한국어에서 Viterbi 형태소 복원

이제승○, 김재훈
한국해양대학교, 컴퓨터공학과 및 해양인공지능융합전공
leeje1231@naver.com, jhoon@kmou.ac.kr

Viterbi Morpheme Restoration in Korean

Je-seung Lee○, Jae-hoon Kim
Dept. of Computer Engineering and Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University

요 약

본 논문은 한국어에서 형태소 복원을 위한 새로운 방법을 제안한다. 일반적으로 기계학습 기반 형태소 분석에서 형태소 복원은 기분석 사전과 약간의 경험규칙을 이용한다. 이와 같은 방법은 모호성을 해결하기 위해 사전에 모든 정보를 저장하는 것이 불가능할 뿐 아니라 단음절 이형태의 모호성을 해결할 수 없을 것이다. 이러한 문제를 완화하기 위해 본 논문에서는 생성된 모호성을 Viterbi 알고리즘을 이용해서 해소한다. 본 논문의 형태소 복원 과정은 기본적으로 기분석 사전과 약간의 경험규칙을 이용하여 형태소 복원 후보를 찾고 여러 후보가 있을 경우(모호성의 생성), 그 결과를 Viterbi 알고리즘으로 이형태를 결정한다. 실험을 위해 모두의 말뭉치(형태 분석)를 사용하고, 평가는 NER 방식으로 평가한다. 그 결과 품사 부착에 대해 96.28%정도의 성능을 보여주었다.

주제어: 형태소 분석, 형태소 복원, Viterbi

1. 서론

형태소는 문장 내에서 최소 의미를 갖는 단위이며, 품사 부착은 형태소에 적절한 품사를 부여하는 것이다. 형태소 복원은 형태소의 표층형(surface form)을 어휘형(lexical form)으로 변환하는 것이다. 일반적으로 문장에서 형태소 단위로 분리한 뒤, 품사 부착을 하고 형태소 복원을 하는 일련의 과정을 형태소 분석이라고 한다. 본 연구에서는 품사 정보를 이용하여 형태소를 추출하는 형태소 복원 과정에 초점을 맞추어 진행한다.

일반적으로 형태소를 분리한 후, 품사를 부착하고 기분석(형태소 원형 복원) 사전을 통해 형태소를 복원한다. 형태소 복원에서 가장 중요한 문제는 모호성 해결이다. 여기서 모호성은 동일한 음절과 품사에 대해 형태소 복원 후보가 여러 이형태로 존재하는 것을 의미한다. 표 1은 모호성의 예를 보이고 있다. 이를 해결하기 위해 표 1과 같은 기분석 사전을 이용하여 형태소가 복원되는 종류마다 품사를 각각 부착하는 방식이 있는데, 이런 경우에는 부착해야 할 품사의 수가 많아져 성능이 떨어질 수 있고, 미등록 문제에 취약한 모습을 보여준다. 본 모델은 기존의 기분석 사전이 가진 단점을 보완하기 위하여 사전과 확률 정보를 바탕으로 Viterbi 알고리즘을 이용하여 형태소의 원형을 복원한다.

본 논문의 2장에서는 형태소 복원 방법의 발전과정과 함께 관련 연구를 소개하고, 3장에서는 Viterbi 형태소 복원에 대해 설명한다. 4장에서는 실험을 통해 모델의 성능을 평가하고, 5장에서는 결론 및 향후 연구의 진행 방향을 제시한다.

표 1. 형태소 원형 복원 사전의 예[1]

음절/품사	형태소/품사
랬/I_VV+B_EP__1	러/I_VV+었/B_EP
랬/I_VV+B_EP__2	래/I_VV+었/B_EP
랬/I_VV+B_EP__3	러/I_VV+았/B_EP
랬/I_VV+B_EP__4	라/I_VV+았/B_EP

2. 관련 연구

한국어의 형태소 복원 방법은 규칙 기반에서 시작해 사전 기반, 기계학습 기반, 심층학습 기반 연구로 발전해왔다. 규칙 기반 연구는 규칙을 통해 적절한 형태소를 찾아내는 방법인데, two-level 모델[2], Head-tail 구분법[3], Tabular 파싱법[4] 등을 통해 형태소 분석을 하는 방법들이 연구되었다.

규칙을 작성하여 형태소 분석을 하기에는 생성해야 하는 규칙과 예외가 너무 많아 개발의 한계가 있는데, 이를 해결하기 위하여 형태소 복원 사전을 이용하는 방법으로 발전하였고, 현재도 활발히 활용되고 있는 방법이다. 형태소 복원 사전을 사용하더라도 여전히 형태소의 모호성 문제를 해결하기는 여전히 많은 문제점이 남아 있었고, 기기의 발전에 따라 기계학습을 이용하는 방법들도 연구되기 시작하였다. Naïve Bayes(NB) 모델을 이용한 연구[5]에서는 기계학습을 이용하여 형태소를 복원하는 것이 기존 모델들과 큰 성능 차이가 없음을 보여주었고, 주의(attention)기반의 sequence-to-sequence 모델으로 형태소 복원 문제를 처리한 연구[6]와 copying mechanism기반의 sequence-to-sequence 모델으로 품사부

착과 형태소 복원 문제를 한번에 해결하는 End-to-end 방식의 분석을 시도한 연구[7]들은 기계학습 중에서도 심층학습을 이용한 대표적인 예고, 현재 대부분의 형태소 복원 문제들은 심층학습을 이용하여 해결한다.

본 논문에서는 심층학습에서 요구하는 방대한 자원과 긴 학습시간을 피하기 위해서 기분석 사전과 Viterbi 알고리즘을 이용한다. 이와 비슷한 연구로 기분석 사전과 lattice-HMM을 이용하여 형태소를 복원한 연구[8]가 있는데, 이 연구에서는 복합 표지를 가진 어절에 대하여 첫 품사와 마지막 품사만을 이용해 HMM(Viterbi)으로 품사와 형태소를 복원한다. 최근 심층학습을 이용하여 품사 부착 성능이 높아졌고, 사전이 어절 단위로 되어 있어 미등록 문제에 취약할 수 있기 때문에 본 논문에서는 부착된 품사를 그대로 사용하며, 음절 단위의 사전을 이용한다. 3.1절부터 이 모델에 대한 자세한 과정을 설명한다.

3. Viterbi 알고리즘을 이용한 형태소 복원

형태소 복원이란 음절별 부착된 품사를 바탕으로 형태소를 만들어내는 것을 뜻하는데, 예를 들어 (됐/B_VV+B_EP)으로 부착되었다면, 적절한 형태소 "되+었"으로 각 태그에 맞는 형태소를 찾는 과정이다. 하지만 표 1과 같이 동일한 음절/품사 쌍에 대해 다양한 형태소가 존재하므로 모호성을 해결할 필요가 있다. 본 논문에서는 두 단계(모호성 생성 단계와 형태소 복원 단계)로 구성된다. 모호성 생성 단계는 형태소 복원 사전과 이형태 생성 규칙을 이용해서 모호성을 생성한다. 형태소 복원 단계는 형태소 음절에 대한 확률 정보와 Viterbi 알고리즘을 이용해서 형태소를 복원한다.

3.1 형태소 모호성 생성 단계

형태소 모호성 생성은 형태소 복원 사전과 이형태 생성 규칙을 이용해서 모호성을 생성한다. 형태소 복원 생성 사전에 포함되지 않은 음절은 이형태 생성 규칙을 통해서 모호성을 생성한다.

1) 형태소 복원 사전 구축

형태소 복원 사전은 학습 말뭉치에서 (음절, 품사) 쌍에 대한 이형태 복원 후보 리스트를 저장해 놓은 사전이다(표 2 참조). 형태소 복원 사전은 기존의 형태소 복원 방법과 같이 학습 말뭉치로부터 표 2와 같은 형태소 복원 사전을 구축한다. 이와 같은 복원 사전을 구축하기 위해서는 표층 음절과 분석 음절의 정렬이 필요하며 본 논문에서는 [9]의 정렬 알고리즘을 수정하여 사용한다.

형태소 복원 사전을 구축할 때 모든 표층 음절과 품사 표지[1]에 대해 이형태 복원 후보(어휘 음절)를 저장하는데, 이는 두 가지 이유가 있다. 첫째는 학습 말뭉치에 존재하지 않을 경우 미등록 음절에 대해서도 복원할 수 있어야 하기 때문이다. 둘째는 단순 표지(예: 표 2에서 음절 '그' 참조)도 모호성이 존재하기 때문이다.

표 2. 형태소 복원 사전

표층 음절	품사 표지	이형태 복원 후보
그	B_VA	그
		긇
러	I_VA+B_EC	렇+으
		르+어
면	I_EC	면

2) 이형태 생성 규칙

표 3은 이형태 생성 규칙의 일부분이다. 표 3에서 규칙 "'이' 추가"는 복합 표지가 VCP로 시작하면 '이'를 추가한다(예: 었/B_VCP+B_EP → 이/B_VCP+었/B_EP). 표 3에서 규칙 "종성 분리"는 복합 표지에 ETM, JKO, JKS가 있을 경우에 표층 음절에서 종성을 분리한다(예: 될/B_VV+B_ETM → 되/B_VV+ㄹ/B_ETM). 표 3에서 규칙 "이중모음 분리"는 복합 표지에 EP가 있을 경우에 이중모음을 분리한다(예: 왔/I_VV → 오/I_VV+았/B_EP).

표 3. 이형태 생성 규칙

규칙	설명
'이' 추가	품사 표지가 VCP일 경우 "이"로 대체하고 나머지 품사 표지에 대해 형태소 후보를 찾는다.
종성분리	ETM, JKO 등의 품사에 대해 원래 음절에서 종성을 분리한다.
이중모음 분리	EP 등의 품사에 대해 이중모음이고 받침이 없거나 "ㅆ"일 때 이중모음을 분리한다.

3.2 형태소 복원 단계

형태소 복원 단계는 3.1절에서 기술한 모호성 생성 단계를 통해서 래티스(lattice)를 구성하고 그 래티스로부터 최적의 경로를 찾는 과정으로 수행된다.

1) 래티스 구성

어절 '그러면'에 대해서 형태소 복원을 위한 래티스 구조는 그림 1과 같다. 그림 1에서 노드(node)는 3.1절에서 기술한 모호성 생성 단계를 통해서 생성한다. 간선(edge)는 2-그램을 이용해서 연결되며 그 가중치는 식 (1)과 같이 정의된다.

$$W(w_i, w_j) = -\log \Pr(w_j \mid w_i) - \log \Pr(w_i \mid t_i) \quad (1)$$
$$= -\log\left(\frac{f(w_i, w_j)}{f(w_i)}\right) - \log\left(\frac{f(t_i, w_i)}{f(t_i)}\right)$$

1) 품사 표지는 단순 표지(예: B_VA)와 복합 표지(예: I_VA+B_EC)가 있다.

가중치 함수 $W(\cdot)$는 최적 경로를 찾기 위해서는 최단 경로 문제(shortest path problem)로 적용해야 하므로 가중치는 로그 확률에 음수를 취하여 사용한다.

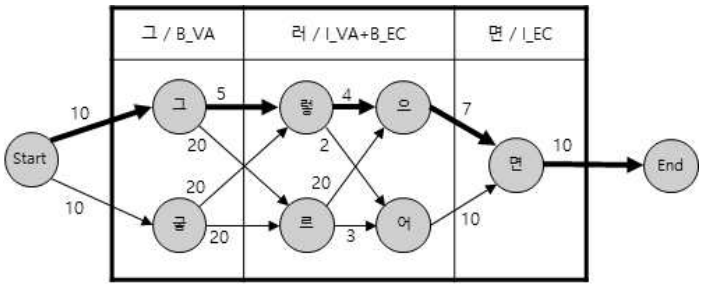

그림 1. 어절 “그러면”의 형태소 복원 래티스

2) Viterbi 형태소 복원

앞에서 언급한 형태소 복원을 위한 래티스 구조에서 최단거리[2]를 찾아서 경로 상에 있는 이형태로 최종적으으

으로 복원한다. 그림 1에서 Viterbi 알고리즘을 통해 구한 최단 경로는 굵은 화살표로 표시된 경로이고, 가중치의 합이 10+5+4+7+10=36으로 최소 가중치를 가진다. 이를 통해 ‘그’, ‘러’, ‘면’이라는 세 개의 음절이 ‘그’, ‘렇+으’, ‘면’이라는 형태소로 복원된다.

4. 실험 및 평가

4.1 실험 환경

품사 부착의 성능을 평가하기 위해 모두의말뭉치(형태분석 말뭉치)[3]를 이용한다. 전체 말뭉치는 총 371,571개의 문장으로 구성되어 있고, 학습 말뭉치와 평가 말뭉치를 문장 단위로 7:3의 비율로 분리하였으며 그 결과는 표 4와 같다.

표 4. 학습 및 평가 말뭉치의 구성

구분	학습 말뭉치	평가 말뭉치
문장 수	260,130	111,441
음절 수	8,521,310	3,640,458
형태소 수	4,550,115	1,942,943
어절 수	2,106,473	900,187

학습 말뭉치와 평가 말뭉치는 각각 260,130 문장(8,521,310 음절)과 111,441 문장(3,640,458 음절)으로 구성된다.

2) https://networkx.org/documentation/stable/reference/algorithms/shortest_paths.html

3) https://corpus.korean.go.kr/

평가 방식은 NER(Named-entity recognition)의 평가 방식[10]을 사용하였다. NER 평가는 네 가지의 평가척도로 나뉘어지는데, 각각의 평가척도를 형태소 분석에 적용하였고, 그 내용은 다음과 같다.

(1) Type : 품사 일치율
(2) Exact : 형태소 일치율
(3) Partial : 형태소 경계 일치율
(4) Strict : 품사 부착율(Type and Exact)

MUC 및 SemEval에서 사용한 평가 방법을 통해 형태소와 품사 표지(strict)를 모두 맞춘 경우 뿐만 아니라 형태소를 맞춘 경우(exact), 형태소 경계 중 일부만 맞춘 경우(partial), 형태소의 표지가 일치하는 경우(type) 등에 대해서 정밀도(precision)와 재현율(recall), F1-점수(F1-score)를 측정한다.

4.2 성능 평가

NER의 평가척도로 제안된 모델을 평가한 결과는 표 5와 같다.

표 5. 품사 부착 및 형태소 복원 실험 결과(%)

eval \ schema		Type	Partial	Exact	Strict
모델 1	정밀도	96.83	98.11	97.24	95.38
	재현율	97.02	98.30	97.42	95.57
	F1점수	96.93	98.20	97.33	95.47
모델 2	정밀도	97.20	97.81	96.83	95.41
	재현율	97.49	98.10	97.13	95.70
	F1점수	97.34	97.95	96.98	95.55
모델 3	정밀도	97.57	98.12	97.26	96.14
	재현율	97.84	98.39	97.53	96.41
	F1점수	97.71	98.26	97.40	96.28

모델 1은 CRF를 이용하여 품사 부착을 진행한 후 본 논문의 형태소 복원 방법을 적용하였으며, 모델 2는 RNN(BiGRU), 모델 3은 RNN(BiGRU)의 마지막 층에 CRF를 더한 BiGRU-CRF를 사용하여 품사를 부착한 후 본 논문의 형태소 복원 방법을 적용한 모델이다. 모델 1과 2를 비교하였을 때, CRF보다 RNN의 Type은 더 높지만, Exact는 더 낮다. 이는 RNN이 단순 표지에 대해 더욱 좋은 표지 부착 성능을 보이지만, 복합 표지에 대해서는 부착 성능이 떨어져 형태소 복원 성능이 CRF(모델 1)보다 낮게 나온 것으로 해석할 수 있으며, 모델 3은 모델 1, 모델 2보다 단일 표지와 복합 표지에 대한 품사 부착 능력이 모두 높다고 평가된다.

5. 결론

기존의 형태소 복원 방식의 단점을 보완하기 위해 본 연구에서는 사전과 빈도수 정보를 바탕으로 Viterbi 알고리즘을 사용하여 형태소 복원을 하는 방법을 사용하였다. 실험 결과 97%의 품사 일치율에 대해 97%의 형태소 일치율과 96%의 품사 부착 성능을 보였다. 형태소 복원의 경우 품사 일치율에 어느 정도 영향을 받는 단점이 있어 추후 형태소 복원 과정에서 형태소 복원 정보를 바탕으로 잘못된 품사를 수정하거나 이형태 생성 규칙을 추가하여 형태소 복원의 성능을 높이는 연구를 진행할 예정이다.

참고문헌

[1] https://tech.kakao.com/2018/12/13/khaiii/

[2] 이성진, Two-Level 한국어 형태소 해석, 한국과학기술원, 전산학과, 석사학위 논문, 1992

[3] 최형석, 이주근. “자연어 어절의 처리 앨고리즘”, 한국정보과학회 학술발표논문집, pp. 36-43, 1984.

[4] 김성용, 최기선, 김길창, “Tabular Parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기”, 한국정보과학회 인공지능연구회 춘계 인공지능 학술발표회 논문집, pp. 133-147, 1987.

[5] 김재훈, 전길호, “NB 모델을 이용한 형태소 복원”, 정보처리학회논문지. 소프트웨어 및 데이터 공학, 제19권, 제3호, pp. 195-200, 2012.

[6] 윤준영, 이재성, “한국어 형태소 분석 및 품사 태깅을 위한 딥 러닝 기반 2단계 파이프라인 모델”. 정보과학회논문지, 제48권, 제4호, pp. 444-452, 2021.

[7] 최병서, 이익훈, 이상구, “신조어 및 띄어쓰기 오류에 강인한 시퀀스-투-시퀀스 기반 한국어 형태소 분석기”, 한국정보과학회 논문지, 제47권, 제1호, pp. 70-77, 2020.

[8] 나승훈, 양성일, 김창현, 권오욱, 김영길, “CRF에 기반한 한국어 형태소 분할 및 품사 태깅”, 한글 및 한국어 정보 처리 학술 대회, pp.12-15, 2012.

[9] 김재훈, 이공주, “한국어 정보처리 : 사례기반 학습을 이용한 음절기반 한국어 단어 분리 및 범주 결정”, 정보처리학회논문지B, 제10권, 제1호, pp. 47-56, 2003.

[10] I. Segura-Bedmar, P. Martínez, and M. H. Zazo, “SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)”, Proc. of the 2nd Joint Conference on Lexical and Computational Semantics and the 7th International Workshop on Semantic Evaluation, pp. 341-350, 2013.