형태소 기반의 한국어 방송뉴스 인식

박영희, 안동훈, 정민화
서강대학교 컴퓨터학과

Morpheme-based Korean broadcast news transcription

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

Abstract

In this paper, we describe our LVCSR system for Korean broadcast news transcription. The main focus is to find the most proper morpheme-based lexical model for Korean broadcast news recognition to deal with the inflectional flexibilities in Korean. There are trade-offs between lexicon size and lexical coverage, and between the length of lexical unit and WER. In our system, we analyzed the training corpus to obtain a small 24k-morpheme-based lexicon with 98.8% coverage. Then, the lexicon is optimized by combining morphemes using statistics of training corpus under monosyllable constraint or maximum length constraint. In experiments, our system reduced the number of monosyllable morphemes from 52% to 29% of the lexicon and obtained 13.24% WER for anchor and 24.97% for reporter.

Ⅰ. 서론

라디오나 방송뉴스는 다양한 분야의 주제를 포함하고, 여러 타입의 음성, 발화 스타일, 주변 잡음과 음악 등을 포함하고 있다. 이러한 다양성은 음성인식을 어렵게 만드는 주 요소들이지만 [1], 이외에도 한국어는 인식단위의 정의에 따라 인식의 어려움이 달라진다. 한국어 형태소는 복합명사와 같이 하나의 형태소로 볼 수도 있고, 더 작은 형태소로 분할할 수도 있다. 이 때문에 형태소 분석 결과에 따라 형태소의 길이와 그로 인한 인식사전의 크기가 크게 달라질 수 있다. 게다가 한국어 형태소는 인식을 어렵게 하는 단음절 형태소들이 많은 부분을 차지하고 있다 (그림 1 실선).

한국어 대어휘 연속음성 인식을 위해서는 형태소를 인식 단위로 하는 것이 일반적이다 [2][3]. [2]의 연구에서는 형태소 자체를 인식단위로 선정하였고, [3]의 연구에서는 형태소 결합을 수행하였으나 어절 내부만을 고려하였을 뿐 아니라, 휴리스틱 정보를 이용한 규칙 기반의 결합을 수행한 이후에 통계적 결합을 수행하여 최적의 형태소 결합이라 할 수 없다.

본 논문에서는 최소 단위가 되도록 형태소 분석을 수행하여 98.8%의 coverage를 갖는 24k 형태소 사전을 생성하였다. 또한 짧은 음성의 인식오류를 줄이기 위하여 짧은 형태소들을 대상으로 형태소를 결합하였다 [4]. 최적의 형태소 결합을 위하여, 지식 기반의 방법과 통계정보를 이용하는 다양한 평가척도들을 비교 실험하였으며, 더불어 단음절 형태소 결합규칙을 추가하여 최적의 결합형태소를 생성하였다. 실험 결과를 보면, 1,000개의 결합형태소를 추가하여 52%의 단음절 형태소를 29%로 감소시켰으며, 앵커와 리포터 각각에 대해 13.24%, 24.97%의 WER를 얻었다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 실험에 사용할 방송뉴스 코퍼스를 소개하고, 3·4장에서는 텍스트 코퍼스의 특징 분석과 결합 형태소 생성, 5장에서는 방송뉴스인식 시스템에 대하여, 6장에서 실험결과를 기술한다.

Ⅲ. 방송뉴스 코퍼스

방송뉴스 인식에 사용한 음성 DB는 97년 2월부터 98년 12월까지의 KBS 9시 뉴스 코퍼스이고, 이들 음성은 배경음악, 주변잡음 등을 포함한 앵커와 리포터의 음성으로 이루어져 있다. 특히 리포터의 음성은 배경잡음이 심하여 음질이 떨어진다. 음성 DB는 앵커의 음성이 6.5시간, 리포터의 음성이 19시간으로 구성되어 있으며, 98년 10월의 음성을 테스트용으로 사용하였다.

텍스트 코퍼스는 방송뉴스 기사를 주로 하며 소량의 신문 기사를 추가하였다. 방송뉴스 텍스트는 인터넷을 통해 수집한 97년 1월부터 99년 2월, 2000년 7월부터 2002년 6월까지의 KBS 9시 뉴스 기사이고, 신문 텍스트는 국어정보베이스II로부터 발췌한 94년, 97년의 조선, 한겨레, 동아일보의 기사이다. 이들 텍스트에 존재하는 기호, 영문자, 숫자 등을 제거 또는 변환한 후, 형태소 분석을 수행하여 16M 형태소 텍스트 코퍼스를 얻었다. 98년 10월의 텍스트는 학습에서 제외시켰다.

III. 형태소 기반의 어휘모델

한국어는 형태소 분석을 어떻게 하느냐에 따라 분석의 결과가 크게 달라진다. 표 1은 형태소 분석의 다른 예이다. 예 I 에서는 한 개의 형태소로 분석되었으나, 예 II에서는 2-3개의 형태소로 분석되는 것을 볼 수 있다. 예 I 의 형태소는 발화 길이가 길어서 인식 단위로는 좋지만, 사전의 크기가 커지고 coverage도 떨어지게 된다. 특히 복합/합성 명사의 경우, 이들을 구성하는 각각의 형태소들은 빈번하게 사용되는 형태소들이다. 반면, 예 II의 형태소들은 발화가 짧기는 해도 coverage가 좋은 작은 사전을 생성할 수 있다.

표 1. 형태소 분석 예

예 I		예 II	
형태소	품사	형태소	품사
방송뉴스	(복합)명사	방송+뉴스	명사+명사
범국민적	(합성)명사	범+국민+적	접두사+명사+접미사
까지와는	조사	까지+와+는	조사+조사+조사
시었습니다	어미	시+었+습니다	어미+어미+어미

그림 1은 학습 코퍼스에서의 형태소의 음절길이 분포 그래프이다. 예 II와 같이 형태소 분석했을 때의 결과를 실선(Baseline)으로 나타내었다. 전체 형태소의 52%가 단음절 형태소로 구성되어 있고, 대부분의 형태소가 4음절 이하임을 알 수 있다. 그러나 이렇게 인식 단위가 짧으면 인식 성능이 떨어지기 때문에 이를 보완하기 위하여, 혼잡도를 줄일 수 있도록 여러 가지 평가척도를 적용하여 결합형태소를 생성하였다 [4]. 그림 1의 점선은 4.3장의 제약조건을 반영하여 생성한 1000개의 결합형태소를 추가했을 때의 형태소 음절길이 분포이다. 단음절은 30% 이하로 감소하였고, 2음절 형태소는 47%로 증가하여 전체 형태소의 음절 길이가 길어진 것을 볼 수 있다. 특히 2, 3음절 형태소가 70% 이상을 차지하고 있다.

베이스라인 사전의 크기를 결정하기 위하여 사전크기와 어휘 coverage 간의 상관 관계를 그림 2에 나타

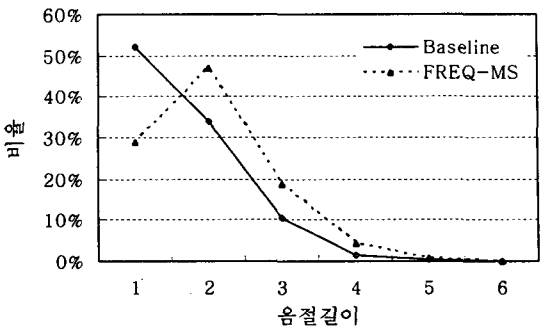

그림 1. 형태소의 음절길이 분포

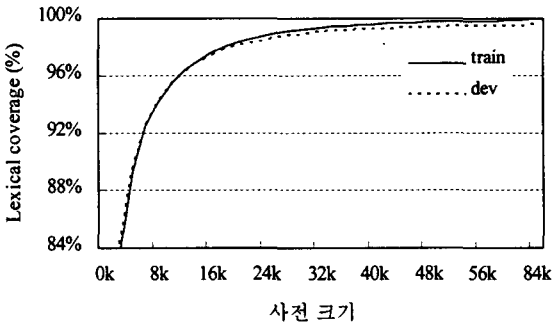

그림 2. 사전크기 vs. coverage(%).

내었다. 이 그래프로부터 98.8%의 coverage를 갖는 24k 형태소 사전을 선정하였다. 결합형태소는 포함하지 않은 사전이다. 충분히 높은 coverage를 가질 뿐만 아니라 일관성 있는 형태소 분석을 바탕으로 생성된 사전이므로, 매일 새로운 어휘가 나오는 방송뉴스라 할지라도 그 새로운 어휘들을 구성하는 기본 어휘들은 이미 사전에 포함되어 있으므로 어휘사전의 크기가 급격히 커지는 일은 없을 것으로 본다.

IV. 결합형태소 자동생성

지식기반 형태소 결합과 평가척도를 이용한 자동생성 방법을 비교·분석하였다 (결합형태소 자동생성 방법에 대한 자세한 설명은 [4] 참조).

4.1 지식기반 형태소 결합

지식기반의 형태소 결합 방법은 본 연구 이전에 수행한 연구[5]로 [3]의 규칙기반 결합과 유사하다. 결합규칙은 단순히 품사 정보만을 이용하여 생성하였고, 조사, 어미, 의존명사, 보조용언 등의 짧은 형태소를 대상으로 하여 결합규칙을 생성하였다. 그러나 이 방법은 상당히 주관적이고 경험에 의존하기 때문에 규칙 생성이 어렵고, 형태소 분석 결과의 정확성이 높아야만 하므로 소용량의 텍스트에는 적용 가능하지만 방송뉴스와 같은 대어휘의 텍스트에는 적용하기 어렵다.

그림 3에서 가장 작은 혼잡도 감소를 보였다.

4.2 평가척도를 이용한 형태소 결합

다음의 세가지 평가척도를 비교·분석하였다. 형태소 v의 빈도수를 $N(v)$, 전체 형태소를 N라 할 때,

- 형태소 쌍의 빈도수 (FREQ): $N(v,w)$
- Mutual information (MI):

$$N(v,w)\log\frac{N(v,w)N}{N(v)N(w)}$$

- Unigram log likelihood의 변화율 (ULL):

$$\sum_{v\in\hat{V}}\hat{N}(v)\log\frac{\hat{N}(v)}{\hat{N}}-\sum_{v\in V}N(v)\log\frac{N(v)}{N}$$

$\hat{N}$: 형태소 결합후의 통계정보

각 평가척도를 이용하여 언어모델 혼잡도를 최소로 하는 형태소 쌍을 선정하고, 학습 코퍼스의 해당 형태소 쌍을 결합하였다. 언어모델 혼잡도의 감소가 최소로 될 때까지 위 과정을 반복한다. 그림 3은 각 평가척도를 이용한 결합형태소를 추가했을 때의 혼잡도 변화 그래프로 형태소만 사용한 경우보다 혼잡도가 많이 감소하는 것을 볼 수 있다. 혼잡도와 인식률 모두에서 FREQ가 가장 좋은 성능을 보인 반면, ULL은 성능 개선이 가장 작게 나타났다.

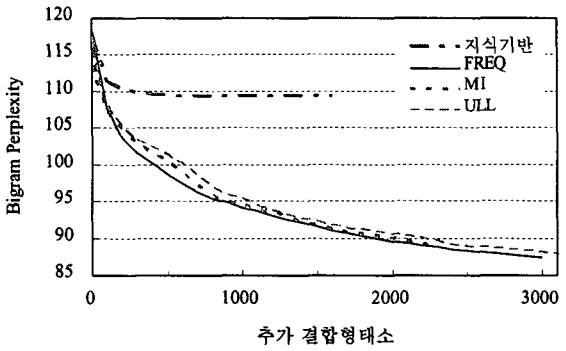

그림 3. 결합형태소 추가에 따른 언어모델 혼잡도

4.3 제약사항을 반영한 형태소 결합

최적의 결합형태소 생성을 위하여 결합형태소 생성 과정에 제약사항을 추가하였다. FREQ 척도를 이용하였다.

첫 번째는 단음절 형태소 제약(FREQ-MS)으로, 형태소 결합 과정에서 적어도 하나는 단음절 형태소일 때만 결합 가능하도록 하여 단음절을 줄이는 데에 초점을 두었다. 두 번째는 형태소 결합길이 제약(FREQ-LEN)으로, 최대 6개까지 형태소가 결합되는 예가 있으나 결합길이를 제한하여 가능한 한 많은 형태소가 결합될 수 있도록 하였다.

이 두 제약사항은 언어모델 혼잡도와 인식 성능 모두를 개선하였다 (표 2).

Ⅴ. 한국어 방송뉴스인식 시스템

5.1 음향 모델 및 언어 모델

이 논문에서 사용된 인식 시스템은 트라이그램을 사용한 1-패스 세미다이나믹(semi-dynamic) 네트워크 디코더에 기반한다. [6]에서 소개된 대로, 이 디코더는 한국어 낭독체 음성인식용으로 개발되었으며, MFCC 기반의 39차 특징벡터를 사용한다. 음향 모델은 2절에서 언급한 음성데이터베이스로부터 연속 HMM을 학습하였다. 각 HMM은 세 개의 상태를 가지며, 각 상태당 8개의 가우시안 혼합분포를 갖는다. 사전은 역시 2절의 데이터베이스로부터, 3절 및 4절에 언급한 형태소 및 결합 형태소 단위로 구성하였다. 언어 모델의 경우, 수정된 Kneser-Ney 감가(discounting) 방법으로 트라이그램 모델을 얻은 후, 엔트로피 프루닝(pruning) 및 2회 이하의 출현빈도를 갖는 이벤트를 모델에서 제외하도록 하였다. 이 트라이그램은 컴팩트 스태틱(compact static)한 인식 네트워크를 구성하기 위해 사용된다.

5.2 1-패스 세미다이나믹 트라이그램 네트워크 디코더

시스템의 디코더는 기본적으로 스태틱 네트워크에서 작동하는 비터비(Viterbi) 디코더이다. 먼저 주어진 언어 모델을 유한 상태 기계(finite state machine)로 표현하여 언어 모델(LM) 네트워크를 구한다. 이 네트워크는 각 LM 히스토리에 해당하는 노드와 그 다음에 나타날 수 있는 단어로 만들어지는 새로운 히스토리 노드를 연결함으로써 만들어진다. 그리고 동일한 히스토리를 갖는 단어노드들마다 그들의 발음열(HMM열)로 변환하여 후자 트리(successor tree)를 구성하여 인식 네트워크를 컴파일한다. 이 인식네트워크는 후자 트리간 꼬리 공유(shared tails) 기법[6]을 이용하여 전체 네트워크의 크기를 줄이도록 하였다. 이러한 공유 기법은 각 트리의 선형 꼬리(linear tail; 단 하나씩만의 child node를 가지는 경로)사이의 동치 관계를 이용한 방법으로, 트리구조의 네트워크의 경우 사실상 오토마타 최소화 결과에 근접한 결과를 얻을 수 있다. 한편, 이전에 구현된 네트워크 생성알고리듬은 프루닝된 히스토리에 대해서도 백오프(backoff) 전이를 위한 루트 노드만으로 구성된 빈 후자 트리를 생성하게 함으로써 네트워크의 비효율성, 특히 꼬리 공유 결과의 효율이 좋지 못하였다 (그림 4(a)). 그러나 이번 시스템에서는 노드가 백오프 아크만 포함하는 경우, epsilon 제거 알고리듬을 적용하여 이러한 비효율성을 해결하였다.

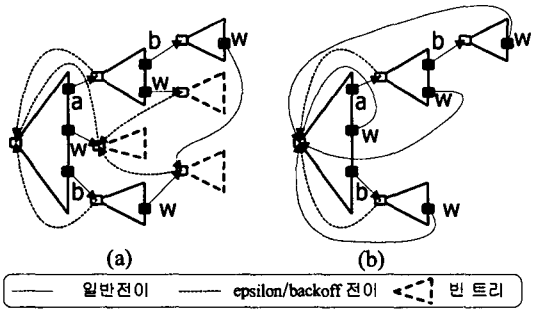

그림 4. 향상된 네트워크 구조.

그림 4(a)는 단어 w가 학습 코퍼스에 충분히 나타나지 못해 잘린 후(cutoff) 해당 후자트리(점선으로 된 트리)가 루트노드만으로 만들어진 상황을 보여준다. 이 경우, 유니그램(히스토리가 없는 상황)의 w에 해당하는 선형 꼬리는 한·두단어의 히스토리를 가지는 트리의 w의 것과 만나지 않기 때문에 공유가 불가능하였다. 그러나 백오프 아크의 가중치를 그 이전 아크에 누적시킨 후 백오프 아크를 제거한다면 훨씬 적은 수의 백오프 아크를 갖는 네트워크 구조를 얻을 수 있다(그림 4(b)). 이러한 향상된 네트워크 구조로 인해 트라이그램 사용시, 꼬리 공유 기법 적용을 통한 크기가 원래의 30~40%수준으로 줄어든 것을 확인할 수 있었다. 이전의 경우[6], 동일한 환경에 대해 약 60%대 수준으로 감소시켰다.

Ⅵ. 실험 결과

인식 실험에 사용한 음성 데이터는 98년 10월의 데이터로, OOV를 포함하지 않은 앵커 181문장, 리포터 180문장으로 구성되어 있다. 인식 대상 음성의 평균길이는 앵커가 21형태소, 리포터가 23형태소이고, 전체 형태소는 3,594와 4,290이다.

표 2. 언어모델 혼잡도 vs. WER (%).

평가척도	혼잡도	앵커 WER	리포터 WER
Baseline	68.2	15.92	27.18
MI	64.0	13.93	26.06
Freq-SM	63.6	13.24	25.73
Freq-Len	63.7	14.22	24.97

표 2는 결합형태소 1000개씩을 베이스라인 24k 사전에 추가했을 때의 언어모델 혼잡도와 인식 결과이다. 베이스라인과 비교할 때, 앵커 음성에 대해서 2.68% (FREQ-MS), 리포터 음성에 대해서는 2.21% (FREQ-LEN)의 WER를 감소시켰다.

베이스라인과 FREQ-MS의 인식 결과를 비교하여 형태소의 음절길이에 따른 인식 에러의 분포 및 감소

표 3. 형태소의 음절길이에 따른 에러 감소

음절길이	Baseline 에러수	FREQ-MS 에러수	감소 에러수
1	813	693	120
≥2	639	594	45
합 계	1,452	1,287	165

에러 수를 표 3에 나타내었다. 단음절 형태소의 에러가 전체 에러에 대하여 72.7%가 감소하여, 본 논문의 접근 방법이 매우 효과적임을 보여준다.

Ⅶ. 결론

본 논문에서는 한국어 대어휘 방송뉴스 인식을 위한 형태소 기반의 어휘모델과 베이스라인 시스템을 소개하였다. 98.8%의 coverage를 갖는 24k 기본 사전을 선정하였고, 단음절 형태소를 줄일 수 있는 최적의 결합형태소를 생성하였다. 실험을 통해, 단음절 형태소를 52%에서 29%로 감소시켰고, 앵커와 리포터 각각에 대해서 2.68%, 2.21%의 WER를 감소시켰다.

감사의 글

본 연구는 과기부 특정연구개발과제(과제번호 M1-0107-01-003) 지원으로 수행되었으며, 실험에 사용된 한국전자통신연구원의 방송뉴스 음성 DB 사용 허가에 감사드립니다.

참고문헌

[1] Langzhou Chen, L. Lamel, G. Adda and J.L. Gauvain, "Broadcast news trinscription in Mandarin," *Proc. of ICSLP*, 2000.

[2] Ha-Jin Yu, H. Kim, J.S. Choi, J.M. Hong, K.S. Park, J.S. Lee, H.Y. Lee, "Automatic recognition of Korean broadcast news speech," *Proc. of ICSLP*, 1998.

[3] Oh-Wook Kwon and Jun Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," *Speech Communication*, 2002. (in press)

[4] 박영희, 정민화, "대어휘 연속음성 인식을 위한 결합형태소 자동생성," *한국음향학회지*, 21권 4호, pp.407-414, 2002.

[5] 이경님, 정민화, "의사 형태소 단위의 연속 음성 인식," *제 15회 음성통신 및 신호처리 워크샵*, 1998.

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