한국어 형태소 복원 확률 모델의 계산 방법 비교

이다니엘[O], 김보겸[†], 이재성[†]
[O]충북대학교 정보산업공학과, [†]컴퓨터교육과
daniellee@cbnu.ac.kr[O], bogyum@cbnu.ac.kr[†], jasonlee@cbnu.ac.kr[†]

Comparison of Calculation Methods for Probabilistic Korean Morpheme Recovery Model

Daniel Lee[O], Bogyum Kim[†], Jae Sung Lee[†]
[O]Dept. of Information & Industrial Engineering, [†]Dept. of Computer Education,
Chungbuk National University

요 약

형태소 복원은 형태소 분석의 한 단계로 문장에 나타난 형태소의 변형 현상을 분석하여 규칙화하고 이를 이용하여 형태소 원형을 복원하는 것이다. 본 논문에서는 형태소 품사 부착 말뭉치로부터 다양한 형태소 변화 규칙을 학습하여 효과적으로 형태소 원형을 복원하기 위한 계산 방법을 비교한다. 이를 위해 계산 모델, 한글 코드, 학습 자료를 다르게 하여 학습하고 그에 따른 성능을 비교 분석한다.

주제어: 형태소 학습, 형태소 원형 복원, 한국어 형태소 분석, 형태소 확률 모델

1. 서론

말뭉치들을 이용한 통계 기반 자연언어 처리시스템은 비교적 짧은 시간에 학습을 통해 많은 규칙을 얻을 수 있어 편리하다[1, 2]. 확률 기반 한국어 형태소 분석 방법은 기 구축된 형태소 품사 부착 말뭉치를 이용하여 모델을 학습하고 자동으로 형태소 분석기를 만들어 낸다[3, 4]. 이는 형태소 품사 부착 말뭉치만 주어진다면 짧은 시간내에 형태소 분석 규칙뿐만 아니라 형태소 사전을 구축하므로 매우 편리하다. 더욱이 실제 사용된 문서에서의 언어 현상을 그대로 학습하여 보다 현실적인 분석이 가능하고, 품사 부착 말뭉치의 형태소 분석 기준이 그대로 반영되므로 다양한 형태소 분석 기준을 쉽게 반영할 수 있다.

확률 기반 형태소 분석은 형태소 분석 단계를 2단계 혹은 3단계로 나누어 처리한다. 2단계는 원형 복원 단계와 분리 및 태깅 단계로 나누고, 3단계 모델은 원형 복원, 형태소 분리, 형태소 태깅의 단계로 나눈다. 이 두가지 모두 첫 단계로 원형 복원 모델이 사용되며, 모델의 성격상 첫 단계의 오류는 다음 단계에서의 최고 성능의 한계(upper bound)로 작용하여 그 성능의 향상이 매우 중요하다.

본 논문에서는 품사 부착 말뭉치를 이용하여 형태소 복원 모델의 성능을 높이기 위해, 여러가지 학습 방법을 비교한다. 형태소 복원 모델은 확률 기반 형태소 분석 모델에서 사용한 모델들을 사용하였다. 또한, 학습 자료를 긍정의 예뿐만 아니라 부정의 예도 반영하여 학습하는 방법을 변형하여 비교하였고, 한글 표기를 다른 코드 체계로 바꾸어 비교하였다.

2. 관련연구

한국어 형태소 복원은 모든 한국어 분석기에서 기본적으로 이루어진다. 이 부분은 형태소 분리와 함께 일어나기도 하고, 형태소 태그 부착과 동시에 일어나기도 한다. 확률 기반 형태소 분석 방법에서는 형태소 복원을 형태소 분리나 태그 부착과는 독립적인 단계로 보고 이를 하나의 모델로 정의하였다[3, 4].

[4]의 모델은 기본적으로 원형 복원 모델을 부분 음절 단위로 처리하였다. 표층형(활용형) W에 대해 어휘형(원형) O가 대응될 경우, 이를 부분 음절렬 k개로 대응시켜 다음과 같은 식 (1)로 복원 모델을 표현할 수 있다. 여기에서 so와 sw는 각각 O와 W에 대한 부분 문자열이다. 특히 대응 표층 문자열이 빈문자열(null)일 경우, 빈문자열은 다음 표층 문자열과 결합하여 빈문자열 규칙이 생성되지 않도록 했다. 변환 규칙 추출은 단순하게 어절의 앞부터 비교하여 달라진 부분을 표시(시작점)하고, 다시 어절의 뒤부터 비교하여 달라진 부분을 표시(끝점)하여 시작점과 끝점 사이의 음절열을 추출하여 확률 계산에 이용하였다.

$$p(O|W) = p(so_{1,k}|sw_{1,k}) \approx \prod_{j=1}^{k} p(so_j|sw_j) \quad (1)$$

[3]의 모델은 복원을 같은 언어간의 번역으로 보고 통계적 기계 번역 방법[5]을 사용하였다. 즉, 복원 방법을 수식(2)와 같이 원형에서 활용형으로의 변환 모델과 원형에 대한 언어 모델로 나누어 처리하였다. 각각의 모델은 어절을 자소로 풀어 쓰고, 변환 모델은 이 자소들

간의 변형된 부분을 중심으로 좌우 자소를 문맥으로 하여 변환 규칙을 추출하였고(식 3), 언어 모델은 원형 자소들의 연결 관계를 모델링하였다(식 4). 모델 학습을 위해서 글자 정렬(align)을 한 후([6]의 프로그램을 글자용으로 수정함), 변화된 부분을 찾아 변환 확률 계산에 이용하였다.

$$p(O|W) = p(W|O)p(O) \quad (2)$$

$$p(W|O) = \prod p(l,w,r|l,o,r) \quad (3)$$

$$p(O) = \prod p(o_{i,i+1}|o_{i-2,i-1}) \quad (4)$$

3. 계산 모델

본 논문에서는 형태소 복원 확률 모델을 효율적으로 계산하기 위해 필요한 여러 요소중 3가지 요소를 살펴보고 이 요소의 영향을 비교 분석한다.

첫째 요소는 계산 모델이다. 아래의 수식 (5)는 기본모델로 일반적인 조건부 확률 계산식에 해당한다. [4]의 연구에서는 계산식에서 특별한 언급은 없었으나, 수식 (5)를 사용했을 것으로 추정된다. [3]의 연구에서는 확률 계산을 기본적인 지역확률과 전역확률로 나누어 계산하였다. 지역확률은 주어진 자소열 혹은 음절열 x에 대해 여러 가능성 있는 자소열 혹은 음절열 y들 중 하나를 선택할 확률이고, 전역확률은 x에서 y로 변환이 일어날 확률이다. 수식 (6)은 기본모델에 전역확률을 곱한 것으로 확장모델이다. (편의상 기본모델은 A, 확장모델은 B로 나타낸다.)

$$p(l,x,r|l,y,r) = \frac{cnt(l,x,r,y)}{1+\sum_{y} cnt(l,x,r,y)} \quad (5)$$

$$p(l,x,r|l,y,r) = \frac{cnt(l,x,r,y)}{1+\sum_{y} cnt(l,x,r,y)} \times \frac{cnt(l,x,r,y)}{1+\sum_{x}\sum_{y} cnt(l,x,r,y)} \quad (6)$$

두번째는 학습 자료 선택 방법이다. 일반적으로 형태소 복원을 하기 위해서는 음운변화가 일어난 문자열을 추출하고 그 자료만을 학습 자료로 사용하여 확률 계산을 하였다[3]. 이를 변화에 긍정적인 자료로 보아 긍정예로만 계산하는 것으로 하고, 반대로 변화가 일어나지 않은 기존의 학습 자료를 검색하여 변화가 일어나지 않는 부정예를 학습에 사용할 수 있도록 했다. 긍정예만을 쓰는 경우, 많은 자료에서 비교적 쉽게 학습 자료를 선별할 수 있어 편리하지만, 변환 확률의 계산은 부정확할 수 있다. 따라서, 이 차이가 어느 정도인지 실제 자료를 통해 검증한다. (긍정예만 학습하는 경우를 P, 부정예를 포함하여 모두 학습하는 경우를 N으로 한다.)

셋째는 한글 표기이다. 한글은 음절 단위, 자소 단위로 표기가 가능하다. 자소 단위 또한 복자음, 복모음을 쓰는가 단자음, 단모음을 쓰는가에 따라 표기 방법이 달라진다. 형태소 복원 모델은 좌우 문맥이나 변환 단위가 제한적이므로, 어떤 한글 표기 방법이 각 계산 모델에 적합한지 검토가 필요하다. 이 연구에서는 한글 음절을 자소에 관계없이 하나의 단위로 처리할 수 있는 2바이트 코드(S로 표시), 한 음절을 초성, 중성, 종성으로 나누고 이를 각각 한 바이트로 표현한 3 바이트 코드(T로 표시), 한글 음절을 초성, 중성, 종성으로 구분하되 중성에서의 복모음을 단모음(혹은 반자음)의 결합으로 표기하고, 종성에서의 복자음도 단자음의 결합으로 표기한 일종의 n 바이트 코드(V로 표시)에 대해 비교한다. 단, 3바이트 코드나 n 바이트 코드의 경우, 초성 'ㅇ'은 생략하였다.

자소 단위 모델(T, V)과 음절 단위 모델(S)는 계산 모델(A 혹은 B)에서 다르게 계산된다. 자소 모델은 식(5)와 식(6)에서 l과 r이 좌측 문맥과 우측 문맥을 나타내는 1개의 자소인 반면, 음절모델에서는 사용되지 않는다. 2장에서 설명한 바와 같이 자소 모델은 [3]의 방법에 따라 글자 정렬(align)방법을 사용하며, 음절 단위 모델은 [4]의 연구 방법을 따라, 어절의 앞과 뒤에서 같은 음절을 제거하여 바뀐 부분을 추출하는 방법을 사용한다. 또한 음절 단위 모델의 경우, 학습 방법 차이로 부정예를 추출하기 어려워, 부정예에 대한 비교는 하지 않는다.

따라서, 모든 실험 방법을 앞에서 정의한 알파벳 조합으로 표현하면 모두 10가지(=2x2x3-2)로, APS, APT, APV, ANT, ANV, BPS, BPT, BPV, BNT, BNV 가 된다.

디코드는 스택 디코딩 방법을 사용하여 확률순으로 출력했다. 학습시 빈문자열(null) 규칙이 발견될 경우, 좌측 혹은 우측의 자소나 음절에 합쳐 규칙을 추출했다. 예를 들어,'가'가 '가아'로 정렬되어 아래와 같이 빈문자열(null)이 '아'와 정렬될 경우, 빈문자열(null)이 '아'로 바뀌는 규칙 대신 '가'가 '가아'로 바뀌는 규칙을 선택했다[4].

(활용형) 가 null
(원 형) 가 아

4. 실험 및 결과

실험은 세종 프로젝트의 문어체 형태소 품사 부착 말뭉치를 이용했다[7]. 사용한 말뭉치 크기는 총 1,250만 어절이며, 이 중에서 따옴표, 마침표, 물음표 등 많은 문장기호를 제거한 후, 약 1,192만 어절의 순수 한글 어절 부분을 사용했다. 말뭉치를 10개로 나누어 3장에서 제시한 계산 모델에 대해 각각 10배수 교차 검증(10-fold cross validation)을 하였다.

평가는 생성된 결과에 대해 정답이 포함된 비율로 계산하였으며 식 (7)과 같이 계산된다. 실험은 각 어절당 10개의 후보를 생성하고 그 정답제시율을 계산한 것이다.

$$\text{정답제시율} = \frac{\text{생성 후보중 정답이 포함된 어절수}}{\text{전체 어절수}} \quad (7)$$

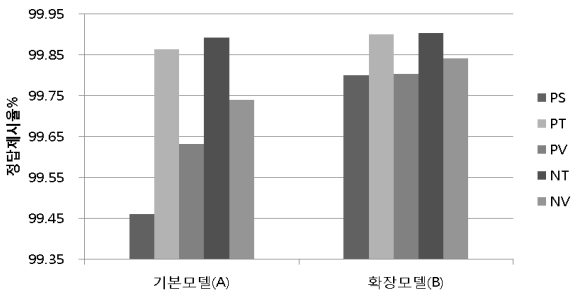
[표1]은 각 방법에 대한 실험 결과와 규칙수이다. 전체적으로 99% 이상의 성능을 보였다. 이는 실제 실험 데이터에서 음운변화가 일어나지 않는 어절이 약 80%를 차지하므로 하한 성능이 비교적 높기 때문이기도 하다. 전체적으로 가장 성능이 높은 것은 99.90%인 BPT와

BNT였다. 더 세밀하게 성능을 비교해보면 BPT에 비해 BNT가 약 0.0001% 뛰어났다. 즉, 비교적 높은 성능에서는 P와 N의 성능차이가 크지 않지만, N이 더 우수함을 보인다.

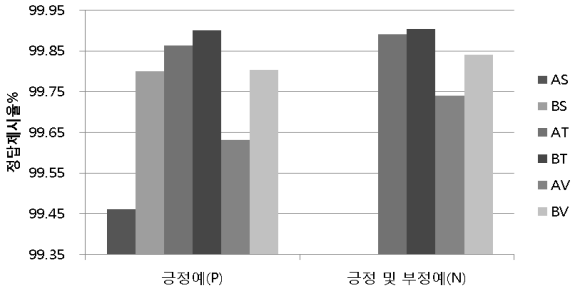
각각의 방법을 분석 요소별로 비교한 그래프가 각각 (그림1), (그림2), (그림3)이다. (그림1)은 기본모델과 확장모델을 비교한 것으로 확장모델이 더 우수함을 알 수 있다. (그림2)는 학습 자료를 비교한 것으로 부정예를 포함한 모델(N)이 대개 더 좋은 성능을 나타냄을 보여준다. 하지만, [표 1]에서 보듯이 규칙수는 더 많이 생성되어 규칙 처리에 드는 비용은 더 많음을 알 수 있다. (그림3)은 한글 처리 방법으로 3 바이트 방법이 가장 우수함을 알 수 있다. 이는 변환 정보를 3 바이트 코드가 비교적 효율적으로 나타낼 수 있음을 뜻한다.

표1. 각 방법에 대한 정답제시율 및 규칙수

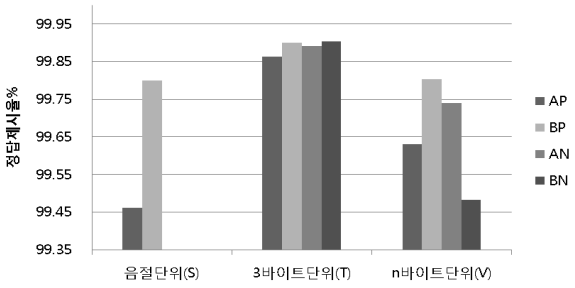
방법	정답제시율	규칙수
APS	99.46%	4123
APT	99.86%	3380
APV	99.63%	2671
ANT	99.89%	4383
ANV	99.74%	3749
BPS	99.80%	4123
BPT	99.90%	3380
BPV	99.80%	2671
BNT	99.90%	4383
BNV	99.84%	3749



(그림1) 기본모델과 확장모델의 비교



(그림2) 학습 자료 차이



(그림3) 한글 표기 차이

5. 결론

형태소 복원 확률 모델은 확률 기반 형태소 분석의 첫 단계로 다음 단계 모델의 성능 한계에 영향을 주기 때문에 전체 시스템의 성능에 중요한 요소이다. 본 논문에서는 형태소 복원 확률 모델의 계산 모델 2가지를 학습 방법과 코드체계를 변화시켜 실험하여 모델의 성능에 미치는 영향을 살펴보았다. 실험 결과, 전역확률을 고려하고, 3바이트 코드 단위를 사용하며, 부정적 학습 자료를 고려하여 확률 계산을 한 방법이 가장 우수한 결과를 보였다. 또한 형태소 복원 모델에서 계산 모델뿐만 아니라, 학습 자료의 선택 방법, 한글 코드의 선택 방법이 확률 모델의 성능에 중요한 영향을 미침을 보였다.

Acknowledgement

본 논문은 지식경제부 산업융합원천기술개발사업의 “웹 인텔리전스를 위한 웹 폭증 데이터 분석형 리스닝 플랫폼용 소셜웹 이슈 탐지-모니터링 및 예측 원천 기술 개발 과제”의 지원으로 개발된 것임.

참고문헌

[1] Manning, C., H. Schutze, Foundations of Statistical Natural Language Processing, The MIT Press, 1999.

[2] E. Charniak, Statistical Language Learning, The MIT Press, pp. 21-38, 1993.

[3] 이재성, "한국어 형태소 분석을 위한 3단계 확률 모델", 정보과학회논문지: 소프트웨어 및 응용, 제 38권 제5호, pp. 257-268. 2011.

[4] 이도길, "한국어 형태소 분석과 품사 부착을 위한 확률 모형", 고려대학교 대학원 컴퓨터학과 박사학위논문, 2005.

[5] P. F. Brown and et al, "The Mathematics of Statistical Machine Translation: Parameter Estimation", Computational Linguistics, vol.19, no.2, pp. 263-311, 1993.

[6] W. A. Gale and K. W. Church, "A Program for Aligning Sentences in Bilingual Corpora" In Using Large Corpora (ed. Armstrong, S.), pp. 75-102. The MIT Press, Cambridge, Massachusettes, London, England, 1994.

[7] 국립국어원, 21세기 세종계획 최종결과물(2009년 12월수정판), 2009.