대용량 연속 음성 인식 시스템에서의 코퍼스 선별 방법에 의한 언어모델 설계

오유리, 윤재삼, 김홍국
광주과학기술원, 정보통신공학과

A Corpus Selection Based Approach to Language Modeling for Large Vocabulary Continuous Speech Recognition

Yoo Rhee Oh, Jae Sam Yoon, and Hong Kook Kim
Dept. of Information and Communications, Gwangju Institute of Science and Technology

{yroh, jsyoon, hongkook}@gist.ac.kr

Abstract

In this paper, we propose a language modeling approach to improve the performance of a large vocabulary continuous speech recognition system. The proposed approach is based on the active learning framework that helps to select a text corpus from a plenty amount of text data required for language modeling. The perplexity is used as a measure for the corpus selection in the active learning. From the recognition experiments on the task of continuous Korean speech, the speech recognition system employing the language model by the proposed language modeling approach reduces the word error rate by about 6.6 % with less computational complexity than that using a language model constructed with randomly selected texts.

Ⅰ. 서론

대용량 연속음성 인식시스템에 대한 연구는 외국에서 활발히 진행되고 있는 반면 한국어 대용량 연속음성 인식시스템에 대한 연구는 크게 미비하여, 영어에 비해 좋은 인식 성능을 보이지 못한다 [1]. 더욱이, 한국어는 단어의 확장성이 커서 단어와 단어 사이에 공백 없이 사용하는 경우가 많다. 그러므로 발음사전 및 언어모델을 생성하는데 어려움이 크다. 일반적으로, 한국어를 포함한 대용량 연속음성 인식시스템의 언어모델을 생성하기 위해서는 많은 텍스트 데이터가 필요하다. 그러나 언어모델을 위하여 수집된 텍스트 중에는 인식시스템의 성능을 향상시키는 것과 저하시키는 것이 혼재한다. 무작위로 수집된 텍스트 코퍼스를 정제 없이 사용하는 경우, 언어모델의 크기가 증가하여 인식시스템의 수행시간이 길어진다. 다시 말해서 적은 텍스트 데이터 량에 의해 학습된 언어모델을 사용할 경우가 상대적으로 많은 데이터 량을 사용하여 구성된 언어모델을 갖는 인식시스템이 좋은 성능을 항상 갖는 것은 아니다. 이러한 이유로 언어모델 설계에 능동 학습(active learning) 기법 [2]을 적용하여, 수집된 텍스트 중에서 인식시스템에 적합한 텍스트를 선택적으로 골라내는 일이 필요하다. 능동 학습 기법은 무작위로 수집한 데이터를 모두 사용하여 학습하는 것이 아니라, 수집된 데이터 중 시스템에 적절한 데이터를 골라내어(selective sampling) 학습함으로써 시스템의 성능을 향상을 도모하는 방법이다.

본 논문에서는 능동 학습 기법을 이용하여 수집된 텍스트를 선별적으로 사용함으로써, 효과적으로 언어모델을 제작하는 방법을 제안한다. 인식시스템의 성능에 긍정적인 영향을 미치는 텍스트를 선별하는 척도로써 perplexity를 사용하였다. 먼저, 텍스트 코퍼스를 적당한 크기의 문장 뭉치들로 쪼개어 언어모델을 생성하고 perplexity를 구한다. 그리고 perplexity가 낮은 문장 뭉치 순서로 언어모델에 적합한지 판단을 하여 선택하는 방법이다.

본 논문의 구성은 다음과 같다. 서론에 이어 제 2 절에서는 언어모델을 적용할 대용량 한국어 연속음성 인식시스템을 소개한다. 그리고 제 3 절에서는 perplexity를 척도로 하는 능동 학습 기법을 이용하여

언어모델을 효과적으로 생성하는 방법을 제안하고, 제 4 절에서 제안한 언어모델 제작 방법을 이용하여 인식 시스템의 성능 평가를 보인다. 마지막으로 제 5 절에서 결론을 맺도록 한다.

Ⅱ. 한국어 연속음성 인식시스템

본 절에서는 언어모델의 성능을 평가하기 위한 대용량 한국어 연속음성 인식시스템을 소개한다.

본 논문에서는 음성정보기술산업지원센터(SiTEC)에서 제공하는 낭독문장 음성 DB (CleanSent01) [3]의 일부를 한국어 연속음성 인식시스템의 학습용 데이터로 사용하였다. CleanSent01은 21세기 세종계획 형태소 분석 균형 말뭉치 10,000 만 어절 중, 형태소 빈도를 고려한 20,217 문장과 PBS 589 문장을 발화한 음성으로 구성되어 있다. 또한 발화 음성 데이터는 방음실 환경에서 제작되었으며, AKG C414-ULS와 Sennheiser 마이크로 동시에 녹음되어 16 kHz의 샘플링레이트, 16 bit 로 저장되었다.

CleanSent01에서 20,806 문장을 남녀 200명이 발화한 음성은 총 200 세트로 구성되어 있다. 여기서 학습용 데이터는 set001 ~ set170 와 set181 ~ set190 이 사용되었으며, 17,006 문장이 발화된 음성으로 총 30,633 개의 서로 다른 단어를 포함한다. 그리고 평가용 데이터는 학습용 데이터에서 사용하지 않은 set171 ~ set180, set191 ~ set200 중에서 일부인 97 문장을 사용하였다.

한국어 대용량 연속음성 인식시스템에는 39 차 특징벡터가 사용되었으며, 이를 위하여 12 차 멜-켑스트럼 계수(MFCC), 로그 에너지를 ETSI에서 제공하는 Front-End 특징 추출 알고리즘 [4]을 이용하여 추출하였고, 1 차, 2 차 미분계수를 사용하였다. 또한 인식 및 학습에 사용된 특징벡터에 켑스트럼 평균 정규화와 에너지 정규화 기법이 적용되었다.

음향 모델을 3-state의 left-to-right 가 사용되었으며, 문맥 독립적이고, 4 개의 혼합밀도와 cross-word 트라이폰 모델을 사용하였다. 또한 음향 모델은 HTK version 3.2 toolkit [5]을 이용하여 학습되었다. 먼저 두 개의 묵음 모델이 포함된 42 개의 단음에 기반을 둔 음향모델에서 시작하여 트라이폰에 기반을 둔 음향모델로 확장한 후, 이진트리를 이용한 상태 공유 단계를 거쳐 상태의 수를 줄였다. 그 결과 인식 시스템은 14,901 개의 트라이폰과 14,485 상태의 음향 모델로 구성된다. 표 1은 한국어 대용량 연속음성 인식시스템을 위하여 사용한 묵음을 제외한 단음을 정리한 것이다. 본 논문에서 사용하는 인식시스템은 트라이폰 음향모델을 사용하기 때문에, 초성 및 종성에 대한 소리는 따로 구분하지 않았다. 예를 들면, 초성의 'ㄱ'과 종성의 'ㄱ'은 소리의 특성이 다르기는 하지만, 하나의 'ㄱ' 모델을 사용한다.

표 1: 한국어 연속음성 인식시스템에서 사용된 묵음을 제외한 40 개의 단음.

모음	단모음	ㅣ(i), ㅔ(e), ㅐ(E), ㅜ(u), ㅗ(o), ㅏ(a), ㅡ(U), ㅓ(v), ㅚ(O)
	이중모음	ㅟ(wi), ㅞ(we), ㅙ(wE), ㅝ(wv), ㅘ(wa), ㅖ(je), ㅒ(jE), ㅑ(ja), ㅕ(jv), ㅛ(jo), ㅠ(ju), ㅢ(xi)
자음		ㅂ(b), ㄷ(d), ㅍ(p), ㅌ(t), ㅃ(B), ㄸ(D), ㅅ(s), ㅆ(S), ㅈ(z), ㅊ(c), ㅉ(Z), ㅁ(m), ㄴ(n), ㄹ(l)

인식시스템에 사용한 발음사전은 CleanSent01 데이터베이스에서 제공하는 사전을 기본으로 하였다. 발음사전에 없는 발음에 대해서는 "표준어 규정"의 표준발음법을 바탕으로 한 발음생성기를 사용하였다. 마지막으로 본 논문에서 한국어 대용량 연속음성의 인식단위는 어절이다.

Ⅲ. 언어모델 설계

서론에서 언급했듯이 대용량 인식시스템의 언어모델을 생성하기 위하여 많은 텍스트 데이터를 사용한다. 그러나 데이터의 량이 많아지면 인식시스템의 수행속도가 증가할 뿐 아니라, 데이터 량에 비례하는 성능 향상을 기대하기는 일반적으로 어렵다. 더구나 인식시스템의 대상이 달라질 때마다 그에 해당하는 텍스트를 새로 수집하는 것은 매우 비효율적이다. 이러한 이유로 무작위로 수집된 텍스트 코퍼스에서 언어모델의 성능을 향상시키는 텍스트들만을 선별하여 언어모델을 생성하는 것이 효과적이다. 본 절에서는 인식시스템의 성능을 향상시키기 위하여 능동 학습 기법에 의해 텍스트 코퍼스에서 선별적으로 텍스트를 추출한 후 언어모델을 설계하는 방법을 제안한다.

본 논문에서는 텍스트 코퍼스에서 텍스트들이 인식시스템에 적합한지를 판단하는 척도로 perplexity를 사용하였다. perplexity는 수식 (1)과 같이 표현되며, $PP(W)$가 perplexity를 나타내고 $H(W)$는 cross-entropy를 나타낸다. 또한 W는 테스트 데이터를, N_W는 테스트 데이터의 크기를 나타낸다.

$$PP(W) = 2^{H(W)} \qquad (1)$$

$$where,\ H(W) = -\frac{1}{N_W} log_2 P(W)$$

일반적으로 perplexity는 언어모델의 성능을 평가하는

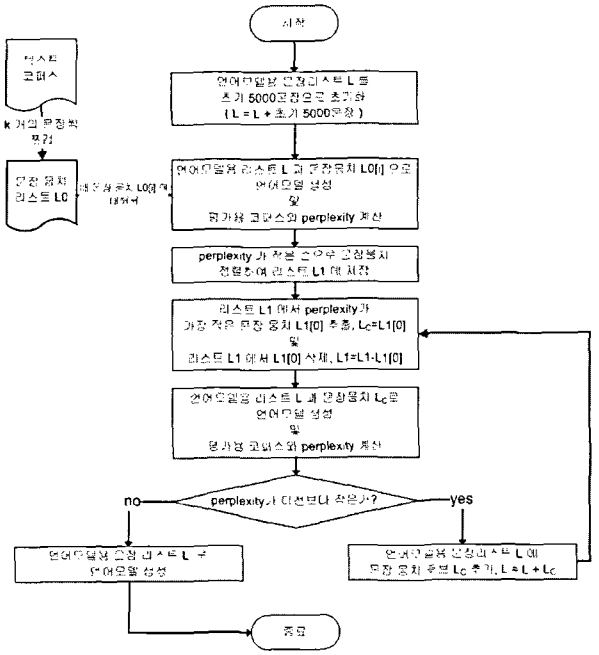

그림 1. perplexity를 척도로 하는 능동 학습 기법에 의해 선별한 후 텍스트 코퍼스에서 대용량 연속음성 인식시스템에 적합한 텍스트를 언어모델을 설계하는 과정.

데 사용될 뿐 아니라, 단어오류율이 perplexity와 수식 (2)와 같이 밀접한 관계가 있다는 연구가 있다 [6].이는 perplexity가 작을수록 단어오류율(%)이 줄어든다는 것이다. 이러한 이유로 본 논문에서 인식시스템에 적합한 텍스트를 찾기 위한 척도로 perplexity를 선택하였다.

$$WER = bPP^{a} \qquad (2)$$

그림 1은 텍스트 코퍼스에서 perplexity를 이용하여 인식시스템에 적절한 텍스트를 선별한 후 언어모델을 생성하는 과정을 나타낸다. 먼저 무작위로 선택한 5,000 문장으로 언어모델용 문장리스트 L을 초기화한다. 또한 텍스트 코퍼스를 x개의 문장을 가지는 문장 뭉치들로 쪼갠다. 그 후, 각각의 문장 뭉치와 초기 언어모델용 5,000개의 문장으로 언어모델을 생성한 후 perplexity를 구한다. perplexity가 작은 순으로 문장 뭉치들을 정렬하여 리스트 L1을 만든다. 그림 1에서 리스트 L1[0]은 리스트 L1에서 색인 0번째의 문장 뭉치로, perplexity가 가장 작은 문장 뭉치를 나타낸다. 다음 과정은 L1[0]의 문장 뭉치를 언어모델에 적합한 텍스트의 후보 L_C로 두고 L1에서 L1[0]을 삭제한 후, 언어모델용 문장리스트 L과 후보 문장 뭉치 L_C로 언어모델을 생성하여 perplexity를 계산한다. 만약, perplexity가 이전보다 작으면 L_C는 인식시스템에 추가

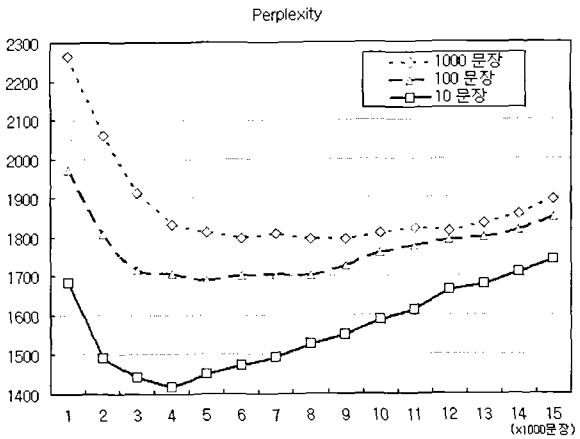

그림 2: 언어모델용 리스트 L과 후보 L_C로 언어모델을 만드는 과정을 15번 반복했을 때 perplexity 비교.

된다. 또한 다음 문장 후보를 찾기 위하여 L1을 검토하는 단계로 돌아간다. 반면 perplexity가 이전보다 크면 L_C는 인식시스템에 적합하지 않다고 판단하여 버리고, L1에 더 이상 인식시스템에 적합한 텍스트가 없으므로 텍스트 선별과정을 종료한다. 최종적으로 언어모델용 문장리스트 L에 모인 텍스트들만으로 언어모델을 생성한다.

Ⅳ. 실험

본 절에서는 제 3 절에서 제안한 언어모델 제작 방법을 실제로 한국어 대용량 연속음성 인식시스템에 적용한 성능 평가를 보인다.

인식시스템은 제 2 절에서 소개한 한국어 대용량 연속음성 시스템을 사용하였고 언어모델은 어절 단위의 back-off bigram을 사용하였다. 또한 텍스트 코퍼스로 ETRI에서 제공하는 "음성인식 언어모델용 텍스트 DB"의 일부를 사용하였다. ETRI 텍스트 코퍼스에서 상관관계가 전혀 없는 텍스트는 제외하였다. 다시 말해, 평가용 텍스트에서 사용된 어절을 전혀 포함하지 않는 텍스들은 제외하였다. 이 과정을 거쳐서 얻은 텍스트들은 총 41,642 문장이다.

초기 언어모델용 5,000 문장은 CleanSent01에서 인식시스템의 학습에 사용된 17,006 개의 발화 문장 중에서 무작위로 선택하였다. 초기 언어모델에 대한 인식시스템의 단어오류율은 30.88%이다. 그리고 문장 뭉치의 크기가 10문장, 100문장, 1,000문장일 때를 나누어 실험하였다.

그림 2는 언어모델용 리스트 L과 후보 L_C로 언어모델을 만들고 perplexity를 구하는 과정을 15번 반복한 결과이고, 그림 3은 각각의 언어모델에 대한 단어오류

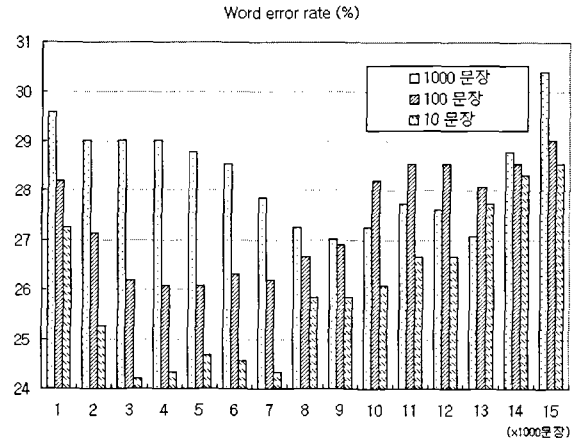

그림 3: 언어모델용 리스트 L과 후보 L_C로 언어모델을 만들고 perplexity를 구하는 과정을 15번 반복할 때 각 언어모델에 대한 단어오류율(%) 비교.

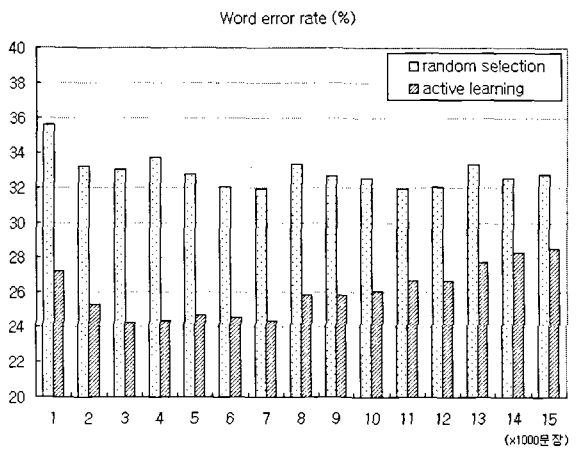

그림 4: 능동 학습 기법으로 텍스트를 선정하여 언어모델을 만든 경우와 무작위로 텍스트를 선정하여 언어모델을 만든 경우에 대한 단어오류율(%) 비교.

율(%)을 구한 것이다. 그림 2와 그림 3으로부터, 언어모델용 리스트 L과 후보 L_C로 언어모델을 만든 후 문장 후보 L_C를 평가하는 척도로 perplexity가 적합함을 알 수 있다. 또한, 문장 뭉치가 1,000 문장인 경우 9,000 문장을 추가하였을 때 27.02%의 단어오류율로서 3.86%의 감소를 보인 반면, 문장 뭉치가 10 문장인 경우 3,000 문장을 추가하였을 때 24.22%의 단어오류율로서 6.66%의 감소를 보였다. 이 결과는 문장 뭉치의 크기가 작을수록 보다 적은 문장으로 높은 단어오류율 감소를 얻을 수 있음을 보인다.

그림 4는 능동 학습 기법을 바탕으로 본 논문에서 제안한 제작 방법에 의한 언어모델(문장 뭉치 크기 = 10 문장)과 무작위로 텍스트를 추가한 언어모델에 대한 인식시스템의 단어오류율(%)을 비교한 결과이다. 이 결과는, 능동 학습을 이용하여 언어모델을 제작하는 경우, 보다 적은 데이터로 높은 단어오류율 감소를 얻을 수 있음을 보인다.

V. 결과

본 논문에서는 대용량 연속음성 인식시스템의 성능 향상을 위하여 능동 기법을 바탕으로 언어모델을 설계하는 방법을 제안하였다. 즉, 언어모델을 위하여 수집된 텍스트 코퍼스에서 인식시스템의 성능을 향상시키는 텍스트만을 추출하여 언어모델을 생성하는 방법을 제시하였다. 또한 언어모델에 적합한 텍스트 여부를 판단하기 위하여 perplexity가 사용되었다. 실험 결과, 제안하는 언어모델 설계 방법을 이용하는 경우, 적은 계산량으로 한국어 연속음성 인식시스템의 단어오류율을 약 6.6% 감소시킬 수 있었다.

Acknowledgement

본 연구는 광주과학기술원 실감방송 연구센터(RBRC)를 통한 정보통신부 대학IT연구센터(ITRC) 사업의 지원, 광주과학기술원 실감콘텐츠 연구센터(ICRC)를 통한 과학기술부 특정연구개발 사업의 지원과 광주과학기술원[GIST] 기관고유사업 지원에 의해 수행되었습니다.

참고문헌

[1] 김봉완, 이용주, "연속 음성 인식을 위한 PTM 음절 모델," 한국음향학회 학술발표대회 논문집, 제 23 권 제 1(s)호, pp. 33-36, May. 2004.

[2] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," Machine Learning, vol. 15, pp. 201-221, 1994.

[3] 김봉완, 최대림, 김영일, 이광현, 이용주, "SiTEC의 공동 이용을 위한 음성 코퍼스의 구축 현황 및 계획," 대한음성학회 말소리, 제 46호, pp. 175-186, Jun. 2003.

[4] ETSI ES 202 050 v1.1.3, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature (2003).

[5] S. Young, *et al*, "The HTK Book (for HTK Version 3.2),"Microsoft Corporation, Cambridge University Engineering Department, Dec. 2002.

[6] Dietrich Klakow, Jochen Peters, "Testing the correlation of word error rate and perplexity," Speech Communication, vol. 38, no. 1, pp. 19-28, Sep. 2002