한국어 품사 및 동형이의어 태깅을 위한 마르코프 모델과 은닉 마르코프 모델의 비교

신준철○, 옥철영
울산대학교 컴퓨터정보통신공학
ducksjc@nate.com, okcy@ulsan.ac.kr

Comparison between Markov Model and Hidden Markov Model for Korean Part-of-Speech and Homograph Tagging

Joon-Choul Shin○, Cheol-Young Ock
Dept. of Computer Engineering & Information Technology, Ulsan University

요 약

한국어 어절은 많은 동형이의어를 가지고 있기 때문에 주변 어절(또는 문맥)을 보지 않으면 중의성을 해결하기 어렵다. 이런 중의성을 해결하기 위해서 주변 어절 정보를 입력받아 통계적으로 의미를 선택하는 기계학습 알고리즘들이 많이 연구되었으며, 그 중에서 특히 은닉 마르코프 모델을 활용한 연구가 높은 성과를 거두었다. 일반적으로 마르코프 모델만을 기반으로 알고리즘을 구성할 경우 은닉 마르코프 모델 보다는 단순하기 때문에 빠르게 작동하지만 정확률이 낮다. 본 논문은 마르코프 모델을 기반으로 하면서, 부분적으로 은닉 마르코프 모델을 혼합한 알고리즘을 제안한다. 실험 결과 속도는 마르코프 모델과 유사하며, 정확률은 은닉 마르코프 모델에 근접한 것으로 나타났다.

주제어: 마르코프 모델, 은닉 마르코프 모델, 형태소, 형태소분석기, 의미 중의성, 동형이의어

1. 서론

하나의 어절이 가지는 의미는 종종 여러 가지일 수 있으며, 대체로 주변 문맥을 통해 한 가지 뜻으로 해석된다. 이것을 의미중의성이라고 하며, 컴퓨터가 이런 중의성을 자동으로 해결할 수 있다면 자연어처리 시스템이 획기적으로 발전할 것이다.

중의성을 가진 어절에 품사와 의미번호를 표시하여 의미를 명확하게 하는 작업을 태깅이라고 한다. 태깅을 위한 품사 태그는 대표적으로 세종태그셋이 있다. 의미번호는 표준국어대사전의 것이 널리 사용되고 있다. 예를 들어 '나는'은 다음과 같이 태깅될 수 있다.

●나__03/NP+는/JX
●나__01/VV+는/ETM
●날__01/VV+는/ETM

예와 같이 중의성을 가진 어절은 다양한 품사와 의미로 해석될 수 있으며, 이는 여러 가지의 분석후보들로 표현된다. 일반적인 자동태깅 절차는 하나의 어절을 분석해서 분석후보들을 만든 다음에, 주변 문맥을 통해 적절한 후보 하나를 선택하는 것이다.

하나의 어절에서 적절한 분석후보들을 만드는 것은 오래전부터 많이 연구되어왔다. 초기에는 규칙기반 알고리즘들이 연구되었으며, 세종말뭉치가 정립된 이후에는 기계학습을 활용한 방법들이 연구되고 있다.

적절한 후보를 선택하는 방법은 세종말뭉치로 기계학습을 하고, 인접한 어절이나 공기어를 활용하여 통계적으로 처리하는 방법들이 연구되어왔다. 그러나 대부분 품사만 결정하거나, 정확률이 낮거나, 이 모든 것을 해결하더라도 속도가 느린 단점을 가지고 있다.

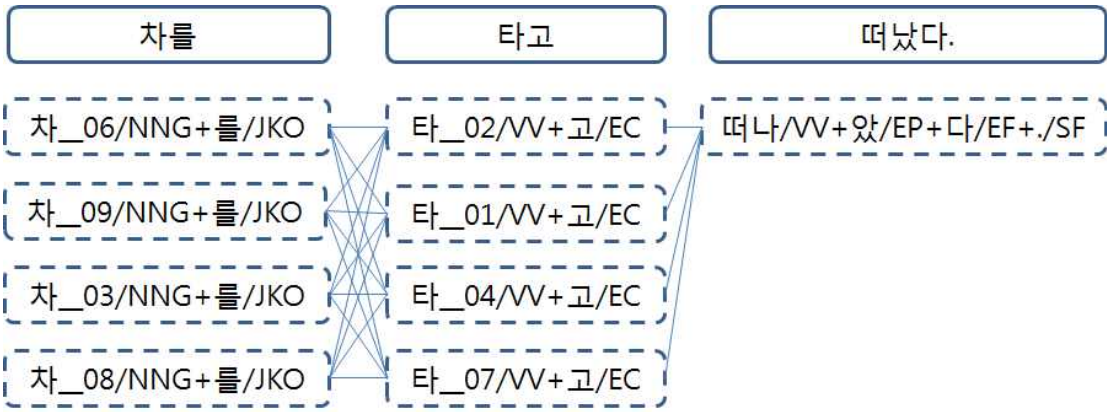
본 논문은 적절한 후보를 선택하는 새로운 알고리즘을 제안한다. 이 알고리즘은 세종말뭉치를 최대한 활용하며, 정확률과 재현율을 최대한 보존하면서 속도를 더 빠르게 하는 것에 목표를 둔다.

2. 관련 연구

자동태깅의 첫 단계는 하나의 어절에서 분석후보들을 생성하는 것이다. 초창기에 어절에서 형태소 원형을 하나씩 복원하는 규칙기반 방법이 주로 연구되었다. 대용량 말뭉치가 정립되기 전에는 기계학습 기반의 분석방법을 적용할 수 없었기 때문이다. 대표적인 연구로 [1]은 다양한 한국어의 불규칙변화를 전산화하여 규칙기반의 분석 방법으로 적용하기 위해 다층 형태론을 제안하였다.

대용량 말뭉치가 정립되고 부터는 기계학습 방법을 활용한 연구가 진행되고 있다. [2]는 대용량 말뭉치에서 기분석 사전과 "부분 어절 기분석 사전"을 만들고, 이를 사용하여 어절을 분석하고 형태소 분석 후보들을 생성하는 알고리즘을 제안한다. 세종말뭉치로 실험한 결과 99.05%의 형태소 분석 재현율(후보 중에 정답이 있을 확률)을 보였으며, 형태소분석을 위한 조합형코드로의 변환 과정, 형태소 분리 및 복원 등의 복잡한 과정이 필요없으며 띄어쓰기 오류에도 안정적인 특징이 있다.

하나의 어절을 분석하여 나타나는 분석 후보들 중에서 가장 적절한 하나의 후보를 자동으로 선택하기 위한 연



[그림 1 Bigram HMM의 예]

구들이 진행되어왔다. 세종말뭉치가 정립되지 않았을 때에는 내부적으로 자체 제작한 소량의 말뭉치를 사용하였다. 이런 경우에는 정확률이 낮았다[3].

소량의 말뭉치에서 은닉 마르코프 모델(Hidden Markov model : HMM)을 기계학습한 연구가 있다[4, 5, 6]. 말뭉치의 양이 적기 때문에 상대적으로 적은 품사 태그를 사용하여 정확률을 높였다. 비록 소량의 말뭉치였지만 초기에 HMM의 유용성을 실험해본 의미 있는 연구들이였다.

[7]은 KAIST에서 구축한 '대한민국 국어정보 베이스'의 품사 부착된 말뭉치를 사용하여 HMM을 학습하였다. 이 방법에서는 수작업으로 구축되는 말뭉치의 특성상 나타날 수밖에 없는 오류에 대해서 저신뢰도 구간 검사를 통해 학습 말뭉치의 신뢰도를 높여 태깅 성능의 향상을 모색하였다.

[8]은 비교적 최근의 HMM 연구로, 약 천만어절의 세종말뭉치를 사용하였다. 우선 어절별로 분석 후보를 생성하기 위해서 CKMA[2]를 사용하였으며, 하나의 후보를 선택하기 위해서 여러 가지의 HMM을 단계별로 적용하였다. 우선 정확률이 높은 모델을 적용해보고, 재현에 실패한 경우(학습된 적이 없는 경우다.) 정확률이 낮더라도 재현율이 높은 모델을 적용하였다. 동형이의어까지 태깅하면서 정확률이 매우 높은 결과를 보였다.

3. 알고리즘

분석 후보에서 의미와 품사가 적합한 하나의 후보를 선택하기 위해서 연구된 방법 중 가장 알려진 것은 HMM이다. HMM은 은닉층(분석 후보, 형태소 원형) 정보간의 전이확률로 최적의 전이열을 찾는 모델이다. HMM 보다 단순한 모델인 마르코프 모델(Markov model : MM)은 오직 관찰되는 정보만을 통해 현재 또는 미래의 은닉 정보를 추측한다.

3.1 은닉 마르코프 모델

HMM은 은닉층을 사용하는 모델이며, 태깅에서는 인접한 어절들의 분석 후보들이 대표적인 은닉층이다. 사람이 수동으로 태깅하지 않은 상태에서는 인접한 어절들이 중의성을 가지면 오직 여러 가지의 분석 후보들만 알 수 있다. 이런 특성 때문에 HMM은 인접한 어절들의 분석 후보가 다양할수록 연산량이 많아진다.

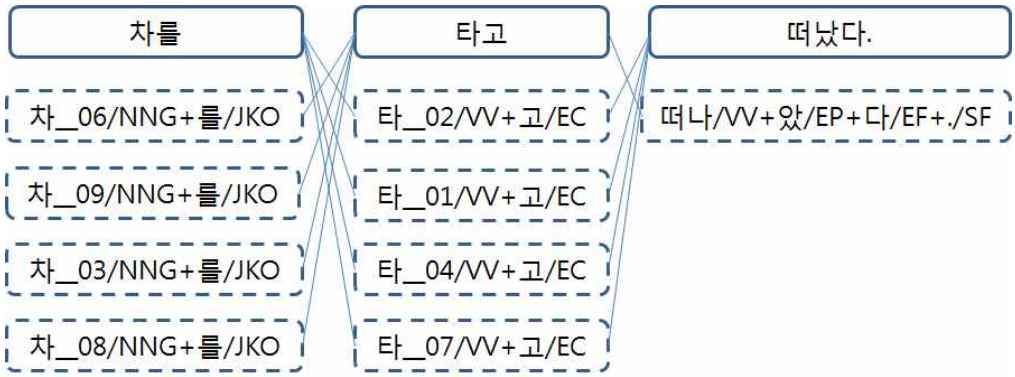
직접적으로 인접한 어절만을 고려하는 HMM을 바이그램(Bigram) HMM이라고 하며, [그림 1]은 Bigram HMM으로 분석 중인 예이다. 3개의 어절로 구성된 문장은 각 어절별로 분석 후보들이 있으며, HMM은 인접 어절의 후보와 자기 어절의 후보 간의 전이확률을 계산하고, 전이확률이 가장 높은 최적열을 Viterbi 알고리즘으로 찾는다.

Bigram HMM의 전이확률은 좌측 어절의 특정 후보와 우측 어절의 특정 후보가 동시에 출현할 확률을 표현한다. [그림 1]안에서 예를 들자면, 첫 번째 전이확률은 좌측 어절이 '차를__06/NNG+를/JKO'이고 우측 어절이 '타__02/VV+고/EC'일 확률을 의미한다. 그러나 실제로 후보 내용 전체를 모델에서 사용하기에는 학습데이터의 용량과 재현율에서 비효율적인 문제가 발생하기 때문에 [8]에서는 인접 어절의 가장 가까운 형태소만을 모델에서 사용하는 모델을 제안한다. 예를 들어 '차를'의 후보들을 처리하는 과정에서 첫 번째로 계산되는 전이확률은 '차__06/NNG+를/JKO'와 '타__02/VV*'가 동시에 출현할 확률이다. '*'는 어절의 나머지 부분이 무엇이 되던 동일한 것으로 간주한다는 의미다.

태깅 시점에서는 좌측 어절에는 후보가 4가지이기 때문에, 정확히 어떤 후보가 정답인지 알 수 없다. 따라서 HMM에서는 모든 가능한 후보 간의 전이확률을 계산해야 한다. [그림 1]에서 차를'과 '타고'의 후보가 각각 4개씩이므로 두 어절 사이에서 구해야하는 전이확률은 총 16가지이고, '타고'와 '떠났다.' 사이의 것을 합치면 총 20개이다.

[그림 1]의 모습은 일부 후보를 생략한 것이며, 실제로는 더 많은 후보가 존재(기분석 사전) 하거나, 생성될 수 있기 때문에 더 많은 전이확률 정보가 필요하다. 각 전이확률을 학습된 정보에서 찾는 과정과, 모든 전이확률들로부터 최적열을 구하는 과정에서 HMM은 MM보다 컴퓨터 자원을 많이 사용해야한다.

3.2 마르코프 모델



[그림 2 Bigram MM의 예]

MM은 오직 관찰된 정보만을 이용하여 알고자 하는 은닉된 정보를 추측한다. 태깅에서는 인접한 어절들과 현재 어절의 표층형이 관찰된 정보고, 현재 어절의 분석 내용이 추측하고자 하는 은닉 정보다.

MM 중에서 직접적으로 인접한 어절만을 고려하면 Bigram MM이라고 하며, [그림 2]는 Bigram MM으로 분석 중인 예이다. 각 어절의 후보는 인접 어절의 표층형과의 전이 확률을 고려한다. '차를'의 후보 4개는 인접 어절인 '타고'와의 전이확률만을 계산하고, '타고'의 후보 4개는 '차를'과의 전이확률을, 그리고 '떠났다.'와의 전이확률 4개를 계산한다. 따라서 그 두 어절 사이에 필요한 전이확률 계산은 총 8개이고, '타고'와 '떠났다' 사이의 것 4개를 합치면 총 12개로 [그림 1]의 것보다 적다.

[8]에서 제안하는 모델과 마찬가지로, 인접한 어절의 표층형 전체를 전이모델에 포함하는 것은 학습용량과 재현율 측면에서 비효율적이기 때문에 일부만을 포함하는 것이 좋다. 1개의 어절만 포함하기엔 인접한 어절을 너무 모호하게 표현하고, 3어절 이상을 포함하면 어절 내용을 이미 거의 다 포함하기 때문에 본 논문에서는 인접한 어절들에 서로 가장 가까운 2음절씩을 학습하는 것을 제안한다.

예를 들어 '타고'와 '떠났다.'사이에서 첫 번째로 구하는 전이확률은 '*타고'와 '떠났*'을 통해서 구한다. 말뭉치에서 위의 각각 2음절을 모두 포함하는 인접한 쌍 어절을 찾고, 해당 어절에 태깅된 형태소가 분석 후보에 포함되는지 확인하는 방식이다. 이 모델을 Bi-Syllable of End and First(BIS-EF)이라고 한다.

3.3 MM과 음운 변동

태깅에서 MM모델이 가지는 문제점은 표층형을 중심으로 전이모델이 구성되기 때문에 동사의 음운변동에 취약하다는 것이다. 예를 들어서 "타고 떠났다."와 "타고 떠난다."는 표층형이 다르기 때문에 구분된다. 둘 다 우측 어절은 '떠나/VV'를 사용하지만 2번째 음절이 '났'과 '난'으로 다르기 때문에 한쪽 문장을 학습하면 다른 한쪽 문장이 입력으로 들어올 때 학습된 내용이 적용되지 못한다. 이런 특성은 HMM을 사용한 방법[8]보다 재현율이 낮아질 요소다.

때문에 본 논문에서 제안하는 BIS-EF모델에서는 일반적인 재현에 실패할 경우에 표층형 대신 형태소 원형을 사용하는 단계를 전이모델에 추가한다. 이 경우 전이확률의 가짓수는 Bigram HMM과 마찬가지로 두 어절의 분석 후보 수의 곱만큼 많아지게 되지만, 오직 두 어절 사이에서 전이확률만 계산하여 결정하기 때문에 최적열을 연산하는 Viterbi과정은 없다.

3.4 MM과 부사절

Bigram을 기반으로 하는 모델은 인접어절을 기준으로 의미를 결정하기 때문에 부사는 악영향을 줄 수 있다. 예를 들어서 "차를 빨리 마셨다."를 처리한다면, '차를'의 의미를 결정지어주는 어절은 '마셨다.' 이지만 부사 '빨리'에 의해 인접하지 못하게 된다. 이런 경우에 BIS-EF는 부사와 함께 그 다음 어절도 마치 인접한 것과 마찬가지로 취급하여 전이확률을 계산한다. HMM에 비하여 이런 튜닝은 MM에서 더 쉽게 구현이 가능하다.

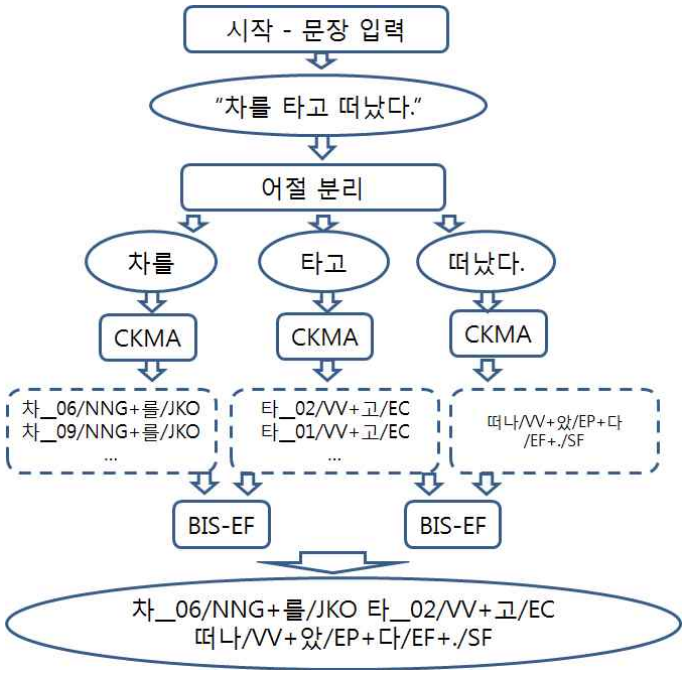
4. 전체 시스템 구성

전체 시스템은 [그림 3]과 같이, 먼저 문장을 입력받으면서 시작한다. 문장은 공백을 기준으로 어절들로 분리되고, 각 어절은 CKMA[2]모듈에 입력되면 분석 후보들이 출력된다. 인접한 어절 쌍 단위로 BIS-EF모듈에 입력되면 각 후보별로 확률점수가 계산되고, 어절별로 가장 높은 점수를 가진 후보가 선택되어 최종 결과로 출력된다.

5. 실험 결과

본 논문의 실험을 위해 사용된 CKMA는 [2]이후 계속된 개선 작업을 통해 일부분이 변경된 것으로 정확률이 조금 더 높다. 때문에 HMM과의 엄격한 비교를 위해 [8]의 실험을 수정된 CKMA와 동일한 말뭉치로 다시 시행하였다.

말뭉치는 일부 오류가 수정된 세종말뭉치를 사용하였



[그림 3 전체 시스템]

으며, 약 1천만 어절을 포함한다. 문장 단위로 잘라서 10문장 중에 1문장씩을 제외하여 테스트셋을 구성하였고, 나머지 9문장씩을 학습에 사용하였다.

[표 1 HMM과 BIS-EF 결과 비교]

	정확률(%)	시간(초)
HMM	96.5393	21.328
BIS-EF	96.3361	9.999

[표 1]은 HMM과 BIS-EF를 비교 실험한 결과다. 둘 다 완전히 동일한 말뭉치와 테스트셋을 사용했으며, 동일한 버전의 CKMA를 사용하여 분석 후보를 생성하였다. 시간 측정은 CKMA에 소모된 시간을 제외한 것으로 순수하게 HMM 또는 BIS-EF에서 소모된 시간만 측정한 것이다.

[표 1]의 실험 결과에 따르면 본 논문이 제안하는 BIS-EF는 HMM에 비하여 약 0.2% 낮은 정확률을 보이지만 속도가 약 2.1배 빠른 것으로 나타난다.

5. 결론

품사 및 동형이의어가 태깅된 형태소 분석 후보에서 적절한 후보 1개를 선택하는 방법으로 HMM은 이미 많이 연구된 모델이다. HMM으로도 높은 정확률을 보이고 있으나, 본 논문에서는 속도를 더욱 개선하기 위해 더 단순한 MM모델에 기반을 둔 BIS-EF를 제안한다.

BIS-EF는 기존 HMM에 비하여 정확률이 0.2% 낮지만, 속도는 2배 이상 빠른 결과를 보여주었다. 따라서 정확률 보다 속도를 필요로 하는 곳에서 더 적합할 것이다.

또한 BIS-EF는 한 가지 전이모델만을 사용하고 있어 개선의 여지가 많이 남아있다. 좀 더 다양한 전이모델을 추가해서 단계적으로 또는 동시에 적용한다면 정확률을 더욱 향상시킬 수 있을 것으로 예상된다.

참고문헌

[1] 강승식, "다층 형태론과 한국어 형태소 분석 모델", 제6회 한글 및 한국어 정보처리 학술발표 논문집, pp.140-145, 1994.

[2] 신준철, 옥철영, “기분석 부분 어절 사전을 활용한 한국어 형태소 분석기”, 정보과학회논문지 : 소프트웨어 및 응용, 제39권, 제5호, pp.415-424, 2012.5.

[3] 김영훈, "한국어에 적합한 효율적인 품사 태깅", 한국콘텐츠학회논문지, 제2권, 제2호, pp.98-102, 2002.6.

[4] 신중호, 한영석, 박영찬, 최기선, "어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사태깅", 한글 및 한국어정보처리 학술대회, p389-394, 1994.10.

[5] 김재훈, 임철수, 서정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사의 태깅", 정보과학회논문지, 제22권, 제1호, pp.136-146, 1995.1.

[6] 강유환, 서영훈, "어절간 주품사 정보와 제약 규칙을 이용한 한국어 품사 태깅 시스템", 제11회 한글 및 한국어 정보처리 학술대회, pp.433-437, 1999.10.

[7] 설용수, 김동주, 김규상, 김한우, "말뭉치 오류를 고려한 HMM 한국어 품사 태깅 시스템", 한국컴퓨터정보학회지, 제15권, 제1호, pp.117-124, 2007.

[8] 신준철, 옥철영, "한국어 품사 및 동형이의어 태깅을 위한 단계별 전이모델", 정보과학회논문지 : 소프트웨어 및 응용, 제39권, 제11호, pp.889-901, 2012.