음성 인식 후처리를 위한 연속 음절 문장의 키워드 추출 알고리즘

조시원*, 이동욱**
동국대학교 전기공학과

Keyword Spotting Algorithm within a Continuous Syllable Sentence for the Post-Processing of Speech Recognition

Shiwon Cho, Dong-Wook Lee
Dept. of Electrical Engineering, Dongguk University

Abstract - 연속적인 음성 인식 결과는 띄어쓰기를 하지 않은 연속 음절 문장들로 이루어져 있다. 본 논문은 음성 인식 후처리 단계에서 연속 음절 문장을 조사/어미 사전을 이용한 어절 생성 과정과 형태소 분석기를 이용하여 어절을 생성한 후 키워드를 추출한다. 실험 결과, 어절 생성기만 적용한 방식보다 제안된 알고리즘의 인식률이 향상되는 것을 확인하였다.

1. 서 론

음성 인식 시스템은 사람의 음성을 인식하여 전자 제품을 동작시키기는 등 차세대 사용자 인터페이스로서 그 활용 범위가 매우 다양하다. 하지만, 음성 인식 시스템이 실생활에서 널리 사용되기 위해서는 주위의 잡음 등 여러 환경 요인들을 극복해야 하며, 높은 인식률도 요구된다. 단순한 음성 신호를 이용하여 음성 인식을 하고 인식된 결과로부터 필요한 정보를 추출하는 과정은 대단히 중요하다. 음성 인식 결과는 띄어쓰기를 하지 않은 연속 음절 문장이기 때문에, 띄어쓰기가 적용된 어절 단위로 다시 분리를 해야 한다. 중국어, 일본어와 같이 띄어쓰기를 하지 않는 언어의 경우, 최장일치법(longest segment method), 음절 바이그램(bigram) 모델 등을 이용하여 단어를 분리하고 있다[1]. 한국어는 어절(語節) 단위로 띄어쓰기를 하기 때문에, 주로 어절 단위로 형태소를 분석하여 단어를 분리한다.

한국어 분석에 관한 연구는 크게 음절 바이그램(bigram) 정보를 이용한 통계적 기법[2]과 형태소 분석 기법[3,4]을 이용한 방법 등이 있다. 통계적 기법은 임의의 두 음절 사이에 공백이 삽입될 확률을 계산하여 미리 정한 임계값과 비교하여 어절 경계 여부를 결정하는 방식이다. 통계적 기법은 학습에 사용된 통계 정보에 의존적이며, 학습 정보와 유사한 문장 유형에서는 성능이 우수하지만, 신조어(新造語)를 비롯하여, 다른 유형의 문장에서는 자료 부족 문제로 정확도가 달라질 수 있다. 형태소 분석 기법은 명사/동사/형용사/조사/어미 등과 같은 형태소 사전 정보를 이용하여 어절을 분리하는 방법이다. 형태소 분석 기법은 사전 정보의 구축과 유지 보수하는데 부담이 크다.

어절을 인식하는 방법은 띄어쓰기 문제와 관련이 있다. 한국어는 일본어나 중국어와 달리 어절과 어절 사이에 공백을 두어 띄어쓰기를 한다. 한국어에서 잘못된 띄어쓰기는 중의성(ambiguity)문제를 발생시킨다. 중의성 문제는 어떤 어절에서 분석 오류가 발생하면, 다음 어절을 분석하는데 영향을 주게 되는 전파오류 (triggered error)를 발생시킨다. 이와 같이 어절 분리에서 오류가 발생하면 형태소 분석을 불가능하게 만들기도 한다.

어절의 경계를 구분하기 위해서는 한국어의 특성을 고려해야 한다. 한국어는 자음과 모음이 결합하여 하나의 음절을 이루며, 명사, 동사, 형용사와 같은 실질 형태소와 조사, 어미와 같은 문법 형태소들과 결합하여 다양한 형태의 어절을 만들어 낸다. 조사와 어미는 어절의 끝에 등장하므로 어절 분리의 기준이 된다. 이와 같이, 조사/어미의 음절 특성을 이용하여 어절의 경계일 가능성이 높은 곳을 추정함으로써 어절을 인식할 수 있다[5].

본 논문은 조사/어미의 음절 특성을 이용하여 음성 인식 후처리를 위한 연속 음절 문장의 키워드 알고리즘을 제안한다.

2. 본 론

2.1 어절 생성

한국어는 띄어쓰기와 관련하여 몇 가지 통계적인 특징을 가지고 있다[6]. 한국어의 어절은 '체언+조사' , '용언+어미' 와 같은 형태로 구성되므로, 조사/어미로 끝나는 곳이 어절의 경계일 가능성이 높다. 본 논문에서 제안하는 어절 생성 알고리즘은 다음과 같다[그림 1].

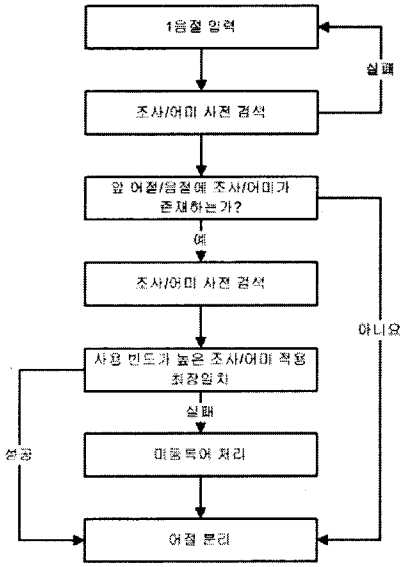

그림 1. 어절 생성 알고리즘

조사/어미 사전에 수록된 음절은 3000여개로 형태소 분석기는 문장에 나타나는 조사/어미 사전과 음절 특성을 이용하여 어절을 분리한다. 본 논문에서는 어절 분리 과정에서 형태소 분석을 동시에 적용하여 분석된 단어의 품사 정보를 같이 표시하였다. 형태소 분석기는 품사 인식, 인식 오류 교정과 분리된 어절의 재결합 과정을 수행한다. 다음 예문처럼 '에서' , '부터' 와 같이 조사/어미가 2개 이상 연속하여 나타났을 경우, 어절 인식 오류가 발생할 수 있다[7].

모터에서부터동력을
→ (모터/명사+에서부터/조사)+(동력/명사+을/조사)
→ (모터/명사+에서/조사+부터/조사)+(동력/명사+을/조사)

이 경우, 어절 인식 오류를 최소화하기 위해 다음과 같은 순서로 조사/어미의 우선순위를 결정하여 어절을 분리한다.

① 사용 빈도가 높은 조사/어미 전용
② 최장 일치 적용

사용 빈도가 높은 조사/어미가 적용된 경우, 다음과 같은 결과를 얻을 수 있다.

모터에서부터동력을
→ (모터/명사+에서부터/조사) (동력/명사+을/조사)

만약 '에서+부터' 에서 분석 실패가 발생했을 경우, '모터에서' 가 미등록어일 가능성이 높기 때문에, 다음과 같이 미등록어 처리를 한다.

모터에서부터동력을
→ (모터에서/미등록어+부터/조사) (동력/명사+을/조사)

어절 생성과 형태소 분석 과정을 거친 다음, 키워드 추출 과정을 수행한다.

2.2 키워드 추출

키워드(keyword)는 화자(話者)의 음성에 포함된 행위의 목적, 대상, 상태, 행위를 의미하는 단어를 말한다. 명사(미등록어 포함), 동사, 형용사에 속하는 단어가 키워드의 대상이 될 수 있다.

어절 인식 과정에서 분리된 각 어절에서 키워드에 해당하는 단어를 추출하는 과정은 다음과 같다[표 1].

라디오를끄고텔레비전을켜라	
(라디오/명사+를/조사) (끄/동사+고/어미) (텔레비전/명사+을/조사) (켜/동사+어라/어미)	어절 분리
(라디오/명사+를/조사) (끄다/동사+고/어미) (텔레비전/명사+을/조사) (켜다/동사+어라/어미)	동사/형용사 원형 복원
라디오 끄다 텔레비전 켜다	키워드 추출

표 1. 키워드 추출 과정

3. 실험 및 결과

본 논문에서는 신문 사설을 대상으로 1만개의 문장을 어절 띄어쓰기가 되어 있지 않은 문장으로 변환하여 실험에 적용하였다. 실험 방법은 띄어쓰기가 되어 있는 원본 문장(A)과 띄어쓰기가 되어 있지 않은 테스트 문장(B)에 제안한 방법을 각각 적용하여 비교 분석하였다.

[표 2]에서 보면 두 개의 테스트 문장에 형태소 해석기만을 적용하였을 경우, 띄어쓰기가 되어 있지 않은 테스트 문장(B)의 정확도가 상대적으로 떨어진다는 것을 볼 수 있다. 테스트 문장(B)에 형태소 해석기를 사용하지 않고, 어절 생성 과정만을 적용하였을 경우 원본 문장(A)의 어절 복원 비율은 76.2%이다[표 3]. 어절 생성기와 형태소 분석기를 같이 적용하였을 경우, 향상된 어절 재현율은 [표 4]와 같이 90.2%이다.

[표 5]는 원본 문장에서 추출한 키워드[표 2]와 [표 4]의 결과에서 추출한 키워드를 비교한 결과이다.

	성공률
원본 문장 (A)	94.2 %
테스트 문장 (B)	58.3 %

표 2. 테스트 문장의 형태소 분석기 적용 결과

	어절 재현율
테스트 문장 (B)	76.2%

표 3. 테스트 문장(B)의 어절 생성기 적용 결과

	어절 재현율
테스트 문장 (B)	90.2%

표 4. 테스트 문장(B)의 어절 생성/형태소 분석기 적용 결과

	키워드 추출 비율
원본 문장 (A)	97.2%
테스트 문장 (B)	87.2%

표 5. 키워드 추출 결과

4. 결 론

음성 인식 후처리 과정을 위하여 연속 음절 문장을 어절 생성기와 형태소 분리기를 적용하여 키워드를 추출하는 방법을 구현하여 그 성능을 확인하였다. 기본적인 어절 분리 방법만을 사용 한 결과보다 형태소 분석기를 같이 적용한 결과의 성능이 좋았다. 향후 성능을 보다 향상하기 위해 미등록어 처리와 어절 재결합 규칙을 보완할 필요가 있다.

[참 고 문 헌]

[1] Nobesawa S., et. el, "Segmenting a Sentence into Morphemes Using Statistic Information Between Words," Proceedings of the 15th International Conference on Computational Linguistics, pp.227-233, 1994.

[2] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 정보과학회논문지, 제23권 제9호, pp.991-1000, 1996.

[3] 최재혁 외 4인, "연속 음성 인식 후처리를 위한 음절 복원 rule-based 시스템과 형태소 분석 기법의 적용", 전자공학회논문지, 제36권, C편, 제3호, pp.47-57, 1999.

[4] 박미성 외 6인, "형태소 분석 기법을 이용한 음성 인식 후처리", 전자공학회논문지, 제36권, C편, 제4호, pp.65-77, 1999.

[5] 최성자 외 2인, "통계 정보를 이용한 한국어 자동 띄어쓰기 시스템의 성능 개선", 한국정보과학회 2004년도 봄 학술발표논문집, 제31권, 제1호(B), pp.883-885, 2004.

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

[7] 신호철, "형태소 분석기를 이용한 자동 띄어쓰기 시스템 구축에 대한연구", 한국어학, 12권1, pp. 167-186, 2000.