한국음향학회지 제21권 제4호 pp. 407~414 (2002)

대어휘 연속음성 인식을 위한 결합형태소 자동생성

Automatic Generation of Concatenate Morphemes for Korean LVCSR

박 영 희*, 정 민 화*
(Young-Hee Park*, Minhwa Chung*)

*서강대학교 컴퓨터학과 음성언어처리연구실
(접수일자: 2002년 4월 2일; 채택일자: 2002년 4월 23일)

본 논문에서는 형태소를 인식 단위로 하는 한국어 연속음성 인식의 성능 개선을 위해 결합형태소를 자동으로 생성하는 방법을 제시한다. 학습 코퍼스의 54%를 차지하고 오인식의 주요인이 되는 단음절 형태소를 감소시켜서 인식 성능을 높이는 것을 목적으로 한다. 품사의 접속 규칙을 이용한 기존의 지식기반의 형태소 결합방법은 접속 규칙의 생성이 어렵고, 학습 코퍼스에 나타난 출현 빈도를 반영하지 못하여 저빈도 결합형태소를 다수 생성하는 경향을 보였다. 본 논문에서 제시하는 방법은 학습데이터의 통계정보를 이용하여 결합형태소를 자동 생성한다. 결합할 형태소 쌍 선정을 위한 평가척도로는 형태소 쌍의 빈도, 상호정보, 유니그램 로그 유도값(unigram log likelihood)을 이용하였고 여기에 한국어의 특성 반영을 위해 단음절 형태소 제약과 형태소 결합길이를 제한하는 두개의 제약사항을 추가하였다. 학습에 사용된 텍스트 코퍼스는 방송뉴스와 신문으로 구성된 7백만 형태소이고, 최빈도 2만 형태소 다중 발음사전을 사용하였다. 세가지 평가척도 중 빈도를 이용한 것의 성능이 가장 좋았고 여기에 제약조건을 반영하여 성능을 더 개선할 수 있었다. 특히 최대 결합 길이를 3으로 할 때의 성능이 가장 우수하여 언어모델 혼잡도는 117.9에서 97.3으로 18% 감소했으며, 형태소 에러율 (MER: Morpheme error rate)은 21.3%에서 17.6%로 감소하였다. 이때 단음절 형태소는 54%에서 30%로 24%가 감소하였다.
핵심용어: 결합형태소, 결합형태소 자동생성, 대용량 연속음성 인식, 언어모델
투고분야: 음성처리 분야 (2.5, 2.7)

In this paper, we present a method that automatically generates concatenate morpheme based language models to improve the performance of Korean large vocabulary continuous speech recognition. The focus was brought into improvement against recognition errors of monosyllable morphemes that occupy 54% of the training text corpus and more frequently mis-recognized. Knowledge-based method using POS patterns has disadvantages such as the difficulty in making rules and producing many low frequency concatenate morphemes. Proposed method automatically selects morpheme-pairs from training text data based on measures such as frequency, mutual information, and unigram log likelihood. Experiment was performed using 7M-morpheme text corpus and 20K-morpheme lexicon. The frequency measure with constraint on the number of morphemes used for concatenation produces the best result of reducing monosyllables from 54% to 30%, bigram perplexity from 117.9 to 97.3, and MER from 21.3% to 17.6%.
Keywords: *Concatenate morpheme models, Large vocabulary continuous speech recognition, Automatic concatenate morphemes generation, Language models*
ASK subject classification: *Speech signal processing (2.5, 2.7)*

책임저자: 박영희 (yhpark@nlpeng.sogang.ac.kr)
127-742 서울시 마포구 신수동 1번지
서강대학교 컴퓨터학과 음성언어처리연구실
(전화: 02-712-8023; 팩스: 02-704-8273)

I. 서 론

대용량 연속 음성 인식을 위한 인식 단위의 선정은 매우 중요한 이슈로 한국어 대용량 연속음성 인식을 위해서는 형태소를 기본 단위로 하는 것이 일반적이다[2,3,9]. 이때 실질 형태소를 제외한 조사, 어미, 접미사, 의존명사, 보조용언 등의 기능성 형태소는 대부분이 단음절로 구성되어 있고, 형태소 경계에서 발생하는 음운변화 현상이 제대로 반영되지 않을 뿐만 아니라 발화구간이 짧은 인식단위어므로 삽입, 삭제와 같은 인식 오류를 빈번히 유발하는 요소이다[9]. 이런 이유로 형태소를 인식단위로 선정할 때 인접한 여러 형태소들을 결합하여 새로운 인식단위를 생성하려는 시도가 진행되어 왔다[2,3,9]. 형태소를 결합하면 형태소간의 음운변화를 반영한 발음열 생성이 가능하고, 언어모델 측면에서는 variable N-gram의 적용 효과를 얻을 수 있다. 인접한 단어들을 결합하는 방법으로는 지식 기반의 방법[2,9]과 평가척도를 이용한 방법[4,7,8,10,11]이 있다. 지식 기반 방법은 각 형태소의 품사 정보를 이용하여 접속규칙을 생성하고 각 접속 규칙에 해당하는 형태소들을 결합하여 결합형태소를 생성하는 반면, 평가척도를 이용하는 방법은 형태소 쌍의 출현 빈도수, 두 형태소의 상호정보 (mutual information), 유니그램 로그 유도값 등의 평가척도를 이용하여 학습 텍스트로부터 통계정보를 얻어내고 그 정보를 이용하여 스코어가 높은 형태소 쌍을 결합한다. [10]의 연구에서는 텍스트의 통계정보 외에 음향학적 특성을 이용한 기준을 도입하였으나 성능의 개선을 얻지 못하였다.

본 연구의 학습에 사용된 학습데이터는 방송뉴스와 신문으로 구성된 7백만 형태소 텍스트 코퍼스이고, 이 중에서 단음절 형태소는 전체의 54%에 달하며 이들은 1,175개의 유일 형태소로 구성되어 있다. 그림 1은 이들 단음절 형태소 각각의 빈도수를 내림차순으로 정렬하여 누적한 분포도이다. 빈도수가 높은 상위 12개 단음절 형태소가 전체 단음절 형태소 출현빈도의 50%를 차지하고, 상위 100개의 단음절 형태소가 전체의 90%를 차지하고 있다. 이러한 단음절 형태소의 특성을 이용하여 모든 단음절 형태소를 다 고려할 필요없이 출현 빈도수가 높은 소수의 단음절 형태소만을 주 대상으로 하여도 상당수의 단음절 형태소 출현을 줄일 수 있다.

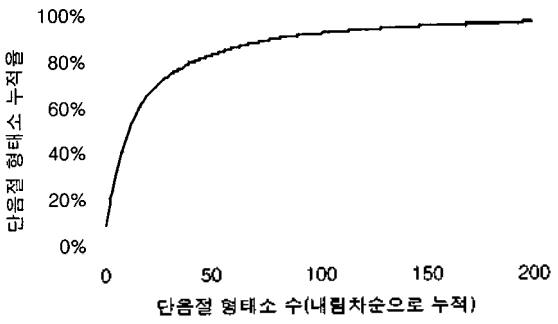

그림 1. 단음절 형태소 누적분포
Fig. 1. Accumulated distribution of monosyllable morphemes.

본 논문에서는 지식기반 방법과 평가척도를 이용한 방법들을 비교·분석하여 한국어에 접합한 생성 방법을 연구하고, 가장 성능이 좋은 생성 방법에 출현 빈도가 높은 단음절 형태소들을 줄일 수 있도록 제약사항을 추가하였다. 이로부터 대어휘 연속음성 인식에 적합한 결합형태소 기반의 언어모델을 생성하고자 한다.

II. 텍스트 코퍼스

실험에 사용한 텍스트 코퍼스는 방송뉴스와 신문의 텍스트가 비슷한 비율로 구성되어 있다. 방송뉴스는 97년 1월부터 99년 2월까지의 KBS 9시 뉴스 기사들로 인터넷과 국어정보베이스 II[1]에서 발취하였다. 신문 텍스트는 모두 국어정보베이스 II에서 발취한 94년, 97년 8월부터 97년 12월까지의 조선, 한겨레, 동아일보 텍스트들이다. 각 텍스트는 전처리를 통하여 필요없는 한자, 영문자, 기호, 숫자들을 제거하거나 변환한 후에 음성 기반의 형태소 분석을 수행하였다[2]. 음성 기반의 형태소 분석은 문자 기반의 형태소 분석과 달리 형태소들의 소리값을 유지하여야 하므로 용언의 불규칙, 축약, 생략 현상에 대한 처리를 하지 않고 원형을 유지하도록 분석하였다. 형태소 분석의 정확성을 위하여 형태소 분석된 결과를 사람이 다시 한번 확인하였다. 복합명사는 분할을 원칙으로 하고, 조사와 어미도 각 기능 단위로 분할하는 등 가능한 최소 단위로 분할하여 미등록어가 최소로 되도록 하였다.

형태소 분석된 전체 텍스트는 7백만 형태소로 구성되고, 이중 45만 형태소를 테스트에 사용하였다. 학습 코퍼스의 형태소 단위 출현 빈도로부터 최빈도 20,660 형태소를 선택하고 다중 발음을 허용하여 베이스라인 사전을 생성하였다. Bigram 언어모델은 CMU-Cambridge SLM[5]을 사용하였으며, Good Turing smoothing 방법을 적용하였다. 베이스라인 형태소 bigram의 혼잡도는 117.9이다.

III. 지식 기반 형태소 결합

의사형태소 생성방법[3]에서는 언어학적 지식을 바탕으로, 형태소들을 결합한 후에도 결합형태소의 품사 카테고리가 변하지 않도록 형태소 품사들간의 결합 규칙을 생성하여, 한 어절 내에서 실질 형태소와 형식 형태소가 분리되도록 어미와 어미를 결합하거나 조사와 조사를 결합하였다 (예: "에_는", "ㅆ_습니다"). 보조 용언의 경우에는 여러 어절이 결합하여 하나의 보조 용언구를 생성하도록 하였다 (예: "ㄹ_수_있", "게_되"). 한국어의 용언구는 대부분이 짧은 형태소로 구성되어 있으면서 여러 어절이 하나의 언절을 구성하여 연속으로 이어서 발화하는 경향이 있다. 다행히도 보조 용언구를 형성하는 형태소의 종류가 적고, 발화 패턴이 유사하므로 결합 규칙의 생성이 용이하고 형태소 간의 음운 변화 반영이 쉬워진다.

이 방법은 인식 후에도 품사 정보를 유지한다는 장점을 가지지만 각 형태소의 품사 정보가 정확해야만 형태소를 결합할 수 있고, 언어학적인 지식을 요구하므로 결합 규칙의 생성이 어렵다. 또한 품사 정보만을 대상으로 결합하므로 각 형태소의 출현 빈도를 반영하지 못하여 저빈도 결합형태소를 매우 많이 생성하는 경향이 있다.

본 연구에서 지식 기반 형태소 결합으로 추가된 결합형태소는 약 4,000개이지만, 이중 10번 이상 나타난 것은 약 1,000개에 불과했다. 나머지 3,000개의 결합형태소는 오히려 데이터 부족 문제를 유발하는 요인으로 작용하는 문제점을 안고 있다. 본 연구에서는 데이터 부족 문제를 줄이고 다른 방법과의 비교를 위해서 출현빈도가 낮은 결합형태소들은 다시 원래의 형태소 단위로 분할하였다. 각 결합형태소의 출현빈도수를 기준으로 하여 빈도수가 높은 순으로 결합형태소를 100개 단위로 순차적으로 추가하였다.

IV. 평가척도를 이용한 형태소 결합

학습 텍스트로부터 형태소 쌍을 자동 선택하여 결합한 후 사전에 추가하면 사전의 크기는 커지겠지만, 그럼에도 언어모델의 혼잡도를 감소시키는 것이 목적이다. 각 평가척도에 의해 선택된 형태소 쌍을 결합하였을 때 학습 텍스트의 로그 유도값을 감소시키는 것은 언어모델의 혼잡도가 감소하는 것을 의미하므로 이를 이용하여 결합할 형태소 쌍을 결정할 수 있다.

(고, 있)과 같이 결합할 형태소 쌍이 결정되면 사전에 새로 생성된 결합형태소 "고_있"을 추가하고 학습 텍스트의 해당 형태소 쌍을 모두 결합한다. 다시 위의 과정을 반복하고, 만약 새로운 형태소 결합 (예: "고_있_다")으로 이전에 결합한 결합형태소 (예: "고_있")의 빈도수가 한계값 이하로 떨어지면 다시 이전 형태소 (예: "고"와 "있")로 분할한다.

4.1. 형태소 쌍 선택 평가척도

평가척도는 연속된 두 단어의 상호관계를 평가할 수 있는 방법으로, 기존 연구에서 사용한 여러 가지 평가척도들을 적용해 보고 한국어 대용량 인식에 가장 적합한 평가척도를 선택한다.

학습 텍스트의 전체 형태소 수를 N, 형태소 u의 빈도수를 $N(u)$, 형태소 쌍 (u, v)의 빈도수를 $N(u, v)$라고 할 때, 실험에 사용한 평가척도는 다음과 같다.

(1) 형태소 쌍의 빈도수 (Freq): $N(u, v)$

(2) 상호정보 (Mutual information, MI):

$$F_{MI}(u, v) = N(u, v) \log \frac{N(u, v) \cdot N}{N(u) \cdot N(v)}$$

(3) 유니그램 로그 유도값 (ULL)의 변화율 ($\triangle F_{UNI}$):

$\tilde{N}$는 형태소 결합 후의 빈도수

$$F_{UNI} = \sum_{w} N(w) \log \frac{N(w)}{N}$$

$$\begin{aligned}
&\triangle F(u, v) \\
&= \tilde{F}(u, v) - F(u, v) \\
&= \tilde{N}(u) \cdot \log \tilde{N}(u) + \tilde{N}(v) \cdot \log \tilde{N}(v) \\
&\quad + \tilde{N}(u_v) \cdot \log \tilde{N}(u_v) - \tilde{N} \cdot \log \tilde{N} \\
&\quad - N(u) \cdot \log N(u) - N(v) \cdot \log N(v) + N \cdot \log N
\end{aligned}$$

위의 세가지 평가척도 외에 [2]에서 우수한 성능을 나타낸 순방향과 역방향 bigram의 기하평균과 [5]의 좌측 확률 (left probability)도 함께 실험하였으나 위의 세 평가척도보다는 언어모델 혼잡도의 감소가 적었으므로, 본 논문에서는 우수한 성능을 보인 (1)~(3)의 평가척도에 대해서만 설명을 하였다.

4.2. 결합형태소 자동생성 방법

결합형태소의 후보 선택에 있어서 계산 시간이나 효율성의 문제를 고려하기 위해 한계 값과 제약사항을 설정하여 하나 이상의 후보를 선택할 수 있도록 한다.

결합형태소 자동생성을 위한 요구사항과 각 단계는 다음과 같다.

(1) 매 단계마다 학습 데이터의 연속된 모든 형태소 쌍의 출현 빈도수, $N(u,v)$에 따라 내림차순으로 정렬하여, 출현 빈도수가 한계값, C_{th} 이상인 형태소 쌍만을 후보로 한다. 실험에서는 $C_{th}=400$으로 설정하였다.

(2) 스코어가 가장 높은 것부터 결합하여 "u_v" 형태의 결합형태소를 사전에 추가하고, 같은 단계에서 형태소 쌍들간의 충돌을 피하기 위해서 $(*, u)$ 형태소 쌍들과 $(v, *)$ 쌍들을 후보에서 제외시킨다.

(3) 상호 정보(MI)를 이용한 방법과 ULL 방법은 $N(u,v)>C_{th}$를 만족하는 형태소 쌍들에 대해 각 평가척도의 스코어를 계산하고, 각 스코어를 기준으로 내림차순으로 정렬한 후, 다음 수식에 의해 $T_{\min}$을 결정하고 $T_{\min}$보다 큰 값을 갖는 형태소 쌍들을 (2)의 규칙을 적용하여 결합한다[12].

$$T_{\min}=(1-p)\max_{u\in V, v\in V} F'(u,v),\quad F':F_{MI}\ or\ \triangle F_{UNI}$$

$T_{\min}$은 F_{MI}와 $\triangle F_{UNI}$의 최소값이고, p는 $T_{\min}$ 계산을 위한 계수이다.

(4) 학습 텍스트에 대해서 추가된 결합형태소를 모두 결합하고, 혼잡도가 감소하지 않을 때까지 위의 과정을 반복한다.

Freq 방법의 경우 각 단계에서 추가하는 결합형태소의 수가 언어모델 혼잡도에 큰 영향을 미치지 않았으므로 수행 속도를 감안하여 추가 결합형태소를 100개로 정하였다.

반면 MI 방법과 ULL 방법은 C_{th}와 p의 설정에 따라 추가되는 결합형태소가 많이 달라지고 언어모델 혼잡도 차이가 많이 나는 양상을 보였다. 계수 p는 각 단계마다 추가하는 결합형태소의 수를 결정하므로 수행속도에 영향을 미친다. 실험에 의해 얻은 MI 과 ULL 방법을 위한 p는 각각 0.3, 0.5일때 최적의 성능을 나타내었다. C_{th}는 추가 대상을 결정하는 역할을 하여 실험에서 두 방법 모두에서 C_{th}가 클수록 언어모델의 혼잡도가 감소하는 현상을 보였다.

표 1. 생성된 결합형태소
Table 1. Generated concatenate morphemes.

최상위 빈도	하_ㄴ, 하_는, 이_ㅂ니다, 해_ㅆ_다, 아_ㄴ, 이_다, 등_이, ㅆ_다, 되_ㄴ, 해_ㅆ_습니다, ㅆ_습니다, 하_고, 등_은, 에_는, 고_있_다
긴형태소	하_고_있_습니다, 되_고_있_습니다, 기자_가_취재_해_ㅆ_습니다, 에_도_불구_하_고, 하_기_로_해_ㅆ_습니다

4.3. 결합형태소 생성 예

평가척도를 이용하여 생성된 결합형태소의 예는 표 1과 같다. 빈도수가 높은 결합형태소들은 빈도수가 가장 높은 것부터 10,000번 이상의 빈도수를 갖는 것들을 차례대로 나열한 것이다. 대부분이 용언과 어미, 조사들의 결합임을 알 수 있다. 또한 결합 길이가 긴 결합형태소들 역시 주로 용언구들로 이루어졌으나 대체로 매우 낮은 빈도수를 보였다.

그림 2의 결과에서 나타난 것처럼 지식 기반 방법보다는 평가척도를 이용한 것의 혼잡도 감소가 크고, 그 중에서도 Freq를 이용한 것이 가장 좋은 성능을 나타내었다.

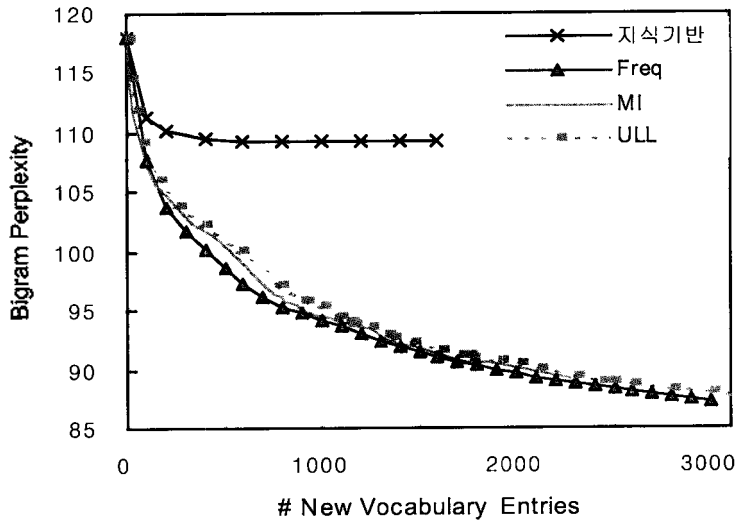

그림 2. 결합형태소 추가에 따른 언어모델 혼잡도
Fig. 2. Bigram perplexity versus the number of concatenate morphemes added to the lexicon.

V. 제약사항을 반영한 형태소 결합

앞에서 설명한 세가지 평가척도들 중 성능이 가장 좋은 Freq를 기반으로, 한국어의 특성을 반영할 수 있도록 다음의 두가지 제약사항을 추가하였다.

5.1. 단음절 형태소 제약

학습 코퍼스의 54%가 단음절 형태소들이며, 이들은 1,175개의 형태소로 구성되어 있다. 또한 전체 단음절 형태소의 50%를 구성하는 것은 단지 12개의 형태소이고, 약 100개의 단음절 형태소가 90%를 차지하였다. 이러한 특성을 반영하여 단음절 형태소만을 대상으로 결합형태소를 생성하면 인식 성능면에서 효과가 있을 것으로 예상된다.

알고리즘의 첫번째 단계에서 C_{th} 제약조건 외에 (u, v) 쌍의 두 형태소 중 적어도 하나는 단음절 형태소인 형태소 쌍만을 결합형태소의 후보로 하는 제약조건을 추가하였다.

5.2. 형태소 결합 길이 제약

가장 긴 결합 길이는 6으로 "되_ㄹ_것_으로_보이_ㅂ니다" 또는 "기자_가_취재_해_ㅆ_습니다"와 같이 매우 긴 형태소가 생성되고 이들은 세 어절에 걸쳐 있다. 그러나 이와 같이 긴 형태소들은 실제로 출현 빈도수가 낮고, 테스트 데이터에도 잘 나타나지 않으므로 인식 성능에 도움을 주지 못한다.

대어휘 인식에 적합한 인식 단위를 만들기 위해서 긴 형태소보다는 적당한 길이의 형태소를 더 많이 만들 수 있도록 형태소 최대 결합 길이에 제약을 주었다. 실험에서 최대 결합 길이를 3 또는 4로 제한하였다.

VI. 결합형태소 기반의 언어모델

지식 기반, Freq, MI, ULL에 의해 생성된 결합형태소를 평가하기 위하여 베이스라인 형태소 사전에 결합형태소를 추가하여 결합형태소 기반의 bigram 언어모델을 생성하여 언어모델 혼잡도를 비교한다. Bigram 언어모델은 Good-Turing backoff smoothing 방법으로 생성하였다.

6.1. 혼잡도 정규화

언어모델 혼잡도는 언어모델의 성능을 평가하기 위해 가장 일반적으로 사용되는 평가수단이지만, 언어모델 혼잡도를 이용하여 평가하기 위해서는 같은 사전과 같은 테스트 데이터를 사용해야만 비교가 가능하다. 각기 다른 사전과 텍스트 데이터를 비교하기 위해서는 언어모델 혼잡도의 정규화가 필요하다.

초기의 테스트 텍스트를 T_w, 형태소 쌍을 결합형태소로 변환한 테스트 텍스트를 T_{cm}이라 하고 T_{cm}으로부터 얻은 언어모델 혼잡도를 PP_{cm}라 하면, T_w의 정규화한 언어모델 혼잡도는 다음과 같이 구할 수 있다[4].

$$\text{Normalized Perplexity},\ PP^* = \exp\left(\frac{N_{cm}}{N}\ln(PP_{cm})\right)$$

N_{cm}은 T_{cm}의 전체 형태소 수이고, N은 T_w의 전체 형태소 수이다.

6.2. Bigram 혼잡도

그림 2는 각각의 결합형태소 생성방법에 의해 생성한 결합형태소를 점진적으로 추가한 bigram 언어모델의 혼잡도 그래프이다.

지식 기반 방법의 혼잡도 감소가 가장 작은 것은 생성된 결합형태소의 출현 빈도수가 매우 낮기 때문이다. 약 200개의 결합형태소 외에는 매우 낮은 출현 빈도수를 보인다. 또한 결합 규칙이 한 어절 안에서 이루어지고 조사, 어미, 보조용언에 대해서만 생성했기 때문이다. Freq 방법의 혼잡도 감소가 가장 크게 나타났으며, MI와 ULL 방법은 Freq보다 혼잡도 감소가 적었다. 언어모델 혼잡도가 감소하도록 매개변수를 조정했을 때 Freq 곡선에 가까워지는 모습을 보였고, 언어모델 혼잡도에 가장 큰 영향을 주는 요소는 출현 빈도수였다. 제약사항을 반영한 Freq-MS (단음절 형태소 제약)와 Freq-Len (형태소 결합길이 제약)은 Freq보다 언어모델 혼잡도가 조금씩 더 감소하는 모습을 보였다. Freq-Len 방법의 경우 형태소 길이를 3으로 했을 때의 성능이 가장 좋았다.

결합형태소를 많이 추가할수록 언어모델 혼잡도가 계속해서 감소하는 현상이 보이지만 실제 인식 성능에도 같은 현상이 보이는지는 알 수 없으므로 추가 결합형태소의 수는 인식 실험을 통해 확인하고자 한다.

6.3. 단음절 형태소 분포

그림 3은 결합형태소 추가에 따라 단음절 형태소가 감

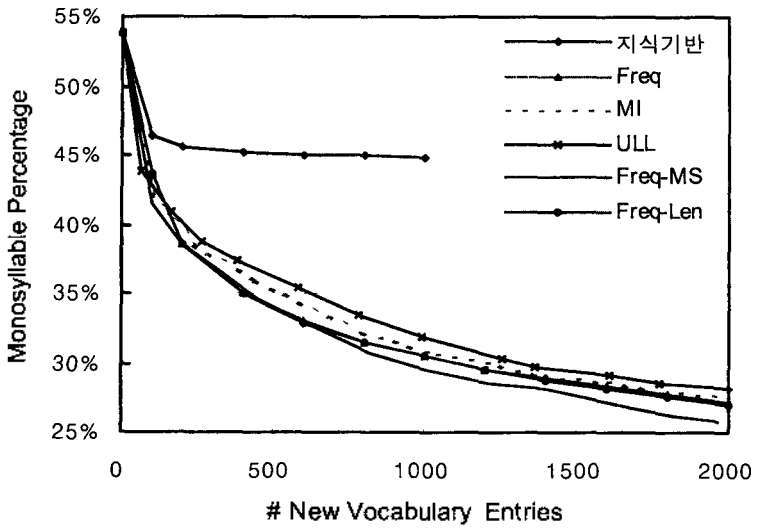

그림 3. 결합형태소 추가에 따른 단음절 형태소 분포
Fig. 3. Percentage of monosyllables versus the number of concatenate morphemes added to the lexicon.

소하는 추이를 보여주는 그래프로 단음절 형태소의 비율이 상당히 감소하고 있음을 알 수 있다. 처음에 54%에 달하던 단음절 형태소가 30% 이하로 매우 급격한 감소를 보였다. 혼잡도와 마찬가지로 제약사항을 반영했을 때 더 많은 감소를 보였다. 뿐만 아니라 빈도수가 높은 형태소들의 많은 부분이 단음절 형태소임도 그래프를 통해 알 수 있다.

VII. 연속음성 인식 실험

7.1. 한국어 연속음성 인식기

결합형태소 기반의 언어모델을 평가하고, 추가할 결합형태소의 적절한 수를 결정하기 위해 HTK (Hidden Markov Toolkit)[6]를 사용하여 인식 실험을 수행하였다. 본 실험에서 사용한 음향모델은 CHMM을 기반으로 하였으며 6개의 Gaussian mixture를 사용하였다.

신문, 방송뉴스, 여러 분야의 서적 등 다양한 분야에서 추출한 문장을 남녀 화자 모두를 포함하도록 낭독체 스타일로 발화한 연속음성을 음향모델 학습 및 인식을 위해 사용하였다. 음향모델 학습에는 18,000문장의 약 30시간 분량을 사용하였다. 테스트용 음성데이터는 테스트 문장의 대표성을 부여하기 위하여 짧은 문장은 배제하였으며, 학습에 사용되지 않은 20화자, 2,000문장 중에서 미등록어가 없도록 임의로 300문장을 선정하였다. 선택된 한 문장은 평균 9.9 어절, 19.5 형태소로 이루어져 있다.

인식단계에서는 형태소 정보를 사용하지 않으며, 지식기반의 결합형태소 생성 단계에서만 사용한다.

7.2. 인식 결과

인식 결과도 언어모델 혼잡도와 같은 이유로 정규화를 필요로 한다. 인식 결과의 정규화를 위해서 인식된 결합형태소를 이전의 형태소 단위로 분할하여 같은 단위로 변환한 후에 인식률을 계산하고 평가하였다.

그림 4는 결합형태소를 추가했을 때 각 결합형태소 생성 방법에 따른 형태소 에러율의 변화곡선으로, 언어모델 혼잡도의 변화 추이에 따라 200, 600, 1,000, 2,000개의 결합형태소를 각각 추가하여 인식 실험을 수행하였다. 인식 성능 역시 언어모델 혼잡도에서와 비슷한 순서를 나타내고 있다. 평가척도들 중에서는 Freq의 형태소 에러율 (MER: Morpheme Error Rate)이 가장 낮고, ULL은 기대치에 못 미치는 결과를 보여주었으며, Freq에 제약사항을 반영한 Freq-Len의 성능이 가장 좋은 것을 알 수 있다. 추가 결합형태소가 적을 때 Freq-Len의 형태소 에러율이 특히 많이 감소하는 것을 볼 수 있다. 또한 언어모델 혼잡도보다 형태소 에러율의 감소가 더 제한적이다.

인식 실험을 위해 추가 결합형태소 수의 결정에 있어서, 결합형태소를 많이 포함할수록 형태소 에러율은 감소하지만 사전 크기가 커지는 것을 고려하여야 하므로 감소가 완만해지는 부분에서 결합형태소 수를 결정하였

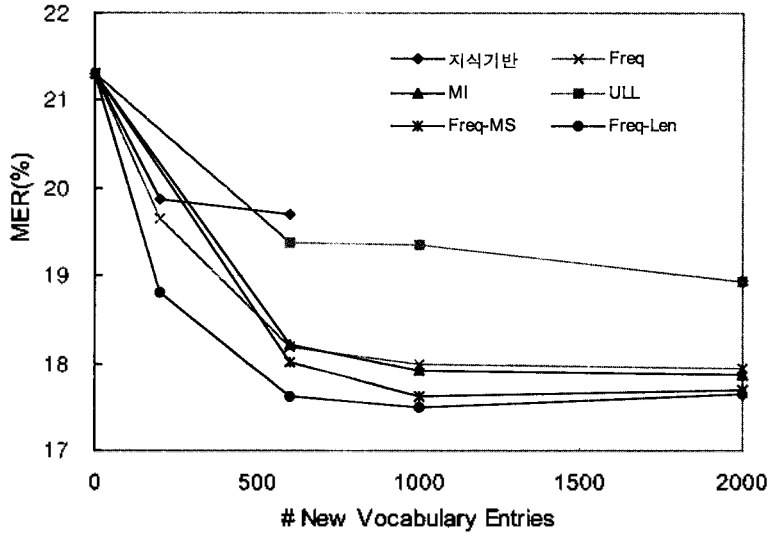

그림 4. 결합형태소 추가에 따른 형태소 에러율
Fig. 4. MER (Morpheme Error Rate) versus the number of concatenate morphemes added to the lexicon.

표 2. 혼잡도 vs. 형태소 에러율
Table 2. Perplexity vs. MER (Morpheme Error Rate).

분 류		PP*	MER (%)
베이스라인		1117.92	21.30
지식 기반		109.40	19.70
제약사항 없을 때	ULL	100.26	19.37
	MI	99.00	18.22
	Freq	97.34	18.19
제약사항 포함	Freq-MS	97.01	18.02
	Freq-Len	97.26	17.61

다. 본 실험에서는 인식 실험결과로부터 추가 결합형태소 수를 600개로 결정하였다.

표 2는 결합형태소를 600개 추가했을 때의 혼잡도와 형태소 에러율을 나타내고 있다. Freq 방법에 형태소 결합 길이 제약을 가했을 때 (Freq-Len) 17%의 형태소 에러율 감소를 보여 가장 좋은 성능을 보여주었다. Freq-MS의 혼잡도가 가장 작게 나타났지만 형태소 에러율은 Freq-Len이 가장 많은 감소를 보여 언어모델 혼잡도가 성능평가를 위한 절대적인 평가 방법이 아님도 확인할 수 있다.

VIII. 결 론

본 논문에서는 한국어 대어휘 연속 음성 인식에 적합한 인식 단위를 결정하기 위하여 결합형태소 기반의 언어모델을 제시하였다. 짧은 인식 단위가 많은 인식오류를 유발하므로 이러한 짧은 인식단위를 줄이도록 결합형태소를 생성하였다. 한국어의 분석을 통해, 형식 형태소들은 대부분이 단음절 형태소로 이루어져 있는 반면 결합 패턴이 일정하고 빈도수가 매우 높다는 특성을 반영하여 단음절 형태소를 54%에서 30% 이하로 감소시켰으며, 언어모델 혼잡도는 117.9에서 97.3으로 18%, 형태소 에러율은 21.3%에서 17.6%로 상대적으로 17% 감소시켰다.

감사의 글

본 연구는 한국과학재단 목적기초연구 (과제번호: 2001-1-30300-003-3)지원으로 수행되었으며, 본 실험에 사용된 삼성 종합기술원의 낭독체 연속음성 데이터베이스 사용 허가에 감사드립니다.

참 고 문 헌

1. 국어정보베이스 II CD-ROM, KAIST, http://kibs.kaist.ac.kr/
2. 박영희, 정민화, "Tagged Word Bigram을 사용한 의사형태소 단위의 한국어 연속음성인식," 한국정보과학회 봄 학술발표 논문집, 1999.
3. 이경님, 정민화, "의사 형태소 단위의 음성언어 형태소 해석," 제 10회 한글 및 한국어 정보처리 학술대회 논문집, 396-404, 1998.
4. C. Beaujard and M. Jardino, "Language modeling based on automatic word concatenations," *Proc. of EUROSPEECH*, 4, 1563-

1566, 1999.
5. P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-CambridgeToolkit," *Proc. of EUROSPEECH*, 5, 2707-2710, 1997.
6. HTK Hidden Markov Model Toolkit, Version 2.2, http://htk.eng.cam.ac.uk/index.shtml
7. Dietrich Klakow, "Language-model optimization by mapping of corpora," *Proc. of International Conference on Acoustics, Speech, and Signal*, 2, 2287-2290, 1998.
8. Hong-Kwang Jeff Kuo, Wolfgang Reichl, "Phrase-based language models for speech recognition," *Proc. of EUROSPEECH*, 4, 1595-1598, 1999.
9. Oh-Wook Kwon, "Performance of LVCSR with morpheme-based and syllable-based recognition units," *Proc. of International Conference on Acoustics, Speech, and Signal*, 3, 1567-1570, 2000.
10. George Saon and Mukund Padmanabhan, "]Data-driven approach to designing compound words for continuous speech recognition," *IEEE Tran. on ASSP*, 9 (4), May 2001.
11. Kazuyuki Takagi, Rei Oguro and Kazuhiko Ozeki, "Effects of words string language models on noisy broadcast news speech recognition," *Proc. of International Conference on Spoken Language Processing*, 1, 154-157, 2000.
12. I. Zitouni, J. F. Mari and K. Smaili, J. P. Haton, "Variable-length sequence language model for large vocabulary continuous dictation machines," *Proc. of EUROSPEECH*, 4, 1811-1814, 1999.

저자 약력

● **박 영 희 (Young-Hee Park)**
한국음향학회지 제21권 제3호 참조

● **정 민 화 (Minhwa Chung)**

1984년 2월: 서울대학교 제어계측학과 (공학사)
1988년 5월: Univ. of Southern California 전기공학과 (M.S.)
1993년 8월: Univ. of Southern California 전기공학과 (Ph.D.)
1994년 9월~현재: 서강대학교 컴퓨터학과 부교수
※ 주관심분야: 음성언어처리, 자연어처리