대화체 연속음성 인식을 위한 언어모델 적응

박영희, 정민화
서강대학교 컴퓨터학과

Language Model Adaptation for Conversational Speech Recognition

Young-Hee Park, Minhwa Chung
Department of Computer Science, Sogang University
E-mail : {yhpark, mchung}@sogang.ac.kr

Abstract

This paper presents our style-based language model adaptation for Korean conversational speech recognition. Korean conversational speech is observed various characteristics of content and style such as filled pauses, word omission, and contraction as compared with the written text corpora. For style-based language model adaptation, we report two approaches. Our approaches focus on improving the estimation of domain-dependent n-gram models by relevance weighting out-of-domain text data, where style is represented by n-gram based tf*idf similarity. In addition to relevance weighting, we use disfluencies as predictor to the neighboring words. The best result reduces 6.5% word error rate absolutely and shows that n-gram based relevance weighting reflects style difference greatly and disfluencies are good predictor.

I. 서론

통계적 언어모델은 음성인식, 정보검색과 같은 다양한 도메인에서 좋은 성능을 보여주고 있지만, 좋은 모델을 만들기 위해서는 같은 영역의 대용량 학습 데이터를 필요로 한다. 대화체 연속음성 인식을 위한 언어모델 생성을 위해서는 대용량의 방송뉴스나 신문기사와 같은 타 도메인 텍스트를 이용하는 것이 일반적이다 [1].

대화체 음성은 방송뉴스나 신문기사와 같은 문어체 텍스트와는 내용(content)과 스타일(style)에서 매우 다른 특징을 보인다. 특히 스타일의 경우는 간투어("어", "음")가 빈번히 사용되고, 줄여서 발화("폴/표를")하거나 생략해서 발화하는 현상("여행삽니다/여행사입니다"), "요"의 빈번한 사용 등의 현상이 대화체에서는 빈번히 나타난다. 이러한 대화체의 특징들 때문에 문어체 텍스트를 그냥 결합하여 사용하는 것은 오히려 성능을 떨어뜨리는 결과를 가져온다 [1][2]. In-domain과 out-of-domain 언어모델을 interpolation하여 언어모델을 적응하는 방법 역시 내용(content)에 대해서는 효과적이지만, 스타일을 반영하기에는 어려움이 있다. 많은 연구들이 이러한 대화체의 특징들을 n-gram 언어모델에 반영하는 것에 중점을 두고 있다. Style과 content의 유사성을 반영하기 위하여 품사/단어 n-gram 통계정보를 가중치로 이용하여 문서 단위로 언어모델을 결합하거나 [1], 모든 단어들을 disfluencies, 토픽 단어, 일반 단어의 카테고리로 분류하고 각 카테고리 별로 가중치를 다르게 하여 언어모델 적응 [3], maximum entropy를 이용하여 언어모델 결합방법 [4] 등이 있다.

본 논문에서는 효과적인 대화체 언어모델의 적응을 위하여 대화체의 스타일을 최대로 반영할 수 있도록 한하는데 초점을 두어 두가지 측면에서 접근하였다. 첫 번째는 n-gram 기반의 tf*idf (Term Frequency Inverse Document Frequency) 유사도를 가중치로 이용하여 문서 단위로 대화체와 타 도메인을 결합하였다. 형태소가 축약되거나 생략되는 것과 같은 대화 현상들은 앞뒤 형태소에 의존적으로 일어나는 현상이지만, 정보검색 기법을 이용한 기존의 연구는 단어의 unigram 정보만을 이용하므로 이러한 현상의 반영이 어렵다는 문제점을 가지고 있다 [1][5]. 이에 대한 해결 방안으로 본 논문에서는 bigram 단어열에 대한 가중치를 구하여 위의 문제를 해결하였다.

게다가 여러 대화현상들 중에서 간투어는 문어체에는 잘 나타나지 않지만 대화체에는 매우 빈번히 나타

나며 인식 성능을 떨어뜨리는 요소이기도 하다 [2][6]. 간투어가 잡음과 같은 비언어적인 요소로 분류되기도 했었지만, 최근에는 주변 단어에 대한 예측 기능이 있다고 알려져 있다 [6]. 본 논문에서는 이와 같은 간투어의 예측 기능을 모델링하여 평가하였다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 실험에 사용한 대화체 코퍼스의 소개 및 특징을 분석하고, 3장에서 본 논문에서 제안하는 언어모델 적응 방법을 설명한다. 4·5장에서는 3장에서 생성한 언어모델을 이용한 인식실험과 결론을 기술한다.

II. 대화체 코퍼스

2.1 코퍼스

대화체 코퍼스는 한국전자통신 연구원의 C-STAR 과제를 위해 98년도에 구축한 여행 계획 영역의 대화체 음성 데이터베이스로, 여행사 직원과 고객의 가상 대화를 녹취한 것이다. 각 대화의 시나리오는 호텔 예약과 교통편 문의 등의 복합적인 내용으로 구성되었다. 총 50 화자, 100 대화, 62,946 어절로 구성되어 있다. 잡음, 간투어, 반복/수정 발화 등의 주석을 달고 표준발음을 전사하는 등의 과정을 수행하였다. 간투어는 전체 어절의 11.2%로 많은 부분을 차지하므로 간투어 모델링의 중요성을 보여준다. 게다가 빈도수가 높은 10개의 간투어가 전체 간투어의 80% 이상을 차지하고 있다 (자세한 사항은 [2] 참조).

다음 단계로 형태소 분석을 수행하였다. 본 논문의 인식 단위는 형태소이므로 대화현상에 대한 분석도 형태소를 기본 단위로 하였다. 형태소 분석 후, 코퍼스의 크기는 103,406 형태소이고, 유일 형태소 수는 2,292이며 모두 발음사전의 어휘로 사용하였다.

타 영역의 데이터로 사용할 텍스트 DB는 약 16M 형태소 크기의 방송뉴스와 신문 기사들이고, 약 42,600개의 문서로 이루어져 있다. 방송뉴스는 약간의 인터뷰를 포함하고 있지만, 위와 같은 대화 현상들이 전사되어있지는 않다.

2.2 한국어 대화체 음성의 대화특징

대화체 음성을 방송뉴스와 같은 문어체와 비교할 때, 잡음, 간투어, 반복/수정 발화 등의 현상을 포함할 뿐 아니라 한국어 대화체에는 다음과 같은 특징들이 빈번히 나타나고 있다.

- "요", "은/는"과 같은 보조사의 빈번한 사용 - 보조사 "요"는 용언구의 뒤에 나타나고, 대화체 코퍼스에서는 4.4%가 나타난 반면, 방송뉴스에서는 0.2%만 관측되었다. 또한 특수 보조사 "은/는"은 "면" 뒤에서만 나타난다. (예: "하구요", "하면은")
- 서술격 조사 "이"의 생략 - "여행삽니다/여행사입니다"와 같이 축약하여 발화하는 경향 때문에 서술격 조사 "이"가 빈번히 생략된다. Force alignment 결과, 약 22%가 생략되었다.
- 연결어미 "하고"가 "와/과"보다 더 많이 사용된다. (예: "이름하고 번호")
- 어미 또는 조사의 축약. (예: "푤/표를", "주십쇼/주십시요")

위의 특징들은 형태소를 기본 단위로 하여 [2]의 결과를 바탕으로 분석된 것으로, 대화현상 중 언어모델에 영향을 주는 것들만을 나열하였다. 이외의 특징들은 다중 발음사전으로 처리하였다 ([2]참조).

기존의 토픽 기반의 언어모델 적응 방법은 기능어는 제외시키는 것이 일반적이지만, 위의 특징들에서 보듯이, 기능어 자체는 중요하지 않지만 대화체의 특징들이 주로 기능어와 관련되어 나타나므로 새로운 언어모델 적응 방법이 필요함을 알 수 있다.

III. 언어모델 적응

한국어의 대화음성의 스타일을 잘 반영할 수 있도록 n-gram 기반의 tf*idf 유사도를 이용하여 타 영역 코퍼스를 결합하고, 간투어의 예측 기능을 모델링한다.

3.1 N-gram 기반 문서 선정

정보검색 기법을 이용한 도메인 간의 유사도를 측정하는 접근 방법[1][5]은 문서에서 각 키워드들의 상대적인 중요성을 weight matrix로 나타낸다. 이러한 문서 분류 기법은 키워드의 출현빈도의 분포를 사용하므로 인접한 단어간의 관련성이나 구조적인 정보를 유사도 측정에 반영할 수 없다. 본 논문에서는 대화체 도메인과 이질성이 큰 타 도메인과의 결합을 효과적으로 수행하기 위하여 n-gram 기반의 tf*idf 유사도를 이용한 새로운 언어모델 적응 방법을 제안한다.

Tf*idf 척도는 벡터 공간 모델을 이용하여 문서를 표현하는데, 본 논문에서는 벡터 모델의 키워드로 단어를 사용하는 대신 bigram 단어열 (w_{n-1}, w_n)을 사용하여 스타일을 반영하였다. 키워드 선정을 위하여, 코퍼스에 나타난 모든 bigram 단어열을 대상으로 inverse document freqeuncy(idf)를 계산하였다.

$$idf(t_k) = \log\left(\frac{N}{n_{t_k}}\right) \quad (1)$$

여기서, t_k를 단어열 (w_{n-1}, w_n)이고, N은 전체 문서 수, n_{t_k}는 단어열 t_k를 포함하는 문서 수를 나타낸다. 모든 문서에 나타나는 키워드는 가중치가 0이 되어,

대화체의 스타일을 반영하는 키워드를 찾아내는데 매우 효과적이다.

각 문서 d_i는 단어열 t_k의 가중치들의 벡터로 표현되고, term frequency tf_{ik}를 단어열 t_k가 문서 d_i에서 출현한 빈도수라고 할 때, 가중치를 계산하는 식은 다음과 같다.

$$wgt(t_k, d_i) = tf_{ik} \cdot idf(t_k), \quad k = 1, \ldots, L \qquad (2)$$

전체 in-domain 코퍼스를 문서 I라고 하면, 두 문서 사이의 유사도 $S(d_i, I)$는 다음과 같은 cosine coefficient를 이용하여 계산한다.

$$S(d_i, I) = \frac{\sum_{j=1}^{L} wgt(t_j, d_i) wgt(t_j, I)}{\sqrt{\left(\sum_{l=1}^{L} wgt(t_l, d_i)^2\right)\left(\sum_{l=1}^{L} wgt(t_l, I)^2\right)}} \qquad (3)$$

방송뉴스와 신문의 모든 문서에 대해 유사도 $S(d_i, I)$을 계산하고 내림차순으로 정렬하면, 유사도는 0과 1 사이의 값을 가지며, 유사도가 높을수록 1에 가까운 값을 가지게 된다. 여기서 구한 유사도를 각 문서의 가중치 $v(d_i) = S(d_i, I)$로 이용한다.

본 논문에서는 in-domain과 out-of-domain의 언어모델을 interpolation하는 대신 각 문서의 적응 가중치 $v(d_i)$를 이용하여 n-gram count를 직접 결합하였다. w_h를 단어 w의 history라고 하면 결합 count $C(w_h, w)$는 다음과 같다.

$$C(w_h, w) = C_I(w_h, w) + \sum_i v(d_i) \times C_O^i(w_h, w) \qquad (4)$$

이때, $C_I(w_h, w)$는 in-domain의 n-gram count이고, $C_O^i(w_h, w)$은 out-of-domain의 i번째 문서의 n-gram count이다.

3.2 간투어 모델

Stolcke [6]은 disfluency를 일반 단어들과 똑같이 취급하여 주변 단어에 대한 예측 기능을 평가하였다. 언어모델 혼잡도의 감소는 적었으나, 간투어가 다음 단어에 대한 예측 기능이 있을 뿐 아니라, 특히 문장의 경계 부분에서 잘 나타나는 특징이 있음을 밝혔다. 또한 여러 시스템들이 간투어를 모델링하여 사용하는데 unigram만을 사용하므로 성능의 개선이 극히 미미했다 [3]. 본 논문에서는 간투어가 주변 단어에 대한 예측 기능이 있다는 것을 기본 가정으로 하여, 이전 연구[6]에서와 같이 간투어를 일반 단어들처럼 취급하여 모델링하고자 한다.

간투어 모델링을 위하여 [6]의 *cleanup* 모델을 이용하였으며, 학습 데이터에서 간투어를 포함한 단어열의 수정이 필요하다. 표 1은 간투어를 포함한 문장의 trigram 단어열을 나열한 것으로, 일반 단어를 위한 trigram count를 위해서는 set 1을 사용한다.

표 1. 간투어 모델링을 위한 trigram 단어열

예문	꽃 이 어 아주 예쁘 다
set 1	꽃 이 아주 / 이 아주 예쁘 / 아주 예쁘 다
set 2	꽃 이 어
set 3	이 어 아주 / 어 아주 예쁘

간투어를 포함하는 set 2와 set 3은 in-domain에만 나타나므로 간투어 모델은 in-domain 데이터만을 대상으로 한다. Set 2와 set 3에 대한 간투어 모델을 다음과 같이 정의하였다.

FP-1: $C(w_h, w) = C_I(w_{ch}, w)$

w_{ch}: 간투어를 포함하지 않는 단어 history

Set 2의 경우만 반영한 모델로 일반 단어의 간투어 예측 기능만을 사용하였다.

FP-2: $C(w_h, w) = C_I(w_{fh}, w) + C_I(w_{ch}, w)$

w_{fh}: 간투어를 포함하는 단어 history

Set 2와 set 3을 모두 포함하는 모델로 간투어의 예측 기능도 모델에 포함시켰다.

Ⅳ. 인식 실험 및 결과

인식 실험은 본 연구실에서 대어휘 연속음성 인식을 위해 개발한 1-패스 세미다이나믹 트라이그램 네트워크 디코더[8]를 사용하였다. 음향모델은 90 대화의 대화체 음성과 약 20시간 분량의 낭독체 음성을 사용하여, 각 상태당 6개의 가우시안 혼합분포를 갖는 연속 HMM 모델을 학습하였다. 대화체 인식을 위해 각각 하나씩의 잡음 모델과 간투어 모델을 추가하였다 [2]. 학습에 사용되지 않은 10개의 대화를 언어모델과 인식을 위한 테스트 셋으로 사용하였다.

표 2. FP-1 모델 - 언어모델 혼잡도 vs. WER (%): CV (in-domain data), BF (brute-force addition of out-of-domain data), uni-TFIDF (기존의 tf*idf weighting), bi-TFIDF (제안한 tf*idf weighting)

가중치	혼잡도	WER(%)
= 0 (CV)	31.94	27.91
= 1 (BF)	63.23	30.36
uni-TFIDF	37.59	27.52
bi-TFIDF	34.96	27.10

표 2는 간투어 모델 FP-1을 사용하고, 가중치가 다른 여러 가지 결합 방법에 의해 생성된 트라이그램 언어모델의 혼잡도와 인식의 결과로 나온 WER(Word Error Rate)이다. 대화체 코퍼스만 사용했을 때(CV)의 언어모델 혼잡도가 가장 작게 나타나지만, 학습 데이터의 양이 너무 적기 때문에 인식 성능은 좋지 못하다. Tf*idf 가중치를 적용한 것의 인식 성능은 모두 우수하고, 특히 본 논문에서 제안한 bi-TFIDF는 CV에 대해서 2.9%, uni-TFIDF에 대해 1.5%의 WER를 감소시켜 가장 좋은 성능을 보여주었다. BF의 결과는 매우 이질적인 도메인의 데이터를 가중치 없이 섞어 사용하였으므로 오히려 성능저하를 보였다.

표 3. FP-2 모델 - 언어모델 혼잡도 vs. WER (%)

가중치	혼잡도	WER(%)
= 0 (CV)	27.78	27.14
= 1 (BF)	52.02	29.22
uni-TFIDF	31.85	26.78
bi-TFIDF	29.92	26.11

표 2와 표 3은 간투어 모델 FP-1과 FP-2를 각각 적용한 결과이다. 두 표를 비교할 때, 각 가중치에 따라 WER가 2.7%-3.8%의 감소로 확실한 성능의 차이를 보여준다. BF의 경우에도 간투어 모델의 효과가 뚜렷이 나타나고 있음을 확인할 수 있다.

실험에서 인식기는 1-패스에 하나의 발화를 인식하는데, 한 발화는 여러 문장으로 이루어져 있으므로 문장 경계에서 간투어의 역할이 더욱 효과적이었다고 볼 수 있다.

결론적으로, 본 논문에서 제안한 n-gram 기반의 tf*idf 유사도를 이용한 언어모델 적응 방법이 가장 좋은 성능을 보였고, 26.11% WER를 얻었다.

V. 결론

본 논문에서는 한국어 대화체 음성을 방송뉴스와 신문 기사들과 비교·분석하여, 특정 보조사의 빈번한 사용, 어미나 조사의 빈번한 생략이나 축약 등의 대화 현상들이 있음을 밝혔다. 또한 효과적인 대화체 언어모델 생성을 위하여 타 도메인 텍스트와 결합하는 과정에 이러한 대화 특징들을 반영할 수 있는 n-gram 기반의 tf*idf 유사도를 이용한 언어모델 적응 방법과 간투어 모델링을 제안하였다.

인식 결과로부터, n-gram 기반의 tf*idf 유사도를 적용하여 간투어 모델 각각에 대해 절대치로 2.9%, 3.8%의 WER를 감소시켰고, 간투어 모델 FP-2를 적용하여 6.5%의 WER를 감소시켰다.

이번 연구에서는 disfluency를 처리하기 위한 모델로 간투어만을 고려하였지만, 차후에는 간투어 이외의 다른 현상들로 각 특징에 맞는 모델링이 필요하다.

또한 좀 더 나은 대화체 음성의 분석 및 처리를 위해서는 잡음, 간투어, 반복/수정 발화 등의 대화현상을 정확하게 전사하는 것이 매우 중요하지만, 아직 한국어 대화체에 대해 목록화된 대화현상이나 전사 규칙 등이 정립되지 않은 상태이다. 대화체 음성과 텍스트 코퍼스 자체도 매우 중요하지만 이들 코퍼스를 제대로 분석할 수 있도록 하는 여러 대화현상들을 어떻게 표기할 것인가 등에 대한 연구도 함께 이루어져야 하겠다.

감사의 글

본 연구는 과기부 국책연구개발사업의 뇌신경정보학 과제 (M1-0107-01-0003) 지원으로 수행되었으며, 실험에 사용된 한국전자통신연구원의 대화체 음성 DB 사용 허가에 감사드립니다.

참고문헌

[1]R. Iyer and Mari Ostendorf, "Relevance Weighting for Combining Multi-domain Data for N-gram Language Modeling,'' *Computer Speech and Language*, Vol.13, pp.267-282, 1999.

[2]박영희, 정민화, "대화체 연속음성 인식을 위한 한국어 대화음성 특성 분석," *한국음향학회지*, 21권, 3호, 2002.

[3]Nobuyasu Itoh, Masafumi Nishimura, and Shinsuke Mori, "A Method for Style Adaptation to Spontaneous Speech by Using a Semi-linear Interpolation Technique," *Proc. of ICSLP*, 2000.

[4]Sanjeev Khudanpur and Jun Wu, "A Maximum Entropy Language Model Integrating N-gram and Topic Dependencies for Conversational Speech Recognition," *Proc. of ICASSP*, 1999.

[5]Milind Mahajan, Doug Beeferman and X. D. Huang, "Improved Topic-dependent Language Modeling using Information Retrieval Techniques," *Proc. of ICASSP*, 1999.

[6]Andreas Stolcke and Elizabeth Shriberg, "Statistical Language Modeling for Speech Disfluencies," *Proc. of ICASSP*, 1996.

[7]Man-hung Siu and Mari Ostendorf, "Modeling Disfluencies in Conversational Speech," *Proc. of ICSLP*, 1996.

[8]Dong-Hoon Ahn and Minhwa Chung, "Compact Subnetwork-based Large Vocabulary Continuous Speech Recognition," *Proc. of ICSLP*, 2002.