지식베이스를 이용한 임베디드용 연속음성인식의 어휘 적용률 개선*

김광호(서강대), 임민규(서강대), 김지환(서강대)

<차 례>

<Abstract>

Vocabulary Coverage Improvement for Embedded Continuous Speech Recognition Using Knowledgebase

Kwang-Ho Kim, Minkyu Lim, Ji-Hwan Kim

In this paper, we propose a vocabulary coverage improvement method for embedded continuous speech recognition (CSR) using knowledgebase. A vocabulary in CSR is normally derived from a word frequency list. Therefore, the vocabulary coverage is dependent on a corpus. In the previous research, we presented an improved way of vocabulary generation using part-of-speech (POS) tagged corpus. We analyzed all words paired with 101 among 152 POS tags and decided on a set of words which have to be included in vocabularies of any size. However, for the other 51 POS tags (e.g. nouns, verbs), the vocabulary inclusion of words paired with such POS tags are still based on word frequency counted on a corpus. In this paper, we propose a corpus independent word inclusion method for noun-, verb-, and named entity(NE)-related POS tags using knowledgebase. For noun-related POS tags, we generate synonym groups and analyze their relative importance using Google search. Then, we categorize verbs by lemma and analyze relative importance of each lemma from a pre-analyzed statistic for verbs. We determine the inclusion order of NEs through Google search. The proposed method shows better coverage for the test short message service (SMS) text corpus.

* Keywords: Vocabulary, Coverage, Embedded speech recognition, Knowledgebase.

* 이 논문은 LG전자의 지원을 받아 수행된 연구임.

1. 서 론

음성 언어 처리 기술에서 괄목할 만한 진보가 있었지만, 임베디드 환경에서의 상용화는 메모리 및 계산 용량의 제약으로 인하여 현재까지 가변어 단어 인식 수준에 머물러 있다. 그러나 short message service (SMS) 음성인식의 경우와 같이 어휘의 수와 문장의 형태가 상대적으로 제한된다면, 음향 모델(acoustic model), 어휘(vocabulary), 언어 모델(language model)의 최적화를 통해서 임베디드 환경에서 연속음성인식기의 구현이 가능하다. 적용률(coverage)을 발성한 단어가 어휘내에 있을 확률로 정의하면, 어휘내의 단어수가 많을수록 적용률은 높아지지만, 언어모델 및 탐색 과정에 의해 요구되는 메모리 용량이 커지게 되므로, 효율적 어휘 구성은 임베디드 환경에서의 연속음성인식기 구현에 있어서 중요한 문제이다.

말뭉치가 주어진 경우, 일반적으로 해당 말뭉치에 대해서 최적화된 어휘를 정하는 방법은 각 단어별로 말뭉치 내에서 사용된 빈도를 구하고, 빈도순으로 어휘를 구성하는 방법이다. 이 방법을 이용하여 영어 신문 자료 텍스트 말뭉치에 대해 총 단어수, 총 어휘수, 어휘수에 따른 적용률이 <표 1>과 같이 측정되었다[1]. 영어의 경우 5,000 단어의 어휘로 얻을 수 있는 최대 적용률은 90% 수준으로 측정되었고, 20,000 단어의 어휘 사용 시에는 97% 수준으로 측정되었다.

<표 1> 신문 자료 텍스트 말뭉치에 대한 총 단어수, 총 어휘수, 어휘수에 따른 적용률[1]

언어	총 단어수	총 어휘수	어휘수에 따른 최대 적용률		
			5K	20K	65K
영어	37.2M	165K	90.6%	97.5%	99.6%

빈도수에 기반하여 최적화된 어휘를 결정하기 위해서 필요한 말뭉치의 크기를 결정해야 한다. 어휘수가 고정되어 있고, 말뭉치를 이용하여 말뭉치 내에서 각 단어별 사용된 빈도에 따라 어휘를 구성하는 경우, 말뭉치의 크기에 따른 적용률의 변화에 대한 실험이 수행되었다[2]. 20,000 단어, 40,000 단어, 60,000 단어의 세 가지의 고정된 크기로 어휘를 설정하고, 신문 자료 텍스트 말뭉치의 단어수를 5M개씩 증가시켜가면서 말뭉치내의 각 단어별 사용 빈도에 따라 어휘를 생성하여, 신문 자료 테스트 자료에 대해서 적용률을 구한 결과, 말뭉치의 크기가 커짐에 따라 적용률은 높아졌지만, 말뭉치의 크기가 30M 단어를 넘어가면서 부터는 적용률의 향상이 미미해짐을 보였다. 어휘의 크기에 따라 수렴되는 적용률은 20,000 단어, 40,000 단어, 60,000 단어 각각 약 96%, 98%, 98.5% 수준이다.

SMS는 개인 사생활에 대한 내용이 많기 때문에, 말뭉치 수집에 많은 비용이 필요하게 된다. 따라서, 단어 빈도를 정확히 추정하기에 필요한 [2]에서 기술된 정도의 말뭉치를 수집하는 것은 매우 어려운 일이다. 만약, 뉴스 자료와 같이 말뭉

치 수집이 용이한 다른 도메인에서 수집한 말뭉치로 부터 SMS에 최적화된 단어 빈도가 효과적으로 생성이 가능하다면, 해당 도메인에서 수집한 말뭉치에서의 단어 빈도로 부터 최적의 어휘를 선정할 수 있다. 그러나 도메인에 따라 같은 크기의 어휘에 대한 적용률은 큰 차이를 보인다. <표 1>에서 보인 바와 같이 [1]의 실험 결과에 따르면 신문 자료 텍스트 말뭉치에 대한 5,000 단어 어휘의 적용률은 90% 정도로 측정되지만, British National 말뭉치의 구어체(spoken) 부분에 대한 5,000 단어 어휘의 적용률은 96.9%로 나타났고, 영국과 아일랜드에서 녹음한 대화들을 전사(transcribe)하여 5M개의 단어로 구성된 CANCODE 말뭉치의 경우에는 5,000 단어 어휘의 적용률은 96.1%로 나타났다[3]. 또한 문어체에 대해서도 도메인에 따라 동일 어휘의 적용률이 많은 차이가 발생한다. [4]에 따르면 문어체 말뭉치로부터 많이 사용되는 2,000개의 단어로 구성된 리스트에 대해서, 학술자료, 신문, 잡지, 소설 각각의 적용률은 78.1%, 80.3%, 82.9%, 87.4%로 정리되어 있다.

단어 빈도를 이용한 어휘 결정 방법의 적용률이 말뭉치에 의존적인 단점을 극복하기 위해서 품사 부착 말뭉치를 이용하여 어휘 적용률을 개선한 방법이 제안되었다[5]. 이 방법에서는 Lancaster-Oslo-Bergen (LOB) 말뭉치[6]에서 정의한 152개의 품사에 대해서 품사에 대응되는 단어들의 어휘 포함 여부의 결정방법에 따라 품사들을 분류했다. 단어들의 어휘 포함 여부가 어휘의 크기에 관련이 없는 101개의 품사에 대해서, 각 품사에 해당하는 모든 단어들을 분석하고, 어휘 크기에 관계없이 어휘에 필수적으로 포함되어야 하는 579 단어를 생성했다. 이 어휘 결정 방법은 SMS 음성인식용 어휘 생성에서 구어체 American National Corpus에서의 단어 빈도를 사용하여 생성된 어휘에 비해서 적용률이 29.2%의 상대적 향상(relative improvement)을 보였다.

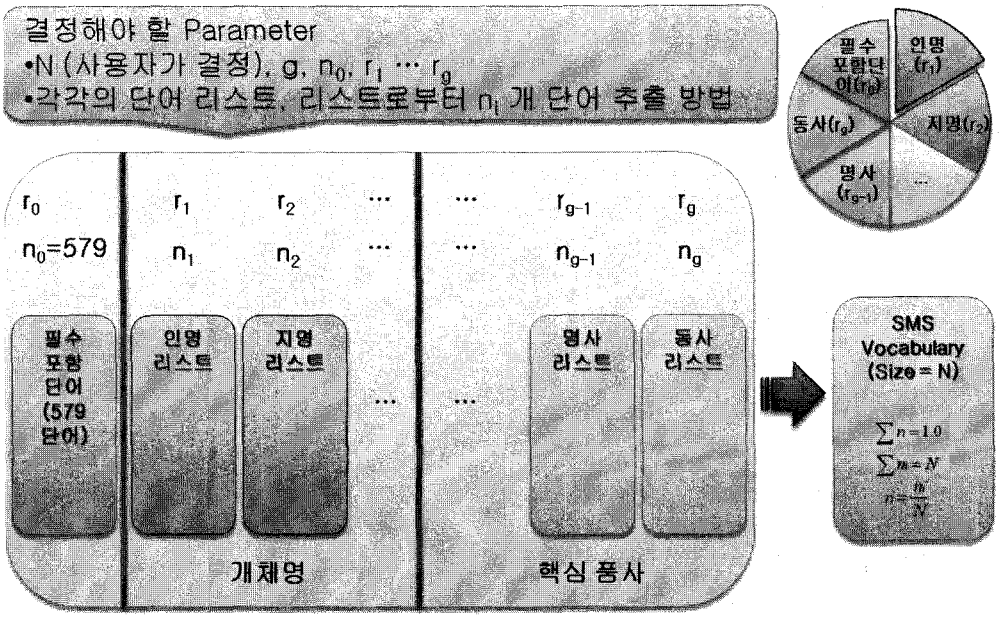
그러나 나머지 51개의 품사에 대해서는 해당 품사를 가지는 단어들의 수가 무제한이기 때문에 (예: 명사, 동사, 개체명), 해당 품사의 단어들의 어휘 반영 여부는 여전히 말뭉치의 단어 빈도에 의존적이다. 이 경우 빈도순으로 어휘를 구성하는 경우와 마찬가지로 유사 단어 중 일부는 어휘에 포함되지만, 나머지 일부는 말뭉치에서 측정한 빈도수가 작아서 어휘에서 제외되는 경우가 발생하여, 제외된 단어는 어떤 경우에도 음성인식이 되지 않는 단점을 가진다. 예를 들어 'Monday'는 어휘에 있는데, 말뭉치에서 'Tuesday'에 대한 빈도가 낮아 어휘에서 'Tuesday'가 제외 된다면 음성인식기의 사용자는 항상 'Tuesday'의 발성 시 음향모델과 언어모델의 정확도와는 관계없이 항상 오인식 결과를 얻게 된다.

본 논문에서는 기존에 제안했던 품사 부착 말뭉치를 이용한 어휘 구성 방법[5]을 기반으로 지식베이스를 이용하여 명사, 동사, 개체명에 관련된 품사에 대해서 말뭉치에 독립적인 어휘 생성 방법을 제안하고, SMS용 연속음성인식에 대해 제안한 방법의 타당성을 평가한다. 2장에서는 기존에 제안한 품사 부착 말뭉치를 이용한 어휘 적용률 개선 방법에 대해서 기술한다. 3장에서는 본 논문에서 제안하는

지식베이스를 이용한 어휘 적용률 개선 방법을 설명한다. 4장에서는 제안한 방법으로 10,000단어 및 15,000단어 어휘를 구성하고, 미국에서 휴대폰을 이용하여 사용자가 직접 작성한 SMS를 대상으로 제안한 방법을 평가한다. 끝으로 5장에서는 결론을 맺는다.

2. 관련 연구

본 장에서는 본 논문에서 제안하는 방법의 기반이 되는 품사 부착 말뭉치를 이용한 어휘 적용률 개선 방법에 대해서 간략히 기술한다. 본 장에서 기술하는 방법의 세부적인 내용은 [5]에 기술되어 있다. <그림 1>은 어휘가 구성되는 과정을 보여주고 있다. 사용자가 어휘의 크기 N을 결정하면, 기 분류되어 있는 n_0개의 어휘 필수 포함 단어가 어휘에 포함된다. 그 후 개체명 및 핵심 품사로 분류되어 있는 g개의 품사들로부터 각각 n_i개의 단어를 선택해서 어휘에 포함하여 어휘를 구성한다.



<그림 1> 어휘 구성 과정

필수 포함 n_0개의 단어 선정 및 g개의 개체명 및 핵심 품사에 속하는 품사 분류 결정은 다음과 같다. LOB 말뭉치의 152개의 품사에 대해서 품사에 대응되는 단어들의 어휘 포함 여부의 결정 방법에 따라 품사들을 크게 4개의 그룹들로 분류 했다. <표 2>는 이들 4개 그룹에 대한 설명이다. 그룹 1에는 품사에 해당하는 모든 단어들이 어휘의 크기에 관계없이 어휘에 포함되어야 하는 be동사, 조동사, 접속사 등이 포함되었다. 그룹 2에는 문법상 중요한 품사들이고 품사별로 해당되는 단어들이 많지 않아서, 해당되는 모든 단어에 대해 어휘의 크기에 관계없이 어휘에 포함되어야하는 품사로는 기수 (태그명: CD, 예: one, two, hundred), 서수 (태그명: OD, 예: first, second) 등이 있다. 같은 방법으로 그룹 1과 그룹 2로 분류되는 품사들에 해당 하는 모든 단어들에 대해서 음성인식 전문가가 어휘 포함 여부를 판별하고, 어휘수에 관계없이 항상 어휘에 포함되는 필수 단어를 분류했다. 분류 결과 579단어가 어휘 크기에 관계없이 모든 어휘에 필수적으로 포함되어야 하는 것으로 조사되었다. 따라서 n_0는 579로 결정된다.

<표 2> 어휘 포함 여부의 결정방법에 따른 품사 분류 및 예시
(예시의 형태: 품사태그-해당단어)

품사 분류	설명
그룹 1 (품사수: 76개)	품사에 해당하는 모든 단어들이 어휘의 크기에 관계없이 어휘에 포함됨 (예: BE-be, BEM-am, EX-there, HV-have, HVD-had,'d)
그룹 2 (품사수: 11개)	문법상 중요한 품사들이고 품사별로 해당되는 단어들이 많지 않아서, 해당되는 모든 단어에 대해 어휘의 크기에 관계없이 어휘에 포함되는 여부를 전문가가 결정함 (예: CD-two,three,hundred,10,45,1000,..., CD-10th,15th,tenth,twenty-first,...)
그룹 3 (품사수: 51개)	개체명 및 핵심 품사(명사, 형용사, 동사, 부사)로서 해당되는 단어들이 많은 품사
그룹 4 (품사수: 14개)	품사에 해당하는 단어들이 어휘에 포함될 가능성이 전혀 없는 경우 (예: 특수 기호)

그룹 3에는 개체명 및 핵심품사(명사, 형용사, 동사, 부사)로서 해당되는 단어들이 많은 품사들은 크게 개체명과 명사군, 동사군, 형용사군, 부사군이다. 그룹 3으로 분류되는 품사에 해당되는 단어의 개수는 각 품사별로 무제한이다. 이들 단어들은 SMS에 사용빈도가 높은 단어들도 있고, 사용될 빈도가 거의 없는 단어들도 있기 때문에 어휘의 크기 및 응용 도메인에 따라서 이들 단어들의 어휘 포함 여부가 달라지게 된다. 개체명 및 핵심 품사들에 해당되는 품사의 수는 총 51이며,

이 값이 g가 된다.

그룹 3으로부터 어휘에 추가하는 단어들의 개수는 전체 어휘의 크기에서 그룹 1과 그룹 2로부터 선정한 필수 포함 단어수(579개)를 뺀 나머지로 결정된다. 그 후, 개체명 리스트와 핵심 품사별 단어 리스트, 품사별 어휘 구성 비율(r_i)은 LOB 말뭉치에서의 빈도 분석을 통해서 결정한 값을 사용한다.

3. 지식베이스를 이용한 어휘 적용률 개선

2장에서 기술한 품사 부착 말뭉치를 이용한 어휘 구성 방법은 그룹 3에 속한 51개의 품사에 대응되는 단어들의 어휘 반영 여부에 대해서 여전히 말뭉치의 단어 빈도에 의존적인 단점을 가지고 있다. 본 장에서는 2장에서 기술한 방법을 기반으로, 지식베이스를 이용하여 명사, 동사, 개체명에 관련된 품사에 대해서 말뭉치에 독립적인 어휘 생성 방법에 대해서 기술한다.

그룹3에 해당하는 품사의 태그는 51개로, 크게 명사 관련 품사들과 개체명 관련 품사들, 동사 관련 품사들로 나누어진다. 3.1절에서는 명사 관련 품사들에 해당되는 단어들의 어휘 반영 방법에 대해서 기술한다. 명사 관련 품사들에 대해서는 Wordnet을 이용하여 등위어(synonym)를 생성하고, 이들 등위어 그룹간의 상대적 중요도를 구글검색을 이용하여 결정한다. 3.2절에서는 동사 관련 품사에 대해 어휘 반영 방법을 설명한다. 동사들을 시제 변화와 인칭 변화를 고려하여 표제어(lemma)에 따라 분류한다. 표제어에 따른 동사군의 어휘 반영 여부는 동사에 대해 분석된 통계치를 활용하여 결정한다. 3.3절에서는 구글검색을 통하여 개체명에 대한 어휘 반영 여부를 결정하는 방법에 대해 기술한다.

3.1. WordNet을 이용한 명사 등위어 생성

Wordnet[7]은 영어에 대한 단어간 연결관계를 나타내는 의미 어휘 목록이다. Wordnet에서 사용하는 품사는 명사, 동사, 형용사, 부사이고, 각 품사에 대해서 의미적 집합으로 표현하고 있다. Wordnet에서는 의미 관계를 각 품사에 대해서 다르게 분류하고 있다. 본 논문에서 이용하고자 하는 명사의 경우, 상위어, 하위어, 등위어, 전체어, 부분어의 의미 관계를 제공한다. 등위어 집합은 개념적, 의미적 유사 단어군으로 동일한 상위어를 가지는 집합이다. Wordnet의 각 품사에 대한 총 단어수와 동의어그룹 수는 <표 3>과 같다.

명사 관련 품사들에 해당되는 단어들의 어휘 반영 방법은 다음과 같다. 우선, Wordnet을 이용하여 명사에 대해서 명사 등위어 단어군을 생성한다. 예를 들어, Wordnet에 'pear'라는 명사에 대한 등위어 단어군을 요청할 경우 Wordnet은 'pear'

<표 3> Wordnet의 총 단어수, 등위어 집합 수

품사	어휘수	등위어 집합 수
명사	117,798	82,115
동사	11,529	13,767
형용사	21,479	18,156
부사	4,481	3,621
합계	155,287	117,659

의 상위어라 할 수 있는 'edible fruit'에 하위어로 속하는 'apple', 'berry', 'pineapple', 'banana' 등과 같이 'pear'가 사용된 문장에서 'pear' 대신 치환이 가능한 단어를 제공한다. 본 논문에서는 명사에 대해서 기존의 방법에서 사용한 명사에 해당되는 단어들을 Wordnet을 이용하여 각 단어별 등위어 단어군(단어에 따라 복수개가 생성될 수 있음)을 생성한다.

기존의 방법에서 사용한 명사에 해당하는 단어들을 Wordnet을 통하여 등위어 단어군을 생성한다. Wordnet을 이용하여 특정 단어에 해당하는 등위어 단어군을 생성하고, 등위어 단어군의 그룹 이름을 상위어로 설정했으며, 동일한 상위어를 가지는 하위어는 등위어 단어군으로 포함된다. 등위어 단어군에 포함된 단어의 어휘 중요도는 지식 베이스를 통해서 결정한다. 지식 베이스로부터 어휘의 중요도를 결정하기 위해서는 지식 베이스의 어휘에 대해 질의(query)한 후에 얻는 결과로부터 원하는 중요도를 추출하는 작업이 요구된다. 질의문은 중요도를 얻고자 하는 어휘로 구성하며, 질의 결과로부터 어휘의 빈도수를 추출한다. 본 논문에서 지식 베이스를 이용한 중요도 결정은, 구글 검색을 이용하여 나오는 문서의 빈도수로 결정한다. 구글 검색을 이용하여 keyword로 어휘를 넣으면, 그 keyword와 관련된 문서와 총 문서의 수를 생성한다. 구글 검색은 수백억개의 문서를 보유하고 있어서 어휘에 대한 빈도수 자체가 매우 높기 때문에 어휘에 따른 중요도를 구할 수 있다.

기존의 방법에서 결정한 명사들에 대해서 구글 검색을 이용한 초기 중요도를 설정한다. 높은 중요도를 가진 명사 단어로부터 시작하여 단어에 대한 Wordnet의 등위어 단어군을 살펴보면서, 기존의 방법에서 결정한 명사들의 중요도와 Wordnet의 등위어 단어군으로부터 나온 단어의 중요도를 비교 분석한다. 기존의 방법에서 결정한 명사에 해당하는 단어의 중요도보다 Wordnet의 등위어 단어군의 단어가 더 높은 중요도를 가진다면, 해당 단어들을 교체한다.

3.2. 표제어별 동사간 중요도 결정

동사는 현재형, 과거형, 완료형 등의 시제와 1인칭, 2인칭, 3인칭에 따라서 그

형태가 변한다. 기존의 품사부착 말뭉치를 이용한 방법은 동사에 대한 어휘 포함 여부를 말뭉치에서의 빈도수로 결정하기 때문에 동사의 변화형에 대한 단어의 어휘 포함 여부가 일정하지 않다. SMS 특성상 한 단어에 대한 현재형(예: run)은 인식 되는 반면 동일한 단어의 과거형(예: ran)이 인식되지 않는 경우 일관성이 떨어지기 때문에 사용자는 혼란을 겪게 된다. 본 장에서는 위의 문제점을 해결하기 위해서 표제어를 도입하여 동사 원형과 변화형을 어휘에 함께 반영하는 방법을 제안한다. 표제어는 사전에서 실린 단어를 기준으로 시제와 인칭, 단수, 복수 등의 동사의 변화형을 아울러 일컫는 말이다.

명사의 경우 Wordnet을 통해 등위어를 생성하고, 중요도를 결정하여 등위어 단위로 어휘에 포함시키는 경우, 몇몇 등위어 집합은 크기가 매우 커서 등위어 집합 내에서도 어휘 포함 여부를 결정해야 하는 경우가 있다. 하지만 동사의 경우 는 원형과 변화형을 모두 포함하더라도 표제어 단위가 크지 않아 어휘를 구성하는 데에 문제가 없다. 따라서 동사에 대한 어휘를 구성하기 위해 본 연구에서는 동사를 표제어별로 그룹지어 분석된 [8]에서의 통계치를 사용하였다. 통계치는 British National 말뭉치에 포함되어 있는 어휘에 대한 분석으로 단어의 원형과 변화형을 표제어 단위로 해당 말뭉치상의 빈도수를 정리해 놓은 통계치 이다. 따라서 동사에 대한 중요도는 표제어의 빈도수를 기준으로 결정하고 표제어가 어휘에 포함될 경우 그에 해당하는 변화형을 최종 어휘에 함께 포함하도록 한다.

3.3. 지식베이스로부터 개체명간 중요도 결정

고유명사는 음성 자료 검색에 있어서 중요한 정보를 가지고 있기 때문에, 어휘 구성에 있어서 고유명사 처리에 대해 더 많은 주의가 요구된다[9]. 그러나 이러한 고유 명사들은 종종 어휘외 단어 (out-of-vocabulary: OOV)가 되고, 따라서 음성인식의 주요한 에러 원인이 된다. [9]에 따르면 65,000 단어의 어휘에서 약 28%의 단어들은 고유명사 또는 약어들이었다.

인명 리스트는 1990년 미국 인구 센서스에서 미국인의 인명에 대해 성, 남성 이름, 여성 이름의 사용 빈도 자료[10]를 이용했다. 인명 이외의 개체명에 대해서는 NYU OAK 시스템[11]에서 사용한 개체명 리스트내의 단어들에 대해서 구글 검색을 통해서 개체명 간의 상대적 중요도를 결정했다.

4. 실 험

테스트 자료는 연구 지원기관으로부터 제공받았다. 이 테스트 자료는 미국현지에서 휴대폰 사용자들이 휴대폰을 이용하여 수신 또는 발신한 SMS를 전사

(transcribe)하여 수집했다. <표 4>는 본 논문에서 사용한 테스트 자료(이하 SMSText_US)의 구성을 보여주고 있다. 총 50명의 휴대폰 사용자가 참여했고, 일인당 20~30개의 SMS를 수집했다. 총 문장수, 총 단어수, 총 어휘수는 각각 2,704개, 13,699개, 2,395개로 집계되었다.

<표 4> 테스트 자료의 구성

자료명	총 문장수	총 단어수	총 어휘수	수집 방법
SMSText_US	2,704	13,699	2,395	50명의 휴대폰에 있는 SMS (20~30개/인)를 전사해서 수집

제안한 방법의 검증을 위해 <표 5>와 같이 6개의 어휘를 구성했다. SMS는 텍스트이지만 구어체 스타일의 문장으로 구성되므로 기존의 어휘는 대표적인 구어체 말뭉치인 American National Spoken 말뭉치[12]로부터 단어 빈도에 따라 생성된 어휘(Voc_ANC_Spoken)와 기존 연구[5]에서 제안한 방법인 품사 부착 말뭉치를 이용하여 생성한 어휘(Voc_POSBased), 본 논문에서 제안한 지식베이스를 이용한 임베디드용 연속 음성인식을 위한 어휘(Voc_KnBased)에 대해서 어휘수가 10K, 15K인 어휘를 생성했다.

<표 5> 제안한 방법으로 구성한 어휘 및 기존 방법으로 구성한 어휘에 대한 설명

어휘명	어휘수	어휘 구성 방법
Voc_ANC_Spoken	10K, 15K	American National Spoken 말뭉치에서 가장 많이 사용된 단어로 구성
Voc_POSBased	10K, 15K	기존 연구[5]에서 제안한 방법을 품사 부착 말뭉치를 이용하여 구성
Voc_KnBased	10K, 15K	본 연구에서 제안한 방법인 지식베이스를 이용하여 구성

4.1. 어휘에 따른 적용률 측정 결과

각각의 어휘를 이용하여 테스트 자료인 SMSText_US에 대해서 적용률을 측정했다. <표 6>은 어휘에 대한 적용률을 보여준다. Voc_ANC_Spoken의 적용률은 10K는 95.09%, 15K는 96.13%가 측정되었고, 품사 부착 말뭉치를 이용한 Voc_POSBased의 적용률은 10K가 96.50%, 15K가 97.42%로 Voc_ANC_Spoken보다 높은 적용률을 보이고 있다. 지식베이스를 이용하여 생성한 Voc_KnBased의 적용률은 10K가 96.88%, 15K가 97.84%로 세 종류의 어휘 중 가장 높은 적용률을 보이

고 있다. 따라서 제안한 방법이 기존의 어휘 구성 방법보다 더 유용한 방법이라는 것을 알 수 있다.

<표 6> 어휘별 SMS 텍스트 자료(SMSText_US)에 대한 적용률

어휘명	적용률(%)	
	어휘 크기 = 10K	어휘 크기 = 15K
Voc_ANC_Spoken	95.09	96.13
Voc_POSBased	96.50	97.42
Voc_KnBased	96.88	97.84

기존의 어휘는 말뭉치에서 등장하는 단어의 빈도수가 해당 단어의 어휘 포함 여부를 결정하기 때문에 어떤 말뭉치를 사용하는가에 따라 단어의 어휘 포함 여부가 달라져서 말뭉치에 의존적이라 할 수 있었다. 본 연구에서 제안한 방법은 말뭉치에 단어의 어휘 포함 여부를 말뭉치에 독립적으로 결정하여 안정적인 적용률을 얻을 수 있다. 또한 어휘를 구성하는 단어들이 품사별로 구분 가능하고, 각 품사별로 다른 어휘 포함 방법을 적용하여 어휘의 전체 단어 수에 따른 체계적인 어휘 구성이 가능하다. 추가로 음성인식기 구현에서 체계적인 어휘 트리(lexical tree)를 구현하는 것이 가능하다. 예를 들어 인명의 소유격('s)에 대해서 이에 대한 음성인식이 가능하게 하기 위해서는 어휘내의 모든 단어에 대해서 's를 붙인 단어를 추가하여 결과적으로 전체 어휘의 수가 두배가 된다. 반면 제안한 방법에서는 인명에 대해서만 's를 붙인 단어를 추가하면 된다.

5. 결론 및 향후 연구

본 논문에서는 기존에 제안했던 품사 부착 말뭉치를 이용한 어휘 구성 방법[5] 기반으로 지식베이스를 이용하여 명사, 동사, 개체명에 관련된 품사에 대해서 말뭉치에 독립적인 어휘 생성 방법을 제안하고, SMS용 연속음성인식에 대해 제안한 방법의 타당성을 평가하였다. 제안한 방법에서는 명사에 대해서 Wordnet을 이용하여 등위어(synonym)를 생성하고, 이들 등위어 그룹간의 상대적 중요도를 구글검색을 이용하여 결정하였다. 동사의 경우 시제 변화와 인칭 변화를 고려하여 표제어(lemma)에 따라 분류하였다. 개체명에 대해서는 구글 검색을 통해서 어휘 포함 여부를 결정하였다.

본 논문에서 제안한 방법으로 구성한 어휘의 적용률은 어휘크기가 15K인 경우 97.84%로 나타났다. 이는 기존의 어휘를 적용했을 때 얻은 적용률(97.42%와 96.13%)보다 높은 수치로써, 제안한 방법이 기존의 어휘 구성 방법보다 더 유용한 방법이라는 것을 알 수 있다.

현재 품사간 어휘 구성비율은 여전히 말뭉치에 의존적이다. 따라서 추후 연구에서는 이들 비율을 말뭉치에 독립적으로 만들고, 어휘 편집을 쉽고 직관적으로 수행할 수 있는 컴퓨팅 환경을 구현 및 보급하고자 한다.

참 고 문 헌

[1] M. Adda-Decker, L. Lamel, "The use of lexica in automatic speech recognition", *Lexicon Development for Speech and Language Processing*, F. van Eynde, D. Gibbon (Eds.), Kluwer Academic, pp. 235-266, 2000.

[2] R. Rosenfeld, "Optimizing lexical and n-gram coverage via judicious use of linguistic data", *Proc. Eurospeech*, pp. 1763-1766, 1995.

[3] S. Adolphs, N. Shemitt, "Lexical coverage of spoken discourse", *Applied Linguistics*, Vol. 24, No. 4, pp. 425-438, 2003.

[4] P. Nation, R. Waring, "Vocabulary size, text coverage and word lists", *Vocabulary: Description, Acquisition and Pedagogy*, N. Schmitt, M. McCarthy (Eds.), Cambridge University Press, pp. 6-19, 1997.

[5] 임민규, 김광호, 김지환, "품사 부착 말뭉치를 이용한 임베디드용 연속음성인식의 어휘 적용률 개선", *말소리*, 제67호, pp. 181-194, 2008.

[6] R. Garside, G. Leech, T. Varadi, "Manual of information for the Lancaster parsed corpus", available at http://khnt.hit.uib.no/icame/manuals/LPC/LPC.PDF.

[7] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[8] G. Leech, P. Rayson, A. Wilson, *Word Frequencies in Written and Spoken English: Based on the British National Corpus*, Pearson ESL, 2001.

[9] R. Ordelman, A. van Hessen, F. de Jong, "Lexicon optimization for Dutch speech recognition in spoken document retrieval", *Proc. Eurospeech*, pp. 1085-1088, 2001.

[10] "Frequently occurring first names and surnames from the 1990 census", available at http://www.census.gov/genealogy/names.

[11] S. Sekine, "OAK system", available at http://nlp.cs.nyu.edu/oak.

[12] R. Reppen, N. Ide, "The American National Corpus: overall goals and the first release", *Journal of English Linguistics*, Vol. 32, No. 2, pp. 105-113, 2004.

접수일자: 2008년 11월 24일
게재결정: 2008년 12월 26일

▶ 김광호(Kwang-Ho Kim)
주소: 121-742 서울시 마포구 신수동 1번지
소속: 서강대학교 컴퓨터공학과
전화: 02) 715-2715
E-mail: kimkwangho@sogang.ac.kr

▶ 임민규(Minkyu Lim)
주소: 121-742 서울시 마포구 신수동 1번지
소속: 서강대학교 컴퓨터공학과
전화: 02) 715-2715
E-mail: lmkhi@sogang.ac.kr

▶ 김지환(Ji-Hwan Kim) : 교신저자
주소: 121-742 서울시 마포구 신수동 1번지
소속: 서강대학교 컴퓨터공학과
전화: 02) 705-8924
E-mail: kimjihwan@sogang.ac.kr