음절 bigram 특성을 이용한 띄어쓰기 오류의 인식

강승식[U]
한성대학교 정보전산학부
sskang@hansung.ac.kr

A Recognition of Word Spacing Errors Using By Syllable Bigram

Seung-Shik Kang[U]
Dept. of Information and Computer Engineering, Hansung University

요 약

대용량 말뭉치에서 이웃 음절간 공기빈도 정보를 추출하여 한글의 bigram 음절 특성을 조사하였다. Bigram 음절 특성은 띄어쓰기가 무시된 문서에 대한 자동 띄어쓰기, 어떤 어절이 띄어쓰기 오류어인지 판단, 맞춤법 검사기에서 철자 오류어의 교정 등 다양한 응용분야에서 유용하게 사용될 것으로 예상되고 있다. 본 논문에서는 한글의 bigram 음절 특성을 자동 띄어쓰기 및 입력어절이 띄어쓰기 오류어인지를 판단하는데 적용하는 실험을 하였다. 실험 결과에 의하면 bigram 음절 특성이 매우 유용하게 사용될 수 있음을 확인하였다.

1. 서론

한글은 자음과 모음을 초성, 중성, 종성의 3개 항목으로 구성하여 하나의 음절을 구성한다. 이러한 음절단위 표기 특성은 한국어 정보처리 시스템을 연구하는데 중요한 요소로서 활용될 수 있다. 음절 unigram은 11,172개 음절이 빈도수에 따라 고빈도 음절과 저빈도 음절, 실제로 거의 사용되지 않는 초저빈도 음절로 구분되는 특성을 기반으로 하고 있다.

이러한 특성은 음절집합을 특정 언어현상에 속하는 것과 그렇지 않은 2개의 집합으로 구분하거나, 특정 언어현상에 속하는 음절들에 대해 빈도수 정보에 의해 구체적으로 각 음절들이 해당 언어현상에 속할 확률을 계산하여 활용하기도 한다.

Unigram 음절 특성은 어떤 음절이 사람의 성씨에 사용되는 정도와 이름에 사용될 확률을 이용하여 3음절 미등록어가 인명인지를 판단하는데 사용될 수 있으며, 특정한 언어현상에 속하는 것과 그렇지 않은 것을 확률적으로 구분할 수 있는 모든 응용분야에서 활용이 가능하다. 한국어 정보처리 연구에서 unigram 음절특성을 이용한 예로는 조사/어미 등 문법형태소에 사용되는 음절집합과 불규칙 용언의 끝음절 특성을 이용하여 형태소 분석후보의 과생성을 방지하여 분석 효율을 증가시키는 방법 등이 있다[1].

2. 관련연구

음절자체의 출현확률만 고려하는 unigram 음절 특성의 제약을 극복하기 위하여 연속된 2음절에 대한 bigram 음절특성으로 확장된다. Unigram의 경우 기억공간 크기가 음절 개수를 기준으로 최대 11,172이며, KS 완성형 한글 코드 집합을 사용할 경우 2,350이다. 그리고 특정 언어현상에만 사용되는 음절집합을 별도로 정의하면 기억공간의 크기를 줄이는 방법도 가능하다.

이에 비해 bigram 음절특성은 기억공간 크기가 약 1억 가지(11,172×1,1172)이고, 현대 한국어에 거의 사용되지 않는 음절들을 제외하여 KS 완성형 코드집합을 기준으로 할 때 2,350×2,350이다. 이러한 기억공간의 제약 때문에 bigram 음절 특성을 실제로 활용하는데 어려움이 있다. 정보검색 시스템에서는 unigram 및 bigram 음절쌍을 색인어로 추출하여 사용하기도 하지만 검색효율이 저하되는 문제가 있다.

형태소 분석에서는 음절 bigram 특성을 '단일어 후보생성 제약조건'으로 활용하여 '단일어 후보'1)를 생성할 것인지를 판단하거나, 조사/어미가 분리될 수 있는지는 판단하는 '형태소 분리 제약조건'으로 적용한 예가 있다[3]. 이 연구에서는 빈도수나 통계적 기법이 아니라 단순히 한글의 연속음절 특성을 이용한 것이다.

1) 명사, 관형사, 부사, 감탄사 등 입력어절 자체가 하나의 형태소로 구성되는 어절

심광섭(1996)은 말뭉치에서 추출한 음절 bigram 빈도수를 이용하여 음절간 띄어쓰기 확률을 계산하는 방법을 제안하였고 이를 자동 띄어쓰기에 유용하게 활용할 수 있음을 보이고 있다[3]. 또한, 음절 bigram 정보는 복합명사 분해시에도 자동 띄어쓰기 문제와 유사한 방법으로 적용되고 있다[4]. 신중호(1997)는 bigram 정보와 동적 프로그래밍 기법을 이용한 어절인식 알고리즘을 제안하였다[5]. 김계성(1998)은 음절정보와 결합규칙을 이용하여 어절 분리 및 재결합 방식에 의한 자동 띄어쓰기 알고리즘을 제안하였다[6]. 강승식(2000)은 조사/어미의 음절특성을 이용하여 띄어쓸 확률이 매우 높다고 판단되는 어절블록을 설정한 후에 어절블록 내에서 형태소 분석기를 이용하여 어절을 인식하는 방법을 제안하였다[7].

3. Bigram 음절정보

한글 bigram 음절쌍과 그 빈도수를 추출하기 위해 1200만 어절 규모의 말뭉치를 구축하였으며, 말뭉치는 표 1과 같이 구성되어 있다.

표 1. 말뭉치의 구성

말뭉치 유형	어절수
신문기사	540만 어절
Krist Collection	370만 어절
KTSET	80만 어절
기 타	210만 어절
합 계	1200만 어절

표 1의 말뭉치는 원시 말뭉치(raw corpus)로서 아래와 같은 특성이 있다.

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

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 291,025개이고, 한글 음절쌍의 개수는 256,189개이다.[2] 이 때 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.

영문자, 숫자, 문장부호 등을 제외하고 순수한 한글 음절쌍 256,189개에 대해 빈도수가 높은 순서로 정렬하여 누적빈도에 대한 백분율을 조사하였다(표 3).[3] 가장 빈도가 높은 음절쌍 10개는 순서대로 '으로/에서/연구/이다/하는/있다/하고/고있/하여/것이'이다.[4]

표 3. 고빈도 음절쌍의 누적백분율[5]

누적 백분율(%)	빈도수	음절쌍 개수
50(50.00)	1941회 이상	2,299개
60(60.00)	1137회 이상	4,057개
70(70.01)	622회 이상	7,171개
80(80.00)	294회 이상	13,269개
90(90.00)	98회 이상	28,651개
95(95.03)	37회 이상	50,406개
99(98.95)	6회 이상	117,765개
99.5(99.52)	3회 이상	156,487개

누적빈도수에 따라 고빈도 음절쌍을 1만개 단위로 끊어서 누적백분율을 계산하여 그래프로 나타내면 그림 1과 같다.

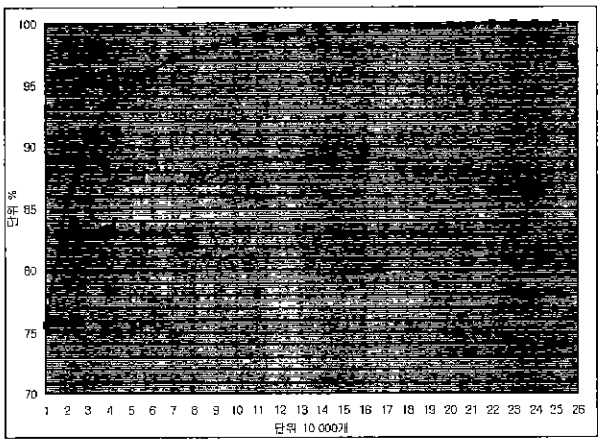

그림 1. 음절 bigram의 누적빈도

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

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

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

5) 고빈도 음절쌍 81,382개(빈도수 14 이상)의 누적빈도 백분율은 97.72%이다.

4. 자동 띄어쓰기와 띄어쓰기 오류어 인식

각 bigram 음절쌍에 대해 공백의 출현위치에 따라 좌공백 빈도, 우공백 빈도, 공백삽입 빈도, 그리고 총 출현횟수를 계산하였다. 음절쌍 <X, Y>에 대한 각 빈도수는 아래와 같이 계산한다.

- 좌 공 백 빈도수 : " XY"의 개수
- 우 공 백 빈도수 : "XY "의 개수
- 사이공백 빈도수 : "X Y"의 개수

4.1 자동 띄어쓰기

(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 $P(x_i, x_{i+1})$를 계산하는 방법은 다음과 같다.

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

$P_R(x_{i-1}, x_i)$: $<x_{i-1}, x_i>$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 2배로 준 것은 사이공백 확률이 좌공백 확률이나 우공백 확률에 비해 기여도가 훨씬 높다고 추정되기 때문이다.[6] $P_M(x_i, x_{i+1})$의 계산식은 아래와 같으며, 좌공백/우공백 확률도 동일한 방법으로 계산하였다.

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

$f_m(x_i, x_{i+1})$: $<x_i, x_{i+1}>$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 1000일 때 좌공백이 500인 경우와 총 빈도수 10일 때 좌공백이 5인 경우의 확률값은 모두 0.5로 계산된다.

(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 $P_M(x_i, x_{i+1})$만 적용하여 임의의 두 음절 사이에 공백삽입 여부를 결정할 때의 임계치는 0.5이다. 이는 $<x_i, x_{i+1}>$의 띄어쓴 빈도와 붙여쓴 빈도를 기준으로 할 때이다.

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

(3) 실험결과

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 어절(11.5K bytes)이다.

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 4는 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[7]

데이터 선택기준(%)	음절쌍 개수	가공된 정답	비가공된 정답
빈도 3이상(99.52)	156,487개	97.7%	94.6%
빈도 6이상(98.95)	117,765개	97.6%	94.4%
빈도14이상(97.72)	81,382개	97.1%	94.0%
빈도37이상(95.05)	50,406개	96.2%	93.3%
빈도98이상(90.00)	28,651개	94.4%	92.0%

4.2 띄어쓰기 오류어의 인식

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 3음절어 "먹을수"에서 '을'과 '수' 사이의 공백삽입 확률을 계산할 때 "수 "의 좌공백 확률이 계산될 수

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

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

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 웹문서에서 수집된 문서에서 일반적으로 자주 나타나는 띄어쓰기 오류어들을 추출하였으며, 실험에 사용된 어절은 279개이다. 실험 데이터의 각 어절들에 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	인식 정확도
빈도 3이상(99.52)	156,487개	82.08%
빈도 6이상(98.95)	117,765개	81.00%
빈도14이상(97.72)	81,382개	77.78%
빈도37이상(95.03)	50,406개	72.76%
빈도98이상(90.00)	28,651개	67.03%

4.3 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 하였다. 임의의 문서에서 515개의 데이터를 수집하여 실험한 결과는 표 6과 같다.

표 6. 라인끝 어절의 자동 띄어쓰기 실험결과

데이터 선택기준(%)	음절쌍 개수	인식 정확도
빈도 3이상(99.52)	156,487개	90.49%
빈도 6이상(98.95)	117,765개	89.90%
빈도14이상(97.72)	81,382개	89.32%
빈도37이상(95.03)	50,406개	88.74%
빈도98이상(90.00)	28,651개	87.77%

5. 결론

1200만 어절 규모의 원시 말뭉치로부터 추출된 한글 음절쌍의 공백빈도수를 이용하여 자동 띄어쓰기 및 띄어쓰기 오류어 인식 실험을 하였다. 156,487개의 음절쌍을 이용했을 때 자동 띄어쓰기 정확도는 97.7%로서 기존의 연구에서 문법형태소의 음절특성 혹은 bigram 정보를 이용하고 형태소 분석기를 이용하는 방법보다 더 높은 정확도를 얻을 수 있었다.
영어 알파벳의 빈도정보 및 bigram 정보는 문서압축 기술 등 다양한 목적으로 활용되어 왔다. 한글의 경우에는 음절빈도 및 bigram 빈도를 자동 띄어쓰기뿐만 아니라 맞춤법 오류의 인식, 철자오류 교정, 대용량 데이터의 효율적인 구축방법 등 한국어 정보처리에 매우 유익하게 활용될 수 있을 것으로 기대된다.

6. 참고 문헌

[1] 강승식, 음절정보와 복수어 단위정보를 이용한 한국어 형태소 분석, 서울대학교박사학위 논문, 1993.
[2] 강승식, "음절특성을 이용한 한국어 불규칙 용언의 형태소 분석", 정보과학회 논문지(B), 22권10호, pp.1480-1487, 1995.
[3] 심광섭, "음절간 상호정보를 이용한 한국어 자동 띄어쓰기", 정보과학회 논문지(B), 23권 9호, pp.991-1000, 1996.
[4] 심광섭, "합성된 상호정보를 이용한 복합명사 분리", 정보과학회 논문지(B), 24권11호, pp.1307-1317, 1997.
[5] 신중호, 박혁로, "음절단위 bigram 정보를 이용한 한국어 단어 인식 모델", 한글 및 한국어 정보처리 학술발표 논문집, pp.255-260, 1997.
[6] 김계성, 이현수, 이상조, "연속 음절 문장에 대한 3단계 한국어 띄어쓰기 시스템", 정보과학회 논문지(B), 25권 12호, pp.1838-1844, 1998.
[7] 강승식, "한글 문장의 자동 띄어쓰기를 위한 어절블록 양방향 알고리즘", 정보과학회 논문지(B), 27권 4호, pp.441-447, 2000.

[부록] 자동 띄어쓰기 실험 예

1. 본 논문 '요약'에 대한 자동 띄어쓰기 실험결과

대용량말뭉치에서 이웃음절간 공기 빈도 정보를 추출하여 한글의 bigram음절 특성을 조사하였다. Bigram음절 특성은 띄어 쓰기가 무시된 문서에 대한 자동띄어 쓰기, 어떤 어절이 띄어 쓰기 오류어인지 판단, 맞춤법 검사기에서 철자 오류어의 교정 등 다양한 응용 분야에서 유용하게 사용될 것으로 예상되고 있다. 본 논문에서는 한글의 bigram음절 특성을 자동띄어 쓰기 및 입력 어절이 띄어 쓰기 오류어인지를 판단하는 데 직용하는 실험을 하였다 실험 결과에 의하면 bigram음절 특성이 매우 유용하게 사용될 수 있음을 확인하였다.

2. 본 논문의 '결론'에 대한 자동 띄어쓰기 실험결과

1200만 어절규모의 원시 말뭉치로부터 추출된 한글 음절쌍의 공백빈도수를 이용하여 자동띄어 쓰기 및 띄어 쓰기 오류어 인식 실험을 하였다 156,487개의 음절쌍을 이용했을 때 자동띄어 쓰기 정확도는 97.7%로서 기존의 연구에서 문법 형태소의 음절 특성 혹은 bigram정보를 이용하고 형태소 분석기를 이용하는 방법보다 더 높은 정확도를 얻을 수 있었다. 영어 알파벳의 빈도 정보 및 bigram정보는 문서 압축기술 등 다양한 목적으로 활용되어 왔다. 한글의 경우에는 음절빈도 및 bigram빈도를 자동띄어 쓰기 뿐만 아니라 맞춤법 오류의 인식, 철자 오류교정, 대용량 데이터의 효율적인 구축 방법 등 한국어 정보처리에 매우 유익하게 활용될 수 있을 것으로 기대된다.

감사의 글

본 논문을 작성하는데 필요한 bigram 데이터를 추출하고, 자동 띄어쓰기 및 띄어쓰기 오류어 인식 실험을 도와 준 한성대학교 정보전산학부 안영훈군께 감사드린다.