제한된 자원을 사용한 한국어 형태소 분석

강상우[O*], 양재철[*], 김학수[**], 서정연[*]
서강대학교, 컴퓨터공학과[*]
강원대학교, 컴퓨터정보통신공학과[**]
swkang@sogang.ac.kr, nova0821@sogang.ac.kr, nlpdrkim@kangwon.ac.kr, seoyj@sogang.ac.kr

Morpheme Segmentation and Part-Of-Speech Tagging Using Restricted Resources

Sangwoo Kang[O*], Jaechul Yang[*], Harksoo Kim[**], Jungyun Seo[*]
Sogang University, Department of Computer science and Engineering[*]
Kangwon University, Program of Computer and Communications Engineering[**]

요 약

한국어 형태소 분석 및 품사 부착에 대한 연구는 지속적으로 이루어져 왔으며 규칙 기반 방법, 통계 기반 방법 등을 중심으로 연구되었다. 본 논문에서는 최근 활용도가 높아지고 있는 모바일 기기에 적합한 한국어 형태소 분석 및 품사 부착 방법을 제안한다. 모바일 기기는 계산 처리 능력과 사용 가능한 메모리가 제한되기 때문에 전통적인 방법을 사용하여 형태소 분석 및 품사 부착을 수행하기에는 한계가 있다. 본 논문에서는 기존의 규칙 기반 형태소 분석 방법인 좌최장일치법을 변형하여 형태소 분석을 수행 하고, 통계적인 방법인 hidden Markov model을 축소하여 형태소 품사 부착을 수행한다. 제안하는 방법은 기존의 hidden Markov model을 사용한 시스템과 유사한 성능을 보여주며 모바일 기기에 적합하도록 소량의 메모리 사용과 월등히 빠른 속도로 형태소 분석 및 품사 부착을 수행할 수 있다.

주제어: 형태소 분석, 품사 부착, 좌최장일치법, HMM, 모바일 기기

1. 서론

형태소 분석이란 주어진 입력문장 또는 어절을 최소 의미 단위인 형태소로 분리하는 작업이다. 이러한 형태소 분석 결과에서 가장 적합한 형태소의 조합과 품사 정보를 선택하는 작업을 품사 부착이라고 한다. 형태소 분석 및 품사 부착의 결과는 정보 검색, 정보 추출, 기계 번역 등 자연어 처리의 여러 응용 분야에서 중요하게 사용된다. 최근 휴대폰, PMP 등의 모바일 기기의 활용도가 높아지고 자연어처리의 응용 분야가 확대되고 있다. 하지만 이러한 기기들은 계산 능력과 사용 가능한 메모리의 부족 등 많은 제약 사항이 따른다. 따라서 모바일 기기에서 자여어 처리를 수행하기 위해서는 제한된 자원에서 효과적인 형태소 분석 및 품사 부착 방법이 필요하다. 본 논문에서는 기존의 규칙 기반 형태소 분석 방법인 좌최장일치법을 변형하여 형태소 분석을 수행하고 통계적인 방법인 Hidden Markov Model (HMM) 을 축소 적용하여 형태소 품사 부착을 수행한다. 제안된 방법은 기본적으로 규칙 기반의 형태소 분석을 사용하기 때문에 분석 속도가 빠르고, 축소된 HMM을 결합함으로서 품사 분석 성능의 저하를 최소화 할 수 있다.

한국어 형태소 분석과 품사 부착에 대한 연구는 80년대부터 지속적으로 연구되어 왔다[1]. 초기에는 주로 규칙에 기반한 방법들[2-4]이 연구되었으나 모든 적용 가능한 규칙을 획득하기 어렵고 많은 비용이 소요되는 단점을 갖기 때문에 이를 극복하기 위하여 통계적인 접근을 시도하였다. 통계적 방법은 통해 대량의 말뭉치로부터 추출한 통계 정보를 이용하여 자동으로 형태소 분석과 품사 부착을 수행한다[5-8]. Hidden Markov Model (HMM)은 품사 부착에 사용되는 대표적인 통계기반 모델이다. HMM을 이용한 방법은 품사 부착 정확도가 높다는 장점을 갖지만, 높은 복잡도와 다량의 통계 정보를 필요로 하기 때문에 규칙 기반 방법과 비교하여 속도가 느리다.

2. 변형된 좌최장일치법을 이용한 형태소 분석

규칙기반 모델인 좌최장일치법[9]은 빠른 분석이 가능하지만 동일 형태의 어절에 대하여 항상 하나의 결과만을 제공하는 단점을 갖는다. 이러한 단점은 동일한 형태소 분리에 대해서 가능한 모든 품사들의 쌍을 결과로 제공함으로 간단히 해결 할 수 있다. 예로 어절 "말한" 에 대하여 좌최장일치법의 분석 결과는 "말/의존명사+한/일반명사" 한 개만을 제시한다. 하지만 올바른 형태소 분석 결과는 "말/의존명사+하/동사파생접미사+ㄴ/관형형전성어미" 이다. 이러한 현상은 좌최장일치법의 특성에 의한 결과로, 올바른 분석 결과인 "하/동사파생접미사+ㄴ/관형형전성어미" 가 긴 형태소인 "한/일반명사" 에 가려져 발생한다. 이러한 현상을 해결하기 위하여 긴 형태소에 의하여 올바른 분석 결과가 가려지는 오류들을 추출하고 이를 부분 결과 사전으로 구성하여 사용한다. 하지만 모든 오류에 대하여 부분 결과 사전을 생성하는 것을 비효율적이기 때문에 부분 결과 사전에 어떠한 부분 문

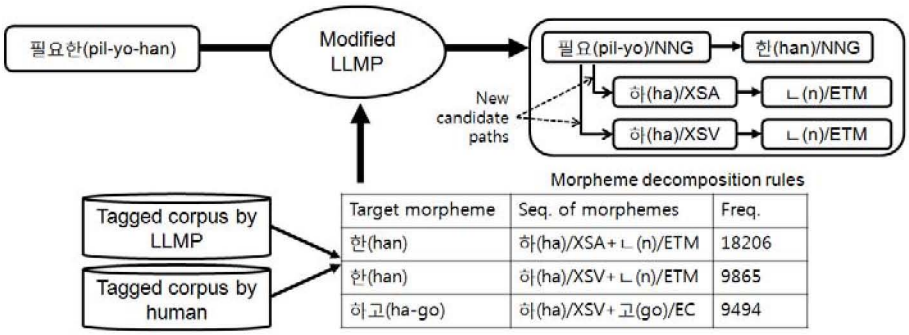

그림 1. 부분 결과 사전 생성 과정의 예

자열과 올바른 형태소 분석 결과를 추가할 것인지 정하는 것을 매우 중요한다. 그림 1은 은 부분 결과 사전을 생성하는 과정을 보여준다. 오류가 발생하는 부분 문자열을 추출하는 과정에서 발생 빈도를 측정하여 발생 빈도수 누적 양의 상위 40%이하인 부분 문자열과 올바른 형태소 분석 결과만을 부분 결과 사전으로 사용한다.

3. 축소된 HMM 기반 한국어 형태소 품사 부착

변형된 좌최장일치법의 형태소 분석 결과는 부분 결과 사전에서 추가된 분석 결과를 제외하기 때문에 동일한 형태소 분리를 갖는다. 이러한 특징을 이용하여 기존의 HMM을 축소하여 한국어 형태소 품사 부착을 수행한다. 제안한 축소된 HMM 기반 한국어 형태소 품사 부착을 위한 확률 모델은 식 1과 같다.

$$
\begin{aligned}
(\hat{W},\hat{T}) &= \arg\max_{W,T} P(W,T \mid S) \\
&= \arg\max_{W,T} \frac{P(W,T,S)}{P(S)} \\
&= \arg\max_{W,T} P(W,T,S) \\
&= \arg\max_{W,T} P(W,T)
\end{aligned}
\qquad \text{(식 1)}
$$

$$
(\hat{W},\hat{T}) \cong \arg\max_{W,T} \prod_{i=1}^{n} P(w_i \mid t_i) P(t_i \mid t_{i-1})
$$

$$
,where\ P'(w_i \mid t_i) =
\begin{cases}
P(w_i \mid t_i), & if\ w_i \text{ has several POS tags} \\
\prod_{j=1}^{k} P(w_i^j \mid t_j) P(t_j \mid t_{j-1}), & if\ w_i \text{ is divided in to } k \text{ morphemes} \\
c & if\ w_i \text{ has only one POS tag}
\end{cases}
$$

식 1에서 변형된 좌최장일치법에 의하여 생성된 형태소 분석 결과는 동일한 형태소 분리를 갖기 때문에 입력 문장은 W로 경계가 고정된다. 기존 HMM에서는 주어진 입력 문장에 대하여 최적의 형태소 순서 열과 형태소 품사 열을 찾아야 하지만, 축소된 HMM에서는 고정된 형태소 순서열 W에 대하여 최적의 형태소 품사 열 T를 찾는다. 또한 축소된 HMM은 기존의 HMM과는 다른 관측 확률을 저장한다. 축소된 HMM에서 관측 확률은 형태소 w_i가 여러 품사로 쓰이는 경우 기존의 HMM과 동일한 확률을 갖는다. 하지만 형태소 w_i가 하나의 품사로만 쓰일 경우, 변형된 좌최장일치법에의한 형태소 분석 결과가 동일한 분리를 갖기 때문에 하나의 품사로만 쓰이는 형태소의 관측 확률은 최적의 형태소 분석 결과를 찾는데 영향을 미치지 않는다. 그림 2의 예에서 "압류"는 일반명사 하나 만을 품사로 갖기 때문에 "압류/일반명사"의 관측 확률은 최적의 형태소 열을 찾는데 영향을 미치지 않는다. 축소된 HMM은 하나의 품사로만 쓰이는 형태소의 관측 확률을 저장할 필요가 없으며, 하나의 품사로만 쓰이는 형태소의 수는 전체 형태소 수의 약 95%에 해당한다. 따라서 통계 정보를 위해 필요한 저장 공간 및 형태소 품사 부착을 수행 할 때 사용하는 메모리의 양이 급격하게 감소한다는 장점을 갖는다.

4. 실험 및 평가

본 논문에서는 학습 및 실험을 위하여 세종 계획 말뭉치를 사용하였다 [10]. 세종 계획 말뭉치는 표 1과 같이 구성되어 있다. 실험을 위하여 말뭉치로부터 형태소 접속 규칙 및 형태소 사전을 추출하여 사용하였으며 uni-gram 관측 확률과 HMM을 위한 형태소 관측 확률 그리고 품사 전이 확률을 Maximum Likelihood Estimator을 사용하여 계산하였다.

표 1. 세종 계획 말뭉치의 구성

Description	Number
Sentence	139,828
Eojeol (space unit)	2,015,860
Morpheme	4,641,546
Tag	46

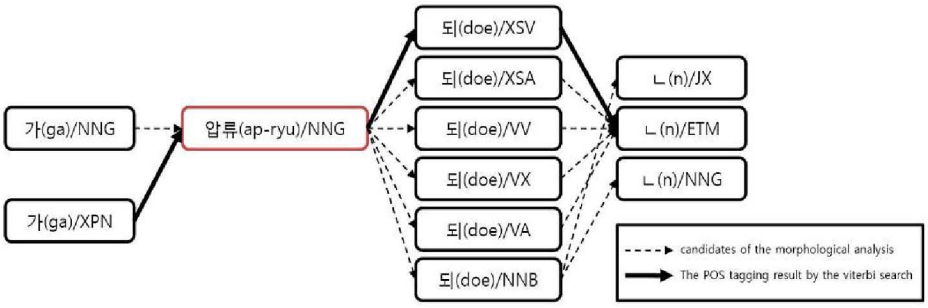

그림 2. "가압류된"의 형태소 품사 부착의 예

시스템의 성능을 평가하기 위하여 재현율, 정확률 그리고 F_1-평가치를 사용하였으며 표 2는 각 시스템의 성능을 보여준다.

표 2. 시스템 성능 비교

System	Recall	Precision	F_1-Measure
LLMP	0.82	0.83	0.82
Traditional HMM system	0.94	0.94	0.94
Proposed system	**0.90**	**0.92**	**0.91**

표 2에서 기존 HMM을 이용한 시스템이 가장 높은 성능을 보여 주었고 제안한 시스템은 약 3% 낮은 성능을 보였다. 제안한 시스템은 변형된 좌최장일치법에서 제시한 형태소 분석 결과를 사용하기 때문에. 모든 가능한 형태소 분석 결과 중 최적의 형태소 품사 부착 결과를 선택하는 HMM을 이용한 방법보다 근소하게 낮은 성능을 보이지만 제안 시스템은 기존의 HMM을 이용한 방법보다 형태소 분석 및 품사 부착 수행 시간이 월등히 빠르고, 적은 저장 공간을 사용하는 장점이 있다 (표 3).

표 3. 메모리 사용량과 응답 속도 비교

	LLMP	Traditional HMM	Proposed system
Memory usage (MB)	1.6	3.0	1.85
Response time (sec/sentence)	0.0154	0.1495	0.0195

5. 결론

본 논문에서는 최근 사용이 급속도로 늘고 있는 모바일 기기에 적합한 한국어 형태소 분석 및 품사 부착 방법을 제안하였다. 빠른 응답 속도를 위해 좌최장일치법을 응용하여 사용하였고 올바른 형태소 분석 결과가 긴 형태소에 가려져 정답으로 생성되지 못하는 단점을 보완하기 위해 부분 결과 사전을 이용하였다. 또한 형태소 품사 부착을 위하여 uni-gram 관측 확률을 이용한 방법과 축소된 HMM을 이용하는 방법을 제안하였다.

감사의 글

본 연구는 지식경제부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041678, 다중영역 정보서비스를 위한 대화형 개인 비서 소프트웨어 원천 기술 개발]

참고문헌

[1]S. Kim, K. Choi and K. Kim, "A Korean morphological analyzer using tabular parsing and connection information," *Proc. of Spring Conference on Artificial Intelligence*, pp.133-147, 1987 (in Korean).
[2]E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," *Computational Linguistics*, vol. 21, no. 4, pp.543–564, 1995.
[3]J. Lovins, "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics*, vol. 11, pp.22-31, 1968.
[4]E. Lee, *An improved method on Korean morphological analysis based on CYK algorithm*, MS thesis, POSTECH, 1992 (in Korean).
[5]K. Church, "A stochastic parts program and noun phrase parser for unrestricted text," *Proc. of Conference on Applied Natural Language Processing*, pp.136-143, 1988.
[6]J. Kim, H. Lim and H. Rim, "Twoply hidden Markov model: A Korean POS tagging model based on morpheme-unit with Eojeol-unit context," *International Journal of Computer Processing of Oriental Languages*, vol. 21, no. 1, pp.5-29, 1998.
[7]E. Charniak, C. Hendrickson, N. Jacobson and M. Perkowitz, "Equations for part of speech tagging," *Proc. of Conference on the American Association for Artificial Intelligence*, pp.784-789, 1993.
[8]D. Lee and H. Rim, "Probabilistic models for Korean morphological analysis," *Proc. of International Joint Conference on Natural Language Processing*, pp.197-202, 2005.
[9]Y. Song, K. Lee and Y. Lee, "Morphological analyzer using longest match method for syntactic analysis," *Proc. of the Morphological Analysis and Tagger Evaluation Contest*, pp.157-166, 1999.
[10]*The National Institute of the Korean Language: Final report on achievements of 21st Sejong project: electronic dictionary*, 2007.