어절 내의 형태소 범주 패턴에 기반한 통계적 자동 띄어쓰기 시스템

(A Stochastic Word-Spacing System Based on Word Category-Pattern)

강 미 영 † 정 성 원 † 권 혁 철 ††
(Mi-young Kang) (Sung-won Jung) (Hyuk-chul Kwon)

요 약 본 논문에서는 형태소 unigram과 한국어 어절을 형성하는 형태소 범주 패턴에 기반하여 어절을 인식하는 한국어 띄어쓰기 시스템을 구현하였다. 기존에 많이 연구된 통계 정보를 이용한 띄어쓰기 모델은 비교적 짧은 시간에 쉽게 구현할 수 있는 장점이 있지만, 한국어의 형태·유형론적 특성 때문에 발생하는 (ㄱ) 자료부족 문제와 (ㄴ) 메모리 크기 문제에 효과적으로 대처하지 못한다. 본 논문은 이 두 문제를 동시에 해결하기 위해 어절을 구성하고 있는 개별 형태소의 통계 정보와 그 형태소의 범주의 통계 정보를 기반으로 하여 띄어쓰기 후보 어절들을 추천한다. 임의의 후보 어절이 최종의 띄어쓰기 단위인 어절이 될 수 있는 확률은 (ㄱ) 해당 후보 어절 내의 각 형태소 확률과 (ㄴ) 해당 후보 어절을 구성하기 위해 그 형태소의 범주가 다른 형태소 범주와 함께 형성하는 패턴 내에서 차지하는 '범주가중치'를 고려하여 구한다. 해당 '범주가중치'는 (ㄱ) 말뭉치로부터 실제로 관찰된 어절의 확률과 (ㄴ) 후보 어절 내의 개별 형태소의 확률과 (ㄷ) 그 범주 가중치에 의해 추정된 어절 확률 사이의 평균 에러(error mean)가 최저가 되는 방향으로 학습하여 얻어진다.

키워드 : 한국어 자동 띄어쓰기, 어절 unigram, 형태소 unigram, 음절 bigram, 형태소 범주 패턴, 통계 정보

Abstract This paper implements an automatic Korean word-spacing system based on word-recognition using morpheme unigrams and the pattern that the categories of those morpheme unigrams share within a candidate word. Although previous work on Korean word-spacing models has produced the advantages of easy construction and time efficiency, there still remain problems, such as data sparseness and critical memory size, which arise from the morpho-typological characteristics of Korean. In order to cope with both problems, our implementation uses the stochastic information of morpheme unigrams, and their category patterns, instead of word unigrams. A word's probability in a sentence is obtained based on morpheme probability and the weight for the morpheme's category within the category pattern of the candidate word. The category weights are trained so as to minimize the error means between the observed probabilities of words and those estimated by words' individual-morphemes' probabilities weighted according to their categories' powers in a given word's category pattern.

Key words : Korean word-spacing; word-unigram; morpheme-unigram; category pattern; stochastic information

1. 서 론

자연어 처리 분야에서는 텍스트 분석과 관련한 작업이 매우 빈번하게 일어난다. 텍스트 분석에서 가장 기초적인 작업은 텍스트로부터 **단어를 식별하고 추출하는 토큰화**(Tokenization)라고 할 수 있다. 토큰의 단위가 되는 단어에 대한 정의는 관점에 따라 다양한데, 본 논문에서는 한국어의 철자법적 관점에서 단어를 띄어쓰기 단위로 보기로 하며, 이를 보통 '어절(語節)'이라 부르므로 본 논문도 이 용어를 쓰기로 한다. 중국어나 일본어와 같이 어절 경계 표지가 없는 언어와는 달리, 한국어

· 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음
† 학생회원 : 부산대학교 컴퓨터공학과
kmyoung@pusan.ac.kr
swjung@pusan.ac.kr
†† 종신회원 : 부산대학교 컴퓨터공학과 교수
hckwon@pusan.ac.kr
논문접수 : 2006년 4월 6일
심사완료 : 2006년 9월 12일

는 철자법으로 어절과 어절 사이에 공백을 두어 띄어쓰기를 하도록 규정하고 있다. 한국어에 있어서 잘못된 띄어쓰기는 중의성(ambiguity)을 유발 시키거나 텍스트 분석에서 잡음(noise)을 일으켜 오히려 토큰화를 방해하며, 가독성을 떨어뜨린다. 이와 같이 한국어에서 띄어쓰기는 텍스트에 대한 사용자 가독성만큼이나 기계 가독성에도 영향을 주는 중요한 요소이다. 문장 내의 띄어쓰기오류는 많은 문법적, 의미적 모호성을 일으키며, 때로는 형태소 분석을 불가능하게 만들기도 한다.

이러한 띄어쓰기를 자동으로 해결하기 위해서는 다음과 같은 한국어의 특성을 고려하여야 한다. 우선 한글은 자모가 한 음소를 표시하는 음소 문자이며 이러한 자모를 음절을 단위로 모아 글자를 만드는 음절문자이다. 띄어쓰기는 일차적으로 바로 이런 음절 사이에 나타난다. 한국어의 음절은 띄어쓰기와 관련하여 몇몇 통계적인 특성을 보인다. 대표적으로 어떤 음절은 어절 경계에 자주 나타나지만, 어떤 음절은 어절 경계에 전혀 나타나지 않는다. 한편, 각각의 어절은 한국어에 고유한 교착어적인 형태론에 의한 특성을 보이는데, 명사, 동사, 형용사 등과 같은 실질형태소 뒤에 조사나 어미와 같은 문법형태소들이 결합하여 다양한 형태의 어절을 생성해 내며, 이러한 어절들이 띄어쓰기의 경계가 된다. 이러한 형태·유형론적인 특성과 더불어 끊임없이 생성되는 신조어들은 통계기반 접근 방법론에서 자료부족 문제를 일으킨다.

한국어의 이와 같은 특성을 고려하여 정확한 띄어쓰기를 자동으로 제시함으로써 한국어 문장을 문법적으로 정확한 어절의 나열로 만들기 위하여, 본 연구의 선행 연구에서는 어절 unigram 기반의 통계적인 띄어쓰기 모델을 제안하였다. 이 연구에서는 통계적 어절 unigram 기반 접근 방식에서 발견되는 자료부족 문제를 해결하기 위해 통계적 기법인 음절 bigram에 기반하여 어절의 경계에 대한 정보를 보완하고 규칙/지식에 기반하여 후보 어절 제시하는 혼합모델을 제안하였지만, 여전히 데이타 부족문제를 모두 해결할 수는 없었다[6]. 더욱이 후보 어절을 확장함으로써 사전의 메모리가 커지는 문제점이 발생하였다. 따라서 본 연구는 (ㄱ) 자료부족 문제와 (ㄴ) 메모리 크기 문제를 동시에 해결할 수 있는 형태소 unigram과 형태소 범주 패턴에 기반을 둔 띄어쓰기 기법을 제시한다.

2. 관련 연구

한국어 띄어쓰기에 관한 연구는 크게 규칙 혹은 지식 기반 접근 방법과 통계적 접근 방법으로 나눌 수 있다. 이전 일부 연구에서는 (ㄱ) 휴리스틱을 이용한 규칙(혹은 지식) 기반 접근 방법을 연구하였으며[4], (ㄴ) 문장을 좌에서 우로 분석하면서 가장 길게 분석된 어절을 띄어쓰기 단위로 선택하는 기법(longest-match strategy, 최장일치기법)을 사용하기도 하였다[2]. 또한, (ㄷ)조사나 어미 등의 음절 특성이나 형태소간의 음절 결합 특성을 고려한 형태소 분석을 기반으로 우선순위를 적용하는 방법도 연구 되었다[8]. 이러한 기존의 규칙(혹은 지식) 기반 연구는 언어적인 특성을 이용함으로써 정확도가 매우 높지만, 명문화된 규칙 이외의 각종 예외 상황이나 언어가 계속 변화함에 따라 나타나는 새로운 규칙을 적용/확장하기 어렵다는 단점이 있다. 반면, 통계기반 접근방법은 규칙을 구축하는 것보다 시간과 노력이 적게 들지만, 구축 결과가 학습 데이타에 치우치며, 데이타 부족 문제가 나타나는 단점이 있다. 대부분의 한국어 통계기반 띄어쓰기 연구는 다음과 같은 한국어의 음절 특성 때문에 n-gram을 사용한다.

2.1 한국어의 음절 특성

일반적으로 음절은 어절에 비하여 그 종류가 한정적이므로 어절에 비하여 자료 부족 문제가 심각하지 않다. 따라서 많은 통계기반 한국어 띄어쓰기 연구는 음절 n-gram을 기반으로 하였으며, 한국어 형태소 및 음운 제약과 '국부적 통사 제약(local syntactic constraints)'으로 인해 나타나는 음절 분포에 기반하여 다소나마 한국어 띄어쓰기 문제의 해결 가능성을 보였다. 다음은 한국어의 어절 경계에서 나타나는 음절의 특성을 일부 나열한 것이다.

- 단어 경계에서 음절 특성과 관련한 음소 제약
  - 종성에 ㅄ, ㅆ, ㄻ, ㄽ 이 있는 음절은 단어 경계가 되는 공백과 같이 나타날 수 없다.
  - 초성에 ㄹ을 포함하는 음절은 외래어를 제외하고는 단어 경계의 왼쪽에 나타날 수 없다.
- 문법적인 형태소, 즉 조사나 선어말 어미와 같은 형태소 내에 나타나는 음절 bigram은 매우 높은 빈도로 출현한다.
  - 에서<PP>[1], 습니<EPP>, 니다<END>
- 문법형태소는 그 종류는 한정적이지만, 교착어적인 특성으로 인하여 매우 높은 생산성을 가진다.
  - 이라 <= 이다<COP> + 라<END>
  - 있다 <= 있다<VX> + 다<END>
- 한글의 어순은 자유롭지만, 국부적인 통사 제약은 존재한다. 즉, 특정 어미, 의존명사, 보조 동사 등은 통사 제약 아래 고정적 음절 패턴을 보인다.
  - …는 # 것… <= {은/는/던/을/를}<ETM> # 것<BN>
  - …고 # 있… <= 고<END> # 있<VX>

1) 이하, 형태소 범주 태그(tag)는 본문의 표 3 참조.

표 1 PNU 말뭉치에서 추출한 상위 10위 음절 bigram

내부 공백 없음			내부 공백 있음			왼쪽 공백 있음			오른쪽 공백 있음		
σ•σ	No	%	σ#σ	No	%	#σ•σ	No	%	σ•σ#	No	%
습•니	250,996	1.71	고#있	305,712	1.03	#있•다	268,474	0.87	으•로#	505,556	1.62
이•라	118,375	0.81	는#것	145,609	0.49	#기•자	164,075	0.53	니•다#	444,540	1.42
입•니	106,189	0.72	지#않	139,250	0.47	#것•이	150,029	0.49	에•서#	392,326	1.26
에•서	76,402	0.52	에#대	129,382	0.44	#있•는	137,684	0.45	했•다#	347,168	1.11
적•으	71,955	0.49	수#있	128,758	0.43	#지•난	117,368	0.38	하•는#	251,286	0.81
자•들	51,822	0.35	고#말	78,793	0.27	#한•국	108,780	0.35	이•다#	229,481	0.74
통•령	48,310	0.33	할#수	75,229	0.25	#것•으	949,42	0.31	있•다#	228,973	0.73
했•습	44,841	0.31	는#이	64,497	0.22	#때•문	80,000	0.26	하•고#	209,739	0.67
기•자	41,468	0.28	에#따	61,583	0.21	#대•한	78,826	0.26	다•는#	149,972	0.48
으•로	39,382	0.27	기#때	55,297	0.19	#한•다	77,974	0.25	있•는#	133,707	0.43

※ σ 는 음절, •는 음절 경계, #은 공백의 위치를 나타냄

표 1은 공백의 위치와 관련하여 PNU말뭉치[2]로부터 추출한 상위 10위 내의 한글만 포함하는 음절 bigram 통계를 보여준다.

2.2 음절 n-gram 모델

띄어쓰기에 대한 기존 연구는 위와 같은 특성들에 기반하여 음절 기반 접근방법을 사용함으로써 비교적 만족할 만한 성과를 얻었다. 한국어의 음절 특성을 이용한 기존 띄어쓰기 연구들([1,3,5]) 중에서 [1]은 음절 bigram 정보를 정제되지 않은 말뭉치로부터 추출하여 두 음절의 왼쪽, 오른쪽, 중간에 공백이 올 확률을 이용하여 특정 위치에 띄어쓰기가 가능한지 추정하였으며, 단어 단위로 76.71%의 정확도와 67.80%의 재현율을 보였다.[3] [5]는 음절 trigram과 은닉 마르코프 모델을 사용하였다. 이 연구에서는 기존의 통계 기반 자동 띄어쓰기 방법에서 이전의 띄어쓰기 상태를 고려하지 않기 때문에 발생하는 문제점을 극복하기 위하여 자동 띄어쓰기를 품사 부착과 같은 분류 문제로 간주하고 은닉 마르코프 모델을 확장한 모델을 제안하였으며, 93.06%의 어절 단위 정확도를 보였다.

이러한 연구들은 n-gram의 차수가 늘어나면 그에 해당하는 타입이 기하급수적으로 늘어나는 n-gram모델의 특성 때문에 음절 bigram을 적용하였지만, 고려하는 주변 음절이 하나밖에 없으므로 성능은 매우 낮다. 이를 극복하기 위하여 음절 trigram을 사용한 연구도 있었다[5]. 이는 상당한 수준의 성능을 보였지만, 사전의 사이즈가 커져 메모리가 많이 소요되는 단점이 있다(표 3 참조). 더욱이 이러한 음절 n-gram 모델은 어절 n-gram모델에 비하여 인식론적인 관점에서 볼 때 덜 직관적이다. 왜냐하면, 어절 띄어쓰기는 어절의 경계를 인식하는 것이지 음절의 경계를 인식하는 것이 아니기 때문이다. 따라서 본 연구의 선행 연구에서는 n-gram의 차수를 늘이지 않고 어절 경계를 인식함으로써 어절 띄어쓰기를 인식론적 관점에서 해결하기 위하여 노력하였다.

2.3 어절 n-gram 기반 접근 방법

어절을 기반으로 하는 한국어 띄어쓰기 연구는 규칙 혹은 지식 기반 접근 방법을 주로 쓰고 있다[3,7]. 통계적 띄어쓰기 모형에서 높은 차수의 n-gram을 적용하기 위해서는 충분한 크기의 말뭉치가 필요하다. 하지만, 한 음절 단위로 고려했을 때 종류가 한정되어 있어 어느 정도 차수를 늘려 나갈 수 있는 음절 n-gram과는 달리 어절의 n-gram은 한국어의 교착어적인 특성상 거의 무한대로 생성될 수 있으므로, n-gram의 차수를 늘여 적용하기 매우 힘들며, 아무리 말뭉치의 크기를 늘리더라도 자료 부족 문제를 해결할 수 없다. 그에 따라, 본 연구의 선행 연구에서 제안한 통계기반 한국어 띄어쓰기 모델은 어절 n-gram 기반 접근을 취했을 때 문제가 될 수 있는 취급 데이타의 개수가 폭발적으로 늘어나는 현상을 방지하기 위하여 어절 unigram을 이용하였다. 문서 내에서 특정 어절이 출현할 확률은 학습 데이타 내에서 어절 unigram의 상대 빈도를 이용하여 추정하였으며 이는 식 (1)과 같다.

2) 음절 bigram과 어절의 빈도 추출을 위해 사용한 부산대학교의 PNU말뭉치는 실제 데이타 내의 오류로 인한 학습 오류를 최대한 줄이기 위하여 띄어쓰기 전문가들에 의해 정제되었다. 2006년 1월 현재 이 말뭉치는 총 33,643,884어절(총 1,950,068개의 중복을 제거한 어절 타입)로 이루어져 있다. 이 말뭉치는 A신문 2년치 기사, B신문의 1년치 기사와 두 방송사의 뉴스 방송 원고 3년치를 정제한 것으로 문어체, 구어체 등 다양한 표현방식의 한글 어휘를 포함한 말뭉치를 이용함으로써 통계 기반 방식이 학습 말뭉치의 영향을 크게 받아 생기는 문제점을 줄이도록 하였다. 총 추출된 어절에는 아라비안 숫자, 로마자, 다양한 특수문자 등이 포함되어 있다. 숫자는 (ㄱ) 소수점을 포함한 숫자와 (ㄴ) 한 자리(0~9), (ㄷ) 두 자리(10~99), (ㄹ) 세 자리(100~999), (ㅁ) 네 자리(1000~9999), (ㅂ) 그 이상의 숫자 단위로 그룹화하고, (ㅅ) IP주소, (ㅎ) 날짜 (년•월•일), (ㅈ) 절 번호(장•절•하위절), (ㅊ) 시간(시•분•패턴으로 유형화 하여 빈도를 조사하였다.

3) 해당 연구에서 어절 단위 성능을 제시하지 않고 있으며 비교를 위해 [8]이 동일한 모델을 구현하여 평가한 성능이다.

$$p(w)=\frac{f(w)}{Tw} \quad (1)$$

Tw = 학습 데이타 내의 총 어절 수; $f(w)$ = 학습 데이타로부터 추출한 어절(w)의 빈도

띄어쓰기가 잘 된 최상의 문장은 문장 내 띄어쓰기 후보가 될 각 어절의 출현 확률이 가장 높은 문장이다.

$$The\ optimal\ sentence = \arg\max_{S} \prod_{k=1}^{n} p(cw_k) \quad (2)$$

cw_k = 문장 내의 k번째 후보어절

어절 unigram 모델에서는 문장 내에 각 어절들의 출현이 서로 독립 사건이라고 가정한다. 교착어적인 특성상 한국어는 어절의 배열이 자유롭지만, 문장 구성상 국부적인 통사 제약이 존재하기 때문에 특정 부분에서 고정 음절 패턴을 발견할 수 있다(2.1절 참조). 이러한 속성을 효율적으로 이용하기 위하여 본 연구에서는 어절 경계에서 출현하는 음절 bigram 내의 띄어쓰기 평가 값인 odds 값을 사용하며 구하는 식은 (4)와 같다. 따라서 띄어쓰기가 잘 된 최적의 문장은 다음 두 가지 파라메터, 어절 cw_k의 확률과 음절 bigram, $\sigma_{k,\lambda}$와 $\sigma_{k+1,1}$ 사이에 띄어쓰기가 가능한지의 평가값인 $odds$ 값를 이용하여 다음 식 (6)과 같이 만들 수 있다. 여기서 $odds$는 말뭉치에서 나타난 임의의 음절 bigram사이의 띄어쓸 확률을 구하는 식 (5)를 기반으로 구해지는데 이는 말뭉치 상에서 해당 음절 bigram을 띄어 쓴 빈도 $f(\sigma_{k,\lambda}\#\sigma_{k+1,1})$와 붙여 쓴 빈도 $f(\sigma_{k,\lambda}\emptyset\sigma_{k+1,1})$ 정보를 이용하여 구한다.

$$cw_k = \sigma_{k,1}\cdots\sigma_{k,\lambda}$$

λ = cw_k 내의 총음절수 (3)

$$odds(\sigma_{k,\lambda},\ \sigma_{k+1,1}) = \frac{p_{innerS}(\sigma_{k,\lambda},\ \sigma_{k+1,1})}{1-p_{innerS}(\sigma_{k,\lambda},\ \sigma_{k+1,1})}$$

$\sigma_{k,\lambda}$ = cw_k의 마지막 음절

$\sigma_{k+1,1}$ = cw_{k+1}의 첫음절 (4)

$$p_{innerS}(\sigma_{k,\lambda},\ \sigma_{k+1,1}) = \frac{f(\sigma_{k,\lambda}\#\sigma_{k+1,1})}{f(\sigma_{k,\lambda}\#\sigma_{k+1,1})+f(\sigma_{k,\lambda}\,\phi\,\sigma_{k+1,1})}$$

k = n이면, $p_{innerS}(\sigma_{k,\lambda},\ \sigma_{k+1,1})$ = 0.5이다 (5)

$$The\ optimum\ sentence = \arg\max_{S} \prod_{k=1}^{n} p(cw_k)\cdot odds(\sigma_{k,\lambda},\ \sigma_{k+1,1}) \quad (6)$$

위의 수식 (6)에 사용한 어절 unigram 모델의 사전 크기는 29.2Mb 이며, 어절 단위 정확도는 93.3%이다. 이 결과는 trigram모델의 정확도 93.06%와 비슷하면서도 trigram 모델의 사전 크기인 63.7Mb보다 그 크기가 반 이상으로 줄어든 것이다.

3. 통계적 단어 unigram모델의 성능 향상

본 연구의 선행 연구에서는 어절 unigram 모델에 음절 bigram을 적용함으로써 높은 성능 향상을 보였지만, 한국어의 교착어적인 특성과 신조어 생성 등의 문제로 인하여 학습 말뭉치에 나타나지 않은 어절의 처리 문제, 즉 자료 부족 문제를 학습 말뭉치를 늘려서는(무한하게 늘일 수는 없다) 해결하기 어려웠다. 또한, 어절 unigram은 음절 trigram보다 작은 크기의 메모리를 차지하지만, 여전히 많은 양의 메모리를 차지하고 있으므로, 이를 효과적으로 처리할 수 있는 방안이 필요하다.

3.1 자료 부족 문제 해결

통계적 접근 방법에서 나타나는 자료부족 문제를 해결하기 위해 본 연구의 선행연구에서는 규칙과 통계를 결합한 혼합 모델을 제안하였다[6]. 후보 어절을 능동적으로 제안하기 위하여 PNU 형태소 분석기[4]를 사용하여 어절 unigram 기반 모델의 자료 부족 문제를 보충하였다. PNU 형태소 분석기는 분석 가능한 형태소들 중에서 최장일치기법을 사용하여 선택된 후보를 능동적으로 확장하고 실질 형태소 범주와 문법 형태소 범주에 서로 다른 가중치를 부여한다. 더욱이, 형태소 분석기는 어근에서 하위 범주가 분리되는 것을 막기 위하여 우선적으로 띄어쓰지 말아야 할 부분에 대한 처리를 수행한다. 이러한 형태소 분석기를 적용한 시스템의 성능은 내부 데이타 98.39%, 외부 테스트 데이타 97.51% 로 내부, 외부 데이타간의 성능차이가 별로 없었다.

3.2 메모리 크기 증가 문제 해결

앞서 기술한 형태소 분석기를 사용한 후보 단어의 능동적인 확장 모델은 높은 성능을 보인다. 하지만, 형태소 분석기를 통한 미등록어 추정 기법은 39.2Mb 크기의 사전이 필요하다(자세한 사전 크기는 표 2 참조). 앞으로 PDA나 휴대전화 등 자원이 한정된 시스템에도 장착될 수 있는 모델 개발이 요구되므로 앞서 기술한 데이타 부족문제와 함께 이러한 시스템에도 사용할 수 있도록 사전 메모리의 크기를 줄이는 방안이 마련되어야 한다.

4. 어절 내 형태소 범주 패턴 기반 띄어쓰기 모델의 구현

4.1 형태소 n-gram을 사용한 띄어쓰기 모델

데이타 부족과 메모리 크기 문제를 모두 해결하기 위하여 본 연구에서 제안하는 모델은 다음과 같은 가정에 기반한다.

정의 1. 형태소는 어절의 직접 구성요소(immediate constituent)이다.[5]

가정 1. 한 어절의 출현 확률은 그 어절에 속한 각

4) 부산대학에서 개발된 형태소 분석기

5) 직접구성요소란 언어학에서 둘 이상의 형태소로 이루어진 구성체를 일차적으로 나누었을 때 나뉘어 나온 각각의 요소를 지칭한다.

표 2 PNU 말뭉치로부터 추출한 통계와 PNU 형태소 분석기에 사용된 통계

		총 개수	메모리 크기	
PNU 말뭉치	어절 unigram	33,643,884		
	어절 unigram 타입	1,950,068		25.1 MB
	음절 bigram	90,235,529		
	음절 bigram 타입	391,732	+ 띄어쓰기 태그	4.1 MB
	음절 trigram	84,239,729		
	음절 trigram 타입	5,116,404	+ 띄어쓰기 태그	63.7 MB
	형태소 unigram	68,771,225		
	형태소 unigram 타입	500,920		5.4 MB
PNU 형태소 분석기	형태소 사전 내의 형태소 표제어	381,443	+ 형태소제약규칙	4.3 MB

형태소의 출현 확률로 추정할 수 있다.

가정 1을 적용하기 위해서는 (ㄱ) 형태소 n-gram의 적용 수준을 정해야 하며 (ㄴ) 통계 어절 사전을 제거하고 난 후, 후보 어절의 출현 확률을 어떻게 구할 것인지를 결정해야 한다. 음절은 bigram 사용에 그리 큰 어려움이 없었지만, 형태소는 그 종류가 어절에 비해서는 적지만, 음절에 비해서는 엄청난 양이므로 형태소 unigram이상은 적용하기 힘들다. 따라서, 본 논문은 **형태소가 각각 독립적으로 출현한다**는 가정 하에 형태소 unigram을 사용한다.

형태소의 확률은 어절의 확률을 측정한 방법과 같이 학습 말뭉치에서 출현하는 형태소의 상대 빈도를 이용하여 측정하며 이는 식 (7)과 같다.

$$p(m)=\frac{f(m)}{Tm}$$

Tm = 학습 말뭉치 내에서 추출한 총 형태소 개수

$f(m)$ = 학습 말뭉치로부터 추출한 형태소(m)의 빈도 (7)

정의 1과 가정 1에 기반하여 어절 w_k의 출현확률은 대상이 되는 어절 내의 모든 형태소의 출현 확률의 곱에 비례한다. 이는 다음 식 (8)과 같다.

$$w_k = m_{k,1}\dots m_{k,\mu},$$

$$p(w_k)\approx\prod_{i=1}^{\mu}p(m_{k,i})$$

$$\approx\prod_{i=1}^{\mu}\frac{f(m_{k,i})}{Tm}$$

μ = 총 형태소 개수.

w_k = 문장 내의 k 번째 어절

$m_{k,i}$ = k번째 어절내의 i 번째 형태소 (8)

$f(m_{k,i})$ = k번째 어절내의 i 번째 형태소의 빈도

사실 식 (8)은 실제 언어의 현상을 반영하지 못한다. 한국어 문장 내에서 어절은 비교적 자유로운 순서로 배열이 될 수 있는 반면, 어절 내에서 형태소의 결합은 제약조건을 가진다. 어절 내 각 형태소는 표 3의 각 형태소 범주에 속하며 이러한 형태소 범주는 한 어절 내에서 일련의 패턴으로 고정되어 있다. 따라서 이러한 언어적 제약 조건을 반영할 수 있는 모델이 필요하다.

4.2 실제 어절과 어절 내의 형태소 범주 패턴 간의 연관관계

각 형태소의 조합 패턴과 그 조합 패턴 내의 개별 요소 간의 결합을 모델링 하기 위하여 정의 1을 어떤 어절 w_k와 형태소 범주와의 상관관계를 명시하는 정의 2로 발전 시킨다.

정의 2. 한 어절은 하나 혹은 여러 개의 형태소로 구성되며, 형태소의 구성은 특별한 범주 패턴을 이루고 있다.

정의 2를 기반으로 본 연구에서는 단일 형태소가 한 어절이 구성하는 경우는 제외하고 2개 이상의 형태소가 한 어절을 구성할 때 어절과 형태소의 관계를 다음과 같이 가정하였다.

가정 2. 어떤 형태소는 어떤 범주 패턴 내의 한 범주에 특정한 가중치를 가지고 속하게 된다.

정의 2와 가정 2에 의해서 어떤 어절의 출현 확률은 그 어절을 이루는 형태소들의 unigram과 해당 형태소가 어떤 범주 패턴에 속해서 어절을 구성할 때의 기여 가중치를 기반으로 계산할 수 있다. 본 논문에서는 이를 위하여 어절을 구성하는 **범주 패턴 내에 각 형태소가 속한 범주의 기여도**를 **기여 가중치**로 두고 학습을 이용하여 이 가중치를 얻어낸다.

4.3 범주 패턴 기반 모델을 위한 기본 개념

이 절에서는 더 자세한 논의에 앞서 한국어의 형태소 범주의 일반적인 특성에 대한 몇 가지 정의를 하겠다. C를 한국어의 일반적인 문법 범주의 집합이라 두면, 본 연구에서 쓰이는 한국어 형태소 범주 C의 전체 항목은 표 3과 같다.

표 3의 '기본 범주'는 일반적으로 사용되고 있는 한국어 형태소들의 범주를 정리한 것이며 '학습 범주'는 본 논문에서 범주 패턴 학습을 위해 기본 범주를 묶거나 그대로 사용한 것이다. '기본 범주'를 대부분 '학습 범주'로 사용하며, 몇몇 경우는 한국어의 특성과 휴리스틱에 기반하여, 아래와 같이 '기본 범주'를 묶어 일반화하여 '학습 범주'로 사용하였다.

표 3 한국어 형태소의 기본 범주 및 '학습 범주'

기본 범주	학습 범주	Tag	기본 범주	학습 범주	Tag
명사	명사	N	호격 조사	〃	PP
기수사			접속격 조사		
의존명사	의존명사	BN	보조격 조사		
수의존 명사			종결형 어미	어미	END
인칭대명사	대명사	PN	연결형 어미		
지시대명사			인용형 어미		
서수사	서수사	NO	명사형 어미	명사화어미	ETN
동사	동사	V	관형형 어미	관형어화어미	ETM
형용동사	형용동사	VA	선어말 어미	선어말어미	EPP
보조동사	보조동사	VX	일반접두사	별도의 범주로 분석하지 않음: 실질형태소 범주에 함께 포함시킴	
일반관형사	일반관형사	MD	수 접두사		
수관형사	수관형사	MDQ	일반 접미사		
부사	부사	ADV	수 접미사		
감탄사	감탄사	III	복수 접미사		
주격 조사	조사	PP	동사화 접미사	동사화 접미사	STV
목적격 조사			형용사화 접미사	형용사화 접미사	STA
관형격 조사			지정사	지정사	COP
부사격 조사					

- 선어말어미(-시-, -옵-/-습-, -었-/-았-, -겠-, 등), 관형형어미, 명사형 전성 어미를 제외한 모든 어미는 하나의 학습 범주로 묶는다. 관형형어미와 명사형 전성 어미는 특정 범주를 다른 범주로 파생시킬 수 있으므로 특별히 '학습 범주'로 독립 시킨다.
- 형용사화 접미사와 동사화 접미사를 제외한 나머지 모든 접두사와 접미사는 기본 범주로 취급하지 않고 기본 범주에 묶어서 취급한다.
- 동사와 형용사는 유사한 형태/통사적인 특성을 보인다 해도, 결합할 수 있는 관형형어미가 범주마다 다르다는 것과 같은 분명한 활용상의 차이를 보이므로 한 학습 범주로 묶지 않았다. 예를 들어 관형형 어미, '-는'은 형용사 어근과는 결합하지 않는다.

표 3의 '학습 범주' 하나하나는 집합 C의 원소가 된다.

$$C=\{c_1,\cdots,c_\alpha\},\ mc_{k,i}\in C$$
$mc_{k,i}$ = k번째 어절(w_k) 내의 i번째 형태소의 학습 범주 (9)

CP를 한국어에서 발견되는 가능한 범주 패턴의 집합이라고 하자. 어절 w_k는 형태소 $m_{k,1}\cdots m_{k,\mu}$로 구성되며, 각 형태소 $m_{k,i}$는 각각 어떤 범주 $mc_{k,i}$에 속하며, 범주 패턴 $mc_{k,1}\cdots mc_{k,\mu}$을 이루고 있다. 따라서 단어의 범주 패턴 $mc_{k,1}\cdots mc_{k,\mu}$은 CP의 원소중의 하나인 cp_j라고 볼 수 있다. 본 연구에서 취급하는 범주 패턴, 즉 CP의 원소는 표 4와 같다.

$$CP=\{cp_1,\cdots,cp_\beta\},\ cp_j=cp_{j,1}\cdots cp_{j,\varepsilon},$$
$$if\ (w_k=m_{k,1}\ldots m_{k,\mu}\ \text{and}\ mc_{k,1}\cdots mc_{k,\mu}=cp_j)$$
$$then\ mc_{k,i}=cp_{j,i}$$

cp_j = k번째 어절(w_k)을 형성하는 형태소범주:

$$mc_{k,1}\cdots mc_{k,\mu} \quad (10)$$

4.4 형태소 unigram과 범주 패턴 기반 띄어쓰기 모델

가정 2와 식 (10)에 의해서 어절은 특정 범주 패턴 cp_j에 속하는 형태소의 연속으로 이루어져 있으며, 각 형태소는 어떤 범주 패턴에 특정 가중치로 출현한다. 따라서 특정 범주 패턴 하에 특정 가중치와 각 형태소의 출현 확률에 기반한 어절의 확률 추정은 실제 단어의 확률에 비례하며 이는 다음과 같은 식 (11)로 나타낼 수 있다.

$$p(w_k)\approx\prod_{i>1}^{\mu}p(m_{k,i})^{wcp_{j,i}}$$
$wcp_{j,i}$ = 범주 패턴 cp_j 내에서의 개별 범주 $cp_{j,i}$ 의 가중치 (11)

식 (2)의 어절 출현 확률은 식 (12)의 범주 가중치를 적용한 형태소 출현 확률로 대체된다. 따라서 어절 unigram 기반 띄어쓰기 모델은 식 (13)의 범주 패턴 기반 띄어쓰기 모델로 다시 쓸 수 있다.

$$EP(cw_k)=\begin{cases}if\ mc_{k,1}\cdots mc_{k,\mu}=cp_j & \prod_{i=1}^{\mu}p(m_{k,i})^{wcp_{j,i}}\\ otherwise & 1/Tw\end{cases} \quad (12)$$

$$\text{최적의 문장}=\arg\max_{S}\sum_{k=1}^{n}\{\log(EP(cw_k))+\log(odds(\sigma_{k,\lambda},\ \sigma_{k+1,1}))\} \quad (13)$$

범주별 결합 가중치를 반영하기 위하여 형태소 출현 확률에 범주별 결합 가중치를 승수로 두었으며, 이에 관하여는 다음 절에서 논하도록 하겠다. 그림 1은 위 식

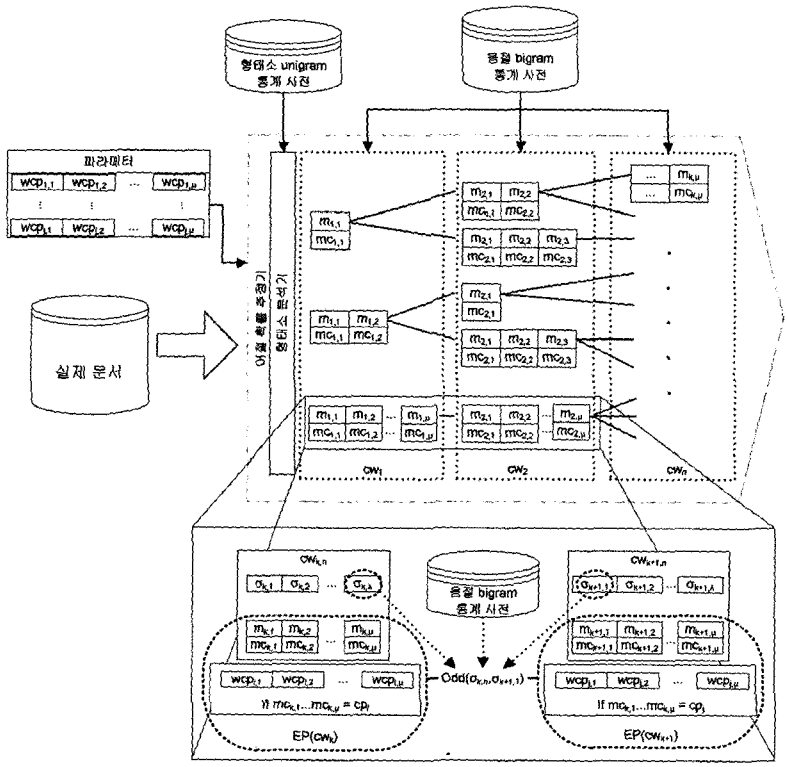


그림 1 범주 패턴 기반 어절 띄어쓰기 처리 과정

(12)와 (13)을 사용한 범주패턴 기반 띄어쓰기 모델의 처리 과정을 도식화한 것이다.

5. 파라메터 학습

5.1 시뮬레이티드 어닐링(simulated annealing)을 이용한 파라메터 학습

파라메터를 설정하기 위하여 각 형태소 범주 패턴별 어절 리스트(CPWL)를 형태소 분석 정보가 있는 PNU 말뭉치로부터 추출하였다. 추출된 형태소 범주는 총 383개 이다. 학습의 효율성을 위하여 추출된 데이타 중 형태소 범주별 어절 리스트 샘플 데이타(SCPWLs)를 추출하였다. 이를 위하여 각 형태소 범주를 어절의 출현 빈도로 정렬하였으며, 샘플 데이타로 설정된 어절의 출현 빈도가 1,000개 미만이면, 모든 데이타를 사용하며, 1,000개에서 10,000개미만이면, 그 중 1,000개를 균등하게 뽑았으며, 10,000개 이상이면 10개 간격으로 한 개씩 균등하게 뽑아서 사용하였다. 이렇게 설정된 샘플 데이타로 최적 파라메터를 얻기 위하여 시뮬레이티드 어닐링(simulated annealing) 알고리즘을 사용하였다. 시뮬레이티드 어닐링 알고리즘은 힐 클라이밍(hill climbing) 알고리즘의 단점을 보완하여, 학습 초기에 나쁜 결과가 제시되는 쪽도 선택함으로써 결과가 국소 최대값들(local maxima)에 빠지는 것을 방지할 수 있다. 그림 2는 파라메터를 학습하기 위한 전체 처리 과정을 도식화 한 것이다.

시뮬레이티드 어닐링에 쓰이는 평가함수와 학습 모형은 (14)와 같으며, 최종적으로 평균 에러를 최소화하는 파라메터, $wcp_{j,i}$가 선택된다. 에러는 (ㄱ) 실제 관찰된 단어의 확률과 (ㄴ) 형태소와 형태소 범주 패턴으로 추정된 단어의 확률 간의 차이의 절댓값으로 측정하였다.

$$\text{평균 에러}(cp_j)=\left[\sum_{\gamma=1}^{Tcp_j}\left|p(w_\gamma)-\prod_{i=1}^{\mu}p(m_{\gamma,i})^{wcp_{j,i}}\right|\right]\Big/Tcp_j$$

$$if\ mc_{\gamma,1}\cdots mc_{\gamma,\mu}=cp_j, \arg\min_{wmcp_{j,i}}\left(\left[\sum_{\gamma=1}^{Tcp_j}\left|p(w_\gamma)-\prod_{i=1}^{\mu}p(m_{\gamma,i})^{wcp_{j,i}}\right|\right]\Big/Tcp_j\right)$$

Tcp_j = 학습에 사용한 범주패턴(cp_j)의 총 개수 (14)

표 4는 범주 패턴 중 가장 빈번하게 나타나는 상위 37개 패턴의 파라메터 설정 결과이다. 이 37개의 패턴은 전체 단어 빈도의 99%를 차지한다.

학습의 편의성과 학습 결과의 적용을 위하여 한국어의 특성을 고려한 몇 가지 휴리스틱을 설정하였다. 이는 전체 시스템의 정확도에 영향을 미치지 않는 범위 내에서 학습을 위한 대상 범주 패턴의 수를 줄이며, 시스템

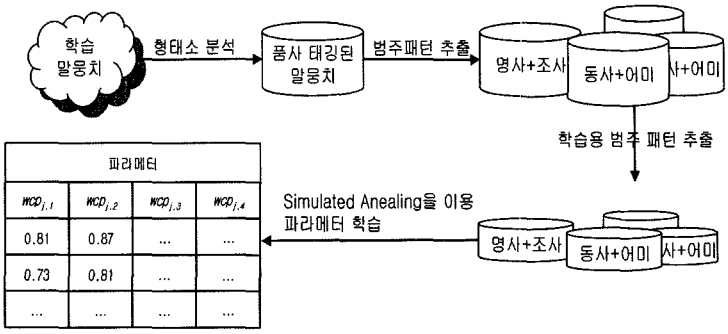


그림 2 파라메터 학습 과정

표 4 각 어절내 범주 패턴의 최적 파라메터

범주 패턴 (cp_j)	파라메터				평균 에러	에러 표준 편차
	$wcp_{j,1}$	$wcp_{j,2}$	$wcp_{j,3}$	$wcp_{j,4}$		
N + PP	0.81	0.87			2.94e-07	3.42e-06
V + END	0.73	0.81			7.02e-07	1.16e-05
V + ETM	0.77	0.62			1.68e-06	2.06e-05
MDQ + BN	0.51	0.73			4.12e-06	8.10e-05
N + STV + END	0.64	0.38	0.70		2.58e-07	1.80e-06
N + STV + ETM	0.72	0.32	0.53		6.43e-07	2.89e-06
BN + PP	0.80	0.52			2.58e-06	3.82e-05
V + EPP + END	0.58	0.44	0.62		5.18e-07	1.43e-05
VA + END	0.75	0.63			9.97e-07	1.85e-05
VX + END	0.62	0.73			4.58e-06	7.96e-05
VA + ETM	0.75	0.48			4.08e-06	4.36e-05
MDQ + BN + PP	0.50	0.64	0.50		7.64e-07	8.33e-06
N + COP + END	0.72	0.00	0.71		1.22e-07	2.43e-06
N + STV + EPP + END	0.58	0.00	0.49	0.62	2.82e-07	3.59e-06
PN + PP	0.69	0.69			3.20e-06	2.96e-05
VX + ETM	0.66	0.56			2.33e-05	1.06e-04
N + COP + ETM	0.81	0.00	0.38		1.67e-07	1.34e-06
N + END	0.68	0.71			1.02e-07	9.51e-07
BN + COP + END	0.65	0.00	0.63		2.29e-06	3.66e-05
N + STA + ETM	0.69	0.00	0.56		1.20e-06	6.89e-06
VX + EPP + END	0.38	0.41	0.65		1.77e-06	1.75e-05
N + STA + END	0.70	0.00	0.65		3.07e-07	2.14e-06
N + ETM	0.96	0.00			1.22e-07	1.47e-06
V + EPP + ETM	0.59	0.34	0.47		2.70e-07	1.92e-06
V + ETN	0.79	0.33			3.60e-07	1.79e-06
VA + EPP + END	0.56	0.27	0.58		3.51e-07	4.39e-06
V + ETN + PP	0.53	0.35	0.59		1.29e-07	9.01e-07
N + COP + EPP + END	0.69	0.00	0.00	0.61	5.40e-08	3.92e-07
N + STV + EPP + ETM	0.60	0.26	0.20	0.37	1.04e-07	3.71e-07
N + STV + ETN	0.67	0.38	0.22		1.95e-07	7.69e-07
BN + COP + ETM	0.68	0.00	0.52		1.17e-06	1.28e-05
N + STV + ETN + PP	0.43	0.37	0.37	0.36	4.73e-08	1.31e-07
VX + EPP + ETM	0.43	0.38	0.45		2.16e-06	6.15e-06
VA + ETN + PP	0.58	0.00	0.72		1.84e-07	1.80e-06
VA+ETN	0.77	0.32			4.73e-07	3.78e-06
BN+END	0.67	0.55			3.85e-07	2.68e-06
VX+ETN	0.64	0.41			3.93e-06	1.15e-05

구현의 일관성을 유지하기 위하여 필요하다.

- 어절 내에 단일 형태소 범주를 가진 패턴 *N, ADV, BN, MDQ, MD, PN, III* 은 곧 어절이다. 따라서 이 패턴들은 따로 학습할 필요가 없으므로, 형태소 범주의 결합 가중치를 고려하지 않는다.

  휴리스틱: 범주 패턴의 결합 가중치는 2개 이상으로 이루어진 범주 패턴에만 적용한다. 하나의 형태소가 한 어절을 이룰 경우에는 그 형태소의 출현 확률은 어절의 출현 확률을 바로 사용한다.

- <PN + END + EPP + ETM >, <V + END + EPP + ETN> and <N + COP + ETN + EPP + ETM>와 같이 전체 학습 코퍼스에서 드물게 출현하는 범주 패턴(본 논문에서는 50개 미만)은 파라메터 학습이 어려우므로 고정 값으로 둔다.

  휴리스틱: 학습이 어려운 패턴에 속하는 어절은 학습 말뭉치 중 한번 나왔다고 가정하여 1/33,643,884 (총 학습 어절 수)의 값을 둔다.

- 한국어 띄어쓰기 규정에 따르면 복합어, 즉 복합명사, 복합용언은 띄어 쓸 수도 있고, 붙여 쓸 수도 있다.

  휴리스틱: 복합어 패턴은 학습하지 않고 나눠진 단위로 취급한다.

5.2 학습 결과

그림 3을 보면 각 형태소의 출현 빈도가 지프의 법칙(Zipf's law)을 따르며 어절도 지프의 법칙을 따른다는 것을 알 수 있다. 더불어 실질 형태소 범주인 명사, 동사와 형식 형태소 범주인 조사, 어미는 다른 분포 양상을 보이는데, 실질 형태소 범주는 어절과 비슷한 분포를 보이는 반면에 형식 형태소 범주는 상위 형태소로 갈수록 빈도가 급격하게 증가하고 있다.

명사 범주는 상위 2.9% 형태소 타입(중복을 제거한 형태소)이 전체 80%의 형태소(토큰, token) 빈도를 차지하는 반면에 조사 범주는 상위 1.3%의 형태소 타입이 전체 80%의 형태소 빈도를 차지하고 있다. 이러한 형태소 범주들은 형태소의 범주의 결합으로 하나의 어절을 형성할 때 차이나는 결합 가중치를 가지게 된다. 이 가중치는 어절 내의 가능한 범주 패턴에 대한 학습으로 얻어지며 형태소 출현 확률에 적용되어 어절의 출현 확률을 추정하는데 이용된다. 그림 4(a)는 각 범주 패턴의 에러 평균과 에러 표준 편차를 보인 것이다. 그림 4(b)는 각 범주패턴 내의 어절 타입별 평균 어절 빈도이며, 그림 4(c)는 어절 빈도 점유율이 80%에서 95%로 변함에 따른 어절 타입의 점유율의 변화를 나타낸 것이다. 다른 결과와의 일관성 있는 경향 표현을 위하여 그림 그림 4(a)의 표준 편차는 편의상 실제 값을 10으로 나눈 것이며 그림 4(c)에서 어절 타입 점유율은 역으로 제시되었다.

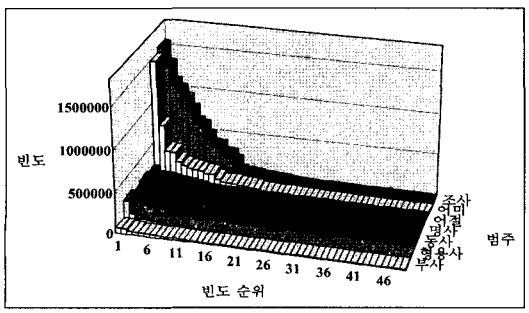


그림 3 범주별 형태소 및 어절 빈도/순위 분포

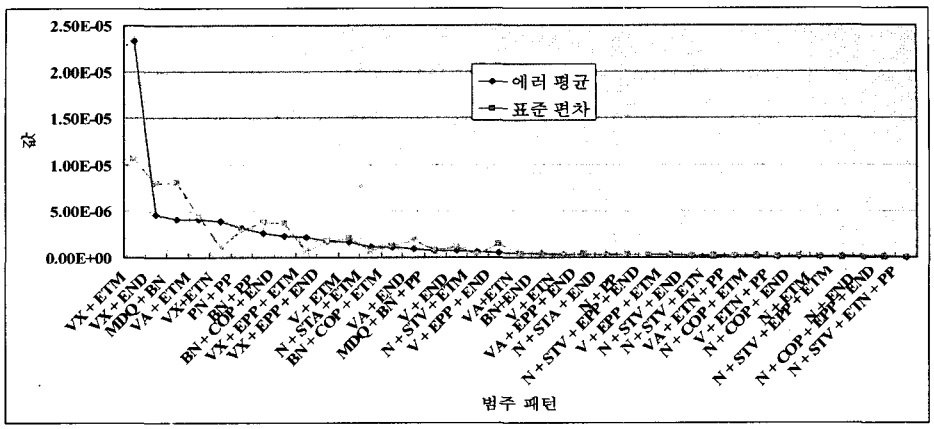
그림 4의 결과로부터 평균 에러와 표준 편차가 클수록 어절 타입의 점유율과 어절수/어절 타입수의 비율이 높다는 것을 관찰할 수 있다.

- 범주 패턴 <VX + ETM>와 <N + PP>에서는 각각 상위 15.89%의 어절 타입과 38.73%의 어절 타입이 각 범주 패턴별 어절 샘플 데이타(SCPWL) 내에서 나타나는 어절 중 95%를 차지한다.
- 범주 패턴 <VX+ETM>와 <N + PP>에서 범주별 어절수/어절 타입수의 비율은 각각 1069.33개와 13.38개였다.

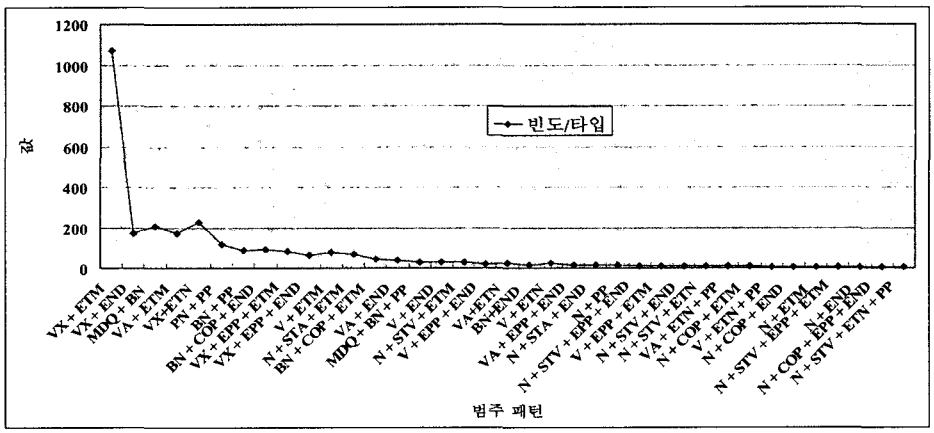
이와 같은 관측결과로부터 패턴 내에서의 각 범주의 기여도를 의미하는 가중치 $wcp_{j,i}$가 효력을 발하고 있음을 볼 수 있다. 끝으로, 그림 5의 여러 도표는 범주 패턴 별로 실제 관측된 어절의 확률과 수식 (11)을 적용한, 즉 형태소 unigram과 범주 패턴 내에서의 해당 형태소의 범주의 가중치를 기반으로 추정한 어절 확률 간의 차이를 보여준다. 그림 5(a), (b), (e)에서는 각 범주 패턴 별 전체적인 경향을 따르지 않는 특이 항목을 관찰 할 수 있다. 이를 조사해 본 결과 그림그림 5(a)는 이런 실제 어절확률을 추출했던 학습 말뭉치에 시사를 다루는 신문 데이타가 많이 포함되어 있기 때문에'대통령'과'기자'와 같은 명사를 포함하는 어절이 특이항목으로 출현하였다. 그림 5(b)에서는 '것', '때문'과 같은 높은 빈도로 나타나는 의존 명사를 포함하는 어절이 이런 현상을 보이고 있으며, 그림 5(e)에서도 높은 빈도로 나타나는 형용사인 '같다', '있다'를 포함하는 어절이 그러하다.

6. 성능평가

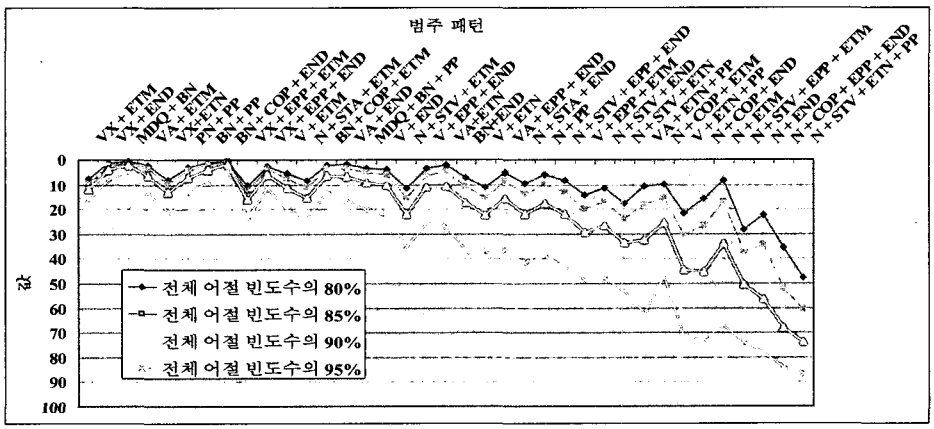
이 장에서는 어절 내 형태소 범주 패턴 기반 모델을 적용하여 구현한 띄어쓰기 시스템의 성능을 측정하고 결과를 분석한다. 본 연구에서는 띄어쓰기의 성능 측정을 위하여 공백이 제거된 성능평가 데이타를 이용한다. 성능평가 데이타는 기존연구[6]에서 사용한 데이타 집합(표 5의 내부1과 외부1, 2)과 추가로 수집한 데이타 집



(a)



(b)



(c)

그림 4 (a) 에러 평균 및 표준편차, (b) 범주별 어절수/어절 타입수의 비율, (c) 80%, 85%, 90%, 95% 어절 중 어절 타입의 점유율

합(표 5의 외부3)으로 구성되어 있다. 본 연구에서 제시한 모델은 후보 어절 사이의 띄어쓰기 여부를 평가하는 것으로 입력 문장의 띄어쓰기를 고려하지 않는다. 이는 이전 연구들의 띄어쓰기에 대한 접근 방법과 유사한 것으로, 입력 문장의 띄어쓰기가 100% 틀렸다는 가정하에 띄어쓰기 문제를 후보위치에서의 띄어쓰기 여부로 단순

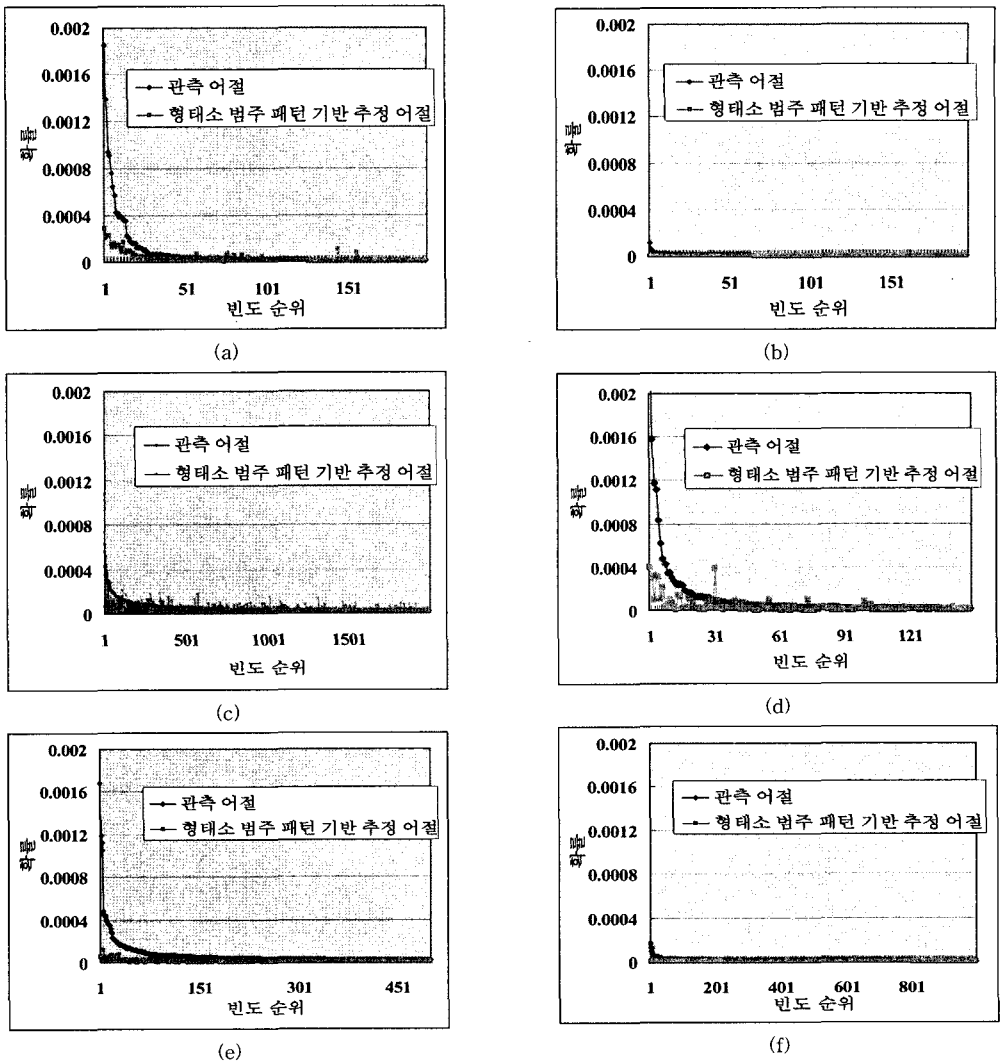


그림 5 실제 관측된 어절 확률과 형태소범주 패턴에 기반하여 추정한 어절 확률 비교: (a) <N+PP>, (b) <BN+PP>, (c) <V+END>, (d) <N+STV+END>, (e) <VA+ETM>, (f) <V+EPP+ETM>

화한 것이다. 따라서, 부분적인 띄어쓰기 평가에서 고려하는 입력 문장 내의 띄어쓰기에 대한 오류 인식 및 오류 수정 정확도의 평가와 부합하지 않는다. 더욱이 앞에서 언급한 바와 같이 띄어쓰기 문제를 단순화하면 띄어쓰기 알고리즘의 적용 범위가 넓어지는 이점이 있다. 예를 들어 띄어쓰기가 거의 되어 있지 않은, 음성 인식의 결과로 만들어지는 텍스트나 길이의 제약을 받는 문서 작성 환경(100자평, 휴대 전화의 문자 메시지 서비스)에서 응용이 용이하다.

그러함에도, 본 논문에서 제안하는 모델은 공백이 제거된 극단적인 형태의 문장뿐만 아니라 일반적인 문장에도 적용할 수 있으므로 그 결과가 일반적으로 사람들이 작성한 문장의 띄어쓰기 정확도보다는 높아야만 할 것이다. 하지만, 표 5에서 제시한 기존의 ETRI 품사 태그 부착 말뭉치와 21세기 세종 말뭉치는 코퍼스는 일반인들의 띄어쓰기 정확도를 살펴보는데 부적합하다고 판단된다. 따라서, 실제 사용자들이 여타의 다른 정제과정을 거치지 않고 생성한 텍스트와 교열전문가 및 교열시스템 등에 의해 정제과정을 거쳤을 것으로 추정이 되는 신문데이타를 적절히 조합하여 외부3과 같은 성능평가 데이타 집합을 설정하였다.

성능의 측정 단위는 어절을 기본 단위로 하며, 시스템의 출력 결과가 얼마나 정확한 어절을 추출해 내느냐를 측정한 **어절 단위 정확도**(Pw)와 실제 정답 어절이 얼

표 6 알고리즘 (13)를 적용한 범주 패턴 기반 모델의 성능 (%)

		ㄱ (*with SWD*)	ㄴ ($wcp_{j,i}$ = 1; *Odds* () = 1)	ㄷ ($wcp_{j,i}$ = 1; *Odds* ())	ㄹ ($wcp_{j,i}$; *Odds* () = 1)	ㅁ ($wcp_{j,i}$; *Odds* ())
내부	*Pw*	96.27	88.94	94.46	96.07	97.57
	Rw	94.64	91.67	95.53	96.09	97.49
	Fw	95.45	90.28	94.99	96.08	**97.53**
외부1	*Pw*	89.52	87.26	92.90	96.24	97.53
	Rw	92.99	91.28	94.88	96.36	97.48
	Fw	91.21	89.23	93.88	96.30	**97.50**
외부2	*Pw*	87.89	88.06	93.29	96.70	98.11
	Rw	92.12	92.21	95.54	97.08	98.19
	Fw	89.96	90.08	94.40	96.89	**98.15**
외부3	*Pw*	91.50	87.38	93.22	95.35	96.26
	Rw	92.08	91.04	94.74	95.65	96.33
	Fw	91.79	89.17	93.97	95.50	**96.29**
외부 평균	*Pw*	89.64	87.57	93.14	96.10	97.30
	Rw	92.40	91.51	95.05	96.36	97.33
	Fw	90.99	89.49	94.08	96.23	**97.31**

표 5 성능평가용 데이타

		문장수	어절수
내부		2,000	25,020
외부1		2,000	13,971
외부2		2,000	17,191
외부3	방송스크립트	600	7,054
	신문데이타	600	8,217
	'100자평'	600	4,304
	합계	1,800	19,575

(ㄱ) 내부 = 전체 학습 말뭉치에서 각 학습 말뭉치가 차지하는 비율에 따라 균등한 비율로 문장을 추출함; (ㄴ) 외부 1 = 21세기 세종 말뭉치로부터 무작위로 문장을 추출함[11]; (ㄷ) 외부 2 = ETRI 품사 태그 부착 말뭉치로부터 무작위로 문장을 추출함[12]; (ㄹ) 방송스크립트 = K방송사의 인터넷 홈페이지에서 제공하는 방송스크립트를 무작위로 추출함; (ㅁ) 신문데이타 = 인터넷에서 서비스되고 있는 5개의 한국 일간지의 기사에서 무작위로 문장을 추출함; (ㅂ) '100자평' = 인터넷에서 제공되고 있는 일간지 홈페이지에서 운영하는 논평 게시판에서 무작위로 문장을 추출함.

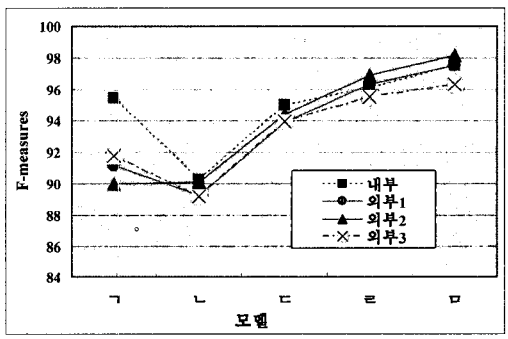


그림 6 파라메터 값 변화에 따른 성능 (어절 단위 F-measure) 변화

마나 정확하게 추출되었느냐는 **어절 단위 재현율**(*Rw*)을 제시한다. 더불어, 띄어쓰기를 많이 하는 시스템일수록 재현율이 높아지며 적게 하는 시스템일수록 정확도가 더 높아지므로 띄어쓰기 시스템의 정확한 성능 평가를 위해서, 이 둘을 합한 **어절 단위 f-measure** (*Fw*)를 종합적인 성능평가로 제시하겠다. 아래 식은 정확도와 재현율을 이용하여 *Fw*를 구하는 식이다.

$$Fw = 2 * (Rw * Pw) / (Rw + Pw) \qquad (15)$$

범주 패턴 기반 통계적 띄어쓰기 기법의 결과는 다음 표 6과 같다.

본 연구에서 제시하는 모델로 구현한 시스템이 자료 부족 문제와 메모리의 크기 문제를 효과적으로 해결했음을 표 6과 그림 6으로 확인할 수 있다. 어절 사전(SWD)을 적용하는 이전 모델을 사용했을 때 외부 데이타보다 내부 데이타에 대해 월등히 높은 성능을 보였던 것에 반해, 본 연구가 제안한 모델을 적용 했을 때 내부 데이타나 외부 데이타에 동시에 높은 성능을 보임을 알 수 있다. ㄴ 모델에서 범주 가중치와 *odds*값을 1로 둔 것은 형태소의 확률만 사용했음을 의미한다. 모델 ㄴ 에서는 범주 가중치를 사용하지 않았지만, 어절 사전만 사용한 ㄱ 모델 보다 결과가 일부 데이타에서는 향상되는 것을 볼 수 있는데, 그 이유는 형태소 분석기가 문장을 분석해 나갈 때 부분적으로 명확한 어절의 경계까지만 분석하며, 불명확한 어절 경계는 취급하지 않는 특성이 반영된 것이다. 어절 내의 범주 패턴 내에서 특정 형태소의 범주 기여도를 반영한 가중치가 성능에 미치는 영향은 ㄴ 모델과 ㄹ 모델의 비교를 통해 알 수 있으며, ㄹ 모델이 ㄴ 모델에 비해 *Fw*기준 6.74%의 성능 향상을 보였다. 최종 성능은 *Fw* 기준 97.31% 이며, 어절 unigram 만 사용한 모델에 비하여 6.33% 나은 성능을 보였다. 표 6에서 외부 3의 띄어쓰기 정확도가 현저하게 낮은 이유는 인터넷 상의 최신 용어나 압축된 표현이 많이 사용된 '100자평'이 포함되어 있기 때문[6]이다.

실제 사용자들의 띄어쓰기와 본 논문에서 제안한 모델의 비교를 위하여 표 5에 있는 외부 3 데이타의 각 항목의 실제 띄어쓰기 정확도와 각 항목을 본 시스템에 적용한 결과를 표 7에서 제시하였다. '100자평'이 다른 항목에 비해서 띄어쓰기 정확도가 현저히 낮은 이유는 '100 자평'의 특성상 사용자가 자신의 생각을 한정된 공간에 압축해서 기술해야 하므로, 의도적으로 띄어쓰기를 잘 지키지 않기도 하고, 짧은 문장의 경우는 모든 문장을 붙여 쓰기도 하기 때문이다. 반면, 신문데이타는 신문사 자체에 기사를 작성을 보조하기 위한 교열시스템을 갖추고 있는 경우가 많으므로 실제 띄어쓰기 정확도가 상당히 높은 것을 볼 수 있다. 이런 여러 데이타의 속성에 따라 본 모델의 적용 결과에 따른 개선 정도가 차이를 보이지만 전체적으로 향상됨을 볼 수 있다.

표 7 실제 사용자들의 띄어쓰기와 본 모델의 띄어쓰기 정확도 비교 (%)

	'100자평'	방송스크립트	신문데이타
실제 사용자들의 정확도	76.08	92.25	95.18
본 모델의 정확도	94.23	95.52	98.15
개선 정도	18.15	3.27	2.97

7. 결론 및 향후 과제

본 연구에서는 형태소 unigram을 이용한 범주 패턴 기반 한국어 자동 띄어쓰기 모델을 제안하였다. 이를 위하여 (ㄱ) 실제 관측된 어절의 출현 확률과 (ㄴ) 형태소와 범주 패턴을 이용하여 추정한 후보 어절의 출현 확률을 간의 차이를 최소화 하는 방향으로 범주 패턴 내 범주 가중치를 학습하여 적용하였다. 그 결과 형태소 unigram의 범주 패턴별 결합특성으로 인하여 순수 어절 unigram모델에 비하여 자료 부족 문제를 효율적으로 해결하였고, 사전 크기를 크게 줄였으며, 띄어쓰기 성능을 대폭 향상 시켰다. 이와 더불어 어절을 형태소 범주 패턴에 따라 나누었을 때 형태소 범주가 한국어 어절 형성에 일정한 기여도를 가지도 참여함을 확인 할 수 있었다. 따라서, 본 논문에서 제안한 모델은 어절의 구성과 직접적인 관련이 있는 형태소 범주 결합 패턴을 이용해서 어절을 구분하므로 보다 직관적이다.

본 연구에서는 범주 패턴내의 각 범주의 가중치를 학습하였다. 그 결과 학습 데이타에 출현하지 않은 어절에 대한 처리 능력을 향상시켰음에도 불구하고, 형태소 각각의 특성을 반영하여 가중치를 학습하지 않음으로 인하여 개별 형태소의 특성을 골고루 고려하지는 못했다. 즉, 같은 범주 패턴 내에 속하는 모든 형태소에는 동일한 가중치를 부여했기 때문에 본문의 그림 5(a), (b), (e)에서와 같이 범주 패턴 내에서 고빈도로 나타나는 형태소의 특성을 제대로 반영하지 못하여 에러평균을 높이는 현상을 가져왔다. 이런 형태소들과 같이 극도로 높은 빈도로 나타나는 형태소를 포함하는 어절은 이런 형태소를 별도의 범주로 학습하는 등의 처리 방안을 연구해야 하겠다.

6) 자세한 사항은 표 7 참조

참 고 문 헌

[1] 강승식, "음절 bigram를 이용한 띄어쓰기 오류의 자동 교정", 음성과학회 논문지, 8권 2호, pp. 83-90, 2001.

[2] 신호철, "형태소 분석기를 이용한 자동 띄어쓰기 시스템 구축에 대한 연구", 한국어학, 12권, pp. 167-186, 2000.

[3] 심광섭, "음절간 상호 정보를 이용한 한국어 자동. 띄어쓰기", 정보과학회논문지: 소프트웨어 및 응용, 23권 9호, pp. 991-1000, 1996.

[4] 심철민, 권혁철, "연어 정보에 기반한 한국어 철자 검사와 교정기의 구현", 정보과학회 논문지: 소프트웨어 및 응용, 23권 8호, pp. 776-785, 1996.

[5] 이도길, 이상주, 임희석, 임해창, "한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델", 정보과학회 논문지: 소프트웨어 및 응용, 30권 4호, pp. 358-370, 2003.

[6] Kang, M.Y., Choi S.W. and Kwon, H.CH., "A Hybrid Approach to Automatic Word-spacing in Korean," Lecture Notes in Computer Science (LNCS) Vol.3029, pp. 284-294, 2004.

[7] Kang, S.S. and Woo C.W., Automatic Segmentation of Words Using Syllable Bigram Statistics. Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, pp. 729-732, 2001.

[8] Kim, S.N., Nam, H.S. and Kwon, H.CH., "Correction Methods of Spacing Words for Improving the Korean Spelling and Grammar Checkers," Proceedings of the 5th Natural Language Processing Pacific Rim Symposium, pp. 415-419, 1999.

[9] Manning, C.D., and Schutze H., "Foundations of Statistical Natural Language Processing," The MIT Press, Cambridge, London, 2001.

[10] Sproat R, Shih, c., Gale, W. and Chang, N. "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," Computational Linguistics, Vol.22 No.3, pp. 377-404, 1996.

[11] 21세기 세종계획 국어기초자료 구축, 문화관광부, 2004.

[12] 한국전자통신 연구원, "ETRI 품사태그 부착 말뭉치 (시험판)", 1999.

강 미 영

1989년 부산대학교 불어불문학과(학사) 1990년 그르노블3대학교 언어과학과(석사). 2000년 파리7대학교 언어학과(박사) 2003년 부산대학교 정보시스템공학과(석사). 2005년 부산대학교 컴퓨터공학과 박사수료.. 관심분야는 자연언어처리, 시멘틱웹, 기계학습

정 성 원

2001년 부산대학교 전자계산학과(학사) 2003년 부산대학교 전자계산학과(석사) 2006년 부산대학교 컴퓨터공학과(박사) 2006년 9월~현재 부산대학교 U-Port 정보기술산학공동사업단 post-doc. 관심분야는 자연언어처리, 정보추출, 정보검색, 기계학습

권 혁 철

1982년 서울대학교 공과대학 전산학(학사). 1984년 서울대학교 공과대학 전산학(석사). 1987년 서울대학교 공과대학 전산학(박사). 1988년~현재 부산대학교 전자전기정보컴퓨터공학부 교수. 1988년~현재 한국어정보과학회 프로그래밍 언어 연구회 운영위원. 1990년~현재 한국어정보과학회 한국어정보처리연구회 운영위원. 1992년~1993년 미국 Stanford 대학 CSLI연구소 연구원. 1992년~1993년 Xerox Palo Alto Research Center 자문위원. 2004년~현재 한국정보과학회 이사. 2006년~현재 한국인지과학회 이사. 관심분야는 자연언어처리, 정보검색, 프로그래밍언어, 인공지능, 시맨틱웹