규칙과 비감독 학습 기반 통계정보를 이용한 품사 태깅 시스템

이동훈°, 강미영, 황명진, 권혁철
부산대학교 컴퓨터공학과 한국어정보처리 연구실
{huni77°, kmyoung, holgabun, hckwon}@pusan.ac.kr

Part-of-Speech Tagging System Using Rules/Statistics Extracted by Unsupervised Learning

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

요 약

본 논문은 규칙 기반 방법과 통계 기반 방법을 동시에 사용함으로써 두 가지 방법의 장단점을 상호 보완한다. 한 문장에 대한 최적의 품사열은 HMM을 기반으로 Viterbi Algorithm을 사용하여 선택한다. 이때 파라미터 값은 규칙에 의한 가중치 값과 통계 정보를 사용한다. 최소한의 일반규칙을 사용하여 구축한 규칙의 적용에 따라 가중치 값을 구하며 규칙을 적용받지 못하는 경우는 비감독학습으로 추출한 통계정보에 기반을 둔 가중치 값을 이용하여 파라미터 값을 구한다. 이러한 기본 모델을 여러 회 반복하여 학습함으로써 최적의 통계기반 가중치를 구한다. 규칙과 비감독 학습으로 추출한 통계정보를 이용한 본 품사 태깅 시스템의 어절 기반 정확도는 97.78%이다.

1. 서 론

한국어는 교착어라는 특수한 성질 때문에 처리에 많은 어려움이 따른다. 또한, 한 어절은 여러 형태소로 구성되므로 영어에 비해서 중의성도 많이 존재한다. 형태소 분석기에서는 한 어절이 가질 수 있는 모든 조합의 형태소 열을 분석하는 것이다. 하지만, 한국어 문장에서 한 어절은 하나의 문법적 기능을 수행하므로, 형태소 분석기에서 제시한 모든 형태소 열 중, 하나를 결정해야 할 필요성이 있다. 각 어절에 대한 형태소 분석 결과들 중, 가장 적합한 형태소 열을 선택하는 작업을 품사 태깅이라 한다. 이것은 자연어 처리의 마지막 단계인 구문 분석 단계에서의 과다한 부담을 줄이기 위해 전처리 단계로 활용될 수 있다[2].

현재 연구된 태깅 시스템은 크게 규칙기반 접근 방법, 통계기반 접근 방법, 그리고 규칙과 통계기반 접근 방법을 통합한 통합 접근 방법으로 분류할 수 있다. 일반적으로 규칙 또는 통계 정보에 기반을 둔 방법은 효율과 성능 면에서 좋은 결과를 내지 못하고 있다[2]. 따라서 최근에는 통합 접근 방법에 대한 연구가 활발히 진행되고 있다. 본 논문은 규칙기반 방법과 통계기반 방법을 효율적으로 통합한 태깅 시스템을 구축하기 위해서 우선 규칙을 적용하고 n회차의 비감독 학습(Unsupervised Learning)으로 추출한 통계정보를 이용한다.

본 논문은 다음과 같이 구성된다. 2장에서는 기존에 연구된 결과를 살펴보고, 3장에서는 규칙 구축 방법에 대해 알아보고, 통계정보와 규칙을 통해서 최적의 품사 열을 결정하는 방법을 소개하고, 4장에서는 실험결과에 대해 설명한다. 끝으로 5장에서 결론 및 향후 과제에 대해서 기술한다.

2. 관련연구

일반적으로 HMM은 이전 하나의 형태소 또는 품사를 참조한다. 이와 같은 순수 HMM기반의 태깅 시스템도 많은 연구도 있지만 이를 확장하여 좀 더 풍부한 문맥 정보를 사용한 태깅 방법도 연구도 있다[3]. 문맥 정보를 많이 고려함으로 인해서 발생하는 자료부족문제를 해결하기 위해서 Simplified back-off smoothing 기법을 사용하였다. 또한, JIM(Joint Independent Model)을 도입하여 정확도를 향상시켰다. 다음은 가장 좋은 성능을 보이는 태깅 시스템 모델이다.

$$\Lambda(C_{[s](K,J)}, M_{[s](L,I)}) \models \Pr(c_{1,u}, p_{2,u}, m_{1,u})$$
$$\approx \prod_{i=1}^{u} \Pr(c_i[,p_i] | c_{i-K,i-1}, p_{i-K+1,i-1}], m_{i-J,i-1}) \times$$
$$\Pr(m_i | c_{i-L,i}[,p_{i-L+1,i}], m_{i-I,i-1})$$

이 시스템의 성능 테스트는 167,115어절로 구성된 'KUNLP 코퍼스'를 대상으로 10-fold cross validation을 사용하였다. 그 결과 $K=2, J=2, L=2, I=2$일 때 96.97%의 성능을 보인다.

HMM 기반 태깅은 문맥 정보만을 고려하므로 임의의 어절이 가지는 특정 형태소 패턴에 대한 품사 중의성을 해결하기가 어렵다. 이를 보완하기 위해 HMM 태깅 이후 후처리로써 오류를 수정하는 연구가 진행 되었다[4]. 이 연구는 미등록어 추정 문제를 기존의 방법과는 달리 형태소 패턴 사전을 구축하고, 조합 가능한 음절 패턴에, 해당 패턴에 가질 수 있는 품사를 포함하여 인코딩하여 사전을 구축한다. 형태소 분석 부분에서 미등록어 발생 시, 이 음절 패턴을 검색하여 초기 품사를 결정한다. HMM을 사용한 태깅에서는 다음과 같은 모델 파라미터를 사용한다.

$$\Pr(m_i,t_i)=\frac{Count(m_i,t_i)}{Count(t_i)} \quad \Pr(t_i,t_{i-1})=\frac{Count(t_i,t_{i-1})}{Count(t_{i-1})}$$

총 41,000개의 형태소에 대해 수작업으로 품사를 부착하고, 30,000개는 HMM학습에 사용하고 나머지 11,000는 성능 테스트에 사용한다. 성능은 HMM만 사용했을 경우, 88.3% HMM과 미등록어 추정을 사용했을 경우 93.1% 끝으로 규칙 기반 오류 교정 루틴을 적용하였을 경우, 94.9의 성능을 보였다.

3. 규칙 및 통계기반 태깅 시스템

3.1 규칙의 구축

규칙은 최소한의 규칙을 구축하였고 일반 규칙과 특수 규칙으로 나뉜다. 일반 규칙의 부여방법은 두 가지로 나눌 수 있다. 구축된 규칙에 의하여 부여되는 가중치의 종류는 두 가지로 나눌 수 있다. 하나는 한 어절에 부여되는 상태 가중치(State Weight)이고, 다른 하나는 두 어절 사이에 부여되는 전이 가중치(Transition Weight)이다. 규칙은 주로 형태소의 패턴이나 위치를 보고 부여할 가중치를 결정하게 되며, 본 시스템에 적용된 규칙의 개수는 약 70여 개이다.

3.2 통계 정보 학습

통계 정보는 비감독 학습으로 원시 말뭉치에 대해 통계정보를 학습하여, 규칙기반 태깅의 정확도를 향상시킬 수 있도록 하였다. 본 시스템에서 사용한 학습 데이터는 신문기사와 방송 원고로 이루어진 33,641,511어절을 대상으로 하였다(표1 참고). 통계 정보는 57개의 품사 집합에 대해 추출하였다. 학습 회차 n회에 대해 그 학습 회차에서 학습 데이터로부터 추출된 전체 형태소 개수를 $N_{total(n)}$이라 하고, 하나의 어절은 N_{ma}개의 형태소 분석 결과 MA를 가진다. 형태소 분석 결과 중 하나인 MA_i는 $N_{mp(i)}$개의 형태소 MP를 가진다. $Tag(MP_j)$는 MP_j의 품사를 의미하고, $F(Tag(MP_j))$는 품사 태그의 빈도를 나타낸다. 학습 회차 n에서 한 어절에 대한 형태소열 MA_i의 확률 $P_n(MA_i)$는 다음과 같다.

$$P_n(MP_j)=\frac{F_{n-1}(MP_j)}{N_{total(n-1)}},1\le j\le N_{mp(i)} \quad (식1)$$

$$P_n(MA_i)=\prod_{j=1}^{N_{mp(i)}}P_n(MP_j),1\le i<N_{ma} \quad (식2)$$

(식2)에서 구해진 통계 정보는 HMM에서 상태 확률(State Probability)로 사용된다. n번째 학습 회차에서 학습데이터의 어절 bi-gram의 개수를 M_{total}이라 하고, 어절 수를 W_{total}라고 한다. 임의의 두 어절 w_{i-1},w_i에 대해, 두 어절이 가지는 형태소 분석결과를 각각MAw_{i-1},MAw_i라고 할 때, 이 어절 bi-gram의 품사 출현 빈도를 $F(Tag(MAw_i)|Tag(MAw_{i-1}))$라고 한다. 이때, 어절 bi-gram의 확률 값은 다음과 같이 정의한다.

$$P_n(Tag(MAw_i)|Tag(MAw_{i-1}))=\frac{F_{n-1}(Tag(MAw_i)|Tag(MAw_{i-1}))}{M_{total}} \quad (식3)$$
$$,1\le i\le W_{total}-1$$

(식3)의 경우, w_{i-1},w_i의 두 어절 사이의 관계를 표현한 것으로써, HMM모델 내에서 한 state인 w_{i-1}에서 다음 state인 w_i로의 전이 확률(Transition Probability)로 사용된다.

한 어절은 N_{ma}개의 형태소분석 결과 MA를 갖는데, 이 중 하나의 MA_i를 결정하는 것은 $n=0$일 때는 규칙만 적용시켜 나온 결과를 사용한다. $n\ge1$의 경우에는 이전 단계에서 구해진 통계자료를 바탕으로 품사를 결정 한 뒤 통계정보를 추출하도록 한다. 학습을 무한히 반복할 수 없으므로 태깅된 모델 구축용 정답 데이터(표 2 참고)에 대해 학습 회차에 따른 성능 변화를 측정하면서 수렴하는 학습 회차를 측정한다.

3.3 가중치 및 통계정보에 의한 최적 품사열 결정

한 문장에 대한 최적의 품사열은 HMM을 기반으로 Viterbi Algorithm을 사용하여 선택한다. 최적 경로를 결정할 때 서로 다른 노드에서 같은 가중치가 결정된 경우, 통계정보에 의한 확률값으로 최적 노드를 선택하도록 한다.

N_{word}개의 어절을 가지는 문장에 대해서 t는 어절 위치 , i는 t번째 어절이 가지는 형태소 분석후보 , j는 $t+1$번째 어절이 가지는 형태소 분석후보를 나타낸다. $N_{ma(t)},N_{ma(t+1)}$는 각각 분석 후보의 개수이다. $W_i(t)$는 t번째 어절까지 계산된 최적의 가중치 값이고, $R_{sw(i)}(t)$는 i번째 후보가 갖는 상태 가중치이다. $R_{tw(ij)}(t)$는 i,j번째 후보 사이에 존재하는 전이 가중치이다. $t+1$번째 어절까지 최적 품사열을 결정하기 위한 가중치값 $W_j(t+1)$은 수식 4와 같이 얻어진다.

$$W_j(t+1)=\max(W_i(t)+R_{sw(i)}(t)+R_{tw(ij)}(t)) \quad (식4)$$
$$,1\le i\le N_{ma(t)},1\le j\le N_{ma(t+1)}$$

다음은 $t+1$번째 어절까지 계산된 최적 품사열을 결정하기 위한 확률 값, $\delta_j(t+1)$를 구하는 식이다.

$$\delta_j(t+1)=\max(\delta_i(t)P(MA_i)P(Tag(MA_i)|Tag(MA_j)) \quad (식5)$$
$$,1\le i\le N_{ma(t)},1\le j\le N_{ma(t+1)}$$

규칙과 통계를 조합한 모델을 이용하여 즉 (식4)와 (식5)를 동시에 사용하여, 최적 품사열을 결정하는 수식은 다음과 같이 나타낼 수 있다.

$$S_{best}(t+1)=\arg max_i\begin{cases}W_j(t+1), W_j(t+1)>W_j(t+1)'\\ \delta_j(t+1) \quad , W_j(t+1)=W_j(t+1)'\end{cases} \quad (식$$

(식6)에서 $W_j(t+1)'$은 t 번째 어절이 가지는 분석 후보들 중 서로 경쟁이 되는 대상을 나타낸다. 따라서$N_{ma(t)}$개의 분석 후보 중, 다음 어절로 전이될 때 같은 가중치를 가지게 되면, 확률값에 의해서 최적 경로를 결정하게 되고, 같지 않으면 가중치가 높은 쪽으로 최적 경로를 결정하게 된다.

4. 실험 및 결과

4.1 결합 모델 실험 데이터

본 연구에서는 비감독 학습으로 형태소별 통계를 추출하기 위해서 신문기사와 TV뉴스 방송 원고로 구성된 표1과 같은 대용량 원시 말뭉치를 사용하였다.

	총 어절의 개수
(A) A신문의 2년분 신문기사	18,959,461
(B) B신문의 1년분 신문기사	9,432,167
(C) TV 뉴스 방송 원고	5,249,883
총 합	33,641,511

표 1. 통계정보 추출을 위한 원시 말뭉치

	총 어절 수
품사정보가 부착된 말뭉치	30,320

표 2. 모델 구축을 위한 답안 데이터

4.2 성능 실험 및 결과

성능 실험에는 소설 및 신문기사 등에서 무작위로 추출한 외부 데이터를 사용하였다. 10000개의 어절로 구성되어 있으며 어절기반으로 정확도를 구하였다.

$$P_{word} = \frac{\text{시스템이 올바르게 반환한 어절 수}}{\text{시스템이 반환한 전체 어절 수}} \times 100(\%)$$

	규칙 기반	결합 모델
어절 기반 정확도	95.51%	97.78%

표 4. 성능 실험 결과

본 논문은 규칙으로 우선 태깅을 한 원시 말뭉치로부터 통계정보를 추출하고 비감독 학습을 5회차 반복함으로써 97.78%의 정확도를 얻을 수 있었다. 6회차 학습부터는 성능이 수렴함을 관찰할 수 있었다.

5. 결론 및 향후 연구

본 논문은 규칙으로 우선 태깅을 한 원시 말뭉치로부터 통계정보를 추출하고 비감독 학습을 반복함으로써 효율적인 태깅 시스템을 구축할 수 있었다. 비감독 학습은 원시 말뭉치를 사용함으로 학습 말뭉치 크기 확장이 용이하며 충분한 통계 데이터를 얻을 수 있다. 본 시스템은 최소한의 규칙과 비감독학습으로 구한 통계 정보를 이용함으로써 bigram 기반이지만 trigram 기반 감독학습 모델과 규칙기반 오류 후처리 모델과 비교할 수 있는 성능을 얻을 수 있었다. 그러나 본 논문은 규칙에 의해 받은 가중치의 크기는 연구하지 못했다. 향후 연구에서는 규칙별로 주어지는 가중치에 대한 더욱 체계적인 연구가 필요하며 규칙기반 방법론과 통계 기반 방법론을 보다 효율적으로 조합하는 모델을 연구해야 한다. 또한, 실질형태소와 형식형태소 간의 통계 기반 가중치 분배에 대해 세분화한 연구도 이루어져야 한다. 끝으로 한국어의 교착어적인 속성에 대한 단서를 충분히 반영한 미등록어 추정 알고리즘도 보완되어야 한다.

<Acknowledegement>
이 논문은 국가지정연구실사업 지원으로 이루어진 것임. (과제번호 M1-0412-00-0028-04-J00-00-014-00)

6. 참고문헌

[1] 임해창, "자연어 처리를 위한 품사 태깅 시스템의 고찰", 정보과학회지 Vol14 No.7, pp.36-57, 1996.07

[2] 김민정, "규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거", 부산대학교 전자계산학과 박사학위 논문, 1997

[3] Sang-Zoo Lee, Jun'ichi Tsujii, Hae-Chang Rim, Hidden Markov Model-Based Korean Part-of-Speech Tagging Considering High Agglutinativity, Word-Spacing, and Lexical Correlativity, Proc. of the 38th ACL, pp. 376-383, 2000

[4] Geunbae Lee, Jeongwon Cha, JongHyeok Lee. Hybrid POS tagging with generalized unknown-word handling. Proceedings of the 2nd international workshop on information retrieval with Asian languages (iral97), Tsukuba-City, Japan, pp43-50, 1997.

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

[6] Eric Brill, "Unsupervised Learning of Disambiguation Rules for Part of speech Tagging," Proc. of the 3rd Workshop on Very Large Corpora, pp. 1-13,1995

[7] Jean-Pierre Chanod, Pasi Tapanainen, "Tagging French - Comparing a Statistical and a Constraint-Based Method", Proc. of the 7th conference of the European chapter of the ACL, Dublin, pp. 149-156,1995