어텐션 쌍방향 LSTM 신경망을 활용한 억양 기반 한국어 방언 자동 식별

이주영[1], 김경화[2], 정민화[1]
서울대학교[1], 대검찰청[2]
excalibur12@snu.ac.kr, savoix@spo.go.kr, mchung@snu.ac.kr

Korean Dialect Identification based on Intonation using an Attention BiLSTM network

Lee Joo Young[1], Kim Kyungwha[2], Chung Minhwa[1]
Seoul National University[1], Supreme Prosecutor's Office[2]

요 약

한국어 방언 식별 연구는 소규모 데이터를 대상으로 방언학과 실험음성학에서 질적 연구로서 꾸준히 이어져 왔으나, 데이터가 부족하였기 때문에 기계학습 방법론을 적용한 방언 자동 식별 연구는 이루어지지 않았다. 본 논문에서는 대검찰청 보유 '한국인 표준 음성 DB'와 국립국어원 방언 DB의 대용량 데이터, 한국어 방언 식별 표지인 억양 특징, 그리고 방언 구간을 감지하여 학습하는 시계열 신경망 어텐션 쌍방향 LSTM 모델을 활용하여 한국어 방언 자동 식별 연구를 진행한다. 음의 높낮이 변화로 표현되는 억양은 F0(기본주파수)를 중심으로 프레임 단위의 여러 음향 특징을 LSTM 모델에서 학습하는 형태로 실현하였고, 연령대 및 발화유형별로 식별 성능을 확인하였다. 성능으로는 두 방언끼리 비교하는 방식으로 학습한 제안 모델이 베이스라인에 비해 평균 25.34%p 성능 향상을 보였다.

Ⅰ. 서론

방언 식별(dialect identification)은 음성으로부터 추출한 음향 패턴을 통해 화자의 방언을 자동으로 추정하는 연구 분야이다. 방언 식별을 언어 식별(language identification)의 하위 분야로 볼 수 있으나, 방언은 동일 언어의 음운 체계 안에서 유사한 언어적 특징을 보이기 때문에 언어 식별보다 연구가 더 까다롭다.

한국어 방언 식별 역시 연구가 까다로운 분야이다. 방언의 특징을 언어학적으로 기술하는 연구는 전통 방언학과 실험음성학에서 지속적으로 이루어져 왔으나 수집된 데이터의 규모가 작아 특정 지역을 대상으로 여러 층위(음운론, 형태론, 통사론 등)의 언어학적 분석을 하거나 인접한 두 지역 방언의 특징 차이를 비교하는 등 소규모의 질적 연구가 중심이었다. 이처럼 데이터가 많지 않았기 때문에 딥러닝 모델을 활용한 자동 식별은 실현이 어려웠다.

음성은 크게 분절적(segmental) 요소와 초분절적(suprasegmental) 요소로 나눌 수 있다. 분절적 요소는 개별 음절(syllable)의 모음과 자음에 해당하고 초분절적 요소는 음성 전체에 얹히는 운율에 해당한다. 운율 특징은 다시 억양(intonation), 리듬(rhythm), 강세(stress)로 나눌 수 있으며, 이 중 억양 특징은 한국어 방언을 구분하는 척도로 사용할 수 있다 [1]. 억양은 음의 높낮이 변화로, 방언 식별 연구에서 실현하기 위해서는 높낮이 변화를 감지하고 학습할 수 있는 모델이 필요하다.

본 연구에서는 억양 정보와 LSTM 계열 인공신경망을 활용하여 한국어 방언 식별 실험을 진행한다. 실험에 사용하는 음성 데이터는 대검찰청의 대용량 코퍼스인 '한국인 표준 음성 데이터베이스'와 국립국어원의 방언 코퍼스이다. 억양 정보를 표현하기 위해 음향 특징으로는 음성의 프레임 단위에서 추출하는 F0, MFCC, 에너지, zero crossing rate 및 이들의 1차 미분값을 합한 30차원 특징 벡터를 사용하며, 음향 특징 벡터를 어텐션 쌍방향 LSTM 모델에 입력하여 학습하는 방식으로 진행한다. 이때, 어텐션 기법은 발화 내 방언 특징이 두드러진 구간을 학습하기 위해 사용한다. 모델의 예측 단계에서는 하나의 모델이 6개 방언을 분류하는 방법, 특정 방언과 나머지 방언으로 나누어 분류하는 방법(One-versus-Rest; OvR), 그리고 두 방언끼리 비교 분류하는 방법(One-versus-One; OvO)으로 나누어 결과를 확인한다.

논문의 나머지 구성은 다음과 같다. Ⅱ장에서는 실험 코퍼스로 사용하는 두 개 데이터셋을 소개하고, Ⅲ장에서는 억양 기반 한국어 방언 식별 실험 설계 과정에 대해 설명한다. Ⅳ장에서는 실험 결과에 대한 분석을 하고, Ⅴ장에서 본 연구를 정리하고 결론을 내린다.

Ⅱ. 실험 코퍼스

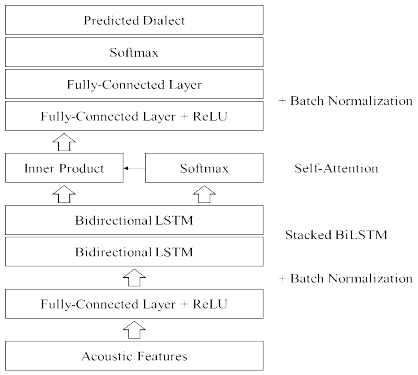
본 연구에서 사용하는 음성 데이터는 크게 대검찰청에서 3년간 수집한 '한국인 표준 음성 데이터베이스'(이하 KSS DB) [2]와 국립국어원에서 수집한 데이터베이스(이하 NIKL DB)이다.

KSS DB는 2,500명 이상의 한국인 화자로 구성된 대용량 음성 코퍼스로 화자별 성, 연령, 지역 정보가 표시되어 있다. 성은 남성과 여성 2 분류, 연령은 20대, 30대, 40대, 50대, 60대 이상의 5 분류, 지역은 수도권, 경남권, 경북권, 전남권, 전북권, 충남권, 충북권, 강원권, 제주권의 9 분류이다. 또한, 화자마다 낭독발화와 자유발화 음성이 있어 다양한 실험이 가능하여, 본 연구에서는 연령대와 발화 유형에 따른 식별 실험에 사용한다.

NIKL DB는 약 150개 음성 파일로 구성된 40시간의 자유발화 방언 데이터로 70세 이상의 고령층 화자로 구성되어 있다. 고령층 화자로 이루어진 코퍼스인 만큼 녹음된 발화에서 드러나는 방언 특색이 매우 짙으며, 따라서 식별 실험에서는 NIKL DB 추가 여부에 따른 식별 성능의 차이를 알아보는 용도로 사용한다.

Ⅲ. 억양 기반 한국어 방언 식별 실험 설계

본 연구의 대상 방언의 수는 6개이다. 이는 경상권, 전라권, 충청권 방언



〔그림 1〕 Attn-BiLSTM 모델 구조도

권의 남북 지역을 각각 합한 것이며, 남북 지역의 방언적 차이가 존재함에도 방언별 데이터 확보와 식별 성능 향상을 위해 분류 가지 수를 9개 방언권에서 6개 방언권으로 축소하였다.

[그림 1]은 한국어 방언 식별에 사용한 어텐션 쌍방향 LSTM (Attn-BiLSTM) 모델 구조도이다. 본 구조의 기본 틀은 감정 분류 실험을 한 [3]에 바탕을 두고 있으나, 자체 실험을 통해 [그림 1]과 같은 형태의 Stacked BiLSTM 층과 Batch Normalization 층을 추가하였다. 어텐션 층은 Stacked BiLSTM 층 바로 위에 배치하여 LSTM 층에서 출력된 벡터에 자기자신의 softmax 가중치 값을 적용하는 self-attention 형태로 만들었다.

음향 특징 또한 [3]를 참고하여 25ms 프레임 단위 F0, MFCC, 에너지, zero crossing rate와 이들의 1차 미분값(first-order derivatives)을 합한 30차원 벡터로 구성하였고, [3]의 음향 특징 셋과 달리 성능에는 영향을 주지 않고 학습 시간만 늘리는 voicing probability는 제외하였다.

[표 1]은 KSS DB와 NIKL DB 데이터셋의 학습과 테스트 데이터 분포이다. 표에서 Utt.와 Spon.은 각각 발화(utterance)와 자유발화(spontaneous speech)를 의미하며, 발화당 길이는 평균 3-5분이다.

실험에 반영한 연령대는 크게 20-40대로 구성된 젊은 층 그룹(young)과 40대 이상의 장노년층 그룹(old)으로 나누었으며, 40대가 양 그룹에 속해 있는 것은 실험 과정에서 40대를 배제하거나 한쪽 그룹에만 속해 있을 때보다 높은 성능을 보였기 때문이다.

프레임 단위 모델 학습 후 테스트 데이터에 대한 평가는 발화 단위로 하였다. 원본 데이터가 3-5분 길이이기 때문에 이를 30초 구간씩 분절하였다. 따라서 실제 테스트 데이터 샘플 수는 [표 1]의 테스트 발화(3-5분 발화) 수보다 많다.

Attn-BiLSTM 모델은 기본적으로 2분류 방식으로 실험하였는데, '특정 방언' vs '나머지 방언'으로 학습한 6개 모델을 사용한 OvR 방식과 '방언 1' vs '방언 2'로 학습한 15개 모델을 사용한 OvO 방식으로 하였다.

〔표 1〕 실험 데이터 분포

Dataset		Train	Test
KSS-DB	# of Utt.	6,218	1,808
Read	Size (hrs)	309.8	67.4
KSS-DB	# of Utt.	2,049	438
Spon.	Size (hrs)	114.8	24.7
NIKL-DB	# of Utt.	151	-
Spon.	Size (hrs)	40.1	-

Ⅳ. 결과 및 분석

[표 2]는 베이스라인 1 [4], 베이스라인 2, OvR 제안 모델, OvO 제안 모델을 각각 사용하였을 때 연령대 및 발화유형별 식별 정확도를 나타낸 것이다. 베이스라인 1은 [4]의 단순 형태의 6분류 LSTM 모델이며 음향 특징은 F0와 MFCC만을 사용했다. 베이스라인 2는 본 연구에서 진행한 실험 중 6분류 실험을 Attn-BiLSTM 모델로 한 것이다. [표 2]의 모델별, 연령대 + 발화유형별 결과를 비교하면, 모든 경우에서 OvO 제안 모델의 성능이 월등히 좋았다. 이는 두 개 방언만 비교하면 되는 모델들로 구성되어 있어 방언 간 분류를 극대화할 수 있었기 때문으로 본다. 이에 반해 OvR 제안 모델은 '나머지 방언'의 데이터 양의 비율이 지나치게 높았기 때문에 모델이 '나머지 방언'으로만 학습하는 경향이 생겨 성능이 오히려 베이스라인들보다 낮게 나타났다. 또한, Attn-BiLSTM 모델이 포함된 실험들은 국립국어원 NIKL DB를 포함한 경우의 정확도도 확인하였는데, 베이스라인 2, OvR 제안 모델, OvO 제안 모델 모두 NIKL DB를 포함하지 않았을 때의 성능보다 포함하였을 때의 성능이 더 높았다. 데이터 양이 추가된 점도 있으나 방언 특징이 짙은 데이터가 추가된 점도 성능 향상의 이유가 될 수 있다. 전체적으로 가장 성능이 좋은 OvO 제안 모델은 베이스라인 1에 비해 평균 25.34%p 정확도가 높았다.

〔표 2〕 한국어 방언 식별 실험 결과

평가 방식	연령대 + 발화유형	발화 단위 정확도(%)
베이스라인 1 [4] 6분류 단순 LSTM	(1) Young + Read	42.06
	(2) Old + Read	46.19
	(3) Young + Spon.	25.85
	(4) Old + Spon.	39.17
베이스라인 2 6분류 Attn-BiLSTM	(1) Young + Read	36.79
	(2) Old + Read	32.71
	(3) Young + Spon.	31.95
	(4) Old + Spon.	29.97
	(5) (4) + NIKL-DB	31.38
제안 모델 Attn-BiLSTM + OvR	(1) Young + Read	25.62
	(2) Old + Read	30.38
	(3) Young + Spon.	26.10
	(4) Old + Spon.	33.82
	(5) (4) + NIKL-DB	35.23
제안 모델 Attn-BiLSTM + OvO	**(1) Young + Read**	**62.34**
	(2) Old + Read	**66.21**
	(3) Young + Spon.	**58.74**
	(4) Old + Spon.	**67.33**
	(5) (4) + NIKL-DB	**68.51**

Ⅴ. 결론

본 연구에서는 한국어 방언 표지인 억양을 프레임 단위 음향 특징과 어텐션 쌍방향 LSTM 학습 방식을 통해 모델링 하여, 대용량 음성 코퍼스에서 한국어 방언 자동 식별 실험을 진행하였다. 발화 내 방언 특징 구간을 잘 학습하기 위해 어텐션 기법을 사용하였고, 연령대 및 발화 유형별로 실험을 나누어 진행하였으며, 예측 방식으로 OvR과 OvO 방식을 Attn-BiLSTM 모델과 함께 사용하는 것을 제안하였다. 모델 성능으로는 OvO 기반 Attn-BiLSTM 모델이 가장 우수하였고, 베이스라인에 비해 평균 25%p 성능이 향상하였다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음" (IITP-2021-2016-0-00464)

참 고 문 헌

[1] 이병운. (1998). 중부방언, 경남방언, 전남방언의 억양에 대한 비교 연구. 우리말연구, 8.

[2] 신지영, 김경화. (2017). 한국인 표준 음성 DB 구축 (II). 말소리와 음성과학, 9(2), 9-22.

[3] Mirsamadi, S., Barsoum, E., Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2227-2231). IEEE.

[4] Lee, J., Kim, K., Lee, K., Chung, M. (2019). Gender, Age, and Dialect Identification for Speaker Profiling. In 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA).