(제 10회 한글 및 한국어 정보처리 학술대회)

연속 음성 문자열에 대한 한국어 띄어쓰기 시스템

김계성*, 이현주**, 김성규*, 최재혁***, 이상조*
* 경북대학교 컴퓨터공학과, ** 경북대학교 국어국문학과, *** 신라대학교 컴퓨터교육과

Korean Spacing System for Continuous Speech Characters

Kye Sung Kim*, Hyun Ju Lee**, Sung Kyu Kim*, Jae Hyuk Choi***, Sang Jo Lee*
* : Dept. of Computer Engineering, Kyungpook National University
** : Dept. of Korean Language & Literature, Kyungpook National University
*** : Dept. of Computer Education, Silla University

kskim@comeng.ce.kyungpook.ac.kr

요 약

대용량의 연속된 음성을 인식하는 데에는 형태소 사이의 음운변동과 언절과 어절 사이의 불일치 등으로 인한 어려움이 따른다. 그러므로 언어학적인 지식을 이용한 자연어 처리 기술과의 결합이 필수적이라 할 수 있다. 본 논문에서는 문장 단위의 연속 음성 문자열을 올바른 어절로 띄어주는 시스템을 제안한다. 먼저 띄어쓰기 발음열 사전을 이용하여 어절의 경계를 추정한다. 이 때 보다 정확한 띄어쓰기 위치를 추정하기 위하여 2음절 이상의 최장 조사·어미와 음절 분리가능빈도가 이용된다. 이렇게 해서 분리된 어절들은 음절 복원기를 거친 뒤, 형태소 분석을 행하여 올바른 어절인지를 검사한다. 분석에 실패한 어절은 띄어쓰기 오류 유형에 따라 교정을 한 후 형태소 분석을 재시도한다. 제안한 시스템을 테스트해 본 결과 96.8%의 정확도를 보였다. 본 시스템은 음운 변동 처리기와 함께 말소리를 음성 그대로 인식하는 인식기의 후처리로 이용할 수 있을 것이다.

1. 서 론

음성 언어는 컴퓨터와 사용자 사이의 가장 쉽고 자연스러운 대화수단으로서, 이에 관한 음성 인식, 음성 합성 등의 연구가 활발히 진행 중이다[1-4]. 음성 인식은 고립 단어 인식과 연속 음성 인식으로 나눌 수 있는데, 현재까지 한국어 음성 인식은 대부분 고립 단어 위주였으며, 연속 음성에 관한 연구는 아직 미흡한 실정이다[2].

연속 음성 인식은 고립 단어 인식과는 달리 언절(말을 하는 단위)와 어절(띄어쓰기 단위)의 불일치, 형태소 사이의 음운 변동 등으로 인하여 처리에 많은 어려움이 따른다[3]. 언절과 어절의 불일치는 자연스럽게 말을 할 때 말이 끊어지는 마디와 글의 띄어쓰기 마디가 일치하지 않기 때문에 나타난다. 이 때 하나의 언절은 여러 개의 어절이 포함할 수 있으므로 음성 인식한 결과를 이용할 수 있기 위해서는 올바른 어절 단위로 분리해 주어야 한다.

연속 음성 인식은 위의 문제점들을 해결하기 위해 한국어의 음성학적 지식, 언어학적 지식을 이용하는 자연어 처리와의 결합을 시도하고 있다. 이러한 시도로 [4]에서는 다이폰 인식기를 기반으로 형태소 분석 결과를 출력하는 음성 언어 처리 모델을 제안하였다. [2]는 대용량의 음성을 음소열로 바꾸어 연속적으로 입력하면서 다음에 올 음소를 예측하고, 이 예측률이 저조한 위치를 음운의 경계라 보고 분절하였다.

현재까지 띄어쓰기에 관한 연구는 문서상에서 나타나는 띄어쓰기 오류를 대상으로 하며, 부분 띄어쓰기 오류를 교정하는 철자교정기에 대한 연구가 대부분이었다[5-7]. 이러한 연구들은 음성인식의 결과로 생기는 언절과 어절의 불일치 문제를 해결할 수 없다.

따라서 본 논문에서는 연속 음성 문자열을 대상으로 하는 띄어쓰기 시스템을 제안한다. 여기에서 연속 음성 문자열이란 문장 단위의 연속된 말소리를 음성 그대로

1) 본 연구는 1997년 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었습니다.

인식한 문자열을 말한다.

본 시스템은 이 연속된 문자열들을 올바른 어절로 분리하기 위해 먼저 띄어쓰기 발음열 사전을 이용하여 어절의 경계를 추정한다. 이 때 보다 정확한 분리 위치를 추정하기 위하여 2음절 이상의 최장 조사·어미와 각 음절의 분리가능빈도를 이용한다. 이렇게 분리된 어절들은 음절 복원기를 거친 뒤 형태소 분석을 행하여 올바르게 분리되었는지를 검사한다. 분석에 실패한 어절은 띄어쓰기 오류 유형에 따라 교정을 한 후 형태소 분석을 재시도한다.

2장에서는 띄어쓰기 발음열 사전과 음절 분리가능빈도를 살펴보고, 3장에서는 제안한 띄어쓰기 시스템의 전체 구성을 알아본다. 그리고 4장에서 실험 및 평가를 하며, 5장에서 결론을 맺는다.

2. 띄어쓰기 발음열 사전과 음절 분리 가능 빈도

본 장에서는 연속 음성 문자열의 띄어쓰기 위치를 추정하기 위해 이용되는 띄어쓰기 발음열 사전과 음절 분리 가능빈도에 대해 살펴보기로 한다.

띄어쓰기 발음열 사전은 한국어 어절 구성의 특징과 음운 변동을 고려한 것이다. 한국어의 어절은 '체언+조사, 용언+어미'가 상당 부분을 차지하며, 관형사, 부사 등은 단독 어절을 형성한다. (단, 부사는 보조사와 결합할 수 있다.) 그러므로 조사, 어미, 관형사, 부사 등의 음절을 어절의 끝을 추정하는데 이용할 수 있다.

또한 한국어는 발음될 때 음소 결합의 제약성, 발음 편의에 의한 자연성, 말의 청취 효과에 따른 명확성 등의 이유로, 한 음절의 초성과 중성, 중성과 종성, 종성과 다음 음절의 초성 사이에 음운의 변동이 일어나게 된다. 다음의 예를 살펴보자.

(1) 공부하고
(2) 학교에 읻꼬
(3) 말하지 안코

(1)-(3)은 "공부하+고", "읻+고", "않+고"가 발음된 형태이다. 모두 어미 '고'가 포함된 어절이지만 앞음절의 종성에 따라 음운이 다르게 발음되고 있음을 볼 수 있다.

따라서 본 논문에서는 어절 간 분할을 위한 정보로 한국어의 이러한 음운 변동 조건을 이용한 띄어쓰기 발음열 사전을 구성하였다. 띄어쓰기 발음열 사전에는 조사·어미, 부사, 관형사 등의 음절에 대해 각각의 음운 변동 조건과 그에 따른 변동된 음운이 수록되어 있다. 사전의 구성은 표 1과 같으며, 약 6000개의 정보를 가진다.

예를 들어 어미 '으나'는 앞음절의 종성 'ㄲ'이 연음될 경우 "끄나"로 발음한다. 따라서 음성문자열 "끄나"는 앞음절 종성 조건으로 "#"을 가진다.

표 1. 띄어쓰기 발음열 사전

음성문자열	앞음절의 종성
가	#
⋮	⋮
고	#,ㄹ
고는	#,ㄹ
⋮	⋮
과	ㄴ,ㄹ,ㅁ,ㅇ
⋮	⋮
꼬	ㄱ,ㄴ,ㄷ,ㅁ,ㅂ
꼬는	ㄱ,ㄴ,ㄷ,ㅁ,ㅂ
⋮	⋮
꽈	ㄱ,ㄷ,ㅂ
꾸나	ㄱ,ㄷ,ㅂ
⋮	⋮
끄나	#
끄나마	#
⋮	⋮
한테	#,ㄴ,ㄹ,ㅁ,ㅇ

('#': 앞음절 종성이 비어있음(fillcode)을 의미함)

음절 분리 가능 빈도는 음절의 특성을 이용한 것이다. 한국어의 조사·어미 음절은 체언이나 용언의 일부가 될 수 있다. 띄어쓰기를 할 때 체언, 용언의 일부가 되는 음절이 조사·어미로 인식되면 분리 위치 추정이 잘못되거나, 두 군데 이상의 분리점이 생겨 모호성을 가지게 된다. 이러한 모호성을 줄이기 위해 각 음절의 특성을 고려한 분리가능빈도를 다음과 같이 정의한다.

$$\text{분리가능빈도}(S) = \frac{Frequency(E)}{Frequency(F)}$$

분리가능빈도(S)는 음절 S가 어절의 끝음절로 사용될 가능성을 나타낸다. Frequency(E)는 어절의 끝음절로 사용될 빈도를 가리키는 것으로 약 50만 어절의 말뭉치에서 추출하였다. Frequency(F)는 체언이나 용언의 첫음절로 사용될 빈도를 말하며, 약 23만 단어가 수록된 어휘사전에서 추출하였다. 이들은 모두 음성 문자열로 변환시킨 뒤 추출하였다. 각 음절의 분리가능빈도는 표 2와 같으며, 약 1500개의 음절에 대한 빈도값을 가진다.

표 2. 음절별 분리 가능 빈도

순위	음절	Frequency(E)	Frequency(F)	분리가능 빈도
1	를	6.997	0.001	6997.00
2	른	1.584	X	1584.00
3	는	9.195	0.006	1532.50
4	떤	0.912	X	912.00
5	믈	1.683	0.002	841.50
⋮	⋮	⋮	⋮	⋮
⋮	에	2.191	0.529	4.142
⋮	⋮	⋮	⋮	⋮
⋮	다	0.366	0.830	0.441
⋮	⋮	⋮	⋮	⋮
⋮	란	0.080	0.026	3.077
⋮	⋮	⋮	⋮	⋮

(X: 어휘사전에서 첫음절로 쓰이지 않음)

Frequency(F)가 'X'인 경우는 어휘 사전에서 첫음절로

사용되지 않기 때문에 '0.001'의 값을 주어 분리가능빈도를 계산하였다. 분리가능빈도가 가장 높은 음절은 '를'이다. 분리가능빈도(를)은 조사 '를'과 앞음절의 종성 'ㄹ'이 연음되어 발음된 '을'을 포함한 값이다. 즉, 음성문자열 '를'은 조사, 어미의 끝음절로 사용될 가능성이 아주 높은 음절이기 때문에 어절의 첫음절이 되지 못하게 한다.

그리고 말뭉치내 사용빈도가 높으면서 여러 어절에 걸쳐 사용되는 "-ㄹ+쑤+[인/일/엄/업-], [-기/끼/키] 위한-, -에 대한-, -에 의한-" 등의 정보도 띄어쓰기에 이용한다.

3. 연속한 음성 문자열 띄어쓰기 시스템

제안한 띄어쓰기 시스템의 전체 구성은 그림 1과 같다.

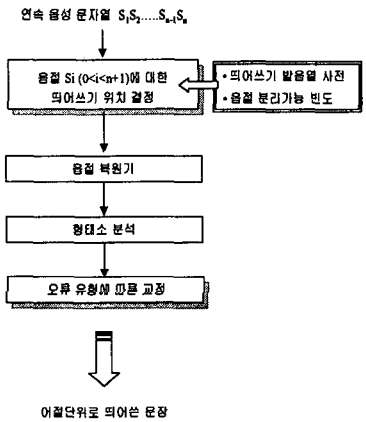

그림 1. 띄어쓰기 시스템의 구성

본 시스템은 문장 단위의 연속된 음성문자열을 입력으로 한다. 먼저 띄어쓰기 발음열 사전과 음절 분리가능빈도를 이용하여 어절 분리 위치를 추정한다. 이 때 한 음절을 사이에 두고 분리 위치가 추정되었을 경우 두 분리 위치 중에 보다 정확한 분리 위치를 결정해야 한다. 한 어절은 조사·어미로 추정되는 1음절로 구성될 가능성이 희박하기 때문이다. 두 분리 위치에서 2음절 이상의 최장 조사·어미가 우선적으로 선택되며, 두 음절 모두 1음절 조사·어미로 추정된 경우에는 음절 분리 가능 빈도를 이용하여 높은 쪽을 분리 위치로 선택한다.

이렇게 분리된 어절들은 음절 복원기를 통하여 음운 변동이 일어나기 전의 상태로 복원되며, 복원된 어절은 형태소 분석을 이용하여 띄어쓰기가 올바른지를 검사한다. 본 시스템이 사용하는 형태소 분석기는 양방향 최장일치법에 의한 형태소 분석을 행한다[8]. 형태소 분석에 실패한 어절은 오류 유형에 따라 교정을 한 후에 형태소 분석을 재시도한다.

그림 2는 띄어쓰기 오류의 유형과 그 교정방법을 보여준다. 문서상에 나타난 오류는 크게 붙띄오류, 띄붙오류, 복합오류로 나뉜다. 어절 A는 n음절($A_1A_2...A_{n-1},A_n$)로, 어절 B는 m음절($B_1B_2...B_{m-1},B_m$)로 구성되어 있으며, ↑는 현재 분석 위치를 보여준다.

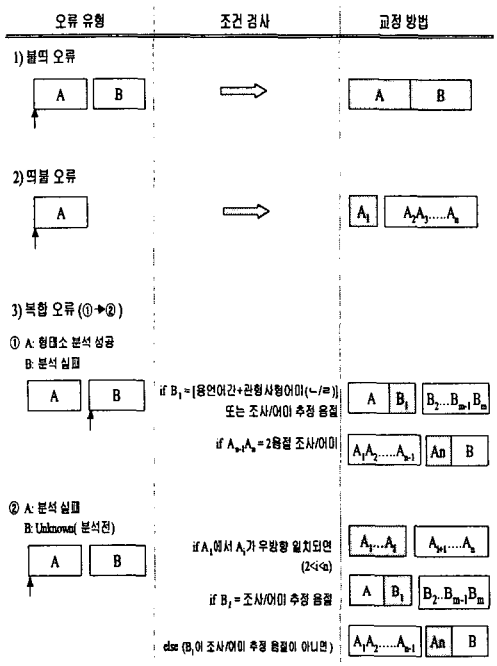

그림 2. 오류 분석 및 교정 방법

붙띄오류는 주로 2음절 이상의 조사, 어미가 한 어절을 구성할 때 나타나며, 띄붙오류는 1음절 명사/의존명사가 포함된 어절에서 많이 나타난다. 복합 오류는 붙띄, 띄붙오류가 복합된 오류로 ①에서 ②의 순서로 교정을 행한다. 다음은 오류 유형별 예와 그 교정방법이다.

오류 예	교정
(ㄱ) 붙띄 오류	
강화도 에는	-> 강화도⌒에는
(ㄴ) 띄붙 오류	
걷가타요	-> 걷∨가타요
(ㄷ) 복합 오류	
① 새로 운정채글	-> 새로⌒운∨정채글
② 되어나 가야	-> 되어∨나⌒가야
③ 일부기 득꿔늬	-> 일부∨기⌒득꿔늬
④ 나라 의역싸가	-> 나라⌒의∨역싸가
⑤ 구겨너 얻씀니다	-> 구겨∨너⌒얻씀니다

이 중 붙띄 오류는 두 어절을 결합시켜 재분석하면 교정이 가능하지만, 결합시에 제약을 두지 않으면 과분석을 초래할 수 있다. 본 논문에서는 이러한 과분석을 막기 위해 결합 음절수에 제약을 둔다. 약 50만 어절의 말뭉치를 분석해 본 결과 한 어절을 구성하는 음절 수는 1~6음절 사이가 98.4%를 차지하였다. 따라서 두 어절 결합 단계는 앞, 뒤 어절의 음절수 합이 6음절 이하일 때만 수행하도록 한다. 단 '-ㄹ+수+있/없-'과 같이 반드시 띄어써야 하는 경우는 어절 간 결합에서 제외시킨다. 표 3은 말뭉치를 통해 얻은 어절의 길이별 빈도를 보여준다.

표 3. 어절의 길이별 빈도

어절길이	빈도	어절길이	빈도
1음절	7.3%	7음절	1.0%
2음절	27.6%	8음절	0.28%
3음절	34.8%	9음절	0.15%
4음절	18.2%	10음절	0.06%
5음절	8.0%	:	:
6음절	2.5%	:	:

제안한 시스템으로 다음 예를 분석해 보자.

"그러나그거슨흉내조차낼쑤엄는이림니다"

㈀ 띄어쓰기 발음열 사전의 일부

음성문자열	앞음절 종성
거	#
그나	#
그나마	#
는	#,ㄴ,ㅁ,ㅇ
러	#,ㄹ
러나	#
러니	#,ㄹ
슨	#
조차	#,ㄴ,ㄹ,ㅁ,ㅇ

㈁ '거'와 '슨'의 분리가능빈도

음절	Frequency(E)	Frequency(F)	분리가능빈도
거	0.194	0.400	0.485
슨	0.372	×	372.0

㈂ 띄어쓰기 과정

① 띄어쓰기 발음열 사전

그러나/ 그거/ 슨/ 흉내조차/ 낼/ 쑤/ 엄는/ 이림니다

② 음절 분리가능빈도

그러나/ 그거슨/ 흉내조차/ 낼/ 쑤/ 엄는/ 이림니다

③ 음절 복원기

- 그러나 : 그러나, 글어나
- 그거슨 : 그거슨, 그것은
- 흉내조차 : 흉내조차
- 낼 : 낼
- 쑤 : 수
- 엄는 : 없는, 업는, 엄는
- 이림니다 : 이립니다, 일입니다, 이림니다

④ 형태소 분석 및 교정

그러나/ 그것은/ 흉내조차/ 낼/ 수/ 없는/ 일입니다

"엄는"은 "없는,업는,엄는"으로 복원이 가능하지만 "-ㄹ/수/없-"의 띄어쓰기 정보를 이용하여 "없는"으로만 복원된다.

4. 실험 및 평가

본 실험은 음성 문자열로 변환시킨 약 10만 어절의 동아일보 사설을 대상으로 하였다. 실험 1은 띄어쓰기 발음열 사전과 음절 분리가능빈도만을 이용한 경우이며, 실험 2는 실험 1에 형태소 분석과 교정을 적용한 경우이다. 테스트해 본 결과 본 시스템의 각 단계별 정확도는 다음과 같다.

표 4. 본 시스템의 정확도

	정확도
실험 1	72.9%
실험 2	96.8%

실험 1은 72.9%의 정확도를 보이며, 교정 단계를 적용시킨 후 띄어쓰기 정확도가 96.8%로 향상되었다.

실험 1은 본 논문에서 제안한 띄어쓰기 발음열 사전과 음절별 분리가능빈도를 이용하는 단계로, 이 단계에서 나타난 27.1%의 띄어쓰기 오류들을 분석하면 다음과 같다.

표 5. 실험 1에서 나타난 오류 분석

오류 종류	오류율
붙띄 오류	5.1%
띄붙 오류	2.3%
복합 오류	19.7%

띄어쓰기 오류를 분석한 결과 27.1%의 전체 오류 중에서 복합 오류가 19.7%를 차지하였다. 실험 1에서 복합 오류가 많은 비중을 차지하는 이유는 체언, 용언의 일부가 되는 음절이 조사·어미 음절로 오인식되어 생긴 모호성 때문이다. 이들은 형태소 분석과 교정 단계를 거치면 대부분 해결된다.

하지만 잘못 분리된 연속된 어절들이 모두 형태소 분석에 성공한 경우는 다음과 같이 오분석을 하게 된다.

(4) 사너퐈도시화에 -> 산업화도/시화

이러한 경우 교정 단계를 거쳐 올바른 분석 후보를 생

성할 수도 있지만 분석에 성공한 어절에 대해서도 모두 교정 단계를 수행하면 많은 과분석을 초래하므로 이 경우에는 교정 단계를 거치지 않는다.

(5) 공부하는데마는 -> 공부하는데/ 많은

"공부하는데"는 "공부하는/데"로 분리되어야 하지만 어미 "-ㄴ데"와 의존명사 "데"의 구분이 모호하기 때문에 오분리된 경우이다. 이와 같이 의존 명사와 그 형태가 같은 조사, 어미, 접미사 등의 구분은 부분•파싱이나 의미 정보를 추가해야만 해결이 가능한 문제이다.

5. 결론

본 논문에서는 문장 단위의 연속 음성 문자열에 대한 띄어쓰기 시스템을 제안하였다. 음운 변동 조건을 가지고 있는 띄어쓰기 발음열 사전과 음절 분리가능빈도를 이용하여 띄어쓰기 위치를 결정하였다. 이렇게 분리된 어절들은 음절 복원기를 통해 음운 변동이 일어나기 전의 상태로 복원되며, 복원된 후보는 형태소 분석기를 통하여 올바른지를 검사한다. 분석에 실패한 어절은 오류 유형에 따라 교정방법을 선택한 뒤에 형태소 분석을 재시도한다. 본 시스템을 테스트해 본 결과 96.8%의 정확도를 보였다.

하지만 잘못 분리된 어절이라 하더라도 형태소 분석에 모두 성공한 경우는 처리할 수 없으므로 이에 대한 연구가 계속되어야 한다. 또한 의존 명사와 형태가 같은 접미사, 조사, 어미 등의 구분은 기존의 철자 교정기에서도 나타나는 문제로 현 단계에서 해결이 어렵기 때문에 부분 파싱이나 의미 정보 등의 연구를 병행하여야 한다.

제안한 시스템은 음운 변동 처리기와 함께 말소리를 음성 그대로 인식하는 인식기의 후처리로 이용할 수 있을 뿐만 아니라 응용분야에 제한을 받지 않는 범용 띄어쓰기 시스템을 구축하는 데에도 이용할 수 있다.

참 고 문 헌

[1] 전재훈, 차선화, 정민화, "형태음운론적 분석에 기반한 한국어 발음 생성", 정보과학회 가을 학술발표논문집, pp. 247-250, 1997.

[2] 이찬도, "음성인식·합성을 위한 한국어 운율단위 음운론의 계산적 연구: 음운단위에 따른 경계의 발견", 정보처리논문지 제4권 제1호, pp. 280-287, 1997.

[3] 이원일, 이근배, 이종혁, "한국어 음성 인식 결과의 선언적 형태소 분석", 제6회 한글 및 한국어 정보처리 학술대회, pp.322-325, 1994.

[4] 김경희, 이근배, 이종혁, "한국어 음성 언어 처리를 위한 음소 단위 인식과 형태소 분석의 결합", 정보과학회 논문지(B) 제22권 제10호, pp.1488-1498, 1995.

[5] 심철민, "어절 간 연관 관계와 오류 유형 추정 규칙에 기반한 한국어 철자 교정기", 부산대학교 전자계산학과 석사학위논문, 1995.

[6] 강승식, "한국어 형태소 분석기 HMM의 형태소 분석 및 철자 검사 기능", 제8회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.246-252, 1996.

[7] 박봉래, 임해창, "코퍼스용 철자 및 띄어쓰기 오류 교정 시스템", 한국어 정보처리 소식 제3권 제1,2호, pp.15-27, 1995.

[8] 최재혁, "양방향 최장일치법 의한 형태소 분석기의 구현", 경북대학교 전자학과 박사학위 논문, 1993.

[9] 문화체육부, 국어 어문 규정집, 대한교과서주식회사, 서울, 1997.