운율어를 이용한 한국어 위치 정보 데이터의 다중 발음 사전 생성

김선희*, 전재훈*, 나민수**, 정민화***
*서울대학교 인문정보연구소, **서울대학교 인지과학협동과정, ***서울대학교 언어학과
{sunhkim, jhjeon, dix39, mchung}@snu.ac.kr

Multiple Pronunciation Dictionary Generation For Korean Point-of-Interest Data Using Prosodic Words

Sunhee Kim*, Je Hun Jeon*, Minsoo Na**, Minhwa Chung***
* Center for Humanities and Information, SNU
** Interdisciplinary Program in Cognitive Science, SNU
*** Department of Linguistics, SNU

요 약

본 논문에서 위치 정보 데이터란 텔레메틱스 분야의 응용을 위하여 웹상에서 수집한 Point-of-Interest (POI) 데이터로서 행정구역 및 지명, 인명, 상호명과 같은 위치 검색에 사용되는 어휘로 구성된다. 본 논문은 음성 인식 시스템을 구성하는 발음 사전의 개발에 관한 것으로 250k 위치 정보 데이터로부터 운율어를 이용하여 불규칙 발음과 발음 변이를 포함하는 가능한 모든 발음을 생성하는 방법을 제안하는 것을 목적으로 한다. 원래 모든 POI 는 한 번씩만 데이터에 포함되어 있으므로, 그 가운데 불규칙 발음을 포함하는 POI 를 검출하거나 발음을 생성하기 위해서는 각각의 POI 하나하나를 일일이 검토하는 방법밖에 없는데, 대부분의 POI 가 복합명사구로 이루어졌다는 점에 착안하여 운율어를 이용한 결과, 불규칙 발음 검출과 다중 발음 생성을 효율적으로 수행할 수 있었다. 이러한 연구는 음성처리 영역에서는 위치정보데이터의 음성인식 성능을 향상하는 데 직접적인 기여를 할 수 있고, 무엇보다도 음성학과 음운론 이론을 음성 인식 분야에 접목한 학제적 연구로서 그 의미가 있다고 할 수 있다.

1. 서론

본 논문에서 위치 정보 데이터(Point-of-Interest: POI)란 행정구역 및 지명, 인명, 상호명과 같은 위치 관련 어휘로 구성된 데이터로서, 이는 텔레메틱스를 비롯한 다른 지명 정보 관련 분야의 응용 프로그램의 개발에 필수적이다. 위치 정보 데이터는 모두 특정 장소를 나타내는 고유명사에 해당하고 새로운 상호명의 출현에 따라 많은 신조어를 포함하는 등, 일반적인 텍스트 데이터와는 다른 특성을 보인다.

구조적으로는 '전망좋은집' 과 같이 수식어를 포함한 명사구의 형태를 보이는 경우도 있고, '홍도야울지마라' 와 같이 하나 이상의 문장으로 이루어진 경우도 있으나 대부분의 경우는 '한국교육문화사' 와 같이 명사들이 결합된 복합명사구로 구성되어 있다. 이러한 특성을 가진 위치 정보 데이터의 가능한 모든 발음을 생성해 내기 위해서는 언급한 여러 특성 가운데 두 가지가 고려되어야 한다. 첫째는 데이터가 고유명사라는 점인데, 이는 일반적으로 많은 불규칙 발음이 포함되어 있을 것으로 예상된다. [1]에 의하면 53,750 문장으로 이루어진 한국어의 일반 텍스트 코퍼스에서 불규칙 발음은 6.67%가 포함되어 있는 것으로 보고되었다. 대부분의 한국어 불규칙 발음이 명사에서 나타나므로[1], 많은 고유명사와 복합명사구로 구성되어 있는 위치 정보 데이터는 이보다 많은 불규칙 발음이 포함되어 있을 것으로 예상된다.

둘째로 대부분의 데이터가 복합명사구로서 2 개 이상의 명사로 구성되어 있어서, 하나의 복합명사구를 구성하고 있는 명사가 다른 많은 복합명사구에도 나타난다. 불규칙 발음을 포함하는 명사가 다른 여러 개의 복합명사구를 형성한다면, 이 불규칙 발음을 포함하는 단어들을 찾아 내야 할 필요가 있다. 또한, 2 개 이상의 명사가 결합되는 경우에는 그 구성 요소인 단어들이 결합할 때 여러 다른 발음 변이가 관찰된다. 따라서 불규칙 발음과 발음 변이 현상을 모델링 하기 위해서는 복합명사구를 그것을 구성하고 있는 하위 구성요소로의 분할이 필요하다.

복합명사구의 하위 구성요소를 추출해 내기 위해서 1 차적으로 형태소 분석기를 사용하는 것을 생각해 볼 수 있는데, 이 경우에는 최소한 두 가지의 문제가 예상된다. 먼저, 대부분의 복합명사구는 두 개 이상의 명사가 결합되어 있어서, 형태소 분석 과정은 결국 명사를 찾아내어 분석해 내는 과정이 되는데, 위에서 언급한 대로 위치 정보 데이터는 많은 고유 명사와 신조어를 포함하고 있어서 기존의 형태소 분석기로는 제대로 데이터를 처리하기가 어려울 것이다. 뿐만 아니라, 한국어의 경우에 실제로 복합명사나 합성어에서 불규칙 발음이 관찰되는데, 이러한 복합명사가 형태소 분석기에 의하여 각각의 명사로 분할되게 되면 불규칙 발음을 포함하는 단어를 제대로 추출해 낼 수 없게 된다.

따라서, 복합명사구를 구성하고 있는 요소를 분할하기 위하여 명사가 아닌 다른 단위에 의한 분할이 필요하게 되는데, 본 논문에서는 [2]에서 제안된 운율어(Prosodic word)의 개념을 이용한다. [2]에 의하면 운율어란 "강세구로 실현될 수 있는 분절음의 최소 연쇄(the minimal sequence of segments which can be prod uced as one Accentual Phrase (AP)" 라고 정의하였는데, 이는 바꾸어 발하면 운율어란 따로 끊어서 발화할 수 있는 분절음의 최소 연쇄, 즉 일종의 '잠재적인 강세구' 라고도 할 수 있다. 예를 들어, '한국교육문화사' 란 단어는 다음과 같이 4 가지의 발음으로 실현될 수 있다. (띄어쓰기는 끊어 읽기, 즉 강세구의 경계를 표시함)

(1) /한국교육문화사/

a. 한국 교육 문화사 [hankuk kyoyuk munhwasa]

b. 한국교육 문화사 [hankuk' yoyuk munhwasa]

c. 한국 교육문화사 [hankuk kyoyu ŋ munhwasa]

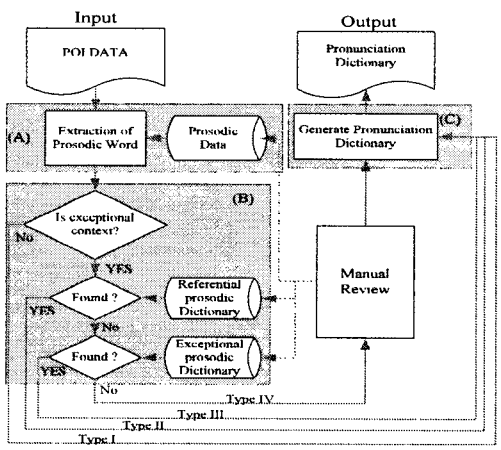
d. 한국교육문화사 [hankuk' yoyu ŋ munhwasa]

따라서, 본 논문은 음성 인식 시스템을 구성하는 발음 사전의 개발에 관한 것으로 웹상에서 수집한 250k 위치 정보 데이터로부터 운율어를 이용하여 불규칙 발음과 발음 변이를 포함하는 가능한 모든 발음을 생성하는 방법을 제안하는 것을 목적으로 한다. 이러한 연구는 음성처리 영역에서는 위치정보데이터의 음성인식 성능을 향상하는 데 직접적인 기여를 할 수 있고, 무엇보다도 음성학과 음운론 이론을 음성 인식 분야에 접목한 학제적 연구로서 그 의미가 있다고 할 수 있다.

본 논문은 먼저, 2 장에서 위치 정보 데이터를 위하여 운율어 기반의 다중 발음 생성 시스템을 제안하고, 3 장에서는 실험 및 그 결과를 제시한 다음, 4 장의 결론으로 마무리한다.

2. 위치 정보 데이터의 다중 발음 생성

<그림 1>은 위치 정보 데이터의 다중 발음을 생성하기 위한 방법을 나타낸 것이다. 이러한 방법은 새로운 데이터가 추가될 때마다 반복되어 실행될 수 있도록 설계된 것으로, 전체 과정은 데이터나 시스템을 이용하는 과정인 (A), (B), (C) 세 부분과 수동 처리 부분(Manual Review)으로 나뉘어져 있다.



<그림 1> 위치 정보 데이터의 다중 발음열 생성 과정: (A) 입력된 위치 정보 데이터를 운율어로 분할; (B) 정의된 사전을 이용하여 불규칙 발음 운율어 검출; (C) 문자음성변환기를 이용하여 다중 발음 생성

(A) 과정은 입력된 위치 정보 데이터를 운율어로 분할하는 과정이다. 초기 운율어 사전은 수작업으로 만들어서 사용하게 되는데, 이때 각각의 운율어는 원 데이터에서의 위치 및 빈도 정보와 함께 수록되고, 수록된 운율어는 다시 어두 운율어 사전(Dic-f), 어중 운율어 사전(Dic_m), 어말 운율어 사전(Dic-r)으로 분리된다. 입력된 데이터를 운율어로 분할하는 알고리즘은 다음과 같다.

1. 입력된 POI에 대해서 Dict_f를 이용해 어두에 사용 가능한 Prosodic Word List(fw_i) 검출하고, Dict_r을 이용해 어미에 사용 가능한 Prosodic Word List(rw_i)를 검출한 다음, 각 리스트에서 $i = 0$ 을 기준으로 Prosodic Word로 나눌지 결정하는 검색을 시작 한다. (단 음절 길이가 가장 긴 단어가 리스트에서 0번째 기록되고, 음절 길이가 긴 순서대로 위치시킨다.)
2. fw_i 와 rw_i 의 단어 경계가 일치하면 출력 List에 기록하고 단계 8로 간다.
3. 각 단어에서 i 가 기록된 리스트 이상이면 탐색을 중단하고 단계 7로 간다.
4. fw_i 와 rw_i 의 단어 경계가 서로 Cross 되어 있으면 $D(fw_i, fw_{i+1})$, $D(rw_i, rw_{i+1})$ 을 이용해 그 값이 작은 단어의 첨자를 $i = i+1$ 한 다음 2단계로 돌아 간다.
5. fw_i 와 rw_i 의 단어 경계가 서로 만나지 않으면 $D(fw_i, fw_{i+1})$, $D(rw_i, rw_{i+1})$ 을 이용해 그 값이 큰 단어를 출력 List에 기록한다.
6. 전체 POI의 나머지 부분에 대해서 Dict_m을 이용해 가능한 Prosodic Word List를 다시 구성한다. 단 앞 단계에서 fw_i 가 선택된 경우 신규 리스트 이름을 fw_i 로, rw_i 가 선택 된 경우 이름을 rw_i 로 하고

$i = 0$ 로 한 다음 2단계로 돌아 간다.

7. 입력 POI가 기존에 정의 되지 않은 Prosodic Word를 포함한 것으로 간주하고, Prosodic Word로 나누지 않고 출력 종료 한다.
8. 출력 List를 Prosodic Word 리스트로 출력하고 종료 한다.

* $F(w_i) = \sqrt{freq(w_i)} \times len(w_i)$ (검색된 단어 w_i 가 하나의 Prosodic Word로 나누어질 Measure 함수. $freq(w_i)$ 는 w_i 가 사전 작성시 사용된 빈도 수, $len(w_i)$ 는 w_i 의 음절 길이)

* $D(w_i, w_{i-1}) = \dfrac{F(w_i) - F(w_{i+1})}{len(w_i) - len(w_{i+1})}$ (긴 단어 w_i 를 포기하고 다음 단어 w_{i+1} 를 선택할 함수)

(B)는 [3]이 제안한 불규칙 발음 검출 방법을 위치 정보 데이터에 적용하여 불규칙 발음 운율어를 검출하는 과정이다. [3]은 일반적으로 예외로 처리되어 사전에 무작위로 추가되는 불규칙 발음이 일정한 환경에서 관찰되는 3 가지의 음운 현상이라고 규명하고, 이러한 연구 결과에 따라 주어진 데이터로부터 불규칙 발음을 검출하는 방법을 제안했다. [1, 3]에 의한 불규칙 발음과 관련된 음운 현상과 그 환경은 다음과 같다.

	p	t	s	c	k	l	V
m							
n						(II)	
N			(I)				
l							
V							
C							(III)

<표 1> 불규칙 발음 음운 현상과 그 환경

(C: a consonant of a consonant cluster; V: a vowel or diphthong)

(I) Lexical Tensification

(II) Nasalization of the lateral

(III) /n/-Epenthesis, Neutralization/ Simplification + Liaison

(A) 과정에서 분할된 운율어는 (B)과정에서 먼저 이 운율어가 불규칙 발음이 나타나는 음운 환경에 해당하는지 아닌지 여부에 따라 먼저 주어진 운율어가 불규칙 발음 환경에 속하는 않는 경우 Type I 로 분류된다. 불규칙 음운 환경에 속하는 운율어인 경우는 다시 정의된 사전에 포함되어 있는지의 여부에 따라 Type II 와 Type III 로 나뉜다. Type II 는 운율어 가운데 불규칙 음운 환경에 속하지만 규칙적인 음운현상을 보이는 운율어들(여기에서는 '참조 운율어' 라고 지칭함)이고, Type III 는 불규칙 발음 운율어이다. 마지막으로 불규칙 음운 환경에 속하지만 참조 운율어도 불규칙 운율어도 아닌 운율어들은 Type IV 로 분류되어 전문가에 의한 수작업으로 Type III 나 Type IV 로 분류되게 된다(Manual Review). (A)와 (B) 과정에서 새로 검출된 모든 종류의 운율어들은 다음 번 작업 이전에 각각 업테이트하게 된다.[6]

이와 같이 (A)와 (B) 및 수동 분류를 통하여 운율어로 분할된 POI 는 최종적으로 (C)과정에서 문자음성변환기를 이용하여 가능한 모든 발음을 생성하게 된다. 여기에서 채용하는 문자음성변환기는 지식기반 시스템으로서, 발음을 생성하기 위한 음소변동 규칙으로는 다음과 같은 10 개의 규칙으로 이루어져 있다[3, 4, 5]: (1) 종성 중화, (2) 자음군 단음화, (3) 장애음 뒤의 경음화, (4) 격음화, (5) 장애음의 비음 동화, (6) 유음화, (7) 이중 비음 동화(장애음의 비음화+/ㄹ/의 비음화), (8) 연음(재음절화), (9) 구개음화, (10) 단모음화.

다음은 데이터 가운데 불규칙 발음 운율어를 포함하는 경우에 가능한 모든 발음을 도출한 예이다.

(2) /즉석김밥나라/

a. 즉썩 김빱 나라 [cIs' ek kimp' ap nala]

b. 즉썩낌빱 나라 [cIs' ek' imp' ap nala]

c. 즉썩 김빰나라 [cIs' ek kimp' amnala]

d. 즉썩낌빰나라 [cIs' ek' imp' amnala]

3. 실험 및 결과

실험은 웹상에서 추출한 250k POI 데이터를 50k 씩 5 개 그룹으로 나누어서 진행하였다. 첫 50k 데이터에서 운율어 분할하고 참조 운율어와 불규칙 운율어를 검출하는 작업은 전문가에 의한 수작업으로 진행되었다. 이 첫번째 단계에서 검출한 운율어와 참조 운율어 및 불규칙 운율어를 이용하여 2 번째 50k 데이터를 <그림 1>에서 제안한 방법에 따라 새로운 운율어, 참조 운율어, 불규칙 운율어를 검출하고, 최종적으로 다중 발음을 생성하였다.

제안한 방법의 성능은 (A), (B) (C) 각각의 과정에서 따로따로 평가될 수 있다. 먼저, <표 1>은 (A)과정에서는 입력된 POI 가운데 기존의 검출된 운율어에 의하여 운율어 분할이 가능한 POI 의 비율을 나타낸다. 이는 비록 새로운 POI 데이터가 입력이 되더라도 운율어를 이용할 경우에 평균 77.6%는 모두 기존의 운율어에 포함하고 있는 것들로서 바로 발음열이 생성될 수 있다는 것을 의미한다.

Training	Test	Detected POI (%)
50k (1st 50k)	50k (2nd)	75.1
100k (1st + 2nd)	50k (3rd)	78.3
150k (1st + 2nd + 3rd)	50k (4th)	78.2
200k (1st + 2nd + 3rd + 4th)	50k (5th)	78.8
Average		77.6

<표 2> 기존 운율어를 이용하여 발음 생성이 가능한 POI (%)

<표 3>은 (B)과정의 성능을 나타내는 것으로 검출된 불규칙 POI 의 비율이다. 위에서 언급한 같은 방법으로 데이터를 50k 씩 증가시켜감에 검출되는 불규칙 POI 비율과 그때의 오류율이다. 불규칙 POI 는 전체적으로 그 비율이 크지는 않지만 증가하는 경향을 보이고, 적은 오류율 (0.43%)로 평균 68.8%의 불규칙 발음 POI 를 검출해 내었다.

Training	Test	Detected Irregular POI (%)/ Error Rate (%)
50k (1st 50k)	50k (2nd)	64.2/0.5)
100k (1st + 2nd)	50k (3rd)	69.2/0.19
150k (1st + 2nd + 3rd)	50k (4th)	69.2/0.48
200k (1st + 2nd + 3rd + 4th)	50k (5th)	72.7/0.53
Average		68.8/0.43

<표 3> 불규칙 발음 POI 검출 비율 (%)

다음 <표 4>는 (B) 과정 이후에 불규칙 환경 운율어로 검출되었으나 Type II 나 III 에 속하지 않고 Type IV 로 분류되어 수동 분류에 들어 가는 POI 의 비율이다. 각 단계에서 평균 22.4%의 POI 가 수동 분류 작업에 해당하게 된다.

Training	Test	Detected POI (%)
50k (1st 50k)	50k (2nd)	24.9
100k (1st + 2nd)	50k (3rd)	21.7
150k (1st + 2nd + 3rd)	50k (4th)	21.8
200k (1st + 2nd + 3rd + 4th)	50k (5th)	21.2
Average		22.4

<표 4> 수동 분류가 필요한 POI (%)

마지막으로 (C)과정에서는 다중 발음 생성 성능을 평가해 보기 위하여 운율어로 분할하지 않은 경우와 운율어로 분할한 경우에 문자음성변환기를 이용하여 생성된 발음 수를 비교하였다. 전자의 경우에 생성된 평균 발음 수는 1.3 개이고 후자의 경우에 생성된 평균 발음 수는 1.6 개이다.

5. 결론

지금까지 웹상에서 수집한 250k 위치 정보 데이터로부터 운율어를 이용하여 불규칙 발음과 발음 변이를 포함하는 가능한 모든 발음을 생성하는 방법을 제안하였다. 원래 모든 POI 는 한 번씩만 데이터에 포함되어 있으므로, 그 가운데 불규칙 발음을 포함하는 POI 를 검출하거나 발음을 생성하기 위해서는 각각의 POI 하나하나를 일일이 검토하는 방법밖에 없는데, 대부분의 POI 가 복합명사구로 이루어졌다는 점에 착안하여 운율어를 이용한 결과, 불규칙 발음 검출과 다중 발음 생성을 효율적으로 수행할 수 있었다.

현재는 음성인식에서 운율어를 이용하여 다중 발음 사전을 만들어 인식률에서 그 효과를 실험하는 연구가 진행 중이다. 향후 실제 음성 데이터를 분석하여 지식 기반으로 생성한 다중 발음과 실제 발음과의 유사도를 비교할 필요가 있고, 나아가서 다른 외국어의 POI 의 경우에도 불규칙 발음을 검출하거나 다중 발음의 생성을 위하여 제안한 방식을 적용할 수 있을 것으로 생각된다.

감사의 글

이 논문은 산업자원부 지원 뇌신경정보학사업의 "뇌정보처리의 인지신경 기전에 기반한 대화형 멀티모달 사용자 인터페이스 개발" 과제의 연구비 지원으로 수행되었습니다.

참고문헌

1. Kim, S., "Phonology of Exceptions for Korean Grapheme-to-Phoneme Conversion", *Proc. Interspeech 2004-ICSLP*, pp1285-1288, 2004.
2. Jun, S.-A., *The Phonetics and Phonology of Korean Prosody: intonational phonology and prosodic structure*, Garland Publishing Inc., New York : NY, 1996.
3. Kim, S., J. Ahn, S.-H. Kim, Y.-H. Lee, "A Korean Grapheme-to-Phoneme Conversion System Using Selection Procedure for Exceptions", *Proc. Interspeech 2004-ICSL*, pp1905-1908, 2004.
4. Jeon, J. H., S. Cha, M. Chung,, J. Park, "Automatic Generation of Korean Pronunciation Variants by Multistage Applications of Phonological Rules", *Proc. of the International Conference on Spoken Language Processing*, pp1943-1946, 1998.
5. Jeon,, J. H., M. Chung, "Automatic Generation of Domain-Dependent Pronunciation Lexicon with Data-Driven Rules and Rule Adaptation," *Proc. Interspeech-2005*, pp1337-1340, 2005.
6. Kim, S., J. H., Jeon, M. Na, M. Chung,, "Irregular Pronunciation Detection for Korean Point-of-Interest Data Using Prosodic Word", *말소리*, 제 57 권, pp123-137, 2006.