단어열 패턴 매칭과 Recurrent Neural Network를 이용한 하이브리드 음성 인식 오류 수정 방법

최준휘○, 류성한, 이규송, 박선영, 유환조, 이근배
포항공과대학교, 컴퓨터공학과
{chasunee, ryush, kyusonglee, sypark322, hwanjoyu, gblee}@postech.ac.kr

Hybrid ASR Error Correction Using Word Sequence Pattern and Recurrent Neural Network

Junhwi Choi, Seonghan Ryu, Kyusong Lee, Seonyeong Park, Hwanjo Yu, Gary Geunbae Lee
Department of Computer Science and Engineering, Pohang University of Science and Technology

요 약

본 논문에서는 단어열 패턴과 리커런트 신경망을 이용한 하이브리드 음성 인식 오류 수정 방법을 제안한다. 음성 인식 결과 문장에서 음성 인식 오류 단어가 발견되었을 경우에 첫째로 단어열 패턴과 그 패턴의 발음열 점수를 통해 1차적 수정을 하고 적절한 패턴을 찾지 못하였을 경우 음절단위로 구성된 Recurrent Neural Network를 통해 단어를 음절단위로 생성하여 2차적으로 오류를 수정한다. 해당 방법론을 한국어로 된 음성 인식 오류와 그 정답 문장으로 구성된 TV 가이드 영역 말뭉치를 바탕으로 성능을 평가하였고, 기존의 단순 단어열 패턴 기반의 음성 인식 오류 수정보다 성능이 향상되었음을 볼 수 있었다. 이 방법론은 음성 인식 오류와 정답의 말뭉치가 필요 없이 옳은 문장으로만 구성된 일반 말뭉치만으로 훈련이 가능하여, 음성 인식 엔진에 의존적이지 않는 강점이 있다.

주제어: 음성 인식 오류 수정, Recurrent Neural Network

1. 서론

음성 인식기는 음성 신호를 인식하여 그 실제 문장이 무엇인지 인식하는 시스템으로써 대화 시스템과 같은 많은 응용 프로그램에서 사용되고 있다. 일반적으로 음성 인식기는 응용 시스템의 컴포넌트로써 독립적으로 작동하며, 일반적으로 쉽게 접근 가능한 Google 음성 인식기와 같이 쉽게 이용할 수 있는 음성 인식기는 해당 응용 프로그램이 원하는 발화들의 언어 모델보다 큰 언어 모델을 이용하고 있어 오히려 해당 응용 프로그램에서 이용하기에 음성 인식 오류율이 높을 수 있다. 또한, 이러한 쉽게 이용할 수 있는 음성 인식기는 음성 인식 자체의 핵심 디코딩 부분이나, 각종 모델을 조정하도록 제공되지 않아 후처리 형태의 음성 인식 오류 수정이 필요하다.

기존의 많은 후처리 방식의 음성 인식 오류 수정 방법론은 음성 인식 결과와 그 본래 문장 쌍으로 구성된 병렬 말뭉치를 필요로 한다 [1], [2], [3]. 정민우 외 2명의 논문 [1]에서는 노이즈 채널 모델(Noisy Channel Model)을 이용하여 음성 인식 오류의 패턴을 검출 하였으고, 해당 모델은 음성 인식 결과와 그 본래 문장 쌍으로 이루어진 병렬말뭉치로 훈련되었다. Rinnger와 Allen의 논문 [2]에서도 언어 모델을 표현하기 위한 비터비 서치(Viterbi Search) 알고리즘과 노이즈 채널 모델을 이용하여 음성 인식 오류의 검출과 수정을 하는 후처리적 방법론을 제안하였는데, 역시 병렬 말뭉치가 필요하였다. Brandow와 Strzalkowski [3]는 규칙 기반의 방법론을 제안하였는데, 이 역시 병렬 말뭉치로부터 음성 인식 오류 수정 규칙을 생성하였다. 하지만 이러한 병렬 말뭉치는 얻기 힘들며, 음성 인식기의 성능과 그 음성 인식 환경에 크게 의존적이다. 예를 들어, 응용 프로그램의 음성 인식기가 변경되었을 경우에 기존에 훈련된 음성 인식 오류 수정 모델은 기존의 환경에 의존적이므로 효과가 없어진다. 이 경우에 음성 인식 오류 수정 모델은 새로운 음성 인식기와 그 환경에 맞추어 다시 만들어진 병렬 말뭉치에서 다시 만들어져야하고, 음성 파일 데이터가 존재하지 않는 경우에는 병렬 말뭉치의 생성조차 불가능해지게 된다.

본 논문에서는 위와 기술한 한계점을 극복하기 위해, 단어열 패턴 매칭 기반의 음성 인식 오류 수정 방법론과 Recurrent Neural Network(RNN) 기반의 방법론을 제안한다. 기존의 방법론과는 다르게, 제안하는 방법론은 필요한 모델들이 위에 기술한 병렬 말뭉치가 필요치 않고 일반적인 문장으로만 훈련 가능하므로, 음성 인식기에 의존적이지 않다.

2. 방법론

2.1. 전체 구조

제안하는 방법론의 구조는 크게 두 부분으로 이루어져 있다. 음성 인식 오류 검출부는 입력 문장으로부터 음성 인식 오류를 검출한다. 음성 인식 오류 수정부는 음성

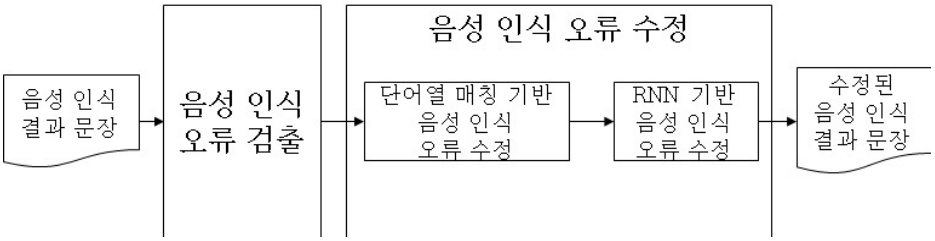

그림 1. 음성 인식 오류 수정 전체 구조

인식 오류 검출부로부터 검출된 오류를 수정한다. 위 구조에서 필요한 모든 모델은 음성 인식 오류가 없는 일반 말뭉치로부터 훈련된다. 본 논문에서는 음성 인식 오류 검출은 기존의 방법론[4]을 이용하고, 음성 인식 오류 수정에 초점을 맞추어 방법론을 제안한다.

2.2. 단어열 패턴 매칭 기반 음성 인식 오류 수정

단어열 패턴 매칭 기반 음성 인식 오류 수정은 일반 말뭉치에서 추출된 단어열 패턴과 그 패턴의 발음열을 바탕으로 수정한다. 단어열 패턴은 3 ~ 5 개의 단어로 구성되어 있으며, 단어는 형태소 단위로 분리된 것이다. 예를 들어 말뭉치에 '나는 사과를 먹었다' 와 같은 문장이 있다고 할 때, 이를 이형 형태소 분석을 통해 '나 는 사과 를 먹 었다' 로 분리하고, 이를 3 ~ 5개의 단어로 나누어 '나 는 사과', '는 사과 를', '사과 를 먹' 등과 같은 패턴을 말뭉치의 모든 문장으로부터 만들어 저장한다. 단어열 패턴을 평가하는데 있어서, 그 단어열의 발음열이 유용한데, 이는 음성 인식 오류가 있는 문장일지라 하더라도 그 발음열은 본래의 문장과 비슷하기 때문이다 [5]. 기본적으로 음성 인식기에서 제공되는 것이 음성 인식 결과 문장만이라고 가정하였기 때문에 발음열 평가를 위해 규칙 기반 Grapheme-to-phoneme(G2P) 모듈을 사용하였다.

시스템에 문장이 입력되면 해당 문장에서 교체할 부분과 교체하지 않을 부분을 결정한다. 음성 인식 오류 검출부에서 검출된 오류와 그 오류 앞 뒤의 단어를 교체할 부분으로 간주하고, 그 외의 단어는 교체하지 않을 부분으로 간주한다. 교체할 부분으로 오류 단어와 그 주위 단어를 함께 간주하는 이유는 오류가 발생하면 언어 모델의 특성으로 인해 그 주위 단어도 충분히 오류가 있을 수 있기 때문이다. 문장에서 모든 교체할 부분에 대해 교체할 부분과 그 주위의 단어들을 바탕으로 단어열 패턴을 구성하고, 그 패턴과 조건이 일치하는 패턴을 말뭉치에서 만들어 기저장된 패턴들 중에서 찾는다. 조건은 교체하지 않을 부분은 완벽히 일치해야하며, 교체할 부분은 다른 단어들이어야 한다는 조건이다. 이렇게 찾아진 패턴들을 발음열 유사도를 바탕으로 점수를 내어 평가하여 가장 점수가 높은 패턴으로 교체한다. 점수 계산 방법은 아래와 같다.

$$Replacing part\ \ Score\ \ s_i = \sum_{j \in C_i} \frac{\sum_{k \in M_j} (l_j - f(p_j, p_k))}{l_j}$$

여기서 C_i는 교체할 부분 i를 포함하여 구성된 패턴들의 집합이고, M_j는 패턴 j와 조건이 일치하여 검색된 패턴의 집합이다. l_j는 패턴 j의 발음열의 길이이고, p_j와 p_k는 패턴 j와 k의 발음열이며, 함수 f는 두 발음열의 레벤슈타인 거리(Levenshtein Distance)를 계산하는 함수이다. 이 수식에 따르면 가장 발음열이 유사하며 가장 많이 검색된 교체부분으로 교체하게 된다.

기술된 방법을 예를 들어 설명하면, '나 는 사각 을 먹 었다' 에서 '사각' 이 검출된 오류 단어라고 할 때, 교체할 부분은 '는 사각 을' 이 되며, 교체하지 않을 부분은 '나', '먹', '었다' 가 된다. 여기서 교체할 부분의 주위 단어를 포함한 패턴을 구성하게 되는데, 해당 패턴은 '나 는 사각 을 먹' 과 '는 사각 을 먹 었다' 와 같은 패턴이 구성되고, 해당 패턴과 조건과 부합하는 패턴을 찾으면 '나 는 사과 를 먹', '나 도 과일을 먹', '는 사과 를 먹 었다' 과 같은 패턴이 검색된다. 이 중에 발음열이 유사하고 많이 나와 점수가 가장 높은 교체 부분은 '는 사과 를' 이 되어 '는 사각 을' 은 '는 사과 를' 로 교체된다. 해당 방법은 오류 수정 방법이 매우 직관적이고 쉬우나, 만족되는 패턴이 없을 경우에는 고치지 못하는 단점이 있다.

2.3. Recurrent Neural Network 기반 음성 인식 오류 수정

음성 인식 오류 검출에서 검출된 음성 인식 오류 부분이 단어열 패턴 매칭 기반 음성 인식 오류 수정 방법으로 수정하려 하였으나, 만족되는 패턴이 없어 음성 인식 오류를 수정하지 못하는 부분이 발생하였을 때를 대비하여 RNN 기반의 음성 인식 오류 수정으로 보조하도록 하였다. 일반적인 Feed-forward Neural Network와 달리 RNN은 연속된 데이터를 바탕으로 다음 데이터의 예측에 매우 특화되어 있는 Neural Network이다. 음성 인식 문장의 오류 부분을 수정하기 위해서 우리는 앞 문맥의 흐름을 고려하여 해당 오류 부분이 원래 무엇이었는지 음절 단위 예측을 통해 알아내고자 하는 것이다. 음절 단위로 예측을 하는 이유는 일반적인 단어 단위의 예측은 단어의 개수만큼으로 입력 및 출력 차원이 커지며, 소수

단어의 예측이 쉽지 않고, 그만큼 데이터의 Sparseness 문제 또한 발생하기 때문이다. 음절 단위를 통해 이를 해결하면, 입력 및 출력 레이어의 차원을 획기적으로 줄이며, 또한 음절 단위로 데이터를 취급하므로, 데이터의 Sparseness 문제 또한 해결할 수 있다.

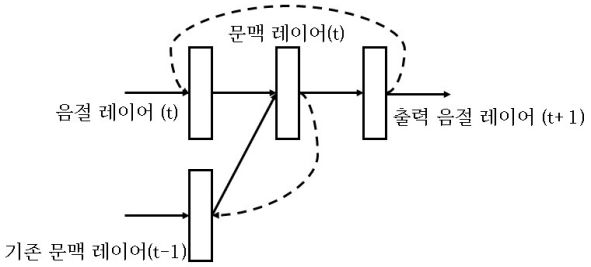

그림 2. RNN 기반 음성 인식 오류 수정을 위한 RNN 구조

음성 인식 오류 수정을 위한 RNN 구조는 일반적인 RNN 기반 언어 모델의 구조와 유사하다 (그림 2) [6]. 훈련 방법으로는 일반적인 역전파(Back-propagation) 알고리즘으로 사용하였다. 문맥 레이어를 구성할 때의 활성 함수는 Sigmoid 함수를 이용하였고, 출력 음절 레이어를 구성할 때의 활성 함수는 Softmax 함수를 이용하였다. 음절 레이어와 출력 음절 레이어는 음절의 1-of-N 코딩을 통하여 구성되었다. 문장의 첫 음절부터 끝 음절까지 신경망에 넣어 훈련하였고, 단어의 끝마다 End 심볼을 넣어 훈련하였다.

예측 시에는 입력 레이어는 음절의 1-of-N 코딩을 통하여 입력되는 레이어와 기존의 문맥 레이어가 합쳐져 구성된다. t=0 일 때의 음절 레이어는 검출된 오류 단어 직전 단어의 마지막 음절이고, 기존 문맥 레이어는 오류로 검출된 단어의 직전 단어까지의 음절들이 축적된 레이어를 사용한다. t=1 부터의 음절 레이어는 예측된 음절 레이어가 들어가며, 직전 t-1의 문맥 레이어가 기존 문맥 레이어로 사용된다. 오류에 해당하는 단어에서 음절을 예측할 때 단어의 End 심볼을 넣어 같이 훈련하였으므로, 그에 따라 End 심볼이 나올 때까지의 음절을 포함하여 예측 단어가 구성된다. 예측 단어가 구성이 완료되면 오류 단어는 예측 단어로 교체된다.

4. 실험 및 결과

표 1. WER 감소율

	WER 감소율 (%)
단어열 패턴 매칭 기반	27.8
RNN 기반	7.7
단어열 패턴 매칭 + RNN 기반	28.2

실험을 위해 음성 인식 결과와 그 정답 문장으로 이루어진 약 6,500 문장쌍의 병렬 말뭉치를 준비하였다. 해당 음성 인식 오류 문장을 생성하기 위해 사용한 음성 인식기는 Hidden Markov Model 기반으로 이루어져 약 300,000 단어의 규모의 언어 모델로 훈련이 되었으며, 단어 단위 오류율(Word Error Rate, WER)은 약 16.43%이다. 제안된 방법론에서 사용되는 모델들을 훈련하기 위한 말뭉치로는 약 29,000 개의 오류가 없는 일반 말뭉치로 훈련이 되었다. 사용된 말뭉치들은 모두 한국어이며, TV 가이드 영역의 말뭉치이다. RNN을 구성하는 입력 및 출력 음절 레이어는 한글 기준으로 약 11,000 차원으로 구성되었으며, 문맥 레이어는 500차원으로 구성하였다.

표 2. 단어열 패턴 매칭 모델의 사용량 대비 WER 감소율

단어열 패턴 매칭 모델 사용량 (%)	단어열 패턴 매칭 기반 방법의 WER 감소율 (%)	RNN 기반 방법의 WER 감소율 (%)
100	27.8	0.4
90	25.0	0.5
80	22.2	1.3
70	19.4	1.5
60	16.7	2.9
50	14.0	4.1
40	11.1	4.8
30	8.2	5.5
20	5.6	6.3
10	2.6	7.5
0	0.0	7.7

제안된 음성 인식 오류 방법을 통해 음성 인식기의 결과 문장의 WER을 감소시켰다(표 1). 단일 방법으로는 단어열 패턴 매칭 기반의 음성 인식 오류 수정 방법론이 RNN 기반의 음성 인식 오류 수정 방법론 보다 효과적이었다. 그러나 단어열 패턴 매팅 기반의 음성 인식 오류 수정 방법론의 단점으로는 패턴이 존재해야만 음성 인식 오류 수정에 강점이 있으며, 훈련 말뭉치가 커감에 따라 그 모델의 크기가 기하급수적으로 커져, 수정 시간이 다소 오래 걸린다는 단점이 있다. 반면 RNN 기반 음성 인식 오류 수정은 훈련 데이터의 양과는 상관 없이 모델의 크기는 일정하며, 훈련되면 될수록 성능이 더 좋아지는 장점을 가지고 있고, 어떤 예측도 음절 단위로는 일정한 예측 시간을 가지고 있다는 장점이 있다. 그러나 그 성능이 단어열 패턴 매칭 기반에 이르기엔 한참 부족하여, RNN 기반의 음성 인식 오류 수정은 단어열 패턴 매칭 기반의 보조적으로 이용하여 결합한 방법이 가장 좋은 성능을 기록하였다.

RNN 기반의 음성 인식 오류 수정의 효과를 단어열 패턴 매칭 모델의 사용량에 따라 평가하기 위해 단어열 패턴 매칭 모델의 사용량을 줄여가며 RNN 기반의 성능을 평가 하였다(표 2). 29,000개의 문장으로 기 구축된 단어열 패턴 매칭 모델에 존재하는 패턴들을 임의로 줄여감에 따라 RNN 기반의 음성 인식 오류 수정 방법론의 성능이 증가하였다. 단어열 패턴 매칭 기반의 음성 인식 오류 수정은 직접적으로 단어열 패턴에 의존하고 있기 때문에, 일반적이지 않은 문장을 이용하는 경우에는 음성 인식 오류 수정이 잘 이루어지지 않을 가능성이 있다.

5. 결론

이 논문에서는 음성 인식기 엔진에 의존적이지 않은 후처리 기반의 음성 인식 오류 수정 방법론을 제안하였다. 먼저 단어열 패턴 매칭 기반의 방법론으로 수정을 하고 수정에 실패하였을 경우에 보조적인 방법으로 RNN 기반의 음성 인식 오류 수정을 수행한다. 본 방법을 통하여, 한국어 TV 가이드 영역의 음성 인식 결과 문장의 오류를 약 28.2% 감소시켰다. 또한 두 모델의 각각의 성능을 구체적으로 평가하기위해 단어열 패턴 매칭 모델의 사용량에 따른 RNN 기반 방법론의 성능을 평가하였다.

본 방법론은 영역에 한정되지 않은 오픈된 음성 인식기를 사용하는 한정된 영역의 응용 프로그램에서 효과를 볼 수 있는 방법론이며, 또한 음성 인식기를 직접적으로 다룰 수 없을 경우에 쉽고 간단하게 음성 인식기의 성능을 높일 수 있는 방법이다. 본 방법론은 영역 적응과 같은 방법론과 유사하고, 음성 인식 결과 말뭉치를 이용치를 이용하지 않는다는 점에서 음성 인식기 엔진에 의존적이지 않으며, 따라서 음성 인식기를 사용하는 어떠한 응용 프로그램에도 적용될 수 있다.

사사

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송연구개발사업의 일환으로 수행하였음. [B0101-15-0307 , 인간 수준의 평생기계학습 SW 기초 연구 (기계학습연구센터)]

참고문헌

[1] Jeong, M., Jung, S., Lee, G.G., Speech recognition error correction using maximum entropy language model, In: Proc. of INTERSPEECH, pp. 2137-2140, 2004.

[2] Ringger, E.K., Allen, J.F., A fertility channel model for post-correction of continuous speech recognition, In: Spoken Language, 1996. ICSLP 96. Proceedings Fourth International Conference on. vol. 2, pp. 897-900, IEEE, 1996.

[3] Brandow, R.L., Strzalkowski, T., Improving speech recognition through text-based linguistic post-processing (May 16 2000), uS Patent 6,064,957.

[4] Choi, J., Lee, D., Ryu, S., Lee, K., Kim, K., Noh, H., Lee, G.G., Engineindependent asr error management for dialog systems, In Intenational Workshop Series on Spoken Dialogue Systems Technology (IWSDS), 2014.

[5] Choi, J., Kim, K., Lee, S., Kim, S., Lee, D., Lee, I., Lee, G.G., Seamless error correction interface for voice word processor. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. pp. 4973-4976, IEEE, 2012.

[6] Mikolov, T., Karafia´t, M., Burget, L., Cernocky`, J., Khudanpur, S., Recurrent neural network based language model, In INTERSPEECH, pp. 1045-1048, 2010.