규칙과 어절 확률을 이용한 혼합 품사 태깅 모델

황명진° 강미영 권혁철
부산대학교 컴퓨터공학과 한국어정보처리 연구실
{holgabun°, kmyoung, hckwon}@pusan.ac.kr

POS-Tagging Model Combining Rules and Word Probability

Myeong-jin Hwang°. Mi-young Kang. Hyuk-chul Kwon.
Korean Language Processing Lab, School of Electrical & Computer Engineering,
Pusan National Universty

요 약

본 논문은, 긍정적 가중치와 부정적 가중치를 통해 표현되는 규칙에 기반을 둔 품사 태깅 모델과, 형태소 unigram 정보와 어절 내의 카테고리 패턴에 기반하여 어절 확률을 추정하는 품사 태깅 모델의 장점을 취하고 단점을 보완할 수 있는 혼합 품사 태깅 모델을 제안한다. 이 혼합 모델은 먼저, 규칙에 기반한 품사 태깅을 적용한 후, 규칙이 해결하지 못한 결과에 대해서 통계적인 기법을 사용하여 품사 태깅을 한다. 본 연구는 어절 내 카테고리 패턴정보에 따른 파라미터 set과 형태소 unigram만을 이용해 어절 확률을 계산해 내므로 다른 통계기반 접근방법에서와는 달리 작은 크기의 통계사전만을 필요로 하며, 카테고리 패턴 정보를 사용함으로써 통계기반 접근 방법의 가장 큰 문제점인 data sparseness 문제 또한 줄일 수 있다는 이점이 있다. 특히, 본 논문에서 사용할 통계 모델은 어절 확률에 기반을 두고 있기 때문에 한국어의 특성을 잘 반영할 수 있다. 본 논문에서 제안한 혼합 모델은 규칙이 적용된 후에도 후보열이 둘 이상 남아 오류로 반환되었던 어절 중 24%를 개선한다.

1. 서론

자동 품사 태깅은 자연언어를 분석하는 방법의 하나로, 여러 품사를 가질 수 있는 형태소나 어절에 적절한 품사를 자동으로 할당하는 것을 말한다. 태깅의 결과는 맞춤법 검사, 자동 띄어쓰기, 자동 색인, TTS, 음성 인식 등 언어 처리 관련 분야의 기반 기술이 될 수 있다.

자동 품사 태깅과 관련한 여러 연구가 있어 왔으나 크게 규칙 기반 모델과 통계 기반 모델, 그리고 이들을 통합한 모델로 나눠볼 수 있다.

본 논문은 규칙 기반 품사 태깅 모델의 장점과 통계 기반 품사 태깅 모델의 장점을 갖추면서도 통계 기반 모델의 단점인 대용량 통계 자료 저장 문제와 자료 부족 문제를 줄임과 동시에, 한국어의 형태론적 특성을 고려한 통합 모델을 제안한다.

2장에서는 관련 연구를 소개하고, 3장에서는 본 논문에서 제안한 모델에 대해 기술하며, 4장에서는 본 논문에서 제안한 시스템의 성능을 평가하고 분석한다. 5장에서는 결론과 향후 연구를 다룬다.

2. 관련 연구

품사 태깅을 위한 모델은 크게 규칙 기반 모델과 통계 기반 모델이 있다. 규칙 기반 품사 태깅 모델은 규칙이 적용된 경우에 대해서 높은 정확도로 품사 중의성을 해결하며, 태깅 결과에 대해 설명이 가능하다는 장점이 있다. 반면 광범위한 언어 현상에 적용할 수 있는 보편적 규칙을 찾기 어렵고 품사 중의성을 해결하지 못하는 경우가 있다. 만들어진 규칙들 간의 경중을 추정하기가 쉽지 않으며, 많은 규칙을 제어하기가 쉽지 않아 견고성이 떨어진다. 규칙 구축 시에는 많은 수작업이 필요하며, 새로운 환경에 대한 적응력이 낮다는 단점도 있다.[1]

반면 통계 기반 품사 태깅 모델은 거의 모든 언어 현상에 적용할 수 있지만, 실세계 언어 현상을 충분히 대표할 수 있는 양과 질의 말뭉치가 존재하지 않으므로 data sparseness 문제에 부딪힌다.[1] 문맥 정보는 길이가 길수록 높은 정확도를 제공하지만 그보다 훨씬 많은 저장 공간을 요구하며 data sparseness 문제 또한 늘어난다.

2.1 HMM과 규칙 혼합 모델

이러한 두 모델의 장점을 취하고 단점을 보완할 수 있는 통합 품사 태깅 모델도 연구되었다. 규칙 기반 품사 태깅 모델[2]과 HMM 기반 품사 태깅 모델[3]을 각각 적용한 후, 규칙 기반 시스템을 따르게 하되 중의성이 해결되지 않는 경우에는 통계 기반 시스템이 제시한 결과를 따르는 방법을 [4]에서 시도하였다. 영어를 대상으로 한 실험이었다. 규칙은 수동 구축한 규칙과 파서의 부수적인 결과를 이용한 모델을 사용하였고, HMM은 bigram 정보를 사용하였고, Baum-Welch algorithm을 이용해 파라미터를 자동 학습하였다.

다른 실험으로는, HMM 기반 품사 태깅 모델로 중의성을 해결한 후 규칙을 이용한 후처리를 통해 태깅 정확도를 높이는 방법을 [5]에서 시도하였다. 형태소를 기반으로 한 한국어가 실험 대상이었다. 후처리에 사용한 규칙은 Brill[6]과 비슷한 방법으로 구축하였다. HMM을 이용한 태깅 결과를 정답 말뭉치와 비교하여 태깅 오류를 수정할 수 있는 변환 규칙을 추가하는 방법이었다. HMM에

사용된 파라미터는 Baum-Welch algorithm을 사용해 자동 학습하였다. 그러나 후보들 간에 경계가 같아야 하는 이 알고리즘의 특성은, 경계가 다른 여러 형태소로 분석되는 한국어 어절의 특성과 맞지 않아서 학습시 정답 말뭉치가 있어야 했다.

2.2 어절 카테고리 패턴 기반 모델

통계 기반 모델에서의 대용량 통계 자료 저장 문제와 data sparseness 문제를 해결할 방법을 [7]에서 찾을 수 있다. 한국어 자동 띄어쓰기에 대해 연구한 이 논문은, 띄어쓰기할 확률을 구하기 위해 카테고리 패턴에 기반하여 어절 확률을 계산하는 모델을 제안하였다.

이 모델은 교착어인 한국어의 특성을 고려한 어절 기반 모델이며, 어절 unigram의 출현 확률을 형태소 unigram을 통해 추정하는 방법을 제시하였다. 이로 인해 data sparseness 문제와 대용량 통계 자료 저장 문제를 크게 줄였다.

한국어에서 문장은 어절 열로 구성되고 어절은 형태소 열로 구성된다. 이 카테고리 패턴 기반 품사 태깅 모델은, 사람이 어절 하나만으로도 그 품사 조합이 무엇인지 알아차린다는 기본 아이디어에서 시작하였다. 즉, 어절의 출현은 그것을 구성하는 각 형태소의 출현 조합으로 미루어 짐작할 수 있다는 것이다.

한국어는 교착어이므로 문법적 관계가 조사나 어미와 같은 형태소에 의해 결정된다. 따라서 문장 내 어절의 순서가 자유롭다. 다만, "*관형형어미 다음에 의존명사나 명사가 온다*"든지 "*타동사 앞에 목적격조사가 와야 한다*"와 같은 부분적인 통사적 제약은 존재한다. 반면 어절 안에서 형태소 사이의 순서는 제약이 강하다. 예를 들어 형식형태소는 홀로 쓰일 수 없고 반듯이 실질형태소와 결합함으로써만 문장 내에 있을 수 있다.

다시 말해서 한국어에서는 어절 내 형태소 간 제약은 아주 강하고 어절 간 제약은 비교적 약하다. 여기서 제약이 강하다는 것은 더욱 견고한 전이 정보를 얻을 수 있다는 말로 풀이된다. 이는 사람이 어절 하나만 보고도, 대부분 품사 조합이 무엇인지 알아차리는 것에 대한 이론적 뒷받침일 수 있다.

이와 같은 한국어의 특성을 반영하여 구축한 카테고리 패턴 기반 모델의 특징은 다음과 같이 요약할 수 있다.

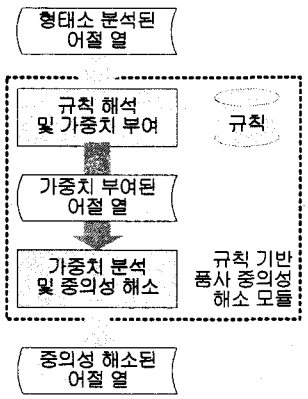
- 형태소 unigram, 카테고리 패턴과 파라미터 등만을 사용하므로 형태소 bigram 이상을 사용하는 여타 통계 기반 모델보다 저장해야 할 통계 자료의 양이 적다.
- 형태소 unigram을 사용하므로 자료 부족 문제가 형태소 unigram 모델만큼 적다.
- 어절 unigram 확률을 계산할 수 있는 모델이므로 규칙 기반 태깅 시스템이 해결하지 못한 중의성을 해결할 때 유용하다.

3. 규칙과 어절 확률을 이용한 혼합 품사 태깅 모델

본 절에서는 규칙 기반 모델과 통계 기반 모델의 장점을 취하고 단점을 보완할 수 있는 혼합 품사 태깅 모델을 제안한다. 이 혼합 모델은 수동 구축한 규칙 기반 모델과 어절 카테고리 패턴 기반 모델을 적용한 두 모듈로 구성된다.

3.1 규칙 기반 품사 중의성 해소 모듈

규칙 기반 품사 중의성 해소 모듈의 구조 및 처리 과정은 그림?과 같다.



[그림?] 규칙 기반 품사 중의성 해소 모듈 구조

입력 데이터는, 형태소 분석된 어절 열(각 어절은 다수의 분석 후보를 가짐)이다. 출력 데이터는, 품사 중의성이 해소된 어절 열(각 어절은 대부분 하나의 분석 후보만 가짐)이다. 규칙 모델의 특성상 일부 어절은 중의성이 해소되지 않을 수 있다.

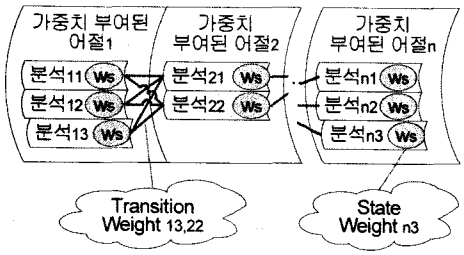
처리 과정은 (1) 규칙 해석 및 가중치 부여와 (2) 가중치 분석 및 중의성 해소로 이루어진다. (1)에서는 규칙 해석을 통해 입력 데이터에 적용 가능한 규칙을 찾고, 규칙에 정의된 가중치를 해당 부분에 할당한다. (2)에서는, (1)에서 부여한 가중치를 각 분석 후보별로 종합하고, 후보별 우선순위를 정한다.

규칙은 수작업으로 구축하였다.

규칙은 2가지 종류로 나누었는데 (1) 어절 간 규칙과 (2) 어절 내 규칙이다. (1)은 앞, 뒤 어절의 형태소나 품사 간에 미치는 영향, 또는 문장 내에서 어절 위치 제약을 분석하여 도출하고, (2)는 어절 내에서의 형태소 간 제약을 분석하여 도출한다.

각 규칙은 가중치를 가진다. 가중치는 긍정 가중치와 부정 가중치가 있으며, 규칙의 중요도에 따라 가중치의 크기가 다르다. 가중치의 크기는 규칙 간 우선순위를 정하기 위해 사용한다. 규칙이 상충하여 정답 품사를 정하지 못할 때가 있기 때문이다. 따라서 본 논문에서는 한국어의 형태 통사론적 특징을 바탕으로 가중치를 5단계로 구분해 놓았다. 규칙 간 중요도를 세세하게 따지는 작업은 시간과 비용이 많이 들고 쉬운 작업이 아니다.

가중치는 두 가지 종류가 있다. 어절의 한 분석 후보에 부여하는 상태 가중치(State Weight)와, 어절 간에 부여하는 전이 가중치(Transition Weight)가 그것이다. 상태 가중치는 좌우 문맥에 상관없이 독립적으로 가중치를 부여할 수 있는 경우이고, 전이 가중치는 좌 문맥 또는 우 문맥의 영향을 고려하여 부여하는 가중치이다. 그림?는 가중치 종류와 그 가중치가 부여된 상태를 보여준다.



[그림 ?] 규칙 기반 모델에서 사용하는 가중치

3.2 어절 카테고리 패턴 기반 품사 중의성 해소 모듈

어절 카테고리 패턴 기반 모듈은 [7]에서, 한국어 자동 띄어쓰기를 위해 제안한 어절 카테고리 확률을 구하는 기법을 사용한다.

이 논문[7]에서는 어절 unigram 출현 확률을, 그것을 구성하는 형태소 unigram 출현 확률과 미리 학습한 카테고리 패턴 파라미터를 이용하여 계산하는 방법을 제시하였다. 아래에 어절 '날다+는'의 확률을 동사 '날다'와 어미 '는' 각각의 빈도를 이용해 계산하는 예가 있다.

$$P(\text{날다+는}) = P(\text{날다})^{cp1} * P(\text{는})^{cp2}$$

P(날다) : 학습 말뭉치에서 나타난 동사 '날다'의 빈도를 전체 형태소 빈도로 나누어 구한 확률
P(는) : 학습 말뭉치에서 나타난 어미 '는'의 빈도를 전체 형태소 빈도로 나누어 구한 확률
cp1, cp2 : 학습을 통하여 미리 구한 것으로서, 카테고리 패턴 '동사+어미'에서[1] 동사와 어미의 파워를 조절하는 파라미터

파라미터 학습은 카테고리 패턴 모델에서 카테고리 패턴의 파워를 조절하는 역할을 하는 cp1, cp2와 같은 파라미터를 찾는 과정이다. 한 카테고리 패턴에 대해 한 set의 파라미터를 구해야 한다. 카테고리 패턴 '동사+어미'의 파라미터를 구하는 식의 일부가 아래에 있다. 카테고리 패턴 '동사+어미'에 속하는 모든 어절에 대해 Error를 최소로 할 수 있는 cp1과 cp2를 구하면 된다. 이 논문[7]에서는 Simulated-Annealing 기법을 이용하여 구했다.

$$Error = P(\text{날다+는}) - P(\text{날다})^{cp1} * P(\text{는})^{cp2}$$

P(날다+는) : 학습 말뭉치에서 나타난 '동사+어미'로 분석되는 어절 '날다+는'의 빈도를 전체 어절 수로 나누어 구한 확률
P(날다) : 학습 말뭉치에서 나타난 동사 '날다'의 빈도를 전체 형태소 빈도로 나누어 구한 확률
P(는) : 학습 말뭉치에서 나타난 어미 '는'의 빈도를 전체 형태소 빈도로 나누어 구한 확률
cp1, cp2 : 카테고리 패턴 '동사+어미'에서 동사와 어미의 파워를 조절하는 파라미터로 탐색 알고리즘 등을 이용해 구한다.

어절 카테고리 패턴 기반 모델의 가장 두드러진 특징은 어절 단위의 확률을 계산할 수 있다는 것이다. 이전 연구들은 한국어의 형태론적 특징을 반영할 적합한 모델이 없어서, 어절 단위 태깅 모델로는 좋은 성능을 도출하지 못하였고 대부분 형태소 기반 모델을 사용하였다.

어절 카테고리 패턴 기반 모델의 방법으로 어절 unigram 확률을 계산하면 학습 말뭉치에서는 출현하지 않은 어절의 확률도 구할 수 있다는 장점이 있다. 따라서 data sparseness 문제도 자연스럽게 완화되고, 이를 해결하기 위한 smoothing 기법을 따로 두지 않아도 된다.

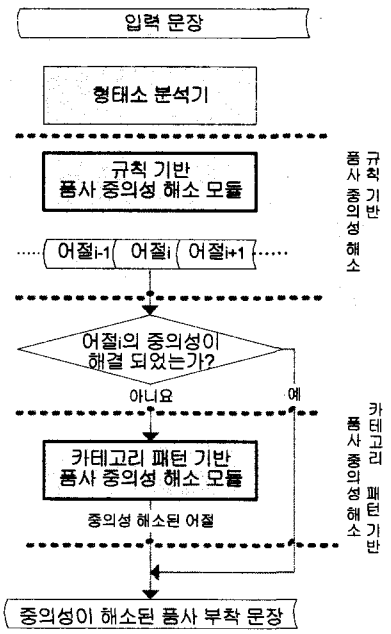
어절 카테고리 패턴 기반 모델은 또한 대용량 통계 자료 저장 문제에도 강함을 확인할 수 있다. 카테고리별 형태소 unigram과 카테고리 패턴별 파라미터만 저장하면 되기 때문이다.

3.3 규칙과 어절 확률을 이용한 혼합 품사 태깅 모델

혼합 모델의 전체 구조및 처리 과정은 그림?과 같다.

처리 과정은 다음과 같다. (1) 규칙 기반 품사 중의성 해소 모듈을 이용해 형태소 분석된 문장에 가중치를 부여하여 중의성을 해소한다. (2) 어절별로 중의성 해결 여부를 판단한다. (3) 중의성이 해결되지 않은 어절은 카테고리 패턴 기반 중의성 해소 모듈에서 어절 확률을 비교하여 해결한다.

1) + : 형태소 경계



[그림 ?] 혼합 모델 구조

수동 구축한 규칙 기반 품사 태깅 모듈은 규칙이 적용된 경우 높은 정확도를 가지지만 단점 또한 가지고 있다. 품사 중의성을 해결하지 못하거나 규칙 간의 경중을 추정하기가 어려운 경우가 그러하다. 예를 들어 아래의 예제 (1)에서는 관형격 조사 '의' 뒤에 명사가 오면 가중치가 부여됨에 따라 후보 품사열 1을 태깅할 수 있지만 예제 (2)에서는 "어[연결형어미] # 오다[보조용언]"[2] 형태소 쌍과 "온[관형사] # 가족[명사]" 형태소 쌍에 동일한 가중치가 부여된다. (연결형어미와 보조용언 간의 통사 결속력만큼이나 관형사와 명사 간의 결속력 또한 강하다.) 따라서 규칙에 의해 하나의 후보 품사열을 선택하기 힘들게 된다.

(1) *악의 축이라고 말했다.*

- 후보 품사열 1
악+의 # 축+이다+라고 # 말하다+ㅆ+다
명사+관형격조사 # 명사+지정사+인용형어미 ...

- 후보 품사열 2
악+의 # 축이다+라고 # 말하다+ㅆ+다
명사+관형격조사 # 동사+인용형어미 ...

:

(2) *여가 시간 증가에 맞춰 온 가족이 참여*

- 후보 품사열 1
...+에 # 맞추다+어 # 오다+ㄴ # 가족+이 # 참여
... 동사+연결형어미 # 보조용언+관형형어미 ...

- 후보 품사열 2
...+에 # 맞추다+어 # 온 # 가족+이 # 참여
... 관형사 # 명사+주격조사 ...

:

4. 실험 및 결과

4.1 실험 데이터

통계기반 태깅 모듈에 도입한 카테고리 패턴 기반 품사 태깅 모델은 신문기사와 TV뉴스 방송 원고를 학습 말뭉치로 이용했고 구성은 다음과 같다.

표 1 학습말뭉치

말뭉치	어절 수
A신문사 2년치	18,967,465
B신문사 1년치	9,400,031
C방송사 TV 뉴스 원고 1년치	2,445,881
D방송사 TV 뉴스 원고 2년치	2,803,411
전체	33,616,788

정확도 평가를 위한 답안 말뭉치는 신문기사에서 무작위 추출한 외부 데이터를 수작업으로 품사 부착하였고 구성은 다음과 같다.

표 2 성능평가 데이터

정답 말뭉치	어절 수
신문기사	32,503어절

정확도 평가에는 다음과 같은 품사 집합을 사용하였다. 문장 부호 및 특수 기호들은 '기호' 집합에 포함해 처리하였다.

표 3 품사 집합

명사	동사	종결형어미	동사화접미사
고유명사	형용사	연결형어미	형용사화접미사
의존명사	보조용언	인용형어미	복수접미사
수의존명사	주격보격조사	명사형어미	관형접미사
인칭대명사	목적격조사	관형형어미	지정사
지시대명사	관형격조사	선어말어미	외국어
양수사	부사격조사	일반접두사	한자
서수사	호격조사	수접두사	도량형단위
부사	인용격조사	일반접미사	화폐단위
감탄사	보조사	수접미사	기호

2) # : 어절 경계

4.2 실험 및 평가 방법

혼합모델에서 사용한 규칙 기반 품사 태깅 시스템은 60여 개의 규칙을 사용하여 97.20%의 성능을 보인다. 또한, 통계기반 태깅 모듈에서 사용한 카테고리 패턴 기반 품사 태깅 기법은, 카테고리 패턴을 69개만 사용하여도 96.04%의 정확도를 가지지만 최적의 성능을 얻기 위해 모든 카테고리 패턴(767개)을 사용하면 96.08%의 성능을 보인다.

모든 평가는 표 2에서 제시한 성능평가 데이터를 이용하였다. 규칙 기반 품사 태깅 모델만을 이용한 평가, 카테고리 패턴 기반 모델만을 이용한 평가, 그리고 이들을 혼합한 모델을 이용한 평가를 각각 하였다.

4.3 실험 결과

실험 결과는 다음과 같다.

표 4 실험 결과

시스템	정확도1
규칙 기반 모델	97.20%
카테고리 패턴 기반 모델	96.08%
혼합 모델	97.53%

품사 부착 말뭉치 구축 시스템과 카테고리 패턴 기반 모델의 혼합을 통하여 0.33%의 정확도 향상이 있음을 확인하였다. 규칙이 적용된 후에도 후보열이 둘 이상 남아 오류로 반환되었던 어절 중 24%를 개선한다.

5. 결론

본 논문은 규칙 기반 품사 태깅 모델과 통계 기반 품사 태깅 모델의 장점을 취하고 단점을 보완할 수 있는 혼합 품사 태깅 모델을 제안하였다. 카테고리 패턴 기반 중의성 해소 모듈에서 타 통계기반 시스템에 비해 월등히 적은 양의 통계 정보를 사용하고, 어절 내 형태소들 간의 결합 패턴을 이용하는 카테고리 패턴기반 모델을 도입함으로써 어절을 인식하여 한국어 형태론적 제약규칙을 최대한 고려하는 태깅을 수행함과 동시에 대용량 통계 자료를 저장해야 하는 문제 등을 해결하였다.

앞으로 카테고리 패턴 기반 품사 태깅 모델의 장점을 유지하면서 어절 간 정보까지 얻을 수 있는 모델을 개발한다면 더욱 견고한 모델을 개발할 수 있을 것이라 판단된다.

<Acknowledgement>
논문은 국가지정연구실사업(과제명: 언어 중심의 지능적 정보처리를 위한 단계적 우리말 분석기술의 개발(M1-0412-00-0028-04-J00-00-014-00))의 지원을 받아 이루어진 것임.

6. 참고문헌

[1] 임해창, 임희석, 이상주, 김진동. "자연어 처리를 위한 품사 태깅 시스템의 고찰", 정보과학회지, 14권, 7호, 36-57 페이지, 1996.

[2] Atro Voutilainen. "A Syntax-Based Part of Speech Analyser.". Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics, pages 157-164, 1995.

[3] Cutting, Doug, Julian Kupiec, Jan Pederson, and Penelope Sibun. "A practical part of speech tagger", In Third ACL Conference on Applied Natural Language Processing, pages 133-140, 1992.

[4] Pasi Tapanainen, Atro Voutilainen. "Tagging accurately - Don't guess if you know", Proceedings of Fourth ACL Conference on Applied Natural Language Processing, pages 47-52, 1994.

[5] Geunbae Lee, Jeongwon Cha, Jong-Hyeok Lee, "Hybrid POS tagging with generalized unknown-word handling", Proceedings of the 2nd international workshop on information retrieval with Asian languages, pages 43-50, 1997.

[6] Eric Brill, "A Simple rule-Based part-of-speech tagger", Proc of the third Conf. on Applied NLP, pages 153-155, 1992.

[7] 강미영, 정성원, 권혁철, "어절 내의 형태소 범주 패턴에 기반한 통계적 자동 띄어쓰기 시스템", 정보과학회지-소프트웨어 및 응용. 2006년 11월 게재 예정.