발음 변이와 개선된 편집 거리를 이용한 음성 인식 후처리

김예진[O1], 박영민[1], 강상우[1], 정상근[2], 이청재[2], 서정연[1]
서강대학교 컴퓨터공학과[1], (주)SK텔레콤[2]
jennyk@sogang.ac.kr, pymnlp@gmail.com, swkang@sogang.ac.kr,
hugman@sk.com, cj0428.lee@sk.com, seojy@sogang.ac.kr,

Post-Processing of Speech Recognition Using Phonological Variables and Improved Edit-distance

Yejin Kim[O1], Youngmin Park[1], Sangwoo Kang[1],
Sangkeon Jung[2], Cheongjae Lee[2], Jungyun Seo[1]
Department of Computer Science and Engineering, Sogang University[1]
Human Machine Interface Technology Laboratory, SK Telecom Inc.[2]

요 약

본 논문에서는 오인식된 고유명사의 후처리 방법을 제안한다. 최근 음성 인식 후처리를 위해 통계적 방법을 이용하는 연구가 활발히 진행되어 왔다. 하지만 고유명사의 음성 인식 후처리는 대용량의 데이터 수집에 많은 비용이 필요하므로 통계적 방법을 효과적으로 적용하기 어렵다. 따라서 본 논문에서는 발음 변이 현상을 고려하여 편집 거리 알고리즘을 개선한 기법을 제안한다. 본 논문에서는 고유명사의 음성 오인식 교정 성능을 검증하였고, 그 결과 P@3의 결과가 비교 모델보다 55%의 성능 향상률을 보였다.

주제어: 음성 인식 후처리, 편집 거리, 음성 인식

1. 서론

최근 음성 인식 시스템이 활용되는 분야는 매우 다양하다. 휴대폰, TV, 자동차등에 음성 인식 기능이 활용되어 IT 기술의 융합이 활발히 시도되고 있다. 그러나 음성 인식 시스템이 실용적으로 활용되기 위해서는 사용자가 만족할 수 있는 수준의 인식률이 요구된다. 음성 인식의 오류는 주변 소음과 같은 환경적 요인 및 음성 신호가 문자로 변환 될 때의 변이 현상 등에 의해 발생한다. 인식된 문자에 오류가 있을 경우 이후의 정보 처리 과정에 다양한 오류를 발생시킬 수 있는 요인이 되기 때문에 많은 후처리 연구가 진행 중에 있다. 그러나 스마트 기기의 기본 기능을 이용하기 위해서 필수적인 고유명사는 음성 인식 성능이 매우 낮고 후처리 성능도 좋지 않다.

기존의 연구는 음성 인식 결과에 포함되어 있는 띄어쓰기 오류를 통계적으로 교정하거나 형태소 분석, 구문 분석 등의 언어적인 요소를 이용한 오타의 교정 모델이 대부분이다[1,2]. 각종 스마트 기기에서 인명, 장소명 등과 같은 고유명사들의 음성 인식 이용은 매우 빈번히 발생한다. 특히 인명과 같은 고유 명사의 경우 통계적인 교정 또는 언어적인 분석을 통한 교정이 어렵다. 그 이유는 고유 명사의 문법적 관계가 뚜렷하게 표현되지 않아 형태소 분석이나 구문 분석 등의 언어 분석이 어렵기 때문이다. 또한 고유 명사는 데이터에서 반복해서 출현할 확률이 낮기 때문에 통계적인 의미 없는 경우가 많다. 인명은 고유 명사 내에서도 반복 출현빈도가 희소하기 때문에 언어적, 통계적 교정이 더욱 어렵다.

본 논문에서는 여러 가지 고유명사 중 인명에 대한 음성 인식 후처리 성능을 향상시키기 위해 빈번히 발생하는 발음 변이 현상을 고려한 기본적인 편집 거리와 다양한 개선된 편집 거리 알고리즘들의 보간법을 적용한 기법을 제안한다. 2장에서는 관련 연구에 대해 소개하고, 3장에서는 발음 변이를 고려한 편집거리 알고리즘을 제안한다. 4장에서는 본 논문에서 제안한 방법을 통한 비교 실험 및 성능 평가를 하고, 5장에서 결론을 맺는다.

2. 관련 연구

최근 음성 인식 후처리에 관한 연구는 1990년대 초반부터 시작되어 최근 급속도로 진행되어 왔다. 김병창의 연구에서는 발음의 변이 현상은 음성 신호가 문자로 변환 될 때의 가장 중요한 부분임을 강조하였다[3]. 노강호의 연구에서는 한글이 갖는 특징을 반영한 편집 거리를 정의하여 음성 인식 후처리를 수행하였다[4]. 이외에도 기존의 편집 거리에서 음소간의 유사도를 정의하고 유사한 단어를 더 정확하게 구분해 내는 알고리즘을 제안한 연구 등이 있다[5]. 이러한 기존 연구들은 대부분 문장 교정을 위한 모델로 교정 시 주로 음소와 음절 단위의 편집 거리를 이용한다[6].

기존의 연구들은 문장 단위의 음성 인식 후처리 방법으로 대개 오인식된 문장과 정답 문장의 일대일 대응 데이터를 필요로 한다. 본 논문에서는 문장에 포함된 고유명사들을 대상으로 하여 고유 명사 사전만을 이용한 후처리 방법을 제안한다. 고유 명사에 특화된 음성 오인식 후처리를 위해 이에 적합한 발음 변이 현상을 적용하고

문자열의 길이가 짧은 고유 명사 발음에 중요시되는 음절 간 연음 관계를 고려한 편집거리 알고리즘을 적용한다.

3. 발음 변이를 고려한 편집 거리 알고리즘

본 논문에서는 인명으로 빈번히 쓰이는 자음들을 발음 변이 현상에 따라 표 1과 같이 분류한다. 편집 거리는 두 개의 단어의 유사도를 측정하여 수치화할 수 있는 알고리즘 이다. 이 알고리즘은 하나의 단어를 다른 하나로 변환하는데 필요한 연산의 최소 거리를 계산한다. 인명의 경우 삭제, 삽입 그리고 대체가 임의적으로 발생하기 때문에 이를 전부 고려한 편집거리 알고리즘을 적용한다.

표 1. 발음 변이 현상 종류

변이 현상	발음 변이 전		발음 변이 후	
	종성	초성	종성	초성
초성 "ㅇ" 무음가	"ㄱ"	"ㅇ"	" "	"ㄱ"
	"ㄴ"	"ㅇ"	" "	"ㄴ"
	"ㄹ"	"ㅇ"	" "	"ㄹ"
	"ㅁ"	"ㅇ"	" "	"ㅁ"
	"ㅂ"	"ㅇ"	" "	"ㅂ"
	"ㅅ"	"ㅇ"	" "	"ㅅ"
자음동화 - 비음화	"ㄱ"	"ㅁ"	"ㅇ"	"ㅁ"
	"ㄱ"	"ㄴ"	"ㅇ"	"ㄴ"
	"ㅇ"	"ㄹ"	"ㅇ"	"ㄴ"
	"ㅁ"	"ㄹ"	"ㅁ"	"ㄴ"
자음축약	"ㄱ"	"ㅎ"	" "	"ㅋ"
	"ㅂ"	"ㅎ"	" "	"ㅍ"
	"ㄴ"	"ㅎ"	" "	"ㄴ"

3.1 발음 변이 현상

사람 이름에 흔히 쓰이는 자음들은 11년간 시대별로 가장 흔한 10개의 남성, 여성의 이름으로부터 추출하였고, 이 자음들 중 발음 변이가 일어나는 현상을 분류하여 표 1에 나타내었다[7].

3.2 편집 거리 알고리즘

기본적인 알고리즘인 Levenshtein의 편집 거리 알고리즘은 수식 (2)의 점화식을 따른다. 단어 $a=a_1...a_n$와 $b=b_1...b_m$ 간의 편집 거리는 d_{mn}에 의해 정의된다[8].

제시하는 모델은 통합적인 Levenshtein의 편집 거리 알고리즘이 적용된 모델이다. 인식된 인명과 인명사전의 각 후보 인명들 간의 음절 Unigram 편집 거리, 음절 Bigram 편집 거리 그리고 자소 편집 거리를 통합한 모델로 후보 인명들 중 N-best 후보를 검색 결과로 추출한다.

최종적인 편집 거리 ED(Edit-distance)는 수식 (1)로 나타내어지며, α,β,γ는 0에서 1사이의 값 $(\alpha+\beta+\gamma=1)$으로 실험을 통해 결정한다.

$$ED=\alpha \cdot ed_{UNI}+\beta \cdot ed_{BI}+\gamma \cdot ed_{PHO}$$

수식 (1)

$$d_{i0}=\sum_{k=1}^{i} w_{DEL}(b_k),\ \text{for}\ 1\le i\le m$$

$$d_{0j}=\sum_{k=1}^{j} w_{INS}(a_k),\ \text{for}\ 1\le j\le n$$

$$d_{ij}=\begin{cases} d_{i-1,j-1} & a_j=b_i \\ \min\begin{cases} d_{i-1,j}+w_{DEL}(b_i) \\ d_{i,j-1}+w_{INS}(a_j) \\ d_{i-1,j-1}+w_{SUB}(a_j,b_i) \end{cases} & a_j\neq b_i \end{cases}$$
$$,\ \text{for}\ 1\le i\le m, 1\le j\le n$$

w_{DEL} : 삭제 시 가중치
w_{INS} : 삽입 시 가중치
w_{SUB} : 대체 시 가중치

수식 (2)

3.2.1 자소 편집 거리

자소 편집 거리를 나타내는 ed_{PHO}는 발음 변이가 적용된 후의 상태만을 고려한다. 인식된 이름의 자소를 분리하고, 후보 인명들 역시 자소를 분리하여 두 인명 간의 자소 편집 거리를 구한다. 이는 문자적으로 비슷한 정도를 수치화하는 역할을 한다.

3.2.2 음절 Unigram 편집 거리

음절 단위의 편집 거리인 ed_{UNI}를 구한다. 편집 거리를 구할 시 인식된 인명과 인명사전의 후보 인명은 발음 변이가 일어난 후의 상태만을 고려한다. 음절은 소리의 단위이므로 음절 Unigram은 하나의 소리를 의미한다. 그러므로 ed_{UNI}은 수식(1)의 각 단계에서 두 음절 간의 소리 차이를 합한 값을 나타낸다. 예로 발음 변이 현상을 고려하기 전, "박영민"은 실제로 더 비슷한 음운을 가지는 "마경민" 보다 "박평민"과의 편집 거리가 더 짧다. 그러나 "박영민"에 발음 변이 현상을 적용하면 "바경민"이라는 음운을 가진다. 그 결과 "바경민"이 "박평민" 보다 "마경민"과의 편집 거리가 더 짧아진다. 이처럼 ed_{UNI}는 음성 인식 후처리 상황을 고려한 음운적 편집 거리를 나타낸다.

3.2.3 음절 Bigram 편집 거리

음절 Bigram 단위의 편집 거리인 ed_{BI}를 구한다. 각 음절 Unigram 편집 거리만으로 측정할 시, 연속되는 소리의 관계가 고려되지 않는 문제가 발생한다. 즉, "김영주"가 "기명수"로 인식 되었을 경우, 음절 Unigram의 편집 거리로만 N-best후보에 "김영주"와 "김응수"가 같은 편집 거리 값을 가지는 후보로 포함되어 있는 경우를 그 예로 들 수 있다. 그러나 음절 Bigram의 편집 거리를 함께 적용한 경우에는 정답인 "김영주"와 편집 거리 값의 차이가 비교적 큰 "김응수"는 N-best 후보에 포함되어 있지 않음을 알 수 있었다. 이는 음절 Bigram 편집 거리가 "김영주" 발음 시작은 "기명",

“김응수”는 “기믕”임을 파악하여, 이어지는 두 음운 차이를 비교하기 때문이다. 음절 Bigram 편집 거리는 위와 같이 음절간의 연음 관계를 고려한 수치를 나타낸다.

4. 실험 및 결과

실험에는 513개의 인명을 사용하였다. 인식기는 구글 음성 인식기를 이용하였고, 테스트를 위한 데이터는 총 인명 513 중 중복 없이 무작위로 선택한 인명을 구글 음성 인식기에 발화하여 100개의 인식된 인명을 수집하였다. 음성 데이터는 TV를 켜 놓은 정도의 노이즈가 있는 환경에서 수집하였다. 그 결과 해당 데이터 중 38개는 정답이었으며, 62개는 오류가 포함된 데이터였다. 실험의 전처리로 실험 시 이용할 모든 데이터에 특수 문자 및 숫자 등은 전부 삭제하였다.

비교 모델은 발음 변이 전의 자소 편집 거리만을 고려한 모델로 표 3의 Baseline으로 표기하였으며, $\alpha=0, \beta=0, \gamma=1$으로 실험 하였다. 비교 모델에 음절 Unigram 편집 거리만 추가된 모델은 “+Uni”로 표기하였으며, $\alpha=0.5, \beta=0, \gamma=0.5$ 으로 설정하였다. 음절 Bigram 편집 거리만 추가된 모델은 “+Bi”로 표기하였고, $\alpha=0, \beta=0.5, \gamma=0.5$ 으로 실험하였다. 본 논문에서 제안하는 궁극적 모델은 비교 모델에 발음 변이를 적용한 Unigram 편집 거리와 Bigram 편집 거리가 모두 추가된 모델로 “+Uni, Bi”로 표기하였으며, $\alpha=0.3, \beta=0.4, \gamma=0.3$으로 설정하였다.

표 3. 비교모델 대비 성능 향상률 측정

P@3 후처리 성능 향상률	Base-line	+Uni	+Bi	+Uni, Bi
발음 변이 적용 전	39%	39%	42%	44%
발음 변이 적용 후	46%	46%	50%	55%

표 4. 모델 별 성능 측정

P@3 후처리 성능	Base-line	+Uni	+Bi	+Uni, Bi
발음 변이 적용 전	77%	77%	80%	82%
발음 변이 적용 후	84%	84%	88%	93%

성능을 측정하는 기준은 테스트를 위해 수집한 데이터인 100개의 인명을 각 모델에 넣었을 때, 결과로 내놓은 N-best의 후보(N=3)에 실제 정답이 있는지의 여부로 판단하였다.

본 논문에서 제안하는 모델을 적용하기 전 음성 인식기 자체의 정답률은 38%였다. 실험 결과는 표 3과 같으며, Precision@3로 후처리 성능 향상률을 나타냈다. 발음 변이 적용 전의 모델과 후의 모델을 비교 시, 전체적으로 발음 변이 적용 후의 모델의 성능이 더 큰 향상률을 보임을 알 수 있다. 이는 인명 음성 인식 시의 특성이 반영된 음운적 편집 거리를 통합적으로 적용한 결과로 판단된다. 음절 Unigram 편집 거리를 추가한 모델과 그렇지 않은 모델의 성능은 발음 변이 적용 전 39% 향상, 적용 후 46%의 성능 향상으로 각 비교 모델(Baseline)과의 성능 향상 차이가 없음을 볼 수 있었다. 반면 음절 Bigram 편집 거리를 추가한 모델은 비교 모델에 비해 약 4~5%의 성능 향상을 보였다. 이는 편집 거리가 두 음절 간 연음을 고려한 결과로 판단된다. 본 논문에서 제안하는 모델은 발음 변이 적용 전의 비교 모델보다 16% 높은 향상률로 가장 좋은 성능을 보였다. 결론적으로 제안하는 모델은 표 4와 같이 P@3 실험에서 93%의 정확도를 보였다.

5. 결론

본 연구에서는 음성 인식을 통한 인명 오인식 결과를 후처리하는 모델을 제안하였다. 한국 인명의 특성을 고려하여, 인명에서 빈번히 발생하는 발음 변이 현상을 정의하였고, 오인식된 음성 인식 결과에 발음 변이 현상을 적용하였다. 그리고 오인식된 결과와 인명사전의 각 인명들 간의 자소 편집 거리, 음절 Unigram 및 Bigram 편집 거리를 통합한 편집 거리를 적용하여 인명사전으로부터 N-best의 후보를 얻는 모델을 제안하였다.

실험 결과를 통해 제안한 방법이 인명의 음성 인식 후처리 성능 향상에 기여함을 알 수 있었으며, 이는 음절 Bigram 편집 거리가 두 음절 간 연음 차이의 정도를 결정하는 특징으로 작용하여 성능 향상에 기여한 것으로 판단하였다.

향후 연구로는 인명에서 빈번히 발생하는 음소를 음운에 따라 그룹화 하여 편집 거리 계산 시에 적용하면 추가적인 성능 향상이 가능할 것으로 보인다. 또한 인명으로 실험 한 본 논문에서 나아가 다양한 고유 명사에도 이와 같은 방법을 이용하여 통계적으로 처리하기 어려운 고유 명사 음성 인식 후처리에도 적용이 가능할 것으로 보인다.

* 본 연구는 지식경제부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041678, 다중영역 정보서비스를 위한 대화형 개인 비서 소프트웨어 원천 기술 개발]

참고문헌

[1] 임동희, 강승식, 장두성, “음성 인식 후처리를 위한 띄어쓰기 오류의 교정”, 한국컴퓨터종합학술대회 논문지(B), 제33권, 제1호, pp.25-27, 2006

[2] 박현재, 박해선, 강원일, 손영선, “문장 성분의 의미 관계를 이용한 한국어 오류 문자 교정 시스템”, 퍼지 및 지능시스템학회 논문지, 제14권, 게1호, pp.28-32, 2004

[3] 김병창, 이원일, 이근배, 이종혁, “한국어 TTS를 위한 무제한 단어 자소열-음소열 변환”, 한국정보

과학회 인간과 컴퓨터 상호 작용 연구회 학술 대회 발표 논문집, pp.319-323, 1998.2
[4] 노강호, 김진욱, 김은상, 박근수, 조환규, “한글에 대한 편집 거리 문제”, 정보과학회논문지 : 시스템 및 이론, 제37권, 제2호, pp.103-109, 2010.4
[5] 노강호, 박근수, 조환규, 장소원, “음소의 분류 체계를 이용한 한글 편집 거리 알고리즘”, 정보과학회논문지 : 시스템 및 이론, 제37권, 제6호, pp323-329, 2010.12
[6] 노강호, 박근수, 조환규, 장소원, “음소의 1차원 배열을 이용한 한글 유사도 및 편집 거리 알고리즘”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제17권, 제10호, pp.519-526, 2011.10
[7] http://ko.wikipedia.org/wiki/한국의_성씨와_이름
[8] Daniel Jurafsky, James H. Martin, Speech and Language Processing. Pearson Education International, Pearson, 2008, pp. 107-111.