방송 뉴스 인식을 위한 언어 모델 적응

김현숙(ETRI), 전형배(ETRI), 김상훈(ETRI),
최준기(ETRI), 윤승(ETRI)

<차 례>

<Abstract>

Language Model Adaptation for Broadcast News Recognition

Hyun Suk Kim, Hyung Bae Jeon, Sanghun Kim, Joon Ki Choi, Seung Yun

In this paper, we propose LM adaptation for broadcast news recognition. We collect information of recent articles from the internet on real time, make a recent small size LM, and then interpolate recent LM with a existing LM composed of existing large broadcast news corpus. We performed interpolation experiments to get the best type of articles from recent corpus because collected recent corpus is composed of articles which are related with test set, and which are unrelated. When we made an adapted LM using recent LM with similar articles to test set through Tf-Idf method and existing LM, we got the best result that ERR of pseudo-morpheme based recognition performance has 17.2 % improvement and the number of OOV has reduction from 70 to 27.

* Keywords : Broadcast News Recognition, Language Model Interpolation, TF-IDF

1. 서 론

방송 뉴스는 정치, 경제, 사회, 문화 등 많은 영역의 뉴스로 구성되고, 특정한 영역의 미리 알 수 없는 즉, 예측하지 못한(unsupervised) 사건에 대한 정보가 다루어진다. 방송 뉴스 인식은 어렵지만, 도전적인 대용량 연속 음성 인식의 과제이다. 청각 장애인은 아나운서의 소리를 듣지 못하고 뉴스 화면만 보면서 내용을 짐작하여야 하므로, 방송 뉴스 인식은 장애인을 위한 방송 자막 시스템에 적용될 수 있다. 또, 자동화된 방송 자막 기술은 TV 및 라디오 방송물, 비디오, 영화 자료 등과 같은 멀티미디어 데이터에서 주제어 추출 및 구간 분리 기술과 접목하여 초대용량의 멀티미디어 데이터를 인덱싱하여 저장 관리하는 분야에서 필수적인 기술이다. 따라서, 방송 뉴스 인식 성능이 향상되어 방송 자막 기술이 안정화된다면 시장 규모는 매우 클 것으로 예측되고, 멀티미디어의 자막, 오디오 인덱싱 등 여러 분야로 확대 적용될 수 있다[1].

방송 뉴스 음성 인식의 성능 향상을 위해 기존에 진행되고 있는 연구는 다음과 같다. 방송 뉴스처럼 테스트 데이터의 영역을 예측할 수 없을 때 언어 모델 적응을 할 수 있는 비교사 언어 모델 적응(unsupervised language model adaptation)에 대한 연구[2][3], LIMSI 독일어 방송 뉴스 전사 시스템의 65k에서 300k로 어휘를 확장하여 성능 향상을 보인 연구[4], 방송 뉴스의 토픽이 변경되는 부분을 다룰 수 있기 위해, 신문 기사를 통한 최신 텍스트를 사용하여 언어 모델의 n-gram 확률을 적응함으로써 성능 향상을 보이는 연구[5]가 진행되고 있다.

뉴스에서는 인명, 지명, 단체명 등 새로 등장하는 어휘가 많으므로, 기존의 코퍼스를 중심으로 만들어진 어휘 사전과 언어 모델은 새로 발생한 사건에 대한 어휘를 포함하고 있지 않을 가능성이 있다. 이렇게 새로 발생한 어휘들은 미등록어(OOV: Out of Vocabulary)로 처리되고, 미등록어가 많을수록 음성 인식 오류가 많아져 방송 뉴스 인식 시스템의 성능을 저하시키게 된다. 따라서, 실생활에서 방송 뉴스 음성 인식 시스템을 계속 사용하기 위해서는, 새로 발생하는 사건에 대한 어휘를 인식할 수 있도록, 기존의 어휘 사전과 언어 모델을 수시로 보완하는 방법이 필요하다. 또, 일정 기간의 방송 뉴스와 신문 기사를 수집하여 코퍼스를 보강할 때는 뉴스 인식 시점을 기준으로 이미 지나간 기사보다는 지속적으로 토픽이 되고 있는 기사를 선택하여 코퍼스를 보강하는 것이 필요하다.

이를 위해 본 논문에서는 방송 뉴스 음성 인식의 성능 향상과 미등록 어휘수를 감소시키기 위해 최근의 방송 뉴스와 신문 기사를 온라인(on-line) 상에서 실시간적으로 수집하고 이에 대한 정보를 방송 뉴스 인식에 필요한 언어 모델과 어휘 사전에 자동으로 반영할 수 있는 온라인 언어 모델 적응 시스템에 대해 제안한다. 또, 수집한 일정 기간의 코퍼스를 그대로 사용한 경우와 지속적인 토픽으로 사용되는 기사를 선택하여 사용하는 경우에 대한 성능 비교를 수행하고, 최신 뉴스에

대한 음성 인식에 적합한 언어 모델 생성 방안을 제안한다. 본 논문의 언어 모델 적응 시스템은 사용자가 편리하도록 GUI 환경을 통하여 최신 기사 수집 및 언어 모델 적응 단계가 통합되어 자동으로 실행될 수 있는 시스템이다.

2. 방송 뉴스와 신문 기사의 특성

2.1. 방송 뉴스와 신문 기사의 문체

방송 뉴스와 신문 기사는 비슷한 시기에 발생한 새로운 사건 사고와 정보를 전달하는 목적으로 쓰여진 문장이지만, 문장의 특성이 다르다[6][7]. 방송 뉴스에서 앵커 방송은 낭독체, 기자의 인터뷰는 대화체로 작성되고, 신문 기사는 산문체로 작성된다. 특히, 종결형 어미부분에서 신문 기사는 "했다, 이다, 있다, 보였다" 등의 어미로 작성되고, 방송 뉴스는 "했습니다, 입니다, 봅니다, 받았습니다" 등의 경어를 나타내는 어미로 끝난다. 또, 방송 뉴스는 음운의 생략이 구어체적이다. 예를 들면, "하여"는 "해, 해서, 했으며, 했고, 했습니다"로 표현하고, "이영호입니다"는 "이영홉니다" 등으로 표현된다. 방송 뉴스와 신문 기사 코퍼스의 특성이 다르기 때문에 방송 뉴스 음성 인식을 위해서는 방송 뉴스 코퍼스는 대규모로 수집하고, 신문 기사는 테스트 날짜에 가장 가까운 기사를 수집하여 언어 모델을 생성하는 것이 효과적이다.

2.2. 방송 뉴스의 신규 발생 어휘 변화

뉴스에서는 인명, 지명, 단체명 등 새로 등장하는 고유 명사 어휘가 많으므로, 최근의 기사를 지속적으로 수집하여, 언어 모델과 어휘 사전에 반영함으로써 OOV 개수를 감소시키고, 음성 인식 성능을 향상시킬 수 있다.

<표 1> 최근 방송 뉴스의 OOV개수 비교

테스트날짜 및 어휘수	사전_A의 OOV 개수	사전_B의 OOV 개수	사전_C의 OOV 개수
8일 (3223)	175	174(0.6%)	145(17.1%)
9일 (3178)	179	173(3.4%)	141(21.2%)
10일(3311)	205	202(1.5%)	170(17.1%)
11일(2271)	110	103(6.4%)	78 (29.1%)
12일(2066)	107	105(1.9%)	82 (23.4%)

<표 1>에서는 최근의 방송 뉴스 코퍼스가 음성 인식의 OOV개수에 어떤 영향을 미치는지 알아보기 위한 실험을 수행하였다. 2003년 10월 8일부터 2003년 10월 12일까지 5일간의 방송 뉴스에 대해 새로 출현하는 어휘수를 측정하기 위하여, 각 테스트 날짜보다 1일전의 동일한 방송 시간에 방영된 방송 뉴스를 날짜별로 수집하였다. <표 1>에서 사전_A에 표기된 수치는 기존의 대규모 코퍼스로 구성한 어휘 사전(vocabulary)과 비교해서 5일간의 방송 뉴스에 새로 출현한 어휘수를 OOV 개수로 측정하였다. 사전_B에 표기된 수치는 기존의 대규모 코퍼스에 테스트 1일전의 방송 뉴스 코퍼스를 추가한 후, 어휘 사전을 생성하고 OOV 개수를 측정하였다. 사전_C는 기존의 대규모 코퍼스에 대한 어휘 사전과 테스트 1일전의 방송 뉴스에 대한 어휘 사전을 적응하여 생성한 어휘 사전이다. 표 1의 결과를 비교해보면, 사전_A와 사전_B의 OOV 개수는 0.6% - 6.4%의 OOV 감소가 있고, 사전_C에서는 사전_A보다 17.1% - 29.1%의 OOV 감소가 있었다. 사전_B보다 사전_C의 OOV 감소 비율이 높은 것을 보면, 신규 기사에 대해 수집한 코퍼스를 기존의 코퍼스에 추가하여 어휘 사전을 생성하는 방법보다는 기존의 어휘 사전에 신규 기사 어휘 사전을 적응시켜 사용하는 방법이 OOV 감소에 효과적임을 알 수 있다.

3. 언어 모델 적응(adaptation) 방법

3.1. 언어 모델 적응

방송 뉴스 기사는 새로 발생한 사건과 정보에 대한 기사를 다루기 때문에 기존의 코퍼스에서 발견할 수 없는 어휘에 대한 정보를 최근의 기사를 수집함으로써, OOV를 해결해야 한다. 이를 위해 최신 기사 코퍼스를 수집하여 언어 모델을 생성하였다. 그러나, 최신 기사 언어 모델은 소규모의 코퍼스로 작성되었기 때문에 어휘 부족으로 인해 일반적인 뉴스에 대한 인식 성능은 저하될 수 있다. 따라서, 기존에 구축되어 사용되는 대규모 방송용 코퍼스로 구축한 언어 모델에 신규 코퍼스 언어 모델을 적응(adaptation)할 수 있어야 한다. 본 연구에서는 일반적인 보간 방법인 선형 보간 방법 (linear interpolation)을 사용하였다. 선형 보간 방법은 아래의 식과 같다[8].

$$P_{interpol}(w_3|w_1, w_2)c = (1-\alpha)P_{old}(w_3|w_1, w_2) + \alpha P_{new}(w_3|w_1, w_2) \quad (1)$$

식(1)의 선형 보간에 따르면, 최신 기사 코퍼스로 구축한 언어 모델에서 관측되지 않는 n-gram에 대해서는 기존의 언어 모델에 가중치를 곱하여 사용하고, 기존의 언어 모델에서 관측되지 않는 n-gram에 대해서는 최신 기사 언어 모델에 가중

치를 곱하여 사용한다. 2개의 언어 모델에 모두 존재하는 n-gram에 대해서는 가중치를 2개의 언어 모델에 각각 곱하여 사용한다. 본 논문에서 선형 보간을 수행하기 위해 사용된 어휘 사전은 최신 기사 코퍼스의 어휘와 기존 코퍼스의 어휘를 통합하여 사용하였다.

3.2. 어휘 사전 구성

음성 인식에서는 사전에 등록되는 어휘가 음성 인식의 기본 단위가 되며, 인식 대상 어휘의 수가 음성 인식 작업의 난이도와 인식 성능을 결정하는 요소가 된다. 대용량의 방송 뉴스 코퍼스가 보유되어 있어도, 언어 모델 생성에 사용되는 어휘 사전은 코퍼스에 고빈도로 출현한 어휘에 대해서만, 어휘 사전이 작성된다. 즉, 코퍼스에서 가장 많이 사용된 어휘에 대해서만 작성되어 빈도가 낮은 어휘는 어휘 사전에서 제외되고, 언어 모델에서는 제외되거나, 또는 unknown으로 표시된다. 기존의 언어 모델이 대규모 코퍼스에 대해 구축되어 있더라도 매일의 사건과 정보를 다루는 방송 뉴스를 음성 인식하기 위해서는, 새로 발생한 사건을 표현한 어휘를 음성 인식 시스템의 어휘 사전과 언어 모델에 수시로 반영할 수 있어야, 실생활에서 사용하기에 적합하다.

최신 기사는 OOV를 줄이기 위해 수집한 코퍼스이기 때문에, 제한된 크기의 어휘 사전에 최신 기사의 어휘를 기존의 기사 어휘 사전에 모두 추가하면, 65k개를 초과하게 된다. 본 논문에서는 최신 기사에서 발생한 어휘를 기존의 어휘 사전에 통합할 때, 최신 기사에서 사용된 어휘가 통합된 어휘 사전에 모두 포함되도록 최신 기사 어휘에 더 큰 가중치를 두고 통합하고, 통합된 빈도수 파일로부터 65k개의 고빈도 어휘를 추출하여 통합된 어휘 사전을 구성하였다.

4. 방송 뉴스 토픽 클러스터링

4.1 방송 뉴스 토픽 추출의 필요성

방송 뉴스 음성 인식과 같은 대어휘 연속 음성 인식에서 사용되는 통계적 언어 모델은 대량의 데이터를 구비할수록 언어 모델의 성능이 향상된다. 그러나, 방송 뉴스는 정치, 경제, 문화 등 다양한 영역과 토픽으로 구성되고, 매일 새로 발생한 사건, 사고, 유용한 정보 등을 기사로 사용하기 때문에, 어제의 기사와 오늘의 기사의 어휘가 상당수 다르다. 그러나, 정치, 경제 기사는 특별한 사건에 대해 계속적으로 다루어지는 토픽 또는 인물이 비슷하므로 수집된 방송 뉴스 코퍼스를 그대로 사용하기보다는 특정한 토픽에 관련된 코퍼스를 추출하여 사용하는 방법

이 필요하다.

본 논문에서는 이렇게 지속적으로 다루어지는 토픽에 대한 어휘를 참조하기 위해, 최근 방영된 일정 기간동안의 방송 뉴스 기사를 수집하여 코퍼스를 보강하였다. 일정 기간의 방송 뉴스 중에서는 이미 기사로서의 효과를 잃어버린 어휘도 포함되어 있고, 계속 기사에 사용되는 어휘도 포함되어 있다. 따라서, 불필요한 기사를 제거하고, 계속 보도되는 기사를 선정할 수 있는 방법이 필요하다. 계속적으로 기사에 사용되는 어휘를 참조하기 위한 방법으로 방송 인식 시점 일자보다 하루 전의 방송 뉴스를 큰 범위의 토픽으로 선정하고, 유사한 방송 뉴스를 추출하기 위한 토픽 기사로 사용하였다. 일정 기간동안 수집한 방송 뉴스와 테스트 당일의 신문 기사 중에서 토픽 기사에 사용된 어휘와 유사한 기사를 추출하기 위해 정보 검색 기법에서 잘 알려진 Tf-Idf 방법을 사용하였다[9].

4.2 유사도(similarity) 측정

신규로 수집한 코퍼스로부터 토픽과 유사한 문서를 추출하기 위해 정보 검색 기법에서 널리 사용되는 Tf-Idf (Term Frequency - Inverse Document Frequency)방법을 사용한다[2][9][10]. 이 방법은 문서를 표현하는 벡터공간모델에 기초를 둔 방법이다. V가 사전의 어휘의 집합이고, 문서 D_i가 벡터 $[w_{i1}, w_{i2}, \ldots, w_{iv}]$로 표현된다면, j번째 단어 요소 $w_{i,j}$ 는 tf와 idf값을 곱해준 값으로 표현된다.

$$w_{ij} = tf_{ij} * \log(idf_j) \tag{2}$$

$$tf_{i,j} = \frac{\text{문서}i\text{에서 나타나는 단어}j\text{의 빈도수}}{\text{문서}i\text{에 나타나는 모든 단어의 수}} \tag{3}$$

$$idf_j = \frac{\text{전체 문서의 수}}{\text{단어}\ j\text{가 나타나는 전체 문서의 수}} \tag{4}$$

이때 임의의 두개의 문서 사이의 유사도(similarity)는 다음과 같은 방법으로 계산된다.

$$sim\ (D_i, D_j) = \frac{\overrightarrow{D_i} \bullet \overrightarrow{D_j}}{\left|\overrightarrow{D_i}\right| \times \left|\overrightarrow{D_j}\right|} = \frac{\sum_{k=1}^{V} w_{i,k} \times w_{j,k}}{\sqrt{\sum_{k=1}^{V} w_{i,k}^2} \times \sqrt{\sum_{k=1}^{V} w_{j,k}^2}} \tag{5}$$

위의 계산식은 두개의 벡터가 이루는 각의 COS 값에 해당한다. 즉, 두 개의 벡

터가 서로 얼마나 많이 틀어져 있는가를 측정하는 평가 방법이다. 유사도 S의 값은 0과 1사이의 값으로 나타난다. 유사도 값이 1에 가까울수록 유사한 토픽이고, 0에 가까울수록 전혀 다른 내용을 포함한 문서이다.

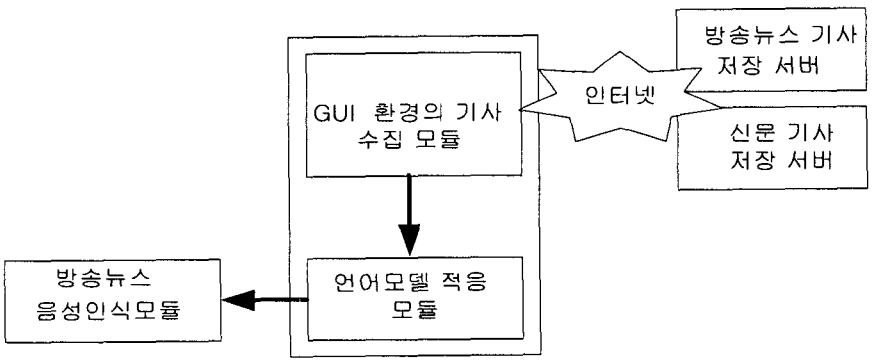
유사도에 따라 문서마다 등급을 설정한 후, 토픽 언어 모델(topic language model)을 생성하기 위해 가장 유사한 문서를 선택하는 과정에서, 몇 개의 유사한 문서를 선택해야 효율적인지에 대한 문제가 발생한다. 상위 몇 개의 문서를 선택하는 것은 토픽의 특성(specificity)을 높이는 반면, 많은 수의 유사 문서를 선택하는 것은 토픽 언어 모델의 강건성(robustness)을 향상시킨다. 따라서, 토픽 언어 모델의 특성과 강건성 사이에는 모순(trade-off)이 존재한다[9].

5. 온라인 언어 모델 적응 시스템

5.1 언어 모델 적응 시스템 구성

방송 뉴스 음성 인식 시스템의 성능을 강건하게 하기 위해서는, 최근의 방송 뉴스와 신문 기사를 온라인상에서 실시간적으로 수집하고 이에 대한 정보를 언어 모델과 어휘 사전에 반영할 수 있는 온라인 언어 모델 적응시스템이 필요하다.

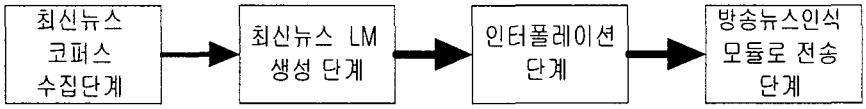
<그림 1>은 언어 모델 적응 시스템의 구성 요소를 나타낸다. 언어 모델 적응 시스템은 GUI 환경의 기사 수집 모듈과 언어 모델 적응 모듈, 방송국에서 제공하는 방송 뉴스 기사 저장 서버와 신문사에서 제공하는 신문 기사 저장 서버, 방송 뉴스가 인식될 수 있는 방송 뉴스 음성 인식모듈로 구성된다.



<그림 1> 언어 모델 적응 시스템 구성

GUI 환경의 기사 수집 모듈에서는 <그림 2>의 최신 코퍼스 수집 단계 기능이 수행되고, 언어 모델 적응 모듈에서는 <그림 2>의 최신뉴스LM 생성 단계 기능과 인터폴레이션 단계 기능이 수행된다.

5.2 적응된 언어 모델 생성 순서



<그림 2> 적응된 언어 모델 생성 순서

언어 모델을 구축하기 위해 먼저 <그림 1>의 GUI 환경을 사용한 기사 수집 모듈을 구동한다. 최신 뉴스 코퍼스 수집 단계에서는 신문 기사를 수집할지 방송 뉴스를 수집할지 GUI 화면에서 클릭하고, 수집할 날짜를 선택한다. 기사를 수집하기 위해 기사가 제공되는 홈페이지를 접근(access) 한 후, 다운로드(download)받는다.

기사 수집이 완료되지 않았으면, 홈페이지 접근과정을 반복해서 수행한다. 기사 수집이 완료되었으면, 수집된 기사에 대하여 영어, 숫자 등을 한글로 변환한다. 숫자 변환 오류, 띄워 쓰기 오류, 맞춤법 오류 등을 수정하고, 의사형태소 태깅을 수행한다. 이런 과정을 수행한 후, 언어 모델 적응 모듈로 전송한다.

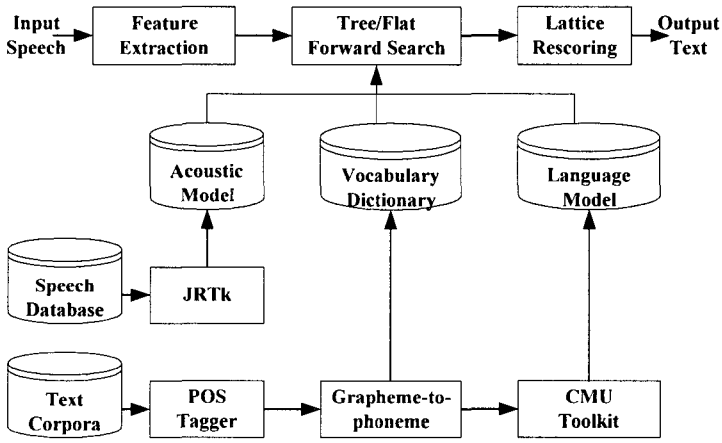
언어 모델 적응 모듈의 최신 뉴스 LM생성 단계에서는 태깅된 어휘들을 1개의 문장 단위로 통합하고 언어 모델을 생성하기 적합한 형태의 코퍼스로 변환한다. 수집된 최신 기사 코퍼스에 대한 어휘별 빈도수를 추출하고, 어휘 사전을 작성한 후, 최신 기사에 대한 언어 모델을 생성한다.

언어 모델 적응 모듈의 인터폴레이션 단계에서는 최신 기사 코퍼스에 대한 어휘 사전과 기존의 코퍼스의 어휘 사전을 통합하여, 인터폴레이션에 사용할 새로운 어휘 사전을 작성한다. 어휘 사전 생성이 완료되면, 기존에 구축되어 있던 대규모 코퍼스에 대한 기존 언어 모델과 신규 언어 모델을 어휘 사전을 적용하여 인터폴레이션하여 최종적으로 방송 뉴스 인식 모듈에서 사용할 적응된 언어 모델을 구축하고 방송 뉴스 음성 인식 모듈로 전송하여 사용한다.

6. 방송 뉴스 음성 인식 실험

6.1. 방송 뉴스 음성 인식 시스템 형상

본 논문에서 사용한 방송 뉴스 인식 실험을 위한 시스템 형상은 <그림 3>과 같다[11]. 음향 모델 훈련을 위해 ETRI-JRTk 툴킷을 이용하고, 통계적 언어 모델링을 구축하기 위해서 CMU-Cambridge toolkit을 사용하였다[12]. 언어 모델 인터폴레이션은 HTK toolkit을 사용하였다[13].



<그림 3> 방송 뉴스 인식 시스템 형상

음성 신호는 16kHz 샘플링, 16 비트로 양자화된 데이터이고 윈도우 사이즈는 16 ms, 프레임간 이동은 10 ms이며, 특징 벡터로는 멜 켑스트럼을 사용하였고, 24차 특징 벡터 생성하기 위해서 LDA (linear discriminant analysis)가 수행되었다. 40개 기본 음소(묵음 포함)를 사용하였으며 각각의 음소는 HMM으로 모델링, 모델의 구조는 3-state left-to-right (no skip path) model로 정의되어 있다. 관측 확률의 경우 조음 현상이 고려된 inter-morpheme에 새논 기반의 음향 모델링을 적용하였다. intraword contexts의 경우, 좌, 우 양방향에 대해서 최대 context 폭은 2, interword contexts의 경우엔 1이다. 3000개의 새논을 사용하였으며 각각의 새논은 16차 Gaussian Mixtures를 갖는 각기 고유의 코드북을 가진다. context clustering을 위해서 모음, 자음, 마찰음 등의 47개의 상세 분할된 음소 카테고리를 결정 트리의 context questions 으로 사용하였다.

인식 단위로는 의사형태소를 사용한다[14]. 의사형태소는 형태소 단위를 수정하여 발음이 유지되도록 하였으며 언어적 지식에 의거하여 짧고 자주 발생하는 형태소를 병합하여 인식 오류를 감소시켰다. 형태소 분석기[15]를 이용하여 텍스트

코퍼스 문장내의 어절을 의사 형태소 단위로 분할하였으며 발음 사전은 형태소 기반의 grapheme-to-phoneme converter[16]를 통해 자동적으로 생성되도록 하였다.

인식된 내용 중 영어 알파벳, 기호, 숫자에 대한 표기는 후처리에서 HTG (hypothesis to grapheme)를 통해 자동으로 복원하여 주었다.

6.2. 실험 환경

방송 뉴스 음성 인식실험에 사용된 음향 모델은 <표 2>와 같이 약 300시간의 방송 뉴스 음성데이타를 수집하여 훈련하였다.

<표 2> 음향 모델 데이터

수집대상 연도	time
KBS(1999~2001) 140일분	약 300시간
SBS(1999~2001) 195일분	

<표 3>과 같이 기존에 보유된 대규모 방송 뉴스 코퍼스를 사용해 언어 모델을 생성하였다.

<표 3> 대규모 방송 뉴스 코퍼스

수집대상 연도	문장수	어절수	byte
KBS(1996~2003)	약230만	약 2,750만	1.23G
MBC(1996~2000)			
SBS(1997~2003)			

OOV를 줄이기 위해 최신 기사로 수집한 코퍼스의 종류는 <표 4>와 같다. <표 3>의 기존의 코퍼스는 대규모 코퍼스이기 때문에 <표 4>의 최신 기사를 대규모 기사에 통합한 후 언어 모델을 생성하면, 최신 기사의 특성이 언어 모델에 잘 표현되지 않는다. 최신 기사에 대한 특성이 표현되고 성능 향상이 있도록 각 최신 기사 코퍼스에 대한 언어 모델을 생성하여 기존의 대규모 코퍼스에 대한 언어 모델에 대해 인터폴레이션을 수행하였다.

<표 4> 최신 기사 코퍼스 종류

구분	신규 수집 대상 코퍼스 설명
최신 기사 1	테스트 set과 동일한 날짜의 신문 기사 전체를 수집
최신 기사 2	테스트 set보다 1일전의 동일한 시간에 방영된 방송 뉴스 대본 수집
최신 기사 3	테스트 set과 동일한 날짜에 미리 방영된 다른 시간의 방송 뉴스 수집
최신 기사 4	테스트 set 의 방송일자를 기준으로 이전 날짜의 1달 동안 방영된 동일한 시간대의 방송 뉴스를 수집

테스트 문장은 앵커, 기자, 인터뷰 기사가 포함된 2003년 10월 14일 KBS 9시 방송 뉴스 239문장을 사용하였다. 음성 인식의 성능을 비교하기 위한 파라미터로 테스트 문장에서 출현한 어휘와 언어 모델에 사용된 어휘를 비교하여, OOV를 조사하였다 또, 테스트 문장에 대한 음절(syllable), 의사형태소 기반(word), 어절(eojeol)에 대하여 인식 성능을 계산하였고, 에러감소율(ERR: Error Reduction Rate)을 계산하였다.

<표 5>는 CMU tool을 사용하여, 인터폴레이션 비율에 따른 퍼플렉서티를 측정한 결과중에서 테스트 날짜별로 가장 낮은 퍼플렉서티를 보인 기존 언어 모델과 신규 기사 언어 모델간의 인터폴레이션 비율을 나타낸다. 신규 기사 언어 모델은 테스트 날짜보다 1일전의 기사로 생성한다.

<표 5> 언어 모델 인터폴레이션 비율 실험

테스트날짜 (2003년)	인터폴레이션 비율		비고
	기존 언어 모델	신규 언어 모델	
10월 8일	0.789	0.211	신규 언어 모델은 테스트 날짜보다 1일 전의 기사 수집하여 언어 모델 생성
10월 9일	0.817	0.183	
10월10일	0.816	0.184	
10월11일	0.757	0.243	
10월12일	0.790	0.210	
10월13일	0.811	0.189	
10월14일	0.777	0.223	

위의 실험 결과를 평균적으로 계산하여, 기존의 언어 모델을 0.8 비율로, 신규 언어 모델을 0.2의 비율로 인터폴레이션 하여 적응된 언어 모델을 생성하였다.

6.3. 실험 결과

<표 6>은 언어 모델을 사용하여 방송 뉴스 테스트 set에 대한 인식 성능을 수행하기 위한 실험 종류를 나타낸다. 실험1은 <표 3>에서 수집한 기존의 대규모 코퍼스에 대한 언어 모델을 사용한 음성 인식 성능이고, 모든 실험의 베이스라인 실험 결과로 사용한다. 실험2-실험9에서는 기존의 대규모 언어 모델과 <표 4>의 4종류의 최신 기사로 구성된 신규 언어 모델을 인터폴레이션하여 적응된 언어 모델을 생성한 후, 음성 인식 실험을 수행하였다.

이를 위해, 실험2에서는 최신 기사1로 구성한 코퍼스에 대한 신규 언어 모델을 생성, 실험3에서는 최신 기사2로 구성한 코퍼스에 대한 신규 언어 모델을 생성, 실험4에서는 최신 기사3으로 구성한 코퍼스에 대한 신규 언어 모델을 생성, 실험5에서는 최신 기사2 와 최신 기사3을 통합한 코퍼스에 대해 신규 언어 모델을 생성, 실험6에서는 최신 기사2, 최신 기사3, 최신 기사4를 통합한 코퍼스에 대한 신규 언어 모델을 생성하였다.

<표 6> 적응된 언어 모델을 사용한 실험

종류	기존 언어 모델 (Pold)	신규 언어 모델(Pnew)			
		최신기사1 (당일 신문)	최신기사2 (전날 방송)	최신기사3 (당일 다른 방송)	최신 기사4 (이전의 1달 방송)
실험1	O	-	-	-	-
실험2	O	O	-	-	-
실험3	O	-	O	-	-
실험4	O	-	-	O	-
실험5	O	-	O	O	-
실험6	O	-	O	O	O
실험7	O	-	O	O	100개 기사
실험8	O	O	O	O	100개 기사
실험9	O	100개 기사	O	O	100개 기사

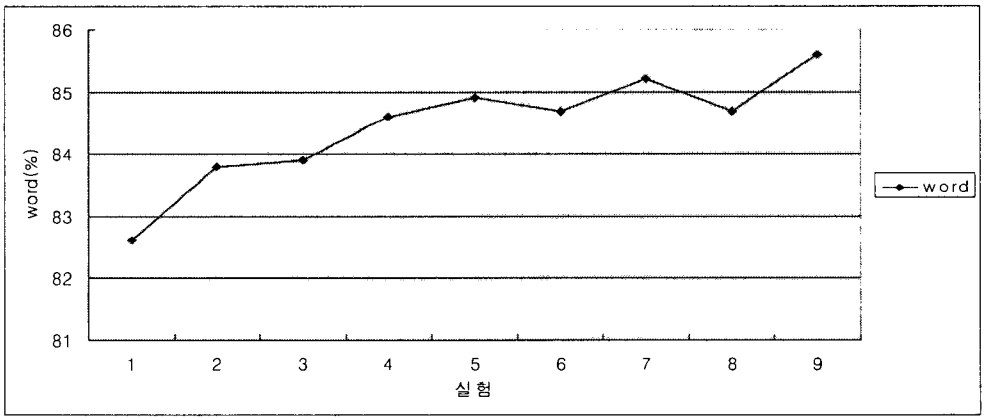
실험7에서는 최신 기사4로부터 100개의 기사만 선택하였다. 왜냐하면, 최신 기사4는 1달간의 방송 뉴스 코퍼스이므로, 계속 방송되고 있는 토픽과 유사한 기사와 이미 지나간 기사로 구성된다. 즉, 테스트 문장의 내용과 관계가 없는 어휘도 있기 때문에 테스트 문장과 유사한 문장을 추출하기 위하여 테스트 1일전의 뉴스와 유사한 문장을 Tf-Idf 방법을 사용하여, 최신 기사 4로부터 가장 유사한 100개의 기사를 추출하였다. 실험7에서는 최신 기사 2, 최신 기사 3과 최신 기사 4로부터 추출한 100개의 유사한 방송 뉴스 기사를 통합한 코퍼스에 대해 신규 언어 모

델을 생성하였다.

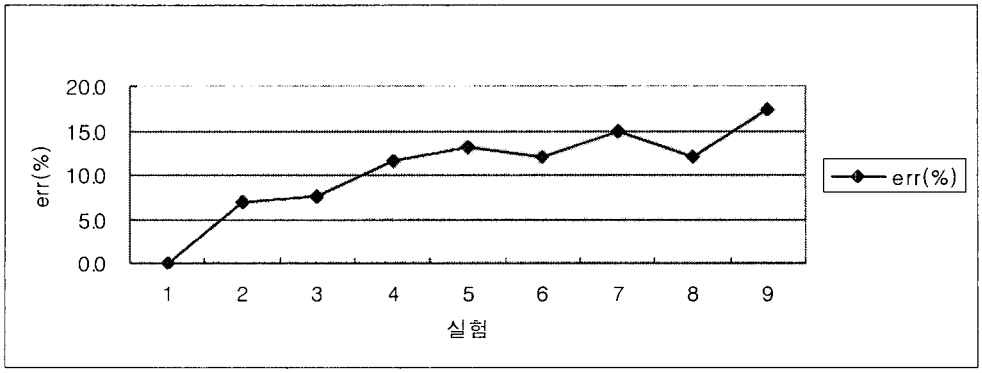
실험8에서는 최신 기사 1, 최신 기사 2, 최신 기사 3과 최신 기사 4로부터 추출한 100개의 유사한 기사를 통합한 코퍼스에 대해 신규 언어 모델을 생성하였다.

최신 기사 1은 테스트 당일의 신문 기사이며, 정치, 경제, 사회 등 다양한 영역으로 구성되어 있고, 방송 뉴스의 토픽과 동일한 기사 또는 동일하지 않은 기사도 포함되어 있다. 최신 기사1로부터 유사한 기사 추출을 위해 Tf-Idf방법을 사용하여, 100개의 신문 기사를 추출하였다. 실험9에서는 최신 기사2, 최신 기사3과 최신 기사4로부터 추출한 100개의 유사한 방송 뉴스 기사, 최신 기사1로부터 추출한 100개의 유사한 신문 기사를 통합한 코퍼스에 대해 신규 언어 모델을 생성하였다.

언어 모델은 음성 인식의 탐색 단계에서 의사형태소 단위로 인식 단위에 영향을 주므로 언어 모델 성능 평가 파라미터인 음절, 의사형태소, 어절 중에서 의사형태소 기반 성능을 중심으로 <그림 4>에 비교하였다.

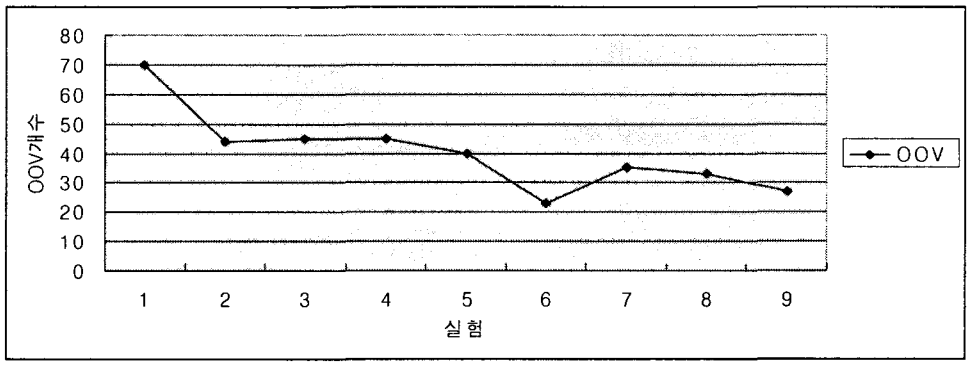


<그림 4> 의사형태소 기반 인식 성능 (%)



<그림 5> 음성 인식 성능 ERR

<그림 5>는 <그림 4>의 의사 형태소 기반 인식 성능에 대한 ERR을 나타낸다.



<그림 6> 음성 인식 실험별 OOV 개수

<그림 6>은 실험에서 사용된 어휘 사전과 테스트 문장의 어휘를 비교하여 OOV 발생 개수를 나타낸다. <그림 4>에서 실험2의 결과는 실험3과 실험4의 결과보다 인식 성능이 낮다. 신문 기사는 산문체이고 방송 뉴스는 낭독체이기 때문에 <그림 6>의 OOV개수는 비슷하지만, 직접적인 성능향상이 낮음을 알 수 있다.

<그림 4>에서 실험3의 결과보다 실험4와 실험5의 결과의 인식 성능이 더 좋다. 방송 뉴스 인식에는 가장 최신의 뉴스를 반영하는 것이 성능향상을 높일 수 있는 조건임을 알 수 있다. <그림 6>에서 실험6은 실험5의 코퍼스보다 최신 기사4가 추가되어 OOV개수는 더 줄었지만 인식 성능은 낮음을 알 수 있다. 이것은 테스트 뉴스와 토픽이 다른 뉴스를 많이 추가하는 것이 인식 성능 향상에 도움이 되지 않음을 알 수 있다. <그림 4>에서 실험7은 실험5와 실험6의 결과보다 성능 향상이 좋다. 코퍼스 크기가 작더라도 토픽이 유사한 문장을 추가하는 것이 좋음을 알 수 있다. 실험8은 최신 기사1을 추가하였어도 실험7의 결과보다 성능이 좋지 않다. 이것은 실험6과 같이 테스트 뉴스와 토픽이 다른 뉴스를 많이 추가하는 것이 인식 성능 향상에 도움이 되지 않음을 알 수 있다. 실험9는 실험8의 결과보다 성능향상이 좋아졌고 가장 높은 성능을 보여 주었다. 코퍼스 크기가 작더라도 토픽이 유사한 문장을 추가하는 것이 좋음을 알 수 있다.

<표 7>에서는 실험1의 기존의 언어 모델만을 사용한 음성 인식 결과 내용과 실험9의 적응된 언어 모델을 사용한 음성 인식 결과 내용을 일부 비교하였다. 적응된 언어 모델을 사용한 경우에, 새로 발생하는 명사 어휘에 대해 인식 성능이 향상되었음을 알 수 있다.

<표 7> 음성 인식 결과 내용 비교

기존 언어 모델	적응된 언어 모델
재선	재신임
당연히	탄핵
교도소	최 도술
신임 ****** 경제 는	신임 투표 문제 는

7. 결 론

기존의 언어 모델이 대규모 코퍼스에 대해 구축되어 있더라도 매일 새로운 사건과 정보를 다루는 방송 뉴스에서는 인명, 지명, 단체명 등 새로 등장하는 어휘가 많이 발생한다. 따라서, 미등록 어휘(OOV)를 줄여 음성 인식 성능을 높이기 위해서, 새로 발생한 사건을 표현한 어휘를 음성 인식시스템의 어휘 사전과 언어 모델에 수시로 반영할 수 있어야 한다. 또, 기존의 코퍼스는 대규모 코퍼스이기 때문에 최신 기사를 대규모 기사에 통합한 후 언어 모델을 생성하면, 최신 기사의 특성이 언어 모델에 잘 표현되지 않는다. 본 논문에서는 최신 기사를 실시간적으로 수집하여 신규 언어 모델을 생성하고, 최신 기사의 특성이 반영되도록 기존 언어 모델과 인터폴레이션할 수 있는 온라인 언어 모델 적응시스템에 대해 제안하였다.

방송 뉴스 음성 인식 실험 결과 신규 언어 모델이 최근 뉴스의 토픽과 비슷한 코퍼스에서 구축되었을 때의 성능이 가장 좋았다. 즉, 일정 기간의 최신 기사 중에서, 계속적으로 뉴스의 토픽이 되고 있는 기사를 Tf-Idf 방법을 사용하여 클러스터링하여 신규 언어 모델을 생성하고, 이를 이용하여 기존의 언어 모델과 신규 언어모델을 인터폴레이션하여 적응된 언어 모델을 생성할 때, 의사형태소 기반 음성 인식 성능 에러 감소율(ERR)이 기존의 성능보다 17.2% 향상되었으며, OOV 개수도 70개에서 27개로 감소하는 효과를 보였다. 또, 음성 인식된 문장에서 내용어인 명사 어휘에 대한 인식 성능 향상이 있었다. 인터폴레이션 비율에서는 기존의 언어 모델을 0.8로 가중치를 주고, 최신 기사 언어 모델에는 0.2로 가중치를 주었을 때가 가장 성능 향상이 좋았다.

참 고 문 헌

[1] 박준, 김승희 et al., "방송 뉴스 자막 처리 시스템 개발", 제17회 *음성통신 및 신호처리 학술대회*(KSCPS2000), 2000.

[2] 최준기, 오영환, "방송 뉴스 음성 인식을 위한 비교사 언어 모델 적응", 제20회 *음성통신 및 신호처리 학술대회*(KSCPS2003), 2003.

[3] L. Chen, J. Gauvain, et al., "Unsupervised Language Model Adaptation for Broadcast News", *Proceedings of the ICASSP*, 2003.

[4] K. McTait, M. Adda-Decker, "The 300k LIMSI German Broadcast News Transcription System", *Proceedings of the Eurospeech*, 2003.

[5] M. Federico, N. Bertoldi, "Broadcast News LM Adaptation using Contemporary Texts", *Proceedings of the Eurospeech*, 2001.

[6] T. Kim, "Lexical Adaptation and Language Model Adaptation For Korean Broadcast News Speech Recognition", Master's thesis, *Dept. of Computer Science*, Sogang University, South of Korea, 2002.

[7] http://www.kbs.co.kr/announcer/html/main_korean.htm.

[8] X. Huang, A. Acero, H. Hon, "Spoken Language Processing", *Prentice Hall*, 2001.

[9] M. Mahajan, D. Beeferman, X.D. Huang, "Improved Topic-Dependent Language Modelling Using Information Retrieval Techniques", *Proceedings of the ICASSP*, 1999.

[10] 신영숙, 정민화, "대어휘 연속음성 인식을 위한 토픽 클러스터링 기반의 언어 모델 성능향상", *한국음향학회 학술 대회*, 2001.

[11] 정의정, 윤승, "Korean Broadcasting News Transcription System with Out-of-Vocabulary Update Module", *한국음향학회 하계학술발표대회 논문집*, 2002.

[12] P. Clarkson, R. Rosenfeld, "Statistical language modeling using the CMU-CAMBRIDGE toolkit", *Eurospeech*, 1997.

[13] S. Young, G. Evermann, et al. "*The HTK Book v3.2*", 2002.

[14] O. Kwon, K. Hwang, and J. Park, "Korean Large Vocabulary Continuous Speech Recognition of Newspaper Articles," *Proc. ICSP99*, pp. 333-336, 1999.

[15] J.-H. Kim, Lexical disambiguation with error-Driven Learning, Ph.D. dissert. *Dept. Computer Science*, Korea Advanced Institue of Science and Technology, 1996.

[16] J. Jeon, S. Cha, et al., "Automatic generation of Korean pronunciation variants by multistage applications of phonological rules," *ICSLP '98*, 1998.

접수일자: 2004년 8월 12일
게재결정: 2004년 9월 9일

▶ 김현숙(HyunSuk Kim)
주소: 305-350 대전광역시 유성구 가정동 161번지
소속: 한국전자통신연구원(ETRI) 음성언어정보연구부 음성인터페이스연구팀
전화: 042) 860-5967
E-mail: hyskim@etri.re.kr

▶ 전형배(HyungBae Jeon)
주소: 305-350 대전광역시 유성구 가정동 161번지
소속: 한국전자통신연구원(ETRI) 음성언어정보연구부 음성인터페이스연구팀
전화: 042) 860-6480
E-mail: hbjeon@etri.re.kr

▶ 김상훈(Sanghun Kim)
주소: 305-350 대전광역시 유성구 가정동 161번지
소속: 한국전자통신연구원(ETRI) 음성언어정보연구부 음성인터페이스연구팀
전화: 042) 860-6480
E-mail: ksh@etri.re.kr

▶ 최준기(JoonKi Choi)
주소: 305-350 대전광역시 유성구 가정동 161번지
소속: 한국전자통신연구원(ETRI) 음성언어정보연구부 음성처리연구팀
전화: 042) 860-6480
E-mail: joonki74@etri.re.kr

▶ 윤승(Seung Yun)
주소: 305-350 대전광역시 유성구 가정동 161번지
소속: 한국전자통신연구원(ETRI) 음성언어정보연구부 음성처리연구팀
전화: 042) 860-5835
E-mail: yunseung@etri.re.kr