음절 bigram를 이용한 띄어쓰기 오류의 자동 교정

Automatic Correction of Word-spacing Errors using by Syllable Bigram

강 승 식*
Seung-Shik Kang

ABSTRACT

We proposed a probabilistic approach of using syllable bigrams to the word-spacing problem. Syllable bigrams are extracted and the frequencies are calculated for the large corpus of 12 million words. Based on the syllable bigrams, we performed three experiments: (1) automatic word-spacing, (2) detection and correction of word-spacing errors for spelling checker, and (3) automatic insertion of a space at the end of line in the character recognition system. Experimental results show that the accuracy ratios are 97.7 percent, 82.1 percent, and 90.5%, respectively.

Keywords : syllable bigram, syllable frequency, automatic word-spacing

1. 서 론

영어에서는 알파벳의 빈도정보 및 bigram 정보는 문서압축 기술 등 다양한 목적으로 활용되어 왔다. 한글은 초성, 중성, 종성이 하나의 음절을 구성하기 때문에 음절 unigram 및 음절 bigram 특성을 한국어 정보처리 관련 연구에 활용할 수 있다. 특히, 한글 11,172 개 음절들은 사용 빈도에 따라 초고빈도(ultra-high frequency), 고빈도(high frequency), 저빈도(low frequency), 초저빈도(ultra-low frequency) 음절로 구분되기 때문이다. 이러한 음절 특성은 음절집합을 특정 언어현상에 속하는 것과 그렇지 않은 집합으로 구분하는 방법과, 특정 언어현상에 속하는 음절 빈도수 정보에 의해 각 음절들이 해당 언어현상에 속할 확률을 계산하여 활용하는 연구가 있었다[1].

또한, 어떤 음절이 인명(성과 이름)에 사용되는 빈도를 이용하여 3 음절 미등록어가 인명인지, 아닌지를 판단하는데 사용할 수 있으며, 미등록어가 외래어인지를 판단하는데 활용할 수도 있다. 즉, 음절 빈도 정보는 음절단위로 특정 언어현상에 속하는 것과 그렇지 않은 것을 확률적으로 구분하는 특성함수(characteristic function)를 정의할 수 있는 모든 응용분야에 적용될 수 있다. 한국어 정보처리 연구에서 음절 특성함수를 이용한 예로는 조사/어미 등 문법형태소에 사용되는 음절집합과 불규칙 용언의 끝음절 특성을 이용하여 형태소 분석후보의

* 국민대학교 자연과학대학 컴퓨터학부

과생성을 방지하는 방법 등이 있다[2, 3].

음절단위의 출현빈도는 연속된 음절간의 공기빈도(co-occurrence frequency)가 반영되지 않기 때문에 그 활용 분야에 제약이 있으며, 연속된 음절쌍에 대한 bigram 음절빈도를 활용할 필요가 있다. 그런데 음절 unigram의 경우 음절수가 최대 11,172 개[1]인데 비해 bigram 음절쌍의 개수는 11,172×11,172 개이고, KS 완성형 코드집합을 기준으로 하더라도 약 550만 개(2,350×2,350)이다. 이러한 기억공간의 제약 때문에 bigram 음절 특성을 실제로 활용하는데 어려움이 있다. 그러나 실제로 한글 문서에 사용되는 음절쌍은 그 중의 일부이므로 희소행렬(sparse matrix) 구현 방법에 의해 음절쌍 빈도정보를 활용할 수 있다.

본 논문에서는 한국어 말뭉치에서 bigram 음절쌍의 빈도를 '공백' 문자의 위치에 따라 계산하고 이를 한글 문장의 자동 띄어쓰기 알고리즘, 띄어쓰기 오류어의 인식, 줄 바꿈 위치에서 공백 삽입 문제에 적용하는 방법을 제안한다. 띄어쓰기 오류어의 인식 및 자동 띄어쓰기 알고리즘은 음성인식 시스템에서 인식된 문장의 띄어쓰기 오류를 자동으로 교정하는데 활용될 수 있다.

2. 관련 연구

한국어 형태소 분석에서는 음절 bigram 특성을 '단일어 후보생성 제약조건'으로 활용하여 '단일어 후보'[2]를 생성할 것인지를 판단하거나, 조사/어미가 분리될 수 있는지는 판단하는 '형태소 분리 제약조건'으로 적용하고 있다[2]. 이 연구에서는 빈도수나 통계적 기법이 아니라 단순히 한글의 음절 특성함수를 이용한 것이다.

심광섭(1996)은 말뭉치에서 추출한 음절 bigram 빈도수를 이용하여 음절간 띄어쓰기 확률을 계산하는 방법을 제안하였고 이를 자동 띄어쓰기에 유용하게 활용할 수 있음을 보이고 있다[4]. 또한, 음절 bigram 정보는 자동 띄어쓰기 문제와 유사한 방법으로 복합명사 분해 문제에 적용되기도 한다[5]. 신중호(1997)는 bigram 정보와 동적 프로그래밍 기법을 이용한 어절 인식 알고리즘을 제안하였다[6].

김계성(1998)은 음절정보와 결합규칙을 이용하여 어절 분리 및 재결합 방식에 의한 자동 띄어쓰기 알고리즘을 제안하였다[7]. 강승식(2000)은 조사/어미의 음절특성을 이용하여 띄어쓸 확률이 매우 높다고 판단되는 어절블록을 설정한 후에 어절블록 내에서 형태소 분석기를 이용하여 어절을 인식하는 방법을 제안하였다[8].

3. Bigram 음절정보의 구축

한글 bigram 음절쌍과 그 빈도수를 추출하기 위해 1,200만 어절 규모의 말뭉치를 구축하였으며, 말뭉치는 표 1과 같이 구성되어 있다. 표 1의 말뭉치는 원시 말뭉치(raw corpus)로서

1) KS 완성형 한글코드를 기준으로 할 때 기억공간의 크기는 2,350이다.
2) 명사, 관형사, 부사, 감탄사 등 입력어절 자체가 하나의 형태소로 구성되는 어절

아래와 같은 특성이 있다.

- 말뭉치는 수집한 상태에서 전혀 가공하지 않았다.
- 띄어쓰기 오류 및 맞춤법 오류가 포함되어 있다.
- 문서작성일 등 한글 문장 이외의 문자들이 포함되어 있다.

표 1. 말뭉치의 구성

말뭉치 유형	어 절 수
신문기사	540만 어절
Krist Collection	370만 어절
KTSET	80만 어절
기 타	210만 어절
합 계	1,200만 어절

표 2. 추출된 bigram 개수

bigram 유형	개 수
<한글, 한글>	256,189
<한글, 영-숫>	15,745
<영-숫, 한글>	15,360
<영-숫, 영-숫>	3,731
합 계	291,025

말뭉치에서 추출된 bigram 개수는 표 2와 같이 291,025 개이고, 한글 음절쌍의 개수는 256,189 개이다.[3] 이때 말뭉치에 나타난 모든 음절쌍이 현대 한국어에서 사용되는 것은 아닐 것으로 추정된다. 그 이유는 말뭉치에는 철자오류로 인해 실제문서에서 사용되지 않는 음절이 포함되었을 가능성이 있기 때문이다. 또한, bigram 빈도가 향후 한글문서에도 그대로 적용되는 것은 아니다. 인명, 회사명, 외래어 등 고유명사와 전문분야의 용어들은 기존의 bigram 특성과 상이한 음절쌍이 사용될 수 있기 때문이다.

표 3. 음절 bigram의 누적백분율

누적 백분율(%)	빈도수(회)	음절쌍 개수(개)
50(50.00)	1,941 이상	2,299
60(60.00)	1,137 이상	4,057
70(70.01)	622 이상	7,171
80(80.00)	294 이상	13,269
90(90.00)	98 이상	28,651
95(95.03)	37 이상	50,406
98(97.72)	14 이상	81,382
99(98.95)	6 이상	117,765
99.5(99.52)	3 이상	156,487

영문자, 숫자, 문장부호 등을 제외하고 순수한 한글 음절쌍 256,189 개에 대해 빈도수가 높은 순서로 정렬하여 누적빈도에 대한 백분율을 조사한 결과는 표 3과 같다.[4] 또한, 표 3의

3) 음절 X, Y에 대해 "XY"뿐만 아니라 "X Y" 유형이 포함되고, 문장부호와 기호 등은 제외하였다.

누적빈도수에 따라 고빈도 음절쌍을 1만 개 단위로 끊어서 누적백분율을 계산하여 그래프로 나타내면 그림 1과 같다. 가장 빈도가 높은 음절쌍을 순서대로 살펴보면 '으로/에서/연구/이다/하는/있다/하고/고있/하여/것이'이다.[5]

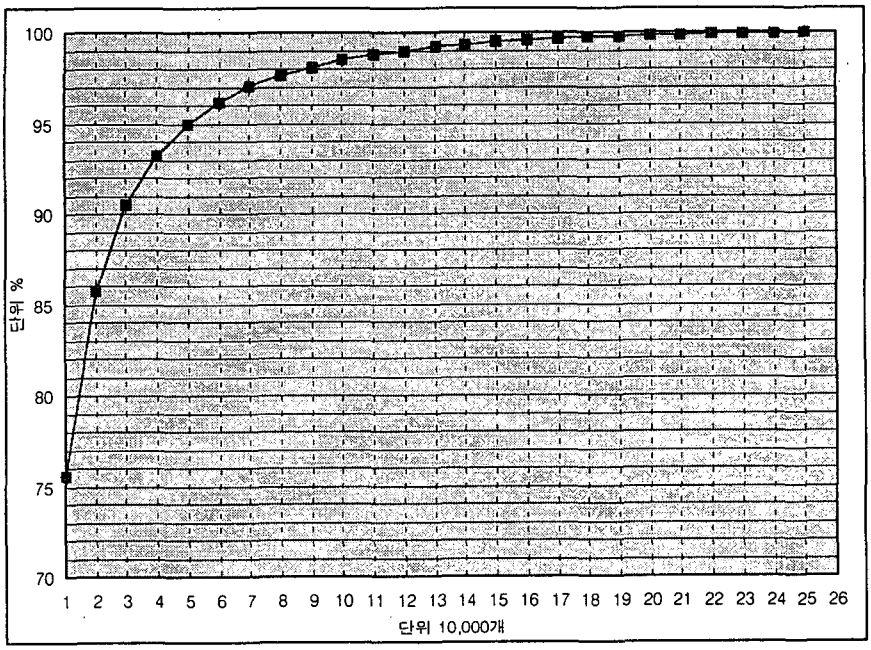

그림 1. 음절 bigram의 누적빈도

4. 음절 bigram을 이용한 자동 띄어쓰기

각 bigram 음절쌍에 대해 공백의 출현위치에 따라 좌공백 빈도, 우공백 빈도, 공백삽입 빈도, 그리고 총 출현횟수를 계산하였다. 음절쌍 <X, Y>에 대해 각 빈도수를 계산하는 구체적인 예는 다음과 같다.

- 좌공백 빈도수 : " XY"의 개수
- 우공백 빈도수 : "XY "의 개수
- 사이공백 빈도수 : "X Y"의 개수

4) 출현빈도 1 또는 2인 음절쌍은 철자오류로 인해 발생했을 가능성이 있으며, 그렇지 않더라도 활용가치가 매우 낮으로 예상된다.

5) '연구'가 10 개의 고빈도 음절쌍에 포함된 이유는 KTSET과 Krist Collection이 논문 데이터로 구성되어 있기 때문일 것으로 추정된다.

4.1 공백삽입 확률 계산식

음절쌍의 공백빈도 정보를 이용하여 자동 띄어쓰기 알고리즘을 구현하기 위한 방법으로 임의의 두 음절 x_1과 x_2 사이에 공백이 삽입될 확률 $P(x_i, x_{i+1})$은 식 (1)과 같이 계산한다. 이 식에서 $P_R(x_{i-1}, x_i)$, $P_M(x_i, x_{i+1})$, $P_L(x_{i+1}, x_{i+2})$는 각각 공백의 위치에 따른 공백확률이다.

$$P(x_i, x_{i+1}) = 0.25*P_R(x_{i-1}, x_i) + 0.5*P_M(x_i, x_{i+1}) + 0.25*P_L(x_{i+1}, x_{i+2}) \quad (1)$$

$P_R(x_{i-1}, x_i)$: $<x_{i-1}, x_i>$의 오른쪽에 공백이 있을 확률
$P_M(x_i, x_{i+1})$: $<x_i, x_{i+1}>$의 중간에 공백이 있을 확률
$P_L(x_{i+1}, x_{i+2})$: $<x_{i+1}, x_{i+2}>$ 왼쪽에 공백이 있을 확률

사이공백 확률 $P_M(x_i, x_{i+1})$의 가중치를 좌공백 확률이나 우공백 확률의 2 배로 준 것은 사이공백 확률이 좌공백 확률이나 우공백 확률에 비해 기여도가 훨씬 높다고 추정되기 때문이다.[6] $P_M(x_i, x_{i+1})$의 계산식은 식 (2)와 같으며, 이 식에서 $f_m(x_i, x_{i+1})$와 $f(x_i, x_{i+1})$은 각각 사이공백 빈도수와 총 빈도수이다. 좌공백 확률과 우공백 확률도 동일한 방법으로 계산된다.

$$P_M(x_i, x_{i+1}) = f_m(x_i, x_{i+1}) / f(x_i, x_{i+1}) \quad (2)$$

$f_m(x_i, x_{i+1})$: $<x_i, x_{i+1}>$의 사이공백 빈도수
$f(x_i, x_{i+1})$: $<x_i, x_{i+1}>$의 총 출현 빈도수

계산식에서 고빈도 음절쌍과 저빈도 음절쌍에 대한 띄어쓰기 기여도는 고려하지 않았다. 즉, 총 빈도수 1,000이고 좌공백이 500인 경우와 총 빈도수 10이고 좌공백이 5인 경우의 확률값은 모두 0.5로 계산하였다.

4.2 임계치 결정 방법

공백삽입 확률에 의해 공백을 삽입할 것인지, 그렇지 않은지를 결정하는 임계치(threshold)는 자동 띄어쓰기 정확도에 많은 영향을 미친다. 임계치가 클수록 붙여쓴 오류가 많아지고, 임계치가 작을수록 띄어쓴 오류가 많아진다. 따라서 최적의 임계치는 띄어쓴 오류와 붙여쓴 오류의 개수가 교차되는 지점의 확률값이다.

사이공백 확률 $P_M(x_i, x_{i+1})$만을 적용하여 임의의 두 음절 사이에 공백삽입 여부를 결정할 때의 임계치는 0.5이다. 이는 $<x_i, x_{i+1}>$의 띄어쓴 빈도와 붙여쓴 빈도를 기준으로 한 것이다. 사이공백 빈도만 적용할 경우 $<x_i, x_{i+1}>$의 띄어쓴 빈도와 붙여쓴 빈도의 차이가 근소한 음절쌍에 대해서는 오류발생 확률이 높아진다. 이 경우에는 좌공백 빈도와 우공백 빈도를 이용하여 오류발생 확률을 줄일 수 있다.

임계치를 변화시키면서 정확도를 계산하는 실험에 의해 정확도가 가장 높은 값을 임계치

6) 좌공백, 우공백, 사이공백 확률의 기여도는 실험적으로 그 가중치를 결정하여야 하나 그 기준이 모호하기 때문에 경험적으로 가중치를 변경하는 실험을 통하여 결정하였다.

로 결정하였는데 그 값은 0.375이다. 따라서 $P_M(x_i, x_{i+1})>0.75$일 경우에는 항상 공백을 삽입하게 된다.

5. 실험 결과

5.1 자동 띄어쓰기 실험

공백삽입 확률 계산식 $P(x_i, x_{i+1})$와 임계치 0.375에 의해 자동 띄어쓰기 실험을 하였다. 자동 띄어쓰기 실험을 위한 데이터 크기는 bigram 정보를 습득하는데 사용되지 않은 말뭉치에서 수집한 1,532 어절이다. 띄어쓰기 정확도를 측정하기 위하여 입력문서 자체(비가공된 정답)와 입력문서를 수정하여 만든 '가공된 정답' 두 가지로 구성하였다. 그 이유는 복합어의 경우에 붙여쓰기와 띄어쓰기가 모두 허용되기 때문이다. 즉, '가공된 정답'은 띄어쓰기와 붙여쓰기가 모두 허용되는 복합어의 경우에 옳은 것으로 간주한 것이다. 표 4[7]는 bigram 데이터 크기별로 '공백 재현율'을 측정한 것으로 '어절 재현율'은 4~10% 가량 낮아질 수 있다. 또한, 임계치를 실험 데이터에 적합한 값으로 결정했기 때문에 1% 가량의 오차가 있을 것으로 추정된다.

표 4. 자동 띄어쓰기 실험결과

데이터 선택기준(%)	음절쌍 개수(개)	가공된 정답(%)	비가공된 정답(%)
빈도 3 이상(99.52)	156,487	97.7	94.6
빈도 6 이상(98.95)	117,765	97.6	94.4
빈도 14 이상(97.72)	81,382	97.1	94.0
빈도 37 이상(95.05)	50,406	96.2	93.3
빈도 98 이상(90.00)	28,651	94.4	92.0

5.2 띄어쓰기 오류어의 인식실험

자동 띄어쓰기 기법을 이용하여 공백이 삽입되어야 할 어절인지를 결정하는 방법에 의해 띄어쓰기 오류어인지 아닌지를 판단하는 실험을 수행하였다. 그런데 자동 띄어쓰기가 문장 혹은 문서 단위로 수행되는데 비해 '띄어쓰기 오류어'의 인식은 어절 단위로 처리된다. 따라서 3 음절어 "먹을수"에서 '을'과 '수' 사이의 공백삽입 확률을 계산할 때 " 수"의 좌공백 확률이 계산될 수 없으므로 기본값(default value)이 주어진다. 띄어쓰기 오류어 인식 정확도를 높이기 위해 음절 X에 대해 "X " 유형에서 우공백 빈도와 " X" 유형의 좌공백 빈도를 구하여 활용할 수 있으나 이 방법을 적용하지 않았다.

실험에 사용된 데이터는 웹 문서에서 수집된 문서에서 일반적으로 자주 나타나는 띄어쓰기 오류어들을 추출하였으며, 실험에 사용된 어절은 279 개이다. 실험 데이터의 각 어절들에

7) 마침표와 쉼표, 물음표, 느낌표 뒤에는 띄어쓰고, 그 외의 다른 문장부호는 붙여쓴다. <한글, 영문자-숫자>는 띄어쓰며, <영문자-숫자, 한글>은 붙여쓰게 하였다.

대해 띄어쓰기 오류어인지, 아닌지를 판단하는 실험을 하였다. 실험에 사용된 bigram 데이터 집합의 크기에 따라 정확도를 측정한 결과는 표 5와 같다.

표 5. 띄어쓰기 오류어 인식결과

데이터 선택기준(%)	음절쌍 개수(개)	인식 정확도(%)
빈도 3 이상(99.52)	156,487	82.1
빈도 6 이상(98.95)	117,765	81.0
빈도 14 이상(97.72)	81,382	77.8
빈도 37 이상(95.03)	50,406	72.8
빈도 98 이상(90.00)	28,651	67.0

5.3 라인 끝 어절의 띄어쓰기 실험

문자인식 시스템에 의한 인식결과는 한 줄의 끝 문자열과 다음 줄의 첫 문자열이 하나의 어절인지, 서로 다른 어절인지를 구별하지 못한다. 이 경우에 bigram과 자동 띄어쓰기 방법을 적용하여 띄어쓰기 실험을 하였으며 그 결과는 표 6과 같다. 표 6에서 사용한 실험 데이터 개수는 515 개이고 이 데이터는 임의의 문서에서 무작위로 추출하였다.

표 6. 라인끝 어절의 자동 띄어쓰기 실험결과

데이터 선택기준(%)	음절쌍 개수(개)	인식 정확도(%)
빈도 3 이상(99.52)	156,487	90.5
빈도 6 이상(98.95)	117,765	89.9
빈도 14 이상(97.72)	81,382	89.3
빈도 37 이상(95.03)	50,406	88.7
빈도 98 이상(90.00)	28,651	87.8

5.4 자동 띄어쓰기 비교 평가

본 논문에서는 사용된 bigram 데이터는 1,200만 어절 말뭉치에서 습득하였는데, 심광섭(1996)은 110만 어절 말뭉치로부터 상호정보를 습득하였다. 또한, 실험 대상 데이터 집합이 동일하지 않기 때문에 공정한 비교를 하기는 쉽지 않다. 자동 띄어쓰기 실험에 대해 각 논문에서 제시된 정확도를 단순히 비교했을 때 본 논문에서 제안한 방법의 정확도는 빈도 3 이상의 데이터를 선택했을 때 가장 높은 정확도를 보이는데 공백 재현율이 97.7%이다(표 4). 이에 비해, 강승식(2000)의 어절블록 양방향 알고리즘의 정확도는 97.3%이고, 심광섭(1996)의 상호정보를 이용한 방법의 정확도는 94.6%로서 본 논문에서 제안한 방법의 정확도가 더 높게 나타났다.

6. 결 론

1,200만 어절 규모의 원시 말뭉치로부터 추출된 한글 bigram 음절쌍의 공백 빈도수를 이용하여 자동 띄어쓰기, 띄어쓰기 오류어의 인식, 줄바꿈 위치에서 공백문자의 인식 실험을 하였다. 156,487 개의 음절쌍을 이용했을 때 자동 띄어쓰기 정확도는 97.7%로서 기존의 연구에서 문법형태소의 음절특성을 이용한 방법, 음절간 상호정보 및 형태소 분석기를 사용한 방법보다 더 높은 정확도를 얻을 수 있었다. 한글의 음절빈도 및 bigram 빈도 정보는 자동 띄어쓰기뿐만 아니라 맞춤법 오류의 인식, 철자오류 교정, 대용량 데이터의 효율적인 구축방법 등 한국어 정보처리에 매우 유익하게 활용될 수 있을 것으로 기대된다.

참 고 문 헌

[1] Kang, S. S. and Kim, Y. T. 1994. "Syllable-based model for the Korean Morphology." *Proceedings of the 15-th International Conference on Computational Linguistics (COLING-94)*, vol.1, 221-226.

[2] 강승식. 1993. *음절정보와 복수어 단위정보를 이용한 한국어 형태소 분석*. 서울대학교 박사학위 논문.

[3] 강승식. 1995. "음절특성을 이용한 한국어 불규칙 용언의 형태소 분석", *정보과학회 논문지(B)*, 22권10호, 1480-1487.

[4] 심광섭. 1996. "음절간 상호정보를 이용한 한국어 자동 띄어쓰기", *정보과학회 논문지(B)*, 23권 9호, 991- 1000.

[5] 심광섭. 1997. "합성된 상호정보를 이용한 복합명사 분리", *정보과학회 논문지(B)*, 24권11호, 1307-1317.

[6] 신중호 · 박혁로. 1997. "음절단위 bigram 정보를 이용한 한국어 단어 인식 모델", *한글 및 한국어 정보처리 학술발표 논문집*, 255-260.

[7] 김계성 · 이현수 · 이상조. 1998. "연속 음절 문장에 대한 3단계 한국어 띄어쓰기 시스템", *정보과학회 논문지(B)*, 25권 12호, 1838-1844.

[8] 강승식. 2000. "한글 문장의 자동 띄어쓰기를 위한 어절블록 양방향 알고리즘", *정보과학회 논문지(B)*, 27권 4호, 441-447.

접수일자 : 2001. 4. 26.
게재결정 : 2001. 5. 31.

▲ 강승식
서울특별시 성북구 정릉동 861-1(우편번호: 136-702)
국민대학교 자연과학대학 컴퓨터학부
Tel: +82-2-910-4800
Fax: +82-2-910-4868
E-mail: sskang@kookmin.ac.kr