한국어 특성과 CRFs를 이용한 자동 띄어쓰기 시스템

이현우(창원대), 차정원(창원대)

<차 례>

<Abstract>

Automatic Word Spacing for Korean Using CRFs with Korean Features

Hyun-Woo Lee, Jeong-Won Cha

In this work, we propose an automatic word spacing system for Korean using conditional random fields (CRFs) with Korean features. We map a word spacing problem into a classification problem in our work. We build a basic system which uses CRFs and Eumjeol bigram. After then, we analyze the result of inner-test. We extend a basic system added by some Korean features which are Josa, Eomi and two head Eumjeols of word extracting from lexicon. From the results of experiment, we can see that the proposed method is better than previous methods. Additionally the proposed method will be able to use mobile and speech applications because of very small size of model.

* Keywords: Automatic word segmentation, Conditional random fields (CRFs).

1. 서 론

음성 인식과 자연어 처리에서 자동 띄어쓰기는 가장 기본이 되는 문제이다. 모든 자연어 처리 시스템이 온전한 띄어쓰기를 가정하고 구현되고, 음성 인식 시스템의 성능 저하 문제는 인식 환경에 따른 잡음 문제도 있지만 그에 못지않게 음성 신호를 문자열로 변환하는 과정에서 여러 오류들이 발생한다. 예를 들어 "아버지가 방에 가신다."와 "아버지 가방에 가신다."와 같이 정상적인 인식은 이루어졌지만, 띄어쓰기로 인해 전혀 다른 의미로 해석될 수 있는 문장도 있으며, "사 십"과 "사십"은 각각 "4 10"과 "40"과 같이 전혀 다른 숫자를 가리킬 수 있다. 자연어 처리 시스템에서도 온전한 띄어쓰기가 되어 있지 않으면 만족할 만한 성능을 얻을 수가 없다. 인터넷에 존재하는 많은 정보들이 비전문가에 의해서 생성되고 있어 띄어쓰기 오류 교정은 중요함이 증가하고 있다.

단어를 띄어 쓰는 영어와는 다르게 한국어는 어절을 단위로 띄어 쓴다. 한국어 띄어쓰기의 제 1원칙으로서 맞춤법 제 2항에는 다음과 같이 규정되어 있다[1].

문장의 각 단어는 띄어 씀을 원칙으로 한다.

여기서 단어는 띄어 씀을 원칙으로 하지만 띄어 쓰는 단위는 어절이다. 위의 원칙에 반해 "숫자와 어울리어 쓰이는 경우에는 붙여 쓸 수 있다.", "단음절로 된 단어가 연이어 나타날 적에는 붙여 쓸 수 있다.", "전문 용어는 단어별로 띄어 씀을 원칙으로 하되, 붙여 쓸 수 있다."와 같이 한국어의 띄어쓰기를 어렵게 하는 예외 항목들이 다수 존재한다.

어느 언어에서나 어려운 일이지만, 위와 같은 예외로 인해 한국어의 경우 단어의 경계를 긋는 문제는 유난히 까다롭다. 특히 맞춤법을 철저히 지켜야 하는 방송, 신문에서도 띄어쓰기, 철자 오류, 비표준어 사용, 외래어 표기 등을 포함한 어문 규정 오류가 <표 1>에서와 같이 64~89%를 차지하고 있으며, 그 중에서 띄어쓰기 오류가 가장 많다[2].

그 중에서도 예외 항목이 없는 항목이 존재한다.

조사, 어미는 그 앞말에 붙여 쓴다.

위와 같은 단어의 경계를 확실히 하는 한국어만의 특성을 사용할 경우, 단어의 구별을 조금이나마 개선시킬 수 있다.

현재 띄어쓰기는 품사 부착과 같은 자연어처리 기술에서 많이 사용되는 hidden Markov model (HMM)과 같은 통계를 기반으로 많은 연구가 이루어지고 있다[3]-[7]. 하지만 본 논문에서는 HMM의 단점인 각 입력열의 독립가정을 해결한 conditional random fields (CRFs)와 한국어만의 특성을 반영한 모델을 만들어 띄어

<표 1> 매체별 띄어쓰기 오류의 유형

	조사 어절	오류 어절	오류 비율	오류 유형		
				어문 규정	어휘	문법/ 문장
방송	42,501	2,569	6.0%	89.53%	6.0%	4.48%
신문	86,974	4,896	5.6%	64.26%	15.2%	20.47%
인터넷	59,748	4,283	7.5%	74.25%	9.27%	16.48%

쓰기 문제를 해결하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 자동 띄어쓰기 연구와 관련된 기존의 연구에 대해 논하고, 3장에서는 한국어 특성과 CRFs를 이용한 모델을 설명하고, 4장에서는 실험 및 평가를 하고, 5장에서 결론을 맺는다.

2. 관련 연구

본장에서는 자동 띄어쓰기에 대한 기존 연구들을 살펴본다. 영어에서는 다양한 방법들이 시도되었다. [8]에서는 정규표현(regular expression)을 이용하여 95%의 성능을 보였다. 그렇지만 문장 기호들이 여러 문맥에 나타남으로써 발생하는 규칙들의 겹침을 막을 수는 없었다. [9]에서는 신경망 네트워크(neural networks)와 결정 트리(decision tree), 그리고 사전을 이용한 방법을 제안하였으며, SATZ(독일어로 '문장')로 알려진 이 시스템은 98.9%의 성능을 보였고, 영어가 아닌 다른 언어에도 적용할 수 있다는 장점이 있다. 그렇지만 한국어에서는 좋은 성능을 보이지 못했다[10].

한국에서도 띄어쓰기 오류에 대한 연구가 많이 시도되었다[3]-[7]. 임의의 두 음절 (x_i, x_{i+1})에 대한 연속한 음절 (x_{i-1}, x_i) 앞에 공백을 삽입하는 좌공백, (x_i, x_{i+1}) 사이에 공백이 사이에 삽입될 사이공백, (x_{i+1}, x_{i+2}) 뒤에 공백을 삽입하는 우공백의 확률을 구하여 임계치 0.375를 넘을 경우 공백을 삽입하는 방식이 제안되었다[3].

그리고 음절 간의 네 가지 통계적 상호정보를 이용하여 띄어쓰기를 삽입하는 방법도 있다. 두 음절 xy가 차례로 어절의 시작에 나타날 통계적 상관관계, xy가 어절의 끝에 나타날 상관관계, xy 사이에 띄어쓰기를 할 상관관계, xy 사이에 띄어쓰기를 하지 않을 상관관계를 구해 적절히 가중치를 주어 선형 결합(linear combination)한 결과, 실험으로 정해진 임계치와 비교해 띄어쓰기 삽입여부를 결정한다. 간단한 접근 방법이지만 학습되지 않은 문장 또는 어절이 등장할 경우 띄어쓰기를 제대로 삽입할 확률이 급격히 떨어진다.

[11]은 규칙과 통계적인 방법을 같이 사용하여 문장 기호가 있는 경우와 없는 경우의 문장 분리 실험을 하여 99.21%와 98.04%의 성능을 각각 보였다.

위의 기존 연구는 띄어쓰기 문제를 문자 간의 패턴이나 간단한 통계로 해결하려고 했지만, [5]는 띄어쓰기에서 공백 삽입 문제를 HMM으로 처리하려는 시도를 보였다. 학습단계에서 "자료부족 문제(sparseness data problem)"를 해결하고자 모델의 파라미터 수를 증가시켜 높은 성능을 보였지만, 띄어쓰기 학습 및 처리 과정에서 파라미터를 만드는 과정이 많은 부담이 된다. 그리고 정확도가 높은 모델과 재현율이 높은 모델이 각기 다른 모델을 가리킨다는 점에서 파라미터 확장을 이용한 범용성이 높은 모델을 만드는 것이 힘들다고 할 수 있다.

[5]와 같이 통계적 모델로 자동 띄어쓰기에 접근한 방법들의 공통된 문제점은 "자료부족 문제"이다. [6]은 이전의 연구들에서 bi-gram과 tri-gram이 uni-gram과 four-gram보다 훨씬 높은 성능을 보인다고 말하고 있다. 그래서 자료부족 문제를 해결하기 위해 학습되지 않은 자료가 출현할 경우, n-gram의 윈도우 사이즈 n을 자동으로 조절할 수 있도록 하여, 자료부족 문제를 해결하였다. 윈도우 크기 n은 bi-gram에서 시작하여 최소 uni-gram, 최대 four-gram까지 확장이 가능하며, (n+1)-gram을 사용하기 위해서는 n-gram보다 (n+1)-gram이 보다 높은 성능을 낼 수 있다는 평가가 있어야 가능하다. 그래서 이 두 모델을 비교하기 위해서는 띄어쓰기에 대한 조건부 확률 $P(t_i|x_{n,i})$와 $P(t_i|x_{(n+1),i})$의 문맥이 주어졌을 때, 이 두 분포를 비교하기 위해 Kullback-Leibler divergence[12]를 이용하여, 그 차이가 임계값을 넘으면 (n+1)-gram으로 확대하거나, (n-1)-gram으로 축소할 수 있다. 그러나 모델을 선택할 때, (n-1)-gram과 같이 축소를 먼저 하느냐, (n+1)-gram과 같이 확대를 먼저 하느냐에 따라 정확도의 차이를 보이는 문제점을 가지고 있다.

[7]에서는 [3], [4], [5], [6]과는 좀 더 다른 방향으로 접근을 하였다. 어절을 이루고 있는 형태소가 하나 혹은 여러 개라는 정의에 맞추어, 한 어절의 출현 확률은 그 어절을 구성하는 형태소들의 출현 확률과 그 형태소의 범주 패턴과 상관관계가 있으며, 어떤 형태소는 어떤 범주 패턴 내의 한 범주에 특정 가중치를 가지고 속하게 된다고 가정하였다. 그래서 학습 말뭉치로부터 학습용 범주 패턴을 추출하여 simulated annealing[13]을 이용하여 가중치를 학습하고, 형태소 분석기를 사용하여 형태소를 추출한 후, 형태소 uni-gram을 사용하여 자동 띄어쓰기 시스템에서 높은 성능을 보였다.

3. CRFs를 이용한 자동 띄어쓰기

기존 자동 띄어쓰기 문제는 문자와 문자 사이에 공백의 삽입 여부를 결정하는 문제로 파악하였다. 따라서 기존의 품사 태깅이나 청킹(chunking)과 같은 방법의

접근법을 사용하지 못하고 문자 간의 패턴이나 간단한 통계로 문제를 해결하려는 시도만 있었다. 본 논문에서는 문자에 바로 태그를 붙이는 방법을 채택함으로써 기존의 자연어 처리 방법을 그대로 적용할 수 있다. 예를 들어 문장 "우리는 우리 말을 사랑한다."의 경우에는 "우/W 리/I 는/I 우/W 리/I 말/I 을/I 사/W 랑/I 한/I 다/I ./I"와 같이 표기된다.

먼저 논문에서 사용한 용어에 대해서 정의한다. 문장 X는 문자 x들의 벡터로 정의된다. <그림 1>에 요약되어 있다. 여기서의 문자는 우리말의 음절, 영문자, 한자, 기호 등을 포함하여 이른다. 예를 들어 '가', 'a', '韓', '!' 등이다.

<그림 1> 용어 정의

$X=< x_1,x_2,\cdots,x_n >$

$T=< t_1,t_2,\cdots,t_n >$

$t=\{W,I\}$

W: 어절 처음, 문장 시작

I: 어절 중간/끝, 문장 끝

3.1. Conditional Random Fields

CRFs는 조건부 확률을 최대로 하는 방향성이 없는 그래프 모델이다[14]. 입력열 $X=x_1x_2\cdots x_n$, 상태열 $T=t_1t_2\cdots t_n$가 주어졌을 때, CRFs에서는 조건 확률로 식 (1)과 같이 정의된다.

$$P(X|T)=\frac{1}{Z_t}\exp\left(\sum_{i=1}^{n}\sum_{k}\lambda_k f_k(x_{i-1},x_i,t,i)\right) \quad (1)$$

여기서 Z_x는 확률값으로 만들어 주는 정규화 값이고 $f_k(x_{i-1},x_i,t,i)$는 자질 함수이다. 또한 λ_k는 각 자질에 대한 가중치를 나타낸다. 자질 함수는 현재 시간 i에 대해 관측열 x에 대해서 전이의 양상을 측정할 수 있다.

매개변수들은 주어진 입력열과 이에 대응하는 상태열에 대한 조건부 확률이 최대화하는 최대 유사도(maximum likelihood)에 의해서 추정된다. 훈련 집합 $\{(t_l,x_l)\}_{l=1}^{N}$에 대해서 다음과 같은 로그 유사도(log-likelihood)를 계산한다.

$$L(\Lambda)=\sum_{l}\log P_\Lambda(x_i|t_i)=\sum_{l}\left(\sum_{i}^{n}\sum_{k}\lambda_k f_k(x_{i-1},x_i,t,i)-\log Z_{t_i}\right) \quad (2)$$

식 (2)를 최대로 하도록 학습한다. 일반적으로 CRFs는 improved iterative scaling

<표 2> 기본모델의 자질. i는 주어진 문장에서의 현재 음절 위치를 가리킨다 (예제는 "나는 아침이 좋다."에서 현재 음절이 '침'일 경우에 대한 자질이다)

자질번호	자질정의	설명	예제
1	x_{i-2}		는
2	x_{i-1}	이전 음절	아
3	x_i	현재의 음절	침
4	x_{i+1}	다음 음절	이
5	x_{i+2}		좋
6	$x_{i-2}/x_{i-1}/x_i$		는/아/침
7	$x_{i-1}/x_i/x_{i+1}$		아/침/이
8	$x_i/x_{i+1}/x_{i+2}$		침/이/좋
9	x_{i-1}/x_i		아/침
10	x_i/x_{i+1}		침/이

(IIS)나 generalized iterative scaling (GIS)[15]를 사용하여 학습한다.

또한 학습 데이터의 과적합(overfitting) 문제를 해결하기 위해서 가우스 사전 평활(Gaussian prior smoothing)[16]을 적용한다.

3.2. 기본모델의 자질 정의

기본모델의 자질을 정의한 <표 2>는 현재 음절로부터 최대 앞으로 두 음절, 뒤로 두 음절을 보게 되어 있다. 이유는 [6]의 이전 연구에서 알 수 있듯이, 자동 띄어쓰기에서는 uni-gram과 four-gram보다는 bi-gram과 tri-gram이 더 높은 성능을 보인다고 하였으므로, 본 논문에서도 이와 같은 점을 반영하였다.

3.3. 기본모델의 내부 실험

기본모델의 내부 실험에서는 학습 말뭉치로 학습을 한 뒤, 기본모델 자신이 학습한 말뭉치로 실험하여, 모델의 특성을 파악하고자 한다.

3.3.1. 내부 실험의 학습 및 실험 말뭉치

내부 실험용 학습 및 실험 말뭉치로는 "세종계획 3차 결과물"[17] 중 하나인 형태소 말뭉치를 사용하였다. 우선 형태소 말뭉치로부터 원시 어절만 추출하여 완벽한 문장으로 완성을 한 다음, 어절의 시작 음절은 'W', 그 외 나머지 음절은 전부 'I'를 부착하여 총 6백 50만 음절로 구성된 학습 및 실험용 말뭉치를 <표 3>과 같이 제작하였다. 단, 원시 말뭉치에서 올바르지 않은 문자가 포함된 문장은 학습

<표 3> 내부 실험용 학습 및 실험 말뭉치의 음절과 어절, 문장 수, 한 문장의 평균 음절, 평균 어절 수, 전체 문서의 붙여쓰기(%)와 띄어쓰기(%)의 비율

음절 수	어절 수	문장 수	평균 음절 수	평균 어절 수	붙여쓰기 비율	띄어쓰기 비율
6,500,000	2,016,750	146,179	44.47	13.80	68.97	31.03

및 실험 말뭉치에서 제외하였다.

3.3.2. 평가 기준

본 논문에서는 띄어쓰기 평가 기준으로 음절 단위의 정확도(P_{sylla}), 어절 단위의 정확도(P_{word}), 어절 단위의 재현율(R_{word})을 사용한다. 음절 단위의 정확도는 정답 문서의 음절과 시스템이 동일하게 띄어 쓴 음절의 비율을 구하며, 어절 단위 정확도는 시스템이 띄어 쓴 결과가 얼마나 정확한지를 측정한다. 어절 단위의 재현율은 정답 문서에 있는 어절과 동일하게 띄어 쓴 어절이 얼마나 되는지 측정한다. 그리고 마지막으로 성격이 다른 정확도와 재현율을 조합하여 전체적인 성능을 나타내기 위해서 F-measure(F_{word})도 평가 기준으로 추가하였다.

$$P_{sylla} = \frac{\text{올바르게 띄어 쓴 음절 수}}{\text{전체 음절 수}} \times 100(\%) \quad (3)$$

$$P_{word} = \frac{\text{올바르게 띄어 쓴 어절 수}}{\text{시스템이 출력한 어절 수}} \times 100(\%) \quad (4)$$

$$R_{word} = \frac{\text{올바르게 띄어 쓴 어절 수}}{\text{정답 문서의 어절 수}} \times 100(\%) \quad (5)$$

$$F_{\beta} = \frac{(\beta+1) \times P_{word} \times R_{word}}{\beta \times P_{word} + R_{word}}, \beta = 1 \quad (6)$$

3.3.3. 기본모델의 내부 실험 결과 및 오류 분석

기본모델의 내부 실험 결과는 <표 4>에 나타나 있다. <표 4>에서 보면 어절 정확도(P_{word})와 음절 정확도(P_{sylla})가 거의 비슷한 성능을 나타내는 것이 특이한 점이다. 이것은 오류가 여러 어절에 나타나는 것이 아니라 특정 부분에서만 집중적으로 발생하고 있다는 것을 말해준다. 우리는 이것을 분석하여 보다 향상된 모델을 작성하고자 한다.

기본모델의 오류를 분석하면 “한글 맞춤법”에서 “명사와 명사는 띄어 쓰지만, 붙여 쓸 수도 있다.”라고 허용한 복합 명사 오류 (36.29%), 명사 내부에서 띄어쓰

<표 4> 기본모델의 내부 실험 평가결과, 단위(%)

모델	기본모델			
평가 기준	P_{sylla}	P_{word}	R_{word}	F_{word}
내부 실험	98.99	98.32	98.54	98.43

기가 발생한 오류 (9.10%), “조사는 그 앞 말과 붙여 쓴다.”라고 명시되어 있는 조사의 오류(0.41%)가 존재한다. 기타 오류는 용언과 관련된 오류인데 대표적인 예제를 보면 [“들추/VV+어/EC+보/VX+니/EC”, “쓰/VV+어/EC+보/VX+았/EP+지만/EC”, “오/VX+아/EC+보/VX+니/EC”]와 같은 “보조용언+어미”오류, [“보상/NNG+받/VV+을/ETM”, “환영/NNG+받/VV+지/EC”, “평가/NNG+받/VV+는다/EF+./SF”]와 같은 “명사+동사”오류, 마지막으로 기호 오류이다. <표 5>는 오류의 비율을 나타내고 있다.

<표 5> 기본모델의 내부 실험 결과에서 각 오류 종류의 비율을 나타낸 표, 단위(%)

오류 종류	단일 명사	복합 명사	명사와 조사	조사와 어미	관형사와 부사	기타 오류
오류 설명	단일 명사 내부에서 띄어 쓴 오류	복합 명사를 띄어 쓰거나 붙여 쓴 오류	명사와 조사를 띄어 쓴 오류	조사와 어미의 다음 어절을 붙여 쓴 오류	관형사와 부사의 다음 어절을 붙여 쓴 오류	기타
비율	9.10	36.29	0.41	1.16	3.38	49.66

<표 6>은 기본모델 내부 실험의 오류 중, “입력 음절”이 조사의 마지막 음절로 사용될 수 있는 경우의 일부이다. 가장 많은 오류가 발생한 “나”에 대한 예제를 보면 [“빠져 나갔다”, “우리 나라”]와 같은 오류와 함께 [“불이 나 좀 봐라”, “살비듬이 나 가려움증”, “전분이 나 오일”]과 같은 오류도 있다. 이와 같은 오류는 조사는 붙여 쓴다고 정해두었지만, 명사 또는 동사의 시작 음절로 사용되어 띄어 쓴 경우도 많다.

<표 7>을 보면 조사의 마지막 음절로 사용될 수 있는 음절의 붙여쓰기와 띄어쓰기 비율을 알 수 있는데, 붙여 쓴 경우는 조사로 사용되었지만 띄어 쓴 경우는 조사이외의 다른 용도로 사용되었다는 것을 알 수 있다. 이러한 말뭉치로 학습할 경우 시스템은 애매한 경우가 발생하여 제대로 띄어쓰기를 할 수 없게 된다. 그러므로 현재 음절이 조사의 마지막 음절이냐 아니냐를 명확히 해주면 <표 7>과 같이 조사의 마지막 음절이지만 띄어 쓴 오류를 해결할 수 있다.

<표 6> 기본모델 내부 실험에서 "입력 음절"이 조사의 음절이나 어절 처음으로 예측한 오류 (Confusion-Matrix, "오류 횟수"는 내림차순 정렬)

순위	입력 음절	정답 태그	예측 태그	오류 횟수	순위	입력 음절	정답 태그	예측 태그	오류 횟수
1	나	I	W	739	16	여	I	W	128
2	가	I	W	726	17	차	I	W	125
3	이	I	W	455	18	들	I	W	95
4	지	I	W	444	19	저	I	W	91
5	만	I	W	268	20	요	I	W	75
6	의	I	W	231	21	은	I	W	61
7	아	I	W	229	22	치	I	W	60
8	과	I	W	224	23	테	I	W	57
9	구	I	W	204	24	뿐	I	W	53
10	다	I	W	185	25	게	I	W	48
11	고	I	W	172	26	라	I	W	36
12	도	I	W	171	27	리	I	W	28
13	서	I	W	154	28	터	I	W	18
14	마	I	W	153	29	든	I	W	18
15	두	I	W	148	30	즉	I	W	18

<표 7> 기본모델에서 조사 마지막 음절로 사용될 수 있는 음절 붙여쓰기(%)와 띄어쓰기(%)의 비율 (<표 6>의 "오류 횟수"에 맞추어 내림차순 정렬)

순위	입력 음절	붙여 쓴 비율	띄어 쓴 비율	출현 횟수	순위	입력 음절	붙여 쓴 비율	띄어 쓴 비율	출현 횟수
1	나	41	59	72,869	16	여	39	61	34,157
2	가	28	72	112,948	17	차	50	50	10,282
3	이	25	75	249,769	18	들	18	82	47,443
4	지	26	74	97,905	19	저	57	43	9,403
5	만	32	68	32,151	20	요	24	76	15,939
6	의	8	92	162,607	21	은	2	98	81,425
7	아	51	49	54,107	22	치	15	85	17,009
8	과	15	86	37,647	23	테	36	64	2,476
9	구	38	62	26,790	24	뿐	64	36	2,586
10	다	9	91	208,228	25	게	5	95	36,711
11	고	11	89	111,865	26	라	3	97	47,412
12	도	12	88	68,120	27	리	2	98	61,590
13	서	12	88	82,357	28	터	10	90	10,137
14	마	55	45	20,022	29	든	6	94	6,323
15	두	56	44	11,631	30	즉	95	5	1,399

<표 8> 모델1의 자질을 정리한 표, 기본모델과 같이 총 10개의 자질을 사용 ("조사-어미 마지막 음절"은 해당 음절이 조사 또는 어미 마지막 음절이면 경우 J, 아니면 N)

자질 번호	자질정의	설명	예제
1	x_{i-2}/조사-어미마지막음절		는/J
2	x_{i-1}/조사-어미마지막음절	이전 음절/N	아/N
3	x_i/조사-어미마지막음절	현재의 음절/N	침/N
4	x_{i+1}/조사-어미마지막음절	다음 음절/J	이/J
5	x_{i+2}/조사-어미마지막음절		좋/N
6	$x_{i-2}/x_{i-1}/x_i$		는아침
7	$x_{i-1}/x_i/x_{i+1}$		아침이
8	$x_i/x_{i+1}/x_{i+2}$		침이좋
9	x_{i-1}/x_i		아침
10	x_i/x_{i+1}		침이

3.4 오류를 반영한 모델1과 모델2

따라서 본 논문에서는 기본모델의 내부 실험에서 얻은 결과를 반영하여 새로운 모델1과 모델2를 제안한다. 모델1은 음절에 부착할 수 있는 태그는 기본모델과 같은 2개(W/I)이지만, "조사-어미의 마지막 음절"이라는 자질(조사-어미의 마지막 음절이면 J, 아니면 N)을 추가하여 만든 모델이며, 자질은 <표 8>과 같다.

모델2는 모델1과 다르게 <표 4>의 내부 실험 성능에서 특정 음절에서 오류가 많다는 것을 알 수 있었으므로, 띄어쓰기와 붙여쓰기를 가장 많이 틀린 음절 중, 상위 100개를 선택하여 이 음절로 시작하는 명사, 동사, 형용사의 앞 2음절을 추출하여 사전을 구성, 말뭉치에서 현재 음절과 다음 음절이 앞 2음절 명사 사전에 존재할 경우 M, 동사 사전에 존재할 경우 D, 형용사 사전에 존재할 경우 H, 두개 이상의 사전에 존재할 경우 C, 어떠한 사전에도 존재하지 않을 경우 N을 추가하여 만들 모델이며, 자질은 <표 9>와 같다.

5. 실 험

4.1. 학습 및 실험 말뭉치

띄어쓰기 오류의 수정에 대한 평가는 표준 평가 말뭉치가 없어, 본 논문에서는

<표 9> 모델2의 자질, 기본모델과 같이 총 10개의 자질을 사용

자질 번호	자질정의	설명	예제
1	x_{i-2}/명사,동사,형용사사전자질		입/N
2	x_{i-1}/명사,동사,형용사사전자질	이전 음절/C	원/C
3	x_i/명사,동사,형용사사전자질	현재의 음절/M	해/M
4	x_{i+1}/명사,동사,형용사사전자질	다음 음절/M	서/M
5	x_{i+2}/명사,동사,형용사사전자질		도/M
6	$x_{i-2}/x_{i-1}/x_i$		입원해
7	$x_{i-1}/x_i/x_{i+1}$		원해서
8	$x_i/x_{i+1}/x_{i+2}$		해서도
9	x_{i-1}/x_i		원해
10	x_i/x_{i+1}		해서

<표 10> 학습 및 실험 말뭉치 전체의 음절, 어절, 문장의 개수. 문장당 평균 음절, 평균 어절의 수와 붙여쓰기(%), 띄어쓰기(%)의 비율 (문장 단위로 학습하기 위해 음절 수에서 ±20개의 오차가 생길 수 있음)

	말뭉치 번호	음절 수	어절 수	문장 수	평균 음절 수	평균 어절 수	붙여쓰기 비율	띄어쓰기 비율
학습	1	500,000	156,736	11,892	42.05	13.18	68.65	31.35
	2	1,000,000	310,151	23,193	43.12	13.37	68.98	31.02
	3	1,500,000	464,406	34,550	43.42	13.44	69.04	30.96
	4	2,000,000	619,143	45,496	43.96	13.61	69.04	30.96
	5	2,500,000	776,062	57,750	43.29	13.44	68.96	31.04
	6	3,000,000	923,484	67,067	44.73	13.77	69.22	30.78
	7	3,500,000	1,082,134	80,592	43.43	13.43	69.08	30.92
	8	4,000,000	1,239,510	95,818	41.75	12.94	69.01	30.99
	9	4,500,000	1,397,097	110,830	40.60	12.61	68.95	31.05
	10	5,000,000	1,549,880	122,359	40.86	12.67	69.00	31.00
	11	5,500,000	1,705,838	132,199	41.60	12.90	68.98	31.02
	12	6,000,000	1,864,163	139,525	43.00	13.36	68.93	31.07
	13	6,500,000	2,016,750	146,179	44.47	13.80	68.97	31.03
외부 실험	0	926,922	288,226	33,354	27.79	8.64	68.91	31.09

[5]과 마찬가지로 외부 실험 말뭉치를 내부 실험에 사용한 "세종계획 3차 결과물"[17]과는 전혀 다른 "ETRI 품사부착 말뭉치"[18]를 사용하였다. 실험용 말뭉치로부터 추출한 원시 어절을 문장 단위로 나누어, 한 문장에서 나온 모든 어절을 붙여 쓴 형태로 가공한 것을 입력으로 하여 띄어쓰기가 수정된 말뭉치를 가공하기 전의 말뭉치와 비교함으로써 전체 음절에 대한 띄어쓰기 상태를 평가하였다.

<표 11> 기본모델, 모델1, 모델2의 음절 정확도, 어절 정확도, 어절 재현률, F-measure, 단위(%)

모델	기본모델				모델1				모델2			
평가 기준	P_{sylla}	P_{word}	R_{word}	F_{word}	P_{sylla}	P_{word}	R_{word}	F_{word}	P_{sylla}	P_{word}	R_{word}	F_{word}
내부 실험	98.99	98.32	98.54	98.43	99.02	98.37	98.57	98.47	99.05	98.40	98.63	98.52
외부 실험	97.32	95.24	96.20	95.72	97.34	95.29	96.19	95.74	97.35	95.28	96.23	95.76

<표 12> 기본모델, 모델1, 모델2의 내부 실험 결과에서 각 오류 종류의 비율을 나타낸 표, 단위(%)

오류 종류	단일 명사	복합 명사	명사와 조사	조사와 어미	관형사와 부사	기타 오류
오류 설명	단일 명사 내부에서 띄어 쓴 오류	복합 명사를 띄어 쓰거나 붙여 쓴 오류	명사와 조사를 띄어 쓴 오류	조사와 어미의 다음 어절을 띄어 쓴 오류	관형사와 부사의 다음 어절을 붙여 쓴 오류	기타
기본 모델	9.10	36.29	0.41	1.16	3.38	49.66
모델1	8.64	35.31	0.41	0.51	3.21	54.72
모델2	8.52	34.38	0.37	0.44	3.13	58.91

그리고 내부 실험에 사용한 말뭉치를 활용하여 말뭉치 크기 별로 성능을 측정하기 위해 50만 음절(말뭉치번호 1)부터, 50만 음절씩 증가시켜 6백 50만 음절(말뭉치번호 13), 총 13개의 학습 말뭉치를 제작하였다.

또한 기본모델의 말뭉치와는 다르게 모델1은 내부/외부 실험에 사용할 학습/실험 말뭉치에 조사-어미 자질이 필요한데, 원본인 형태소 말뭉치에서 품사를 이용하여 조사-어미 자질을 추가하지 않고 외부 조사-어미(2음절 이상) 사전을 이용하여 부착하였다. 이는 차후 외부 조사 사전을 사용하여 시스템 구현 시 동일한 성능을 얻기 위함이다.

4.2. 기본모델, 모델1, 모델2 실험 및 토의

<표 11>는 기본모델, 모델1, 모델2의 성능을 나타낸 표이다. 기본모델을 이용한 내부 실험의 오류를 분석하여 모델1, 모델2에 반영하였다. 그로 인해 F_{word}를 기준으로 모델1은 0.04%, 모델2는 0.09% 증가하였다. 모델2에 비해 모델1의 성능 향상

<표 13> 모델2의 학습용 말뭉치 1번의 첫 번째 문장

1	약	N	W	11	피	N	I	21	다	C	I
2	속	M	I	12	츕	N	I	22	리	N	I
3	장	M	W	13	에	N	I	23	고	N	I
4	소	M	I	14	재	N	W	24	있	N	W
5	인	M	I	15	옥	N	I	25	었	N	I
6	신	M	W	16	이	N	I	26	다	N	I
7	라	N	I	17	먼	N	W	27	.	N	I
8	호	N	I	18	저	N	I	28	"	N	W
9	텔	N	I	19	와	N	W	29	하	H	I
10	커	N	W	20	기	C	W	30	여	N	I

<표 14> 실험에서 발생한 띄어쓰기 오류 유형 및 예제 (*가 부착된 예제는 "허용할 수 있는 띄어쓰기")

오류 유형	정답	예측
조사는 그 앞 말과 붙여 쓴다.	불이나 좀 봐라. 살비듬이나 가려움증 전분이나 오일	불이 나 좀 봐라 살비듬이 나 가려움증 전분이 나 오일
의존 명사는 띄어 쓴다.	아무 것도 아니다. 대부분 이 같은 오류를 범한다.	*아무것도 아니다. *대부분 이같은 오류를 범한다.
보조 용언은 띄어 쓴다.	불이 꺼져 간다. 내 힘으로 막아 낸다.	*불이 꺼져간다. *내 힘으로 막아낸다.
명사는 단어별로 띄어 쓴다.	서울 대학교 사범 대학 영화 사업자	*서울대학교 사범대학 *영화사업자

이 작은 이유는 모델1은 단순히 2음절 이상의 조사-어미 사전으로 조사-어미를 이용한 단어 경계 정보만 제공하는 반면에, 모델2는 앞 2음절 명사, 동사, 형용사을 사용하여 조사-어미보다는 좀 더 다양한 단어 경계 정보를 제공하기 때문이다. 이는 <표 12>의 "단일명사", "복합명사", "조사와 어미"의 오류 비율이 감소했다는 것을 통해서도 알 수 있다.

하지만 기본모델에 사전을 이용하여 단어 경계 정보를 추가하였음에도 불구하고, 성능 향상 폭이 적은데, 이는 단순히 현재음절과 다음음절을 이용하여 명사, 동사, 형용사의 시작임을 결정하기에는 여전히 애매성이 존재하기 때문이라고 분석된다. <표 13>에서 모델2가 사용한 학습용 말뭉치번호 1의 첫 번째 문장을 보면 "약속장소인"의 "속장"과 "소인"이 명사의 시작으로 표기되어 있는데, 실제로 "속장"과 "소인"이라는 명사가 존재한다. 그리고 문장 끝 부분을 보면 두 음절 "하여"가 형용사의 시작임에도 불구하고 바로 앞 음절의 기호 때문에 올바르지 못한 어절 경계 정보가 부착되었다. 이와 같은 문제는 2음절이 아닌 앞 3음절과 같이 크

<표 15> 기본모델, 모델1, 모델2의 음절 정확도, 어절 정확도, 어절 재현률, F-measure, 단위(%)

학습 말뭉치 번호	기본모델				모델1				모델2			
	P_{sylla}	P_{word}	R_{word}	F_{word}	P_{sylla}	P_{word}	R_{word}	F_{word}	P_{sylla}	P_{word}	R_{word}	F_{word}
1	95.66	92.62	93.48	93.05	95.67	92.71	93.43	93.07	95.75	92.83	93.56	93.19
2	96.14	93.44	94.22	93.83	96.18	93.50	94.27	93.88	96.22	93.60	94.30	93.95
3	96.44	94.06	94.51	94.28	96.45	94.10	94.52	94.31	96.50	94.18	94.58	94.38
4	96.60	94.22	94.90	94.56	96.63	94.25	94.96	94.60	96.66	94.31	94.99	94.65
5	96.73	94.42	95.11	94.76	96.74	94.44	95.12	94.78	96.77	94.48	95.18	94.83
6	96.82	94.67	95.13	94.90	96.83	94.74	95.10	94.92	96.87	94.77	95.19	94.98
7	96.95	94.82	95.41	95.11	96.96	94.82	95.44	95.13	96.99	94.88	95.48	95.18
8	97.08	95.02	95.63	95.32	97.10	95.05	95.66	95.35	97.09	95.04	95.64	95.34
9	97.17	95.28	95.64	95.46	97.18	95.32	95.63	95.47	97.19	95.27	95.70	95.48
10	97.26	95.51	95.69	95.60	97.26	95.50	95.68	95.59	97.26	95.47	95.72	95.59
11	97.31	95.55	95.83	95.69	97.32	95.56	95.82	95.69	97.33	95.56	95.85	95.70
12	97.32	95.29	96.13	95.71	97.33	95.29	96.15	95.72	97.33	95.31	96.13	95.72
13	97.32	95.24	96.20	95.72	97.34	95.29	96.19	95.74	**97.35**	**95.28**	**96.23**	**95.76**

기를 늘려서 애매성을 없앨 수 있으나, 자료 부족 문제를 고려하여 2음절로 고정하였다.

<표 15>는 말뭉치 크기에 따른 기본모델, 모델1, 모델2의 외부 실험 성능 평가를 기록한 표이며, 모델2 13번 말뭉치에서 최고 성능을 기록하고 있다.

학습 말뭉치가 50만 음절일 경우, 기본모델과 모델1, 모델2의 성능을 95% 신뢰수준에 검증해보면 의미가 있음을 알 수 있었다. 그러나 이 결과는 13번째 학습인 650만 음절 문서에서는 의미가 없는 것으로 나타났다. 이것은 추가된 한국어 특성이 학습 문서가 작을 경우에는 효과적으로 기여하지만 학습 문서가 많아지면 학습 문서가 추가된 한국어 특성을 포함하기 때문으로 생각된다.

마지막으로 [5]의 경우 본 논문과 동일한 학습 및 실험 말뭉치를 사용하고 있으며 평가 단위로 동일한 단위인 음절 단위의 정확도(P_{sylla}), 어절 단위의 정확도(P_{word}), 어절 단위의 재현율(R_{word})를 사용하고 있어 비교가 가능하다. 그리고 학습 및 실험 말뭉치가 다르지만 [3], [6], [7]과도 <표 16>에서 비교해 보았다.

[7]의 경우는 형태소 분석기를 사용하며 실험 코퍼스의 크기가 작기 때문에 (2,000문장, 17,191어절, 52,686음절) 정확한 비교를 하기 어렵다. 그렇지만 동일한 환경에서 비교한 ?와는 상대 비교를 할 수 있을 것이다. [5]의 결과와 비교하면 음절 성능은 크게 차이가 나지 않으나 어절 성능은 상당한 차이가 난다. 이것은 학습되지 않은 음절에 대한 대응 능력이 HMM보다 CRFs가 뛰어나다는 것을 알 수 있다.

<표 16> [5]의 복합명사를 고려하지 않은 실험과 모델1, 모델2, [3], [6], [7]의 성능 비교, 단위(%): [5]와 본 논문의 학습 및 실험 말뭉치, 평가 기준은 동일함, (*)은 실험 말뭉치의 어절 개수만 존재함

평가 기준	실험 말뭉치 크기	P_{sylla}	P_{word}	R_{word}	F_{word}
모델1	926,922 음절	97.34	95.29	96.19	95.74
모델2	926,922 음절	97.35	95.28	96.23	95.76
[5]	926,922 음절	97.48	89.79	88.28	89.02
[3]	2,599 어절(*)	97.70	-	-	-
[6]	1,252,368 음절	94.71	-	-	-
[7]	52,686 음절	-	97.82	97.84	97.83

5. 결론 및 향후 연구

본 논문에서는 조사정보, 음절 시작정보 등 한국어 특성정보와 CRFs를 이용한 자동 띄어쓰기 모델을 제안하였다. (1) 기본모델의 내부 실험을 통해 오류 유형을 파악하여, (2) 다음 모델인 모델1, 모델2에 반영한 결과, (3) 어절 경계에 대한 정보가 추가되면서 띄어쓰기 오류가 감소하여 성능이 F-measure 기준으로 95.72%에서 95.76%로 0.04% 증가하였으며, "허용할 수 있는 띄어쓰기"중 "명사는 단어별로 띄어 쓴다."에서 복합명사처럼 "명사를 붙여 쓸 수 있다."라고 허용할 경우 모델2의 성능이 음절정확도 97.35%, 어절정확도 95.50%, 어절재현률 96.41%, F-measure 95.92%로 0.2% 증가하였다.

본 논문의 실험결과에서 조사-어미 정보, 명사, 동사, 형용사 시작 음절 정보와 같은 어절의 경계를 확실히 알 수 정보가 있으면 성능 향상에 도움에 된다는 것을 알 수 있었다. 또한, 말뭉치의 형태소 정보를 이용하지 않고 단순히 외부 사전을 이용해 조사-어미 정보를 부착하여 추가학습의 필요성을 줄였다.

하지만 조사-어미 정보, 명사, 동사, 형용사 시작 음절 정보와 같은 어절의 경계를 확실시 할 수 있는 정보만 사용해서는 띄어쓰기의 성능 향상이 미미했으며, 이는 단순히 음절정보만 가지고는 자질을 추출하는데 한계가 있음을 알 수 있다.

추가적으로 본 시스템은 47 MB(UTF-8 기준)를 사용하므로 전처리 등으로 사용하기를 원하는 음성언어처리 응용에 폭넓게 활용할 수 있을 것으로 생각된다.

앞으로는 단순히 음절 정보가 아닌 좀 더 다양한 정보를 활용하여 띄어쓰기 성능을 향상시킬 수 있는 자질을 추출해야 하며, 기호와 저빈도 음절 및 "의존 명사의 띄어쓰기", "보조 용언의 띄어쓰기", "복합 명사와 같은 단어 띄어쓰기"와 같은 오류를 처리할 수 있는 방안을 연구할 예정이다. 또한 학습 모델의 크기를 더 줄여 모바일 환경에 적합한 자동 띄어쓰기 시스템을 개발할 예정이다.

참 고 문 헌

[1] 국립 국어 연구원, *한국 어문 규정집*, 2001.

[2] 정희원, "공공 부문 언어의 문제점과 개선방안", *새국어생활*, 제13권, 제2호, http://www.korean.go.kr/nkview/nklife/2003_2.html, 2003.

[3] 강승식, "음절 bigram을 이용한 띄어쓰기 오류의 자동 교정", *음성과학*, 제8권, 제2호, pp. 83-90, 2001.

[4] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", *정보과학회논문지(B)*, 제23권, 제9호, pp. 991-1000, 1996.

[5] 이도길, 이상주, 임희석, 임해창, "한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델", *정보과학회논문지: 소프트웨어 및 응용*, 제30권, 제4호, pp. 358-371, 2003.

[6] 박성배, 태윤식, 박세영, "Self-organizing n-gram model for automatic word spacing", in *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp 633 - 640, 2006.

[7] 김미영, 정성원, 권혁철, "어절 내의 형태소 범주 패턴에 기반한 통계적 자동 띄어쓰기 시스템", *정보과학회논문지: 소프트웨어 및 응용*, 제33권, 제11호, pp. 965-978, 2006.

[8] G. Grefenstette, "What is a word, what is a sentence, problems of tokenisation", in *Proc. Conference on Computational Lexicography and Text Research*, pp. 79-87, 1994.

[9] D. D. Palmer, M. A. Hearst, "Adaptive multilingual sentence boundary disambiguation", *Journal of Computational Linguistics*, Vol. 23, No. 2, pp. 241-267, 1997.

[10] K. Jeon, *Rule-based Sentence Segmentation for Korean Texts*, M.S. Thesis, Department of Information and Technology, Pohang University of Science and Technology (in Korean), 1996.

[11] J. Shim, D. Kim, J. Cha, G. K. Lee, J. Seo, "Integrated multi-strategic Web document pre-processing for sentence and word boundary detection", *Journal of Information Processing and Management*, Vol. 38, No. 4, pp. 509-527, 2002.

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

[13] V. Cerny, "Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm", *Journal of Optimization Theory and Applications*, Vol. 45, No. 1, pp. 41-51, 1985.

[14] J. Lafferty, A. McCallum, F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data", *Proc. International Conference on Machine Learning*, pp. 282-289, 2001.

[15] S. A. Della Pietra, V. J. Della Pietra, J. Lafferty, "Inducing features of random fields", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 4, pp. 380-393, 1995.

[16] A. L. Berger, S. A. Della Pietra, V. J. Della Pietra, "A maximum entropy approach to natural language processing", *Computational Linguistics*, Vol. 22, No. 1, pp. 39-71, 1996.

[17] 문화관광부, *21세기 세종계획 형태소 분석 말뭉치 구축 지침*, 1999.

[18] 한국전자통신 연구원, *품사 부착 말뭉치 구축 지침서*, 1999.

접수일자: 2008년 2월 21일
게재결정: 2008년 3월 22일

▶ 이현우(Hyun-Woo Lee)
주소: 641-773 경남 창원시 사림동 9번지 소나무5길 국립창원대학교
소속: 국립창원대학교 컴퓨터공학과 자연어처리연구실
전화: 055) 213-3810
E-mail: ggamsso@changwon.ac.kr

▶ 차정원(Jeong-Won Cha) : 교신저자
주소: 641-773 경남 창원시 사림동 9번지 소나무5길 국립창원대학교
소속: 국립창원대학교 컴퓨터공학과
전화: 055) 213-3818
E-mail: jcha@changwon.ac.kr