절을 분해하기 위해, 별도의 음절 복원 기분석 사전에 기반한다. 복합 음절의 경우에는 음절 단위로 처리하므로, 미등록 복합 음절이 거의 발생하지 않는다는 장점이 있어, 별도의 미등록 복합 음절 처리 과정까지는 제안되지는 않았다.

3. 형태소 분할 및 태깅을 위한 결합 방법

본 논문에서 비교하고자하는 형태소 분할 및 품사 태깅의 기본 절차는 [10]의 다음 3단계로 이루어진 형태소

분석 과정이다.

1. 형태소 분할:
   예) <나, 는, 학교, 에, 갔, 다>
2. 품사 태깅:
   예) <나/NP, 는/JX, 학교/NNG, 에/JKB, 갔/VV~EP, 다/EF>
3. 복합형태소 분해:
   예) <나/NP, 는/JX, 학교/NNG, 에/JKB, 가/VV, 았/EP, 다/EF>

위 3단계 절차중 본 논문에서 결합하고자 하는 과정은 형태소 분할과 품사 태깅의 두 단계이다. 다음 절에 linear-chain CRF 및 Semi-CRF에 기반한 두 결합 방법을 기술한다.

3.1 Linear-chain CRF

Linear-chain CRF [2]에 기반한 결합 모델은 음절 기반 태깅으로, [16]와 유사하게, 분할과 품사 태그 정보를 조합하여 태깅의 출력 레이블로 사용한다. 예를 들어, 입력문 "학교에갔다"에 대해 출력 레이블<B-NNG,I-NNG, B-JKB, B-VV~EP, B-EF>와 같이 출력 결과를 구성하는 것이다. 레이블에서 B는 해당 음절에서 새로운 형태소가 시작한다는 것을, I는 해당 음절이 이전 음절의 형태소 단위에 포함된다는 것을 의미한다. {B,I} 태그의 의미에 따라 해석하면 다음의 분할/태깅 열을 얻는다.

<학교/NNG, 에/JKB, 갔/VV~EP, 다/EF>

여기서 "갔"은 복합형태소로, VV~EP는 *복합태그*가 된다, 복합태그는 [10]의 정의에 따르는데, VV~EP는 복합형태소의 시작형태소의 태그가 VV이고 마지막형태소의 태그가 EP를 취한다는 뜻이다. 복합형태소 분해 단계는 linear-chain CRF 및 Semi-CRF 공통으로 사용하는 것으로 3.4절에서 자세히 설명한다.

3.2 Semi-CRF

Linear-chain CRF에서는 품사 태그에 {B, I}의 분할 태그를 덧붙여서 음절의 레이블 집합을 구성함으로써, 분할과 태깅을 동시에 수행한다. 그러나, Semi-CRF [5]는 구조적 분류를 통해 순차열의 출력 구조를 분할과 태깅 정보를 동시에 내포하도록 구성함으로써, 디코딩 과정에서 동시에 분할과 태깅을 수행할 수 있도록 하는 방법이다. 형식적으로 기술하기 위해, 먼저 일반적인 구조적 분류 문제 (structured output problem)를 살펴보자. 주어진 입력문 $\boldsymbol{x}$에 대해 분할/태깅 결과가 $\boldsymbol{y}$라고 하면, 구조적 분류 문제는 다음과 같이 형식화 된다.

$$\boldsymbol{y}^* = \arg max_{\boldsymbol{y}}(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})) \quad (1)$$

Semi-CRF [5]는 구조적 분류 문제의 특수한 예인데, 형식적으로, 입력 $\boldsymbol{x}$는 순차열로 $x_1 \cdots x_n$로 기술되고. 이때 x_i는 주어진 입력의 i번째 단어 (또는 음절)을 n은 음절열의 길이로 $|\boldsymbol{x}|$로도 표현된다. $\boldsymbol{y}$는 세그먼트열로 $y_1 \cdots y_m$으로 기술되는데, 각 $y_{i=}(s_i,e_i,t_i)$로 s_i는 i번째 세그먼트의 시작 위치 (start position)를, e_i는 끝 위치 (end position)를, t_i는 레이블 (label)을 가리키며, m은 세그먼트열의 길이로 $|\boldsymbol{y}|$로도 표기된다.

예를 들어, 입력문 "학교에갔다"에 대한 세그먼트 결과 $\boldsymbol{y}$ = <(1,2,NNG), (3,3,JKB), (4,4,VV~EP), (5,5,EF)>는 linear-chain CRF의 출력 레이블 <B-NNG,I-NNG,B-JKB,B-VV~EP, B-EF>에 대응이 된다.

3.3 Semi-CRF와 linear-chain CRF의 관계

Semi-CRF와 linear-chain CRF는 다음의 관계를 갖는다.

1. Semi-CRF는 linear-chain CRF를 특수한 경우로 포함한다.

형식적으로 Semi-CRF는 linear-chain CRF의 일반화된 형태로, 각 세그먼트 y_i의 최대 길이를 1로 제한하면 Semi-CRf는 linear-chain CRF와 형식적으로 등가가 된다.

2. 분할/태깅 결합 모델의 1차 Linear-chain CRF에 대해, 디코딩 결과가 등가가 되는 Semi-CRF가 존재한다.

여기서 linear-chain CRF는 {B,I}에 기반한 분할/태깅의 결합 모델로 가정하고, 자질벡터를 다음 두 가지로 분류하자.

1. $\boldsymbol{\phi_i}(t,\boldsymbol{x})$: 위치 i의 레이블 t와 입력문 $\boldsymbol{x}$의 함수로 표현되는 모든 자질 벡터의 결합된 형태. 이에 대응되는 자질 벡터는 $\boldsymbol{w_i}(t,\boldsymbol{x})$로 기술한다.
2. $\phi(t,t')$: 레이블 t에서 t'로의 전이 자질로 scalar이다. 이에 대응되는 자질 가중치는 $w(t,t')$로 기술한다.

위의 자질을 사용하는 linear-chain CRF와 등가가 되는 Semi-CRF은 다음의 자질을 사용한다.

1. $\boldsymbol{g}(i,j,t,\boldsymbol{x})$: 시작 위치 i와 끝 위치 j에 대응되는 품사 태그 t에 대한 자질 벡터로, $[\boldsymbol{\phi_i}(B-t,\boldsymbol{x}) \cdots \boldsymbol{\phi_j}(I-t,\boldsymbol{x})]$로 표현된다 (여기서 ;은 여러 개의 벡터를 하나의 벡터로 결합하는 연산자이다). 여기서, $B-t$는 태그 t의 시작 음절을, $I-t$는 태그 t의 중간 또는 끝 음절을 의미하는 음절 단위의 레이블이다. 대응되는 자질 가중치 벡터는 $\boldsymbol{h}(i,j,t,\boldsymbol{x}) = [\boldsymbol{w_i}(B-t,\boldsymbol{x}) \cdots \boldsymbol{w_j}(I-t,\boldsymbol{x})]$로 정의한다.

2. $g(i,j,t,i',j',t')$: 세그먼트 (i,j,t)와 (i',j',t')의 전이 자질 벡터로, $|j-i+1|=1$이면 $\phi(B-t,B-t')$을, 그렇지 않으면 $\phi(I-t,B-t')$로 정의된다. 이에 대한 자질 가중치 벡터는 $h(i,j,t,i',j',t')$로 $|j-i+1|=1$이면 $w(B-t,B-t')$ 그렇지 않으면 $w(I-t,B-t')$ 가 된다.

위의 linear-chain CRF와 semi-CRF는 분할/품사 태깅 결과가 등가임을 쉽게 보일 수 있다. 즉, 이론적으로, 분할/태깅 문제에서 1차 semi-CRF는 1차 linear-chain CRF를 특수한 자질을 사용한 경우로 포함하고 있음을 알 수 있다.

3.4 복합 형태소 분해

Semi-CRF와 Linear-chain CRF의 복합형태소 분할은 [12]의 방법을 따른다. [12]에서는 복합형태소 처리는 기분석에 패턴에 기반을 두며, 기분석 패턴은 두 종류로, 복합형태소에 대한 기분석과 *복합기능형태소* (compound functional morpheme)으로 나뉜다. 예를 들면, 복합형태소의 입력으로 "살려줬/VV~EP "이 주어졌다고 하자. 먼저, "살려줬/VV~EP"이 복합형태소 기분석 패턴에 있는지를 파악한다. 만약, 복합형태소 기분석 패턴에 없다면, 다음 두 단계로 나뉘어 분할이 된다. 먼저, 첫 번째 단계에서는 "살리/VV"를 내용형태소로 분해하고, "어줬/EP"를 복합기능형태소 분해한다. 이때, EP는 복합기능형태소에 대한 복합태그로서, 가장 마지막 단위 형태소의 태그로 지정하였다. 두 번째 단계에서는, 복합기능형태소 "어줬/EP"에 대해서 기분석패턴에 기반하여 분해를 시도한다. 기분석을 참조하여, "어줬/EP: 어주/VX+었/EP"를 후보로 얻어내어, 최종적으로 해당 복합형태소에 대해 "살리/VV+어주/VX+었/EP"의 분해 결과를 얻게 된다. 위의 두 과정을 요약하면 다음과 같다.

살려줬/VV~EP ➔ (내용형태소와 기능형태소 분해)
살리/VV+어줬/EP ➔ (기능형태소 상세 분해)
살리/VV+어주/VX+었/EP

3.5 자질 벡터

앞에서 설명한 두 가지 결합 방법에 대한 자질 벡터를 정리하면 다음과 같다.

Linear-chain CRF에서 자질 벡터

[10]의 형태소 분할을 위한 자질벡터와 동일하며, 다음과 같이 기술된다.

표 1. Linear-chain CRF에서 사용하는 자질 유형

Feature symbol	Desciption
$C_{-2}, C_{-1}, C_0, C_1, C_2$	1음절 (uni-char) 정보
$C_{-1}C_0, C_0C_1, C_1C_2$	2음절 (bi-char) 정보
$C_{-2}C_{-1}C_0, C_{-1}C_0C_1,$ $C_0C_1C_2$	3음절 (tri-char) 정보
$S_{-2}, S_{-1}, S_0, S_1, S_2$	1음절 띄어쓰기 정보
$S_{-1}S_0, S_0S_1, S_1S_2$	2음절 띄어쓰기 정보
$S_{-2}S_{-1}S_0, S_{-1}S_0S_1,$ $S_0S_1S_2$	3음절 띄어쓰기 정보

Semi-CRF에서 자질 벡터

Semi-CRF에서 사용하는 자질 표기를 위해, C_i를 i번째 위치에서 표층음절을, C_i^j는 $C_i \cdots C_j$를, S_i를 i번째 위치 바로 이전에 띄어쓰기가 있었는지 여부를, S_i^j는 $S_i \cdots S_j$를 가리키는 표기라고 하자. 세그먼트 (i,j,t)에 대해서 Semi-CRF에서 사용하는 자질은 다음과 같다.

표 2. Semi-CRF에서 사용하는 자질 유형

Feature symbol	Desciption
$C_i \cdots C_j$	세그먼트에 대한 표층정보
$C_{i-2}C_{i-1}C_i^j, C_{i-1}C_i^j,$ $C_i^jC_{j+1}, C_i^jC_{j+1}C_{j+2}$	세그먼트와 주변 1-2음절에 대한 공기정보
$S_i \cdots S_j$	세그먼트에 대한 띄어쓰기 정보
$C_{i-2}, C_{i-1}, C_i,$ C_j, C_{j+1}, C_{j+2}	1음절 (uni-char) 정보
$C_{i-2}C_{i-1}, C_{i-1}C_i, C_iC_{i+1},$ $C_{j-1}C_j, C_jC_{j+1}, C_{j+1}C_{j+2}$	2음절 (bi-char) 정보
$C_{i-2}C_{i-1}C_i, C_{i-1}C_iC_{i+1},$ $C_iC_{i+1}C_{i+2}, C_{j-2}C_{j-1}C_j,$ $C_{j-1}C_jC_{j+1}, C_jC_{j+1}C_{j+2}$	3음절 (tri-char) 정보
$S_{i-2}, S_{i-1}, S_i, S_j, S_{j+1}, S_{j+2}$	1음절 띄어쓰기 정보
$S_{i-2}S_{i-1}, S_{i-1}S_i, S_iS_{i+1},$ $S_{j-1}S_j, S_jS_{j+1}, S_{j+1}S_{j+2}$	2음절 띄어쓰기 정보
$S_{i-2}S_{i-1}S_i, S_{i-1}S_iS_{i+1},$ $S_iS_{i+1}S_{i+2}, S_{j-2}S_{j-1}S_j,$ $S_{j-1}S_jS_{j+1}, S_jS_{j+1}S_{j+2}$	3음절 띄어쓰기 정보

4. 실험 결과

두 가지 결합 모델의 형태소 분할 및 품사 태깅 성능 평가를 위해, 세종 품사 부착 말뭉치를 이용하였는데 이는 총 253,884개의 문장, 1,008,925개의 어절로 구성된다. 성능 평가 측도로 형태소 단위의 F-measure, 어절 정확률 (EA)를 사용하였으며, 이중 80%를 학습데이터로 20%는 테스트데이터로 사용했다. Linear-chain CRF와 Semi-CRF의 가질 가중치 벡터는 모두 [1]의 Averaged perceptron을 사용하여 학습하였다. 전체 학습데이터에서 최대 반복횟수를 총 30회로 제한하였다.

표 3 Linear-chain CRF와 Semi-CRF의 성능 비교

방법	평가 단계	F-measure	EA
CRF	분할만	98.78%	98.42%
	분할+태깅	97.61%	96.14%
	형태소 분석 전과정	**97.23%**	**95.21%**
Semi-CRF	분할만	97.84%	98.16%
	분할+태깅	96.59%	95.73%
	형태소 분석 전과정	96.83%	94.51%

표 3은 Sejong데이터에서 사용한 성능 결과를 보여준다. 표 3에서 CRF에 대한 결과는 linear-chain CRF의 성능 추치를 보여준다. 표 3에서 보다시피, 3.5에서 사용된 자질 집합에 대해서, linear-chian CRF가 Semi-CRF보다 전 과정에서 다소 높은 성능을 보여주었다. Semi-CRF가 이론적으로 linear-chain CRF를 포괄함에도 불구하고, 본 연구에서는 linear-chain CRF의 성능을 뛰어넘지 못했다. 3.5의 자질의 형태로는 Semi-CRF의 성능이 한계를 보여준다는 것을 의미한다.

Linear-chain CRF의 경우 상태 전이의 개수가 일정하는 점이 세그먼트의 개수에 따른 선호나 편향성 없이 학습될 수 있도록 하는 장점이 된 것으로 보인다. 물론, 이러한 결과는 본 연구에서 사용되는 자질 집합으로 국한된 것이므로, 두 방법에 대한 비교의 최종 결론을 도출하기 위해서는 보다 다양한 자질 집합에서 엄밀한 성능 비교를 해야 할 것이다.

5. 결론

본 논문은 형태소 분할과 품사 태깅을 위한 결합 모델로 linear-chain CRF와 Semi-CRF를 비교했다. 실험 결과, Semi-CRF가 이론적으로 linear-chain CRF를 포괄함에도 불구하고, 본 연구에서는 linear-chain CRF의 성능을 뛰어넘지 못했다. 본 연구는 Semi-CRF의 적용의 초기 연구로, 향후 개선의 여지가 남아있음을 물론이다. 후속 정교한 비교 연구를 통해, 위의 이슈에 대해 보다 엄밀한 실험을 수행할 예정이다. 학습 단계에서, Linear-chain CRF와 등가가 되는 고유 자질을 고정시키고, Semi-CRF에서 추가로 사용되는 자질만을 선택적으로 가중치 학습하는 것도 Semi-CRF의 성능 개선을 위한 한 방법이 될 것이다.

참고문헌

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

[2] Lafferty, J., Mccallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In ICML (Vol. 2001).

[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] Sarawagi, S., & Cohen, W. W. (2004). Semi-Markov Conditional Random Fields for Information Extraction. In NIPS.

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

[7] 강승식. (1994). 다층 형태론과 한국어 형태소 분석 모델. 한글 및 한국어정보처리 학술대회 (HCLT) (pp. 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] 나승훈, 양성일, 김창현, 권오욱, 김영길. (2012). CRF에 기반한 한국어 형태소 분할 및 품사 태깅. 한글 및 한국어 정보처리 학술대회

[11] 나승훈, 김창현, 김영길. (2013). CRF기반 한국어 형태소 분할 및 품사 태깅에서 두 단계 복합형태소 분해 방법. 한글 및 한국어 정보처리 학술대회

[12] 신준철, & 옥철영. (2012). 기분석 부분 어절 사전을 활용한 한국어 형태소 분석기. 한국정보과학회 논문지, 39(5), 415-424.

[13] 심광섭. (2013). 음절단위의 한국어 품사 태깅에서 원형 복원. 정보과학회논문지: 소프트웨어 및 응용, 40(3), 182-189.

[14] 심광섭. (2011). 형태소 분석기 사용을 배제한 음절 단위의 한국어 품사 태깅. 인지과학, 22(3), 327-345.

[15] 이재성. (2011). 한국어 형태소 분석을 위한 3단계 확률 모델. 정보과학회논문지, 38(5).

[16] 이창기. (2013). Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅. In 한국컴퓨터종합학술대회 (pp. 604-606).

[17] 홍진표, & 차정원. (2008). 어절패턴 사전을 이용한 새로운 한국어 형태소 분석기. 한국컴퓨터종합학술대회 논문집 (Vol. 35)