정보검색 기법과 동적 보간 계수를 이용한 N-gram 언어모델의 적응

최준기(KAIST), 오영환(KAIST)

<차 례>

<Abstract>

N-gram Adaptation Using Information Retrieval and Dynamic Interpolation Coefficient

Joon Ki Choi, Yung-Hwan Oh

The goal of language model adaptation is to improve the background language model with a relatively small adaptation corpus. This study presents a language model adaptation technique where additional text data for the adaptation do not exist. We propose the information retrieval (IR) technique with N-gram language modeling to collect the adaptation corpus from baseline text data. We also propose to use a dynamic language model interpolation coefficient to combine the background language model and the adapted language model. The interpolation coefficient is estimated from the word hypotheses obtained by segmenting the input speech data reserved for held-out validation data. This allows the final adapted model to improve the performance of the background model consistently. The proposed approach reduces the word error rate by 13.6% relative to baseline 4-gram for two-hour broadcast news speech recognition.

* Keywords: Language model adaptation, Language model, adaptation corpus, Dynamic interpolation coefficient, Linear interpolation, Speech recognition.

1. 서 론

은닉 마르코프 모델(HMM: Hidden Markov Model)과 비터비(Viterbi) 탐색을 기반으로 하는 연속음성인식에서 N-gram 언어모델은 성공적으로 사용되고 있다. 그러나 N-gram은 자연언어의 복잡성을 표현하기에 단순한 구조를 가지고 있으며 훈련용 코퍼스에 대한 의존성을 가지고 있기 때문에 하나의 N-gram으로 여러 영역을 동시에 표현할 수 없다는 단점을 가지고 있다[1]. 이러한 단점을 극복하기 위한 여러 방법 중 N-gram 적응은 소규모의 적응코퍼스를 이용하여 특정 영역을 강조하여 표현하는 방법으로 언어모델과 연속음성인식의 성능을 동시에 향상시킬 수 있다[2].

N-gram 적응은 크게 나누어 적응코퍼스의 획득과 기본 언어모델과의 병합이라는 두 개의 문제로 나누어질 수 있다. 기존의 적응 코퍼스를 구하는 방법을 분류하면 다음과 같다. 먼저 인식대상 음성의 영역을 미리 분류하고 각 영역에 맞는 학습자료를 준비하여 사용하는 방법이 있다[15]. 그러나 일반적인 경우 영역별 학습자료를 미리 구하는 것은 매우 어렵다. 따라서 N-best 리스트나 격자(lattice)와 같은 음성인식의 중간 결과를 적응 코퍼스로 사용하려는 시도가 있었다[14][16]. 이 방법은 추가 자료를 요구하지 않는 장점이 있으나 연속음성인식의 중간 결과는 많은 오류를 포함하고 있기 때문에 적응코퍼스가 자연언어의 특징을 충분히 반영하지 못하는 단점을 가지고 있다. 이러한 문제를 극복하기 위해 음성인식의 중간 결과에서 인식하고자 하는 음성의 영역을 추정할 수 있는 정보를 추출한 다음, 월드 와이드 웹이나 외부의 대용량 코퍼스를 정보검색 기법을 사용하여 적응 코퍼스를 구하는 방법이 제안되었다[17][18][19]. 그러나 일반적인 정보검색은 사용자의 질의에 대해 의미론적으로 가장 유사한 문서를 검색하는 것이 목표이기 때문에 N-gram을 위한 최적의 적응 코퍼스를 구하는 방법과 차이가 있을 수 있다. Bigi의 연구에서는 언어모델에 기반한 정보검색 기법[4]을 사용하여 적응 코퍼스를 구하는 방법이 제안되었다[3]. 본 논문에서는 Bigi의 연구에서 사용된 유니그램 모델을 확장하여 바이그램과 트라이그램 모델을 사용하여 적응 코퍼스를 구하였다. 구해진 적응 코퍼스는 유니그램 모델로 구한 코퍼스보다 언어모델 적응에 적합함을 실험적으로 증명하였다.

그리고 본 논문에서는 적응 언어모델의 또 다른 문제인 기본 언어모델과의 병합 문제해결을 위해서 음성인식의 중간 결과를 이용하여 언어모델간의 선형 보간 병합의 보간 계수를 동적으로 구하는 방법을 제안한다. 기존의 보간 계수를 구하는 방법으로는 평가자료와 유사한 검증자료(held-out validation data)를 이용하여 동적 언어모델 병합 가중치를 구하거나[5][6], 고정된 정적 가중치를 사용하는 방법이 널리 사용되었으나[19][7], 이러한 방법은 검증자료에 의존하는 단점이 있다. 본 논문에서는 언어모델 병합의 동적 보간 계수의 정확한 추정을 위해서 인식 대상

음성을 언어모델의 변화에 대한 민감도에 따라 구간을 나누고 해당 구간의 단어 후보열을 검증자료로 사용하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 언어모델 적응에 적합한 적응 코퍼스를 구하는 방법에 대해 서술하며 3장에서는 본 논문에서 제안한 동적 보간 계수를 구하는 방법에 대해 설명한다. 그리고 한국어 방송뉴스인식 실험 결과와 결론을 맺는다.

2. 정보검색 기법을 이용한 적응코퍼스의 획득

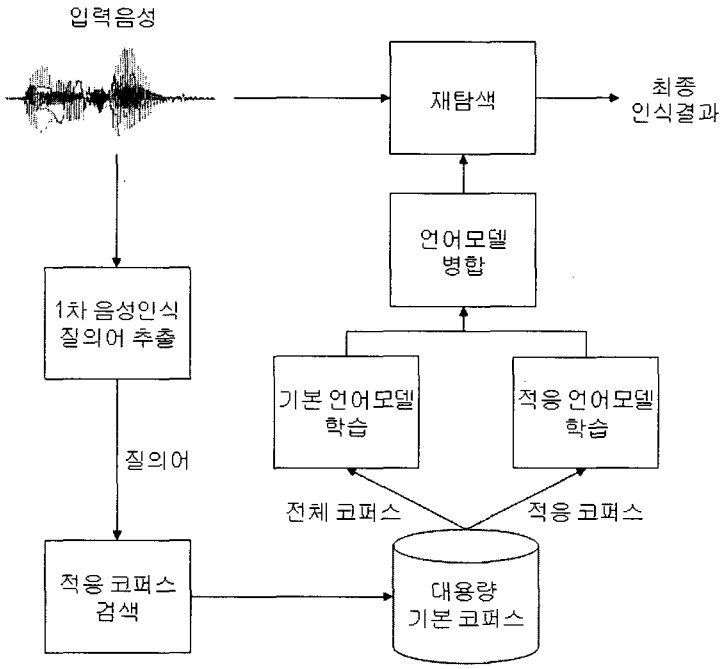
2.1 음향학적 신뢰도 기반 질의 추출

인식대상 음성의 영역이 잘 알려지지 않았을 경우 N-gram 적응을 위한 적응코퍼스는 정보검색 기법을 이용하여 새로운 추가 데이터를 외부에서 검색하여 구축할 수 있다. 그러나 본 논문에서는 순수한 N-gram의 적응 효과를 관찰하기 위하여 기본 언어모델을 작성할 때 사용한 텍스트, 즉 시스템이 미리 보유하고 있는 코퍼스만을 이용하여 적응코퍼스를 구축하였다. 또한 N-gram 적응과정 중에 새로운 어휘의 추가도 금지하여 미등록 어휘의 단순 추가로 인한 성능향상 효과를 제거하였다. <그림 1>에서는 추가 코퍼스 없이 기본 코퍼스만을 이용하여 언어모델 적응을 수행하는 과정을 보여주고 있다.

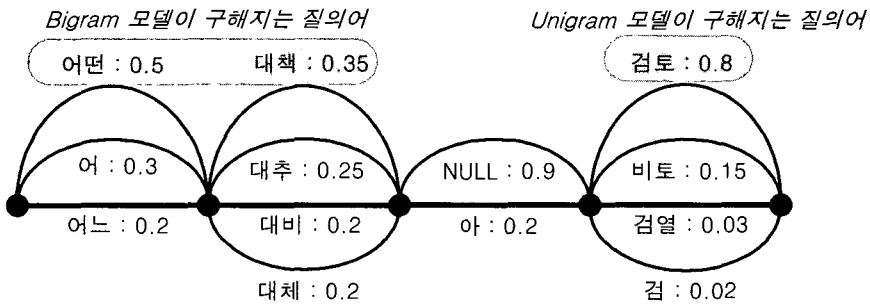
<그림 1>에서처럼 음성인식의 중간결과로부터 추출된 정보, 즉 인식 후보 단어열이나 후보 문장은 정보검색의 질의로 사용된다. 정보검색 기법을 사용하여 구축된 적응 코퍼스는 입력문장과 유사한 주제와 유사한 문형을 갖으면서도 자연언어의 특징을 잃지 않는다는 점에서 음성인식의 중간결과를 직접 적응코퍼스로 사용하는 방법에 비해 장점을 가진다. 그러나 음성인식의 중간결과의 오류는 여전히 남아있기 때문에 음성인식의 오류가 질의에 포함되어 잘못된 적응코퍼스를 검색할 수 있다. 이러한 음성인식의 오류는 적응 언어모델의 성능을 하락시키는 요인이 된다. 본 논문에서는 음성인식 오류의 영향을 줄이기 위해서 음성인식의 중간결과에서 영역정보를 추출할 때 단어 수준의 음향학적 신뢰도를 사용하였다. 즉, 일정 수준 이상의 음향학적 신뢰도를 가지는 단어들을 수집하여 적응코퍼스를 구하기 위한 질의로 사용하였다. 단어수준의 음향학적 신뢰도는 음향학적 사후확률을 사용하였다. 이 사후확률은 Mangu가 제안한 모호한 네트워크(confusion network)의 링크 사후확률(link posterior probability)로부터 구해지며 식은 다음의 식 1과 같다[8].

$$p(l|X) = \frac{\sum_{Q_l} p(q, X)}{p(X)} \quad (1)$$

위의 식 1에서 X는 음성의 관측벡터이며 $p(q,X)$는 격자(lattice)의 경로 q의 결합확률(joint probability)이다. 그리고 Q_l은 모호한 네트워크의 링크 l을 지난 모든 경로이다. 모호한 네트워크(confusion network)의 예제와 이로부터 질의를 추출하는 과정은 <그림 2>에 도식화되어 있다. 질의 추출을 위한 음향학적 신뢰도의 임계값은 실험적으로 결정된다. 링크 사후확률은 0부터 1까지 정규화되기 때문에 임계값을 쉽게 결정할 수 있다.



<그림 1> 추가 코퍼스 없는 언어모델 적응 시스템



<그림 2> 음향학적 신뢰도를 고려한 음성인식의 중간결과 정보 추출

2.2 언어모델기반 정보 검색

본 논문에서 적응 코퍼스를 구성하기 위하여 사용한 정보검색 기법은 언어모델에 기반한 정보검색 기법을 사용한다[21][4][22]. 이 정보검색 기법에서는 질의의 언어모델과 검색 대상 문서의 언어모델의 유사도를 검색 모델로 사용한다. 언어의 생성(generation) 관점에서 보면 특정한 언어모델은 어휘 셋과 고유한 확률 값을 이용하여 문장을 생성할 수 있다. 여기서 특정한 언어모델로 생성한 문장은 반대로 해당 언어모델이 잘 표현할 수 있는 문장이다. 따라서 검색 대상 문서의 언어모델로 문장을 생성했을 때 생성된 문장의 언어모델과 질의의 언어모델이 유사하면 질의와 검색 대상 문서는 유사하다고 말할 수 있다[21]. 따라서 검색 대상문서의 언어모델과 질의의 언어모델의 유사도를 측정하는 방법이 검색 모델로 사용될 수 있다. 언어모델에 기반한 정보검색 기법은 질의의 N-gram 분포와 유사한 문서를 검색할 수 있기 때문에 다른 정보검색 기법보다 N-gram 적응에 적합하다.

검색된 문서와 질의의 유사도는 문서 d의 언어모델 M_d과 질의 q의 언어모델 M_q의 Kullback-Leibler 거리로 표현되며 식 2와 같다.

$$Distance(M_q||M_d) = \sum_w p(w|M_q)\log\frac{p(w|M_q)}{p(w|M_d)} \tag{2}$$

식 2에서 $p(w|M_q)$와 $p(w|M_d)$에 일반적인 언어모델 개념을 도입하여 표현하면 아래의 식 3과 같다.

$$p(w|M_d) = \begin{cases} p_s(w|M_d) & w \in d \\ \alpha_d p(w|M_c) & otherwise \end{cases} \tag{3}$$

위의 식 3에서 $p_s(w|M_d)$는 문서 d에서 출현한 단어 w의 언어모델 확률을 나타내며 단어의 출현빈도로 결정된다. α_d는 미등록 어휘의 언어모델 값을 조절하기 위해 문서 d에 의해 결정되는 상수이다. 그리고 $p(w|M_c)$는 전체 코퍼스에 대한 언어모델 값이다. 식 3에서 언어모델을 구하기 위한 평탄화(smoothing) 방법으로는 Katz 평탄화 방법을 사용하였다. 일반적으로 언어모델에 기반한 정보검색에서는 Dirichlet 평탄화가 더 많이 사용되나[4] 본 논문을 위한 실험에서는 Katz 평탄화 기법이 근소하게 좋은 성능을 보여주었기 때문에 적응 언어모델을 작성하는데 사용하였다. 식 2에 식 3을 대입하여 전개 과정을 거치면 질의와 검색 대상 문서의 언어모델의 유사도는 다음의 식 4와 같이 표현될 수 있다.

$$\sum_{w \in d \cap q} p(w|M_q) \log \frac{p_s(w|M_d)}{\alpha_d p(w|M_c)} + \log \alpha_d \quad (4)$$

Bigi의 연구에서는 적응 코퍼스를 구하기 위해여 식 4와 같이 유니그램만을 검색모델로 사용하였다. 그러나 본 논문에서는 유니그램의 문맥을 확장하여 바이그램과 트라이그램까지 검색모델로 사용하였다. 텍스트 기반의 정보검색에서 검색모델의 확장의 유효성에 대해서는 Miller의 연구에서 언급된바 있다[22]. 본 논문에서 정보검색의 결과로 얻고자 하는 문서는 궁극적으로 현재 인식 중인 음성의 N-gram 분포와 유사한 분포를 갖는 문서이기 때문에 의미적으로 유사한 문서보다도 N-gram의 분포가 더 유사한 문서를 검색하는 것이 더 효과적이다. 질의로 추출된 인식 후보 단어 열에서 2개의 단어가 연결되어 인식된 경우 바이그램을 구하며 연결된 단어가 3개 이상인 경우 트라이그램을 구한다. 검색 모델을 바이그램과 트라이그램으로 확장했을 때 순위 결정함수(ranking function)은 각각 식 5와 식 6처럼 유도될 수 있다.

$$\sum_{w_i w_{i-1} \in d \cap q} p(w_i|M_q, w_{i-1}) \log \frac{p_s(w_i|M_d, w_{i-1})}{\alpha_d p(w_i|M_c, w_{i-1})} + \log \alpha_d \quad (5)$$

$$\sum_{w_i w_{i-1} w_{i-2}} p(w_i|M_q, w_{i-1}, w_{i-2}) \log \frac{p_s(w_i|M_d, w_{i-1}, w_{i-2})}{\alpha_d p(w_i|M_d, w_{i-1}, w_{i-2})} + \log \alpha_d \quad (6)$$

위의 식 5와 6을 이용하여 질의와 문서의 N-gram을 비교하고 일정 임계치 이상의 유사도를 가지는 문서를 모아서 적응코퍼스로 사용한다.

3. 언어모델 병합의 동적 보간 계수

특정 영역에 대하여 최적화된 적응 언어모델은 3장에서 서술한 방법으로 구해진 적응 코퍼스를 이용하여 구축된다. 그러나 적응 코퍼스만을 이용해서 구한 적응언어모델은 적은 학습 자료로 인한 자료 희귀성 문제를 가지기 때문에 기본 언어모델과 병합하여 사용하는 것이 일반적이다[2]. 본 논문에서는 구현이 용이하며 효율적인 선형보간법을 이용하여 두 개의 언어모델을 병합하였다. 최대 사후확률 추정(MAP:Maximum A Posteriori) 기반의 언어모델 병합 기법 역시 선형보간법으로 표현될 수 있다[14]. 선형보간법의 식은 다음과 같다.

$$\widehat{p}(w|h) = \lambda p_a(w|h) + (1-\lambda)p_b(w|h) \quad (7)$$

식 (7)에서 p_a는 적응 언어모델이며, p_b는 기본 언어모델이고, λ는 두 언어모델에 대한 보간 계수이다. 이 보간 계수 λ는 기본 언어모델과 적응 언어모델간의 가중치를 결정하는 변수가 되며 최종적으로 병합된 언어모델의 성능을 결정하고, N-gram의 백 오프(back-off) 모델 보간 계수와 유사하게 동작한다. 보간 계수를 결정하는 기존의 방법은 추가적인 검증자료(held-out validation data)를 사용하는 방법이 일반적이다.

본 논문에서는 추가적인 검증자료를 이용하지 않고 입력음성의 인식 중간결과를 검증자료로 직접 사용하여 동적 보간 계수를 구하였다. 입력음성의 중간결과를 사용하기 위해 먼저 본 논문에서 주로 인식하고자 하는 방송뉴스 문장의 특징을 살펴보자. 방송뉴스의 대본은 대부분 문법에 맞는 문장들로 이루어져있다. 문법에 맞는 자연언어 문장의 특성상 한 문장에서 영역정보를 포함하고 있는 부분과 영역정보를 포함하지 않은 구간으로 나눌 수 있다. 다음의 예시 문장을 보자

> 특히 수능시험을 점수제에서 등급제로 바꾸는 방안에 대해서는 공청회에서 회의적인 반응이 많았습니다.

위의 예시 문장은 '사교육비 경감 대책과 수능시험의 개선'이라는 주제를 가지는 기사의 한 문장이다. 위의 문장에서 "수능시험을 점수제에서 등급제로"라는 부분은 해당 주제의 내용을 담고 있는 것을 알 수 있다. 그러나 "공청회에서 회의적인 반응이 많았습니다." 라는 내용은 굳이 해당 주제가 아니더라도 많이 사용될 수 있는 표현으로 비교적 영역정보를 많이 포함하고 있지 않은 것을 알 수 있다. 따라서 영역정보가 많이 포함된 구간에서는 영역 특화된 적응 언어모델을 보다 가중해야 하며 그렇지 않은 구간에 대해서는 보다 큰 규모의 코퍼스에서 얻어진 기본 언어모델을 가중하여 사용해야 함을 알 수 있다.

그러나 일반적으로 자연언어 문장에서 영역정보를 포함하고 있는 부분을 정확하게 고르는 일은 쉽지 않다. 이는 영역정보를 정확하게 정의할 수 없기 때문이며 영역정보를 사용자가 구분할 수 있도록 잘 정의해도 음성인식의 중간단계에서 영역을 정확하게 추정하는 작업은 매우 복잡한 자연언어처리 기법을 요구하기 때문이다. 따라서 본 논문에는 적응 언어모델이 잘 표현할 수 있는 단어 후보 열이 있는 음성의 분할 구간에 대해서는 적응 언어모델을 가중하고 반대로 기본 언어모델이 잘 표현할 수 있는 구간에 대해서는 기본 언어모델을 가중하는 방법을 사용하였다. 즉 영역 정보를 많이 포함하고 있는 구간을 영역 특화된 적응 언어모델로 잘 표현되는 구간으로 근사한다.

정확한 검증자료의 추정을 위하여 입력음성의 분할이 매우 중요하다. 본 논문에서는 언어모델 값의 변화에 따른 인식 결과의 민감도를 입력음성의 구간 분할의 기준으로 사용하였다. 기본 언어모델과 적응 언어모델이 주어지고, 이 두 언어모델로 음성인식의 중간결과를 각각 재정렬(rescoring)한 후 인식 후보 단어의 순위가 서로 다르게 변화하는 구간과 서로 동일하게 변화하거나 변화하지 않는 구간으로 분할한다. 두 개의 언어모델로 재정렬한 N-best 리스트를 비교하기 위해서 Levenshtein 정렬을 사용하여 인식 후보 단어의 변화를 관찰한다[20]. 이러한 분할 기준은 음성인식의 후처리인 발화검증(utterance verification)에서 널리 사용되는 신뢰도 중 하나인 음향학적 안정도(acoustic stability)와 유사한 의미를 가진다[9]. 음향학적 안정도는 언어모델의 스케일을 변화시키면서 재정렬하여 인식결과의 변화를 관찰하며, 본 논문에서 사용한 언어모델의 변화에 대한 민감도는 언어모델을 바꾸면서 재정렬하여 인식결과의 변화를 관찰하는 차이점이 있다. 입력음성을 분할하는 과정을 서술하면 다음과 같다.

1. 음성인식 중간결과 재정렬
   1.1 기본 언어모델로 음성인식 중간결과를 재정렬하여 N-best 리스트를 얻는다. (N<10)
   1.2 적응 언어모델로 음성인식 중간결과를 재정렬하여 N-best 리스트를 얻는다. (N<10)
2. Levenshtein 정렬 및 구간분할
   2.1 두 개의 N-best 리스트에서 각각의 순위에 맞는 쌍의 문장후보끼리 정렬한 뒤 서로 다른 부분을 찾는다.
   2.2 모든 문장 후보 쌍에 대해서 항상 같은 부분을 고르고 남은 부분끼리 모아서 음성구간을 분할한다.

위의 과정을 그림으로 도식화하면 <그림 3>과 같다. 실제 알고리즘에서는 각각의 언어모델로 재정렬된 N-best 리스트를 결합하는 것이나, 이해의 편의를 위하여 격자로 표현하였다. Kalai[10]의 동적 보간 계수와는 달리 음성의 한 분할 구간에 대해 하나의 보간 계수를 부여하기 때문에 적응 언어모델과 기본 언어모델이 동시에 가중되는 경우가 없다. 위와 같은 방법으로 입력 음성을 분할하면 분할된 구간은 언어모델의 변화에 따라 음성인식의 결과가 다르게 되는 인식 후보 단어 열을 포함하게 되고, 각 인식 후보 단어 열에 대해 언어모델의 정보량, 즉 해당 단어 열의 언어모델 값을 측정하면 어떤 언어모델을 가중해야 하는지 알 수 있다.

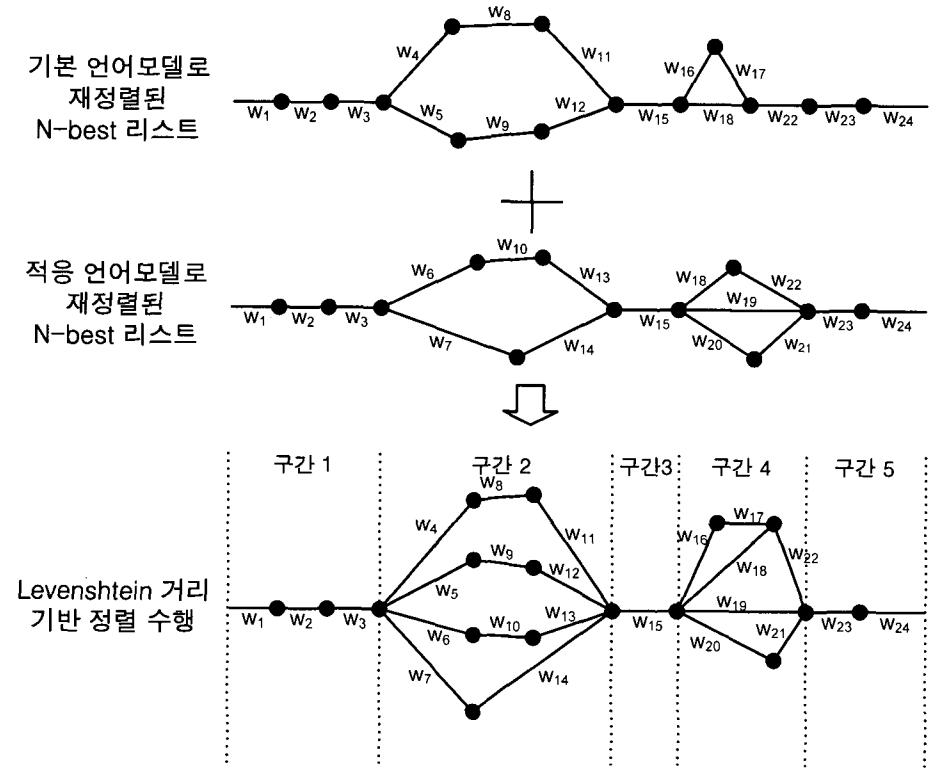
구해진 검증자료를 이용하여 언어모델의 정보량을 측정하고 언어모델의 보간 계수를 구하는 방법은 기대최대화(EM:Expectation Maximization) 알고리즘에 기반한

방법을 사용하였다[11]. 알고리즘의 내용은 다음과 같다. 먼저 각 언어모델을 은닉 마르코프모델의 상태라고 가정하고 검증자료로 사용되는 후보 단어 열을 관측벡터라고 간주하면 보간 계수는 초기 상태에서의 각 언어모델을 표현하는 상태로의 천이확률로 볼 수 있다. 따라서 보간 계수 λ에 대한 기대최대화 알고리즘의 갱신식은 은닉 마르코프모델의 천이확률 학습식과 유사하게 구할 수 있으며 다음의 식과 같이 표현된다. 식 (8)은 기대최대화 알고리즘의 λ_n에서 λ_{n+1}으로의 갱신식이다.

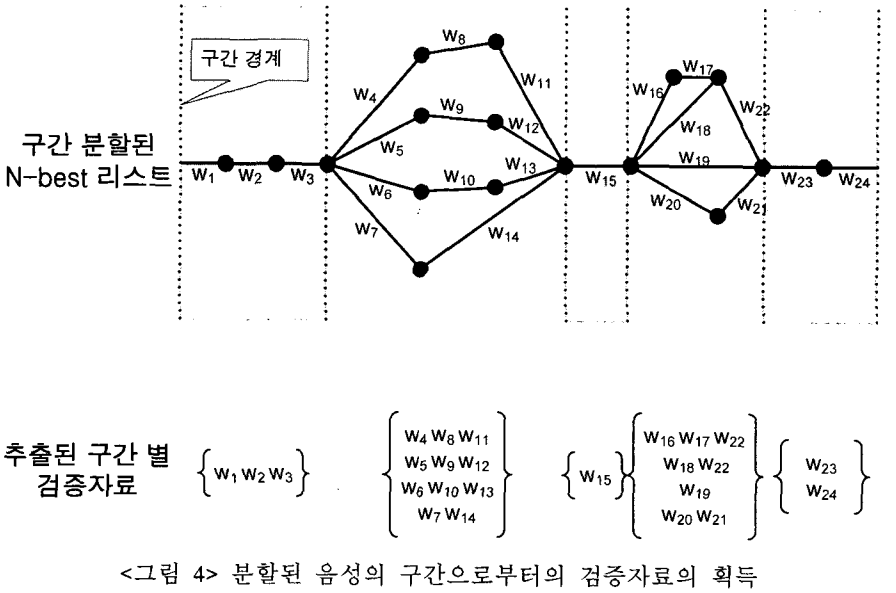
$$\lambda_{n+1} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\lambda_n(i)p_a(w_m|h_m)}{\lambda_n p_a(w_m|h_m)+(1-\lambda_n)p_b(w_m|h_m)} \tag{8}$$

식 8에서 M은 검증자료의 길이이며, $p_a(w_m|h_m)$은 m번째 단어 w_m에 앞서는 단어들의 이력 h_m이 주어졌을 때 적응 언어모델을 의미하며 $p_b(w_m|h_m)$는 같은 문맥과 이력에 대한 기본 언어모델 값을 표현한다. 일반적으로 $n=5\sim6$회의 반복 수행으로 수렴된 보간 계수를 구할 수 있다. 이와 같은 방법으로 추정된 보간 계수는 적응 언어모델로 잘 표현할 수 있는 단어가 검증자료에 포함된 구간에 대해서는 적응 언어모델을 가중하며, 반대로 기본 언어모델로 잘 설명될 수 있는 단어들이 검증자료에 많이 주어지면 기본 언어모델을 가중한다.

구간별로 분할된 음성의 중간 결과를 이용하여 검증자료를 구하는 과정은 <그림 4>에 도식화되어 있다. 모든 과정을 거친 뒤 분할된 음성의 각 구간에 적합한 보간 계수가 결정되며, 음성인식의 중간 결과인 격자에 음성의 구간별 보간 계수를 적용하여 N-gram 재정렬 과정을 거치면 동적으로 적응 N-gram의 가중치가 변화하여 적용된다.



<그림 3> 언어모델의 민감도에 따른 입력음성의 분할



<그림 4> 분할된 음성의 구간으로부터의 검증자료의 획득

4. 실험결과

4.1 한국어 방송뉴스 인식기

본 논문에서는 한국어 방송뉴스 인식기를 사용하여 제안한 적응 언어모델의 성능을 평가하였다. 트라이폰 기반의 연속 확률분포 은닉 마르코프 음향모델을 학습시키기 위하여 300 시간 분량의 방송뉴스 데이터를 사용하였다. 방송뉴스는 KBS, MBC, SBS의 3 개 방송국의 저녁 뉴스 데이터를 사용하였으며 여러 가지 잡음환경이 포함되어 있다. 음성처리를 위하여 13차 멜 켑스트럼 계수(MFCC: Mel Frequency Cepstral Coefficient)와 이에 해당하는 차분(delta)과 가속(delta-delta)의 총 39차 특징 벡터를 추출하였으며 잡음처리에 강인한 인식을 위하여 전역 켑스트럼 평균 차감(CMS: Cepstrum Mean Subtraction)과 성도길이 정규화(VTLN: Vocal Tract Length Normalization)를 수행하였다. 훈련된 은닉 마르코프 모델은 총 3786개의 상태들로 구성되었으며 각 상태는 16개의 가우시안 혼합으로 표현된다.

교착어로서의 한국어는 어근과 어미로 나누어져 많은 활용형이 가능하기 때문에 영어와 같이 단어 단위의 사전 구축이 어렵다. 따라서 본 논문에서는 의사형태소 단위의 발음 사전과 인식단위를 사용하며 모든 인식단위에는 품사태그가 부착되어 있다. 발음 사전은 총 64,997개의 의사형태소로 구성되었으며 발음변환 툴을 이용하여 자동으로 구성되었다. 언어모델의 학습을 위하여 총 223M 의사형태소 분량의 신문기사(조선일보, 동아일보)와 51M 의사형태소 분량의 방송 대본을 사용하였다. Katz 평탄화를 이용한 기본 4-gram은 SRI 언어모델 툴킷을 이용하여 구하였다[12]. 탐색 알고리즘은 tree-trellis 알고리즘을 사용하였다.

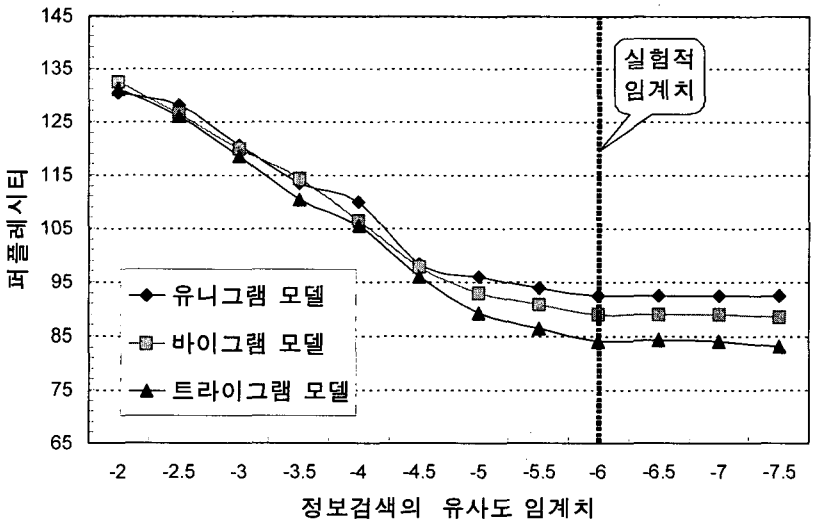
음성인식의 중간결과에서 질의를 추출하기 위한 신뢰도의 임계치 결정과 검색 문서의 유사도 임계치 결정, 그리고 기존의 보간 계수를 구하는 방법의 구현 비교 실험을 위한 검증자료는 3일 분량의 방송뉴스를 사용하였다. 최종 평가자료는 학습자료와 충분한 시간 간격을 가지도록 선정되어 학습자료에 평가자료와 동일한 주제나 기사가 실리지 않도록 하였다. 학습자료로 사용된 언어모델은 1998년 1월부터 2003년 1월까지 수집된 자료이며 평가자료는 2003년 10월에 수집되었다.

평가자료는 SBS와 KBS의 두 개의 방송국에서 수집된 4일 분량의 방송 뉴스이다. 평가자료는 간단한 휴리스틱 규칙과 가우시안 혼합모델(GMM: Gaussian Mixture Model) 분류기를 이용하여 자동으로 총 39개의 꼭지기사(story)와 459개의 문장으로 분할되었으며 각 꼭지기사는 동일한 주제를 다루고 있다. 각 꼭지기사에 대해 음성인식을 수행한 중간결과로부터 적응 코퍼스를 구하기 위한 질의를 추출하였으며 평균 305.7개의 단어가 질의로 추출되었다. 모든 형태소를 질의에 포함하는 것이 조사나 어미 등 빈번하게 사용되는 기능적 형태소를 제외하는 것보다 약간 더 나은 성능을 보여주었다. 따라서 음성인식 중간 결과 중 음향학적 신뢰도

가 일정 임계치 이상인 단어들은 질의에 포함되었다. 언어모델은 꼭지 기사단위로 작성되어 하나의 꼭지기사가 동일한 적응 언어모델로 재정렬 되었다.

4.2 적응코퍼스의 수집 실험

먼저 본 논문에서 제안된 적응 코퍼스 수집 방법의 유효성을 검증하기 위해 수집된 적응 코퍼스에 대한 퍼플렉시티(perplexity) 실험을 검증자료에 대해 수행하였다. 검색된 문서와 질의의 유사도가 일정 임계값 이상인 문서들을 모아서 적응 코퍼스를 작성하게 되는데 이 유사도와 적응 언어모델의 성능간의 관계를 보기 위하여 유사도를 변형하면서 적응 언어모델의 퍼플렉시티 성능을 측정하였다. 결과는 <그림 5>와 같다. <그림 5>에서 볼 수 있듯이 유사도 -6.0 근처에서 적응 언어모델의 성능이 수렴하였다. 또한 본 논문에서 사용한 N-gram 언어모델을 이용한 정보검색 기법의 유효성을 검증하기 위하여 유니그램, 바이그램, 그리고 트라이그램 모델의 성능을 비교하였으며 모델의 확장이 적응 코퍼스의 퍼플렉시티 성능을 향상시킴을 볼 수 있다.



<그림 5> N-gram 언어모델을 사용하여 구한 적응 코퍼스를 이용하여 작성된 언어모델의 퍼플렉시티 성능. 검색된 문서가 질의와 임계치 이상의 유사도를 가질 경우에만 적응 코퍼스로 간주되며 유사도 -6.0에서 적응 코퍼스의 성능이 안정되는 것을 알 수 있다.

다음에는 다른 정보 검색 기법과 언어모델에 기반한 정보 검색기법의 성능을 비교하는 실험을 수행하였다. 본 논문에서 사용한 방법과 벡터 모델링에 기반한 BM25 가중 방법[13]을 비교한 결과는 <표 1>과 같다. BM25 가중 방법에서 검색된 문서의 유사도 임계값 역시 실험적으로 구하여 최적의 적응 코퍼스를 구하였다. 본 실험을 위한 기본 언어모델과의 보간 계수는 검증 자료에 대해 최적의 성능을 보이는 정적 보간 계수(0.8)를 구하여 사용하였다.

<표 1> 기본 언어모델과 적응 코퍼스 수집 방법에 따른 적응 언어모델의 음성인식 성능 실험 결과

		퍼플렉시티	단어오류율
기본 4-gram		136.23	16.84
적응코퍼스 수집방법	BM25 가중 방법	89.04	15.62
	유니그램 모델 기반 검색	92.40	15.74
	바이그램 모델 기반 검색	88.94	15.49
	트라이그램 모델 기반 검색	84.18	15.33

<표 1>에서 볼 수 있듯이 음성인식 실험에서도 N-grma을 사용한 언어모델에 기반한 정보 검색 기법이 BM25 방법보다 좋은 인식 성능을 보여주었다. 총 39개의 꼭지기사에 대해서 perplexity를 개별적으로 구한 뒤 분산분석(ANOVA)를 수행하면 유의수준 0.01에서 언어모델에 기반한 정보검색이 BM25보다 향상된 언어모델을 생성함이 유의함을 알 수 있었다.

4.3 동적 보간 계수를 이용한 적응 언어모델 실험

다음으로 본 논문에서 제안된 동적 언어모델 보간 계수를 이용한 적응 언어모델의 성능 평가를 수행하였다. 먼저 언어모델의 변화에 따라 입력음성의 구간을 분할하는 방법의 유효성을 검증하기 위하여 분할된 입력 음성의 구간 중 언어모델의 변화에 민감하지 않은 구간의 변화를 관측하였다. 그 결과 언어모델 변화에 민감하지 않은 구간은 적응 언어모델의 보간 계수를 0부터 1까지 변화시켜도 후보 단어 열의 순위가 변하지 않음을 확인할 수 있었다.

그리고 본 논문에서 제안한 동적 보간 계수를 검증하기 위해 여러 가지 다른 방법으로 보간 계수를 구하고 음성인식 실험을 통해 성능을 비교하였다. 먼저 검

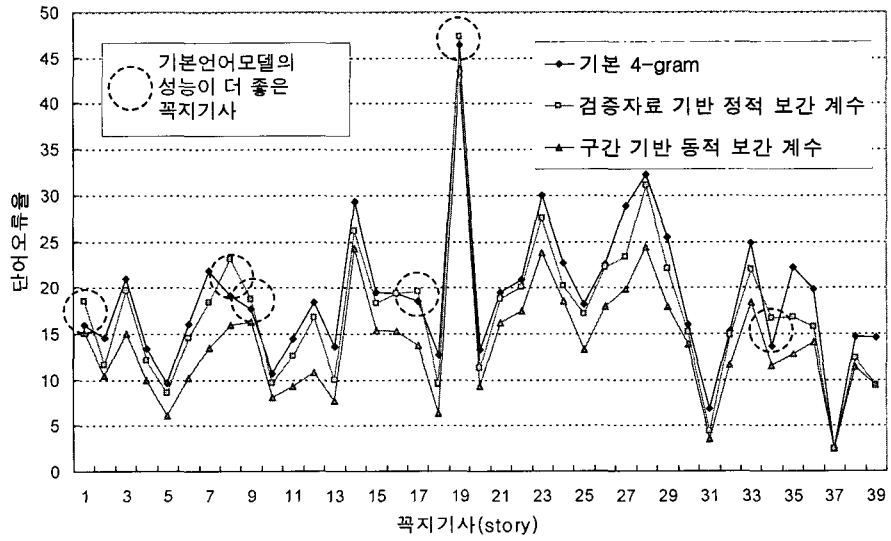
증 자료를 이용해서 최적의 정적 보간 계수를 구하였으며(검증자료 기반 정적 보간 계수)[7][19], 해당 단어의 과거의 함수로 결정되는 동적 보간 계수를 구하였다(검증자료 기반 동적 보간 계수)[5][6], 그리고 음성의 구간을 분할하지 않고 전체 인식 후보열에 대해서 온라인 알고리즘을 통해서 구한 동적 보간 계수를 사용한 적응 언어모델의 인식성능을 비교하였다(온라인 동적 보간 계수)[10]. 실험 결과는 <표 2>와 같다.

<표 2> 언어모델 병합의 동적 보간 계수 비교 실험

보간 계수 유형	단어오류율
검증자료 기반 정적 보간 계수	15.33
검증자료 기반 동적 보간 계수	15.50
온라인 동적 보간 계수	14.02
제안된 구간 기반 동적 보간 계수	13.24

<표 2>에서 볼 수 있듯이 검증자료를 사용하여 구한 보간 계수에 비해서 음성 인식의 중간결과를 사용하는 방법이 향상된 성능을 보여주었다. 그리고 온라인 동적 계수 방법에 비해서 제안된 분할된 음성 구간에 기반한 동적 보간 계수가 성능이 더 좋았으며, 이는 분할된 음성의 구간의 모든 단어 후보열에 대해서 동일한 보간 계수를 적용하는 방법이 각 단어 후보열에 최적인 보간 계수를 부여하는 방법보다 성능이 좋음을 보여준다. 본 논문에서 제안한 정보검색 기법과 동적 보간 계수를 이용하여 기본 언어모델의 인식 성능에 비해 단어 인식률 성능은 절대 3.6% 향상되었으며 기존의 정적 보간 계수를 사용하는 방법에 비해 상대적으로 13.6% 향상되었다.

그리고 제안한 동적 보간 계수를 사용한 적응 언어모델이 기본 언어모델의 성능을 일관되게 향상시키는 것을 관찰하기 위하여 평가자료의 꼭지기사별 성능을 측정하였으며 결과는 <그림 6>과 같다. 제안한 적응 언어모델(구간 기반 동적 보간 계수)이 최적의 정적 보간 계수를 사용한 적응 언어모델에 비하여 항상 좋은 성능을 보여주는 것을 알 수 있다. 정적 보간 계수를 사용한 적응 언어모델은 전체 39개의 꼭지기사 중 6개의 꼭지기사에 대해 기본 언어모델보다 저하된 음성인식 성능을 보여주었다. 문장 별 성능의 변화는 <표 3>과 같다. <표 3>에서 볼 수 있듯이 제안한 동적 보간 계수는 적응 언어모델의 성능 저하 오류를 대부분 줄일 수 있었다.



<그림 6> 꼭지기사별 적응 언어모델과 기본 언어모델의 성능비교

<표 3> 기본 언어모델 대비 적응 언어모델의 문장별 성능변화

	정적 보간 계수	동적 보간 계수
성능향상	128	193
성능변화 없음	256	262
성능저하	75	4
합	459	

5. 결 론

본 논문에서는 추가적인 텍스트 자료가 없을 경우 정보검색 기법을 사용하여 기존 텍스트 코퍼스로부터 N-gram 적응에 효과적인 적응 코퍼스를 획득하는 방법을 제안하였다. 또한 언어모델에 기반한 정보검색에서 트라이그램 모델을 사용하여 일반적으로 사용자가 만족하는 정보검색의 답이 아닌 N-gram 분포가 유사한 문서를 검색하는 방법이 보다 성능이 좋은 적응 코퍼스를 얻을 수 있음을 확인하였다. 그리고 본 논문에서는 간단한 선형 보간 언어모델 병합에서 사용될 수 있는 동적 보간 계수를 제안하였다. 입력 음성의 중간결과를 언어모델의 변화에 민감한 구간과 그렇지 않은 구간으로 나누어 동적 보간 계수를 구간 별로 구하였다. 제안

한 방법은 인식대상 음성의 특정 구간에서 특정 언어모델이 우세하게 나타날 경우 그 구간에서 해당 언어모델을 가중하는 방법이 인식 후보 단어 열 각각에 대해 최적의 보간 계수를 구하는 경우에 비해 성능이 좋음을 실험적으로 증명하였다.

본 논문에서는 기본 언어모델의 성능 저하가 없는 적응 언어모델의 구현을 위한 N-gram만의 병합 방법을 제안하였으나, 음성인식의 중간 결과를 나누어 여러 개의 언어모델로 각각의 신뢰도를 구한 뒤 구간에 대한 새로운 적응 언어모델의 가중치를 결정하는 방법은 향후 청크 결합 정보나 구문 분석 결과와 같은 상위 자연언어 지식을 기본 언어모델의 성능을 저하시키지 않으면서 적응 언어모델로 사용하는 연구로 확장될 수 있다. 또한 본 논문에서는 언어모델에 대한 민감도만을 신뢰도로 사용하였으나 음향모델이나 전처리에서 제공할 수 있는 다양한 음향학적 정보를 모두 고려하여 연속음성인식의 후처리로 사용할 수 있다.

참 고 문 헌

[1] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here", *Proc. IEEE*, vol. 88, no. 8, pp. 1270-1278, 2000.

[2] J. R. Bellegarda, "An overview of statistical language model adaptation", in *Proc. ISCA Adaptation Methods in Automatic Speech Recognition*, pp. 165-174, 2001.

[3] B. Bigi, Y. Huang, and R. De Mori, "Vocabulary and language model adaptation using information retrieval", in *Proc. ICSLP*, pp. 1361-1364, 2004.

[4] C. Zhai, and J. Lafferty, "A study of smoothing methods for language models applied to adhoc information retrieval", in *Proc. ACM SIGIR'01*, pp. 334-342, 2001.

[5] R. Kneser, and V. Steinbiss, "On the dynamic adaptation of stochastic language models". Vol. 2, pp. 586-589, in *Proc. ICASSP*, 1993.

[6] M. Weintraub, Y. Aksu et al., "LM95 project report: fast training and portability", *Technical Report 1*, Center for Language and Speech Processing, Johns Hopkins University.

[7] I. Bulyko, M. Ostendorf, and A. Stolke, "Getting more mileage from web text sources for conversational speech language modeling using class dependent mixtures", *CL-HLT*, pp.7-9, 2003.

[8] L. Mangu, *Finding Consensus in Speech Recognition*, Ph. D. Thesis, Johns Hopkins Univ. 2000.

[9] F. Wessel, R. Schluter et al., "Confidence measures for large vocabulary continuous speech recognition", *IEEE TRANS. on Speech and Audio Processing*, Vol. 9, No. 3, pp. 288-298, 2001.

[10] A. Kalai, S. Chen et al., "On-line algorithm for combining language models", in *Proc. ICASSP*, pp. 745-748, 1999.

[11] F. Jelinek, "Self-organized language modeling for speech recognition", in *Readings in Speech Recognition*, Alex Waibel and Kai-Fu Lee. Morgan Kaufmann, pp. 450-506,

1989.

[12] A. Stolke, "SRILM - an extensible language modeling toolkit", in *Proc. ICSLP*, Denver, pp.901-904, 2002.

[13] S. E. Robertson, S. Walker et al., "Okapi at TREC-4", *NIST TREC-4*, pp. 73-96, 1995.

[14] M. Bacchiani, and B. Roark, "Unsupervised language model adaptation", in *Proc. ICASSP*, pp.224-227, 2003.

[15] G. Adda, M. Jardino, and J. L. Gauvain, "Language modeling for broadcast news transcription", in *Proc. Eurospeech*, pp.1759-1760, 1999.

[16] R. Gretter, and G. Riccardi. "On-line learning of language models with word error probability distributions", in *Proc. ICASSP*, 2001.

[17] M. Federich, and N. Bertoldi, "Broadcast news LM adaptation using contemporary texts", in *Proc. Eurospeech*, pp. 239-242, 2001.

[18] X. Zhu, and R. Rosenfeld, "Improving trigram language modeling with the world wide web", in *Proc. ICASSP*, pp. 533-536, 2001.

[19] 김현숙, 전형배, 김상훈, 최준기, 윤승, "방송 뉴스 인식을 위한 언어 모델 적응", *말소리*, 대한음성학회 학회지, 제 51호, pp. 99-115, 2004.

[20] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", *Doklady Akademii Nauk SSSR*, Vol. 163, No. 4, pp. 845-848, 1965 (Russian). English translation in *Soviet Physics Doklady*, Vol. 10, No. 8, pp. 707-710, 1966.

[21] J. Ponte, and W. B. Croft, "A language modeling approach to information retrieval", in *Proceedings of the ACM SIGIR*, pp. 275-281, 1998.

[22] D. H. Miller, T. Leek, and R. Schwartz. "A hidden Markov model information retrieval system", in *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 214-221, 1999.

접수일자 : 2005년 11월 30일
게재결정 : 2005년 12월 23일

▶ 최준기(Joon Ki Choi)
주소: 305-701 대전광역시 유성구 구성동 한국과학기술원
소속: 한국과학기술원 전산학과 음성인터페이스연구실
전화: 042) 869-3556
E-mail: jkchoi@speech.kaist.ac.kr

▶ 오영환(Yung-Hwan Oh) : 교신저자
주소: 305-701 대전광역시 유성구 구성동 한국과학기술원
소속: 한국과학기술원 전산학과 음성인터페이스연구실
전화: 042) 869-3516
E-mail: yhoh@cs.kaist.ac.kr