論文
5-7-6

# 한국어최적상호명코퍼스설계에관한연구
# (AStudyontheoptimaltextcorpusforcompanynames)

이선정(Sun-Jung Lee)[1]

## 요약

본 논문에서는 114 안내시스템에 저장되어있는 서로 중복되어 있지 않는 1,566,943개의 상호명 코퍼스에서 이 코퍼스의 특징을 가장 잘 표현 해 줄 수 있는 최적 코퍼스를 설계하였다. 최적 코퍼스를 구하기 위해 두 단계의 방식을 택한다. 일 단계는 기본코퍼스에 존재하는 트라이폰이 모두 나타내는 최소의 단어 셋을 구하는 최적 음소균형 코퍼스 셋이고, 다음 단계는 기본코퍼스에 존재하는 트라이폰의 빈번도를 고려하는 최소의 단어 셋을 구하는 음소 분포코퍼스 셋을 설계하였다. 실험 결과 최적 음소 균형 셋으로 8,699 단어가 선정되었으며 최적 음소 분포 균형 셋으로 16,783 단어가 선정되었다. 이러한 최적 코퍼스는 음성 및 합성 시스템을 위한 음성데이터베이스를 구축 할 때 이용된다.

## Abstract

In this paper, we obtain an optimal corpus that can represent its characteristics very well from the baseline corpus which consists of unique 1,566,943 names among company names in a directory assistance service (114). Two kinds of optimal solutions are considered to obtain the optimal corpus. The first solution is to find phonetically balanced corpus (PBC), which are the minimum set including all possible triphones in the baseline corpus. The second solution is to find the phonetically distributed corpus (PDC), which is a minimum set representing the frequency characteristics of triphones in the baseline corpus. We can obtain 8,699 words as the PBC and 16,783 words( similarity measure R = 0.92) as PDC, respectively. These corpora can be used for the development of speech recognition and speech synthesis.

논문접수 : 2004. 7. 3.
심사완료 : 2004. 7. 25.

1) 정회원 : 시립인천전문대학 전자계산학과 부교수

## 1.서론

현재 음성인식기술을 이용한 응용 서비스 중에서 매우 어려운 분야로 평가되고 있는 분야 중 하나로 사람 이름을 인식하거나 회사 상호명을 인식하는 분야가 있다. 특히 상호명은 증권시장에 상장되어 있는 약 1000 여 개의 상호명 이외에도 매우 다양한 기업체가 존재한다. 현재 서울 지역을 중심으로 114 안내 시스템에서 제공되고 있는 상호명은 약 백 오십만 개정도로 이루어져 있다고 한다. 특히 상호명의 경우는 최근 다양한 외래어를 상호로 많이 채택하기 때문에 다양한 음성 및 음운 현상의 연구를 필요로 하고 있다.

최근에는 음성언어처리 기술이 발전함에 따라 대용량 단어의 음성인식, 음성합성기술에 대한 연구가 진행되고 있다. 이러한 대용량 단어가 인식되기 위해서는 훈련 데이터가 필요하고 가능한 다양한 사람들의 발음도 포함되어야 한다. 그러나 150 만개의 상호명을 모두 녹음 할 수는 없으므로 가능한 적은 상호명 개수로 150 만개의 음성 및 언어적인 특징이 포함되어 있어야 한다. 이러한 과정을 거쳐 구해진 상호명 리스트를 최적 코퍼스라고 부른다.

대표적인 최적 코퍼스 설계에 사용되는 알고리즘이 Greedy 알고리즘이다[1]. 이 알고리즘은 대상 코퍼스에 존재하는 기본 유니트가 모두 포함되도록 단어를 선정하는 방식이다. 이 알고리즘에 확률 값 및 빈번도를 추가하는다양한 알고리즘이 개발되었다[2][3][4]. 최근에는 음운환경뿐만 아니라 억양의 특징도 포함되도록 최적 코퍼스를 구하는 방안이 제안되고 있다[5]. 또한 한국어 이름의 특징을 가장 잘 나타내어 줄 수 있는 이름 텍스트 코퍼스에 대해서 연구가 수행된 적이 있다[6].

본 논문은 한국어 상호의 특징을 가장 잘 나타내 줄 수 있는 상호 텍스트 코퍼스를 구하는 방안에 대한 것이다. 이를 위해 먼저114 상호 데이터베이스에서 상호에 사용되는 기본 유니트가 모두 포함되도록 설계하는 방식과 기본 유니트의 빈도를 고려하는 음소 분포셋을 구하는 방식을 사용한다[1]. 먼저 2장에서 상호명 구문 코퍼스에 대해 설명하고 3장에서는 최적 상호명 코퍼스 설계 알고리즘을 설명한다. 4장에서는 상호명 구문 코퍼스를 이용하여 구한 최적 상호명 코퍼스 실험결과에 대해서 논하고 5장에서 결론을 맺는다.

## 2.상호명구문코퍼스

상호명 구문 코퍼스란 국내에서 사용하고 있는 상호명을 문자로 전사한 것을 의미한다. 그러나 이 상호명은 항상 끊임없이 변하므로 지속적인 수집 및 관리가 필요하다. 더구나 상호명의 수는 매우 많으므로 이러한 상호명중에서 음성, 음운학 정보를 가장 잘 나타내어 주는 상호명을 구하는 것이 필수적이다. 본 연구에서는 서울지역을 중심으로 114안내 시스템에 등록되어 있으면서 동일한 상호명을 배제하고 난 이후의 코퍼스를 기본 코퍼스로 선정한다. 기본 코퍼스를 분석하기 위하여 먼저 트라이폰을 구하였다. 트라이폰이란 각각의 음소는 좌, 우측의 음소에 따라 음가가 달라지므로 음성인식, 음성합성 연구의 기본 단위로 사용되고 있다. 그러므로 상호 구문 코퍼스를 발음형태의 음소로 바꾸고 그 후에 상호코퍼스 내에 존재하는 트라이폰을 가장 잘 나타내어 주는 상호명을 구하면 된다.

서울 지역의 114 안내 서비스에 등록된 동일하지 않는 상호명은 약 156만 명이 된다.이것을 발음형태로 전사하면 총 트라이폰 수는 2천5백만 개가 되며 동일하지 않는 트라이폰 개수는 2만 1천개가 된다. 트라이폰을 구할 때 사용되는 기본 음소개수는 64개로 하였다. <표. 1>에서는 114 상호 코퍼스의 분석 결과가 나타나 있다. <표 1>에서 보면 1,566,943개의 상호명을 구성하고 있는 트라이폰 수는 25,009,147개가 있으며 21,145개의 트라이폰으로 1,566,943개의 상호명을 표현할 수 있다는 것을 알 수 있다. 그러므로 이 트라이폰을 모두 포함하며

또한 상호명 코퍼스내의 트라이폰 빈번도와유사한 비율로 존재하는 트라이폰이 있는 상호명 코퍼스 셋을 구하는 것이 매우 중요하다. 이러한 셋을 구하면 음성합성, 음성인식의 훈련을 위해 모든 상호 명의음성 데이터베이스 대신에 이코퍼스를 사용할 수 있으므로 매우 빠르고 쉽게 개발할 수 있게 된다.

| 상호명수 | 1,566,943 |
|---|---|
| 트라이폰 수 | 25,009,147 |
| 트라이폰 종류 | 21,145 |

<표. 1> 114 상호명 코퍼스
<Table 1> 114 Company Names Corpus

## 3. 최적상호명코퍼스설계알고리즘

최적 코퍼스란 음운현상을 표현해주는 최소 코퍼스를 말한다. 앞 장에서 설명한 114 상호 코퍼스는 157만 개 정도의 다른 상호명으로 구성되어 있어 처리하기가 어렵다. 최적 코퍼스란 157만 개 상호명의 음성, 음운 특징을 나타내는 최소 코퍼스 셋을 구하는 것이다.
최소 코퍼스 셋을 구하는 방법은 두 가지가 있다. 먼저 기본적인 음운 특징이 포함되는 최소 단어 셋을 구하는 방식이 있으며 두 번째는157만 코퍼스 내에 존재하는 음운 특징과 유사한 확률분포를 갖는 최소 단어 셋을 구하는 방식이다. 첫 번째 방식은 음소균형 코퍼스(phonetically balanced corpus) 셋이라고 말하고 두 번째 방식은 음소 분포 코퍼스(phonetically balanced corpus) 셋이라고 한다. 두 번째 방식에 의한 코퍼스는 첫 번째 방식의 결과를 가지고 알고리즘을 적용하여 구한다.

### 3.1 음소균형상호명코퍼스

음소 균형 코퍼스 셋을 구하기 위해서는 상호명 선정을 위한 음운 특징의 조건이 필요하다. 음성인식을 위한 음소 균형 코퍼스 셋을 구하기 위한 음운 특징의 조건은 다음과 같다.

조건 1) 음소 균형 코퍼스를 구성하고 있는 상호명 코퍼스에는 157만 개의 상호명에 존재하는 모든 트라이폰인 21,145개가 존재해야 한다.
조건 2) 새로운 상호명을 추가할 때 이미 선정된 상호명에 존재하지 않는 트라이폰 수가 많은 것을 우선적으로 선정한다.
조건 3) 157만 개 상호 코퍼스에 존재하는 21,145개의 트라이폰 중 빈번도가 적은 트라이폰이 포함된 상호명이 가급적 먼저 선정되도록 한다.

위의 조건에 맞는 상호명을 선정하는 상세한 흐름도는 이전 논문에 기술되어 있다[6]. 이 방식의 핵심은 21,145개의 트라이폰을 모두 포함하고 있는 최소의 상호명 개수를 선정하는 것이다. 따라서 조건 2)를 표현하기 위하여 기존에 존재하지 않는 트라이폰 수를 나타내는 값을 스코어로 표현하고 또한 조건 3)을 표현하기 위하여 1/ (빈번도 수)를 스코어로 정의하여 " 조건 2)의 스코어 + 조건 3)의 스코어 " 값을 기준으로 가장 높은 값을 갖는 단어를 선정한다. 이러한 선정을 통하여 21,145개의 트라이폰이 모두 구해지면 새로운 단어의 선정을 포기한다. 이러한 방식으로 구한 단어 리스트가 음소 균형 상호명 코퍼스이다. 이 코퍼스의 특징은 잘 사용되지 않은 트라이폰을 갖고 있는 단어가 우선적으로 선정되므로 외국어, 특이한 단어로 구성되어 있다는 특징이 있다.

### 3.2 음소분포상호명코퍼스

음소 분포 상호명 코퍼스 셋은 상호명에 존재하는 트라이폰의 분포도를 고려하여 선정된 최소 단어리스트로서 157만개의 상호명에 존재하는 트라이폰의 분포도와 음소 분포 상호명 코퍼스 내에 존재하는 트라이폰의 분포도가 유사하여야 하며 또한 사용되는 트라이폰의 개수도 동일 해야 한다. 이러한 코퍼스를 구하기 위해 앞 절에서 구한 음소 균형 코퍼스 셋을 초기값으로 하며 대용량 코퍼스에 속해 있는트라이폰의 분포도와 유사한 분포도를 갖도록 알고리

즘이 결정되어야 한다. 이러한 코퍼스 셋을 구하기 위한 음성, 음운 특징의 조건은 다음과 같다.

조건 1) 157만개의 상호명 코퍼스에 존재하는 빈번도가 높으면서 음소 균형 상호명 코퍼스 내에서는 빈번도가 낮은 트라이폰이 우선적으로 선택되어야 한다.
조건 2) 조건 1)에 의해 음소 분포 코퍼스 셋에 포함된 트라이폰은 다음 상호명 단어를 선택할 때 적게 선정이 되도록 트라이폰의 스코아 값이 변경되어야 한다.
조건 3) 조건 2)에서 스코아 값이 변경될 경우에 157만개의 상호 코퍼스에 존재하는 빈번도가 높은 트라이폰은 적게 변경되어야 한다.

위 조건에 따라 상호명이 선정될 경우 157만개의 상호명에 존재하는 트라이폰 빈번도와상기 알고리즘에 의해 구해진 음소 분포 코퍼스 내에 존재하는 트라이폰의 빈번도와의 유사도가 먼저 정의되어야 한다. 매 단어를 추가 할 경우 유사도를 검사해서향상되지 않으면 상기 조건에 의해 구해진 스코아 중다음 스코아 값을 갖는 단어를 선정한다. 이때 사용되는 유사도(R)는 다음과 같이 정의된다.

$$R = \frac{\vec{v}_c \bullet \vec{v}_d}{|\vec{v}_c| \bullet |\vec{v}_d|} = \cos(\theta) \quad (1)$$

$$\vec{v}_c = [n_c(1), \cdots, n_c(L)] \quad (2)$$

$$\vec{v}_d = [n_d(1), \cdots, n_d(L)] \quad (3)$$

여기서 $n_c(i)$ 은 114 상호명 코퍼스에 존재하는 $i$ 번째 트라이폰의 빈번도이며 $n_d(i)$ 는 음소 균형상호명 코퍼스에 존재하는 $i$ 번째 트라이폰의 빈번도이다. 또한 $L$ 은 상호명 코퍼스에 존재하는 중복되지 않는 트라이폰 종류이다. 위 식은 현 코퍼스에 존재하는 트라이폰의 빈번도가 대용량 코퍼스에 존재하는 트라이폰의 빈번도와 얼마나 유사한지를 나타내는 척도로서 유사도 값이 크면 클수록 음소 균형 상호명 코퍼스에 존재하는 상호명 개수가 늘어나며 트라이폰 분포도도 157만개에 존재하는 트라이폰 분포도와 비슷하게 된다.

## 4. 실험결과

### 4.1 음소균형상호명코퍼스(PBC)

3.1 장에서 설명한 방식으로 157만개의 114 상호명 코퍼스로부터 구한 음소 균형 상호명 코퍼스를 구한 결과가 <표 2>에 나타나 있다. <표 2>를 보면 157만개의 상호명 코퍼스에 존재하는 모든 트라이폰을 포함하는 최소 음소 균형 코퍼스는 8,699개의 상호명으로 구성되어 있으며 이때 사용되는 중복을 포함한 트라이폰 개수는 91,146개 이며 중복을 뺀 트라이폰의 종류는 21,145 인 것을 알 수 있다. <표 1>과 비교하면 음소 균형 코퍼스에는 156만개 상호명에 포함되어 있는 트라이폰 종류의 0.05% 인 8천 7백 개의 상호명 만으로 동일한 트라이폰 종류가 포함되고 있으며 156만 개에 있는 트라이폰 수도 2천 5백만 개에서 0.36% 로 줄어든 9만 1천 여 개로 이루어져 있다.

| 상호명수 | 트라이폰 개수 | 트라이폰 종류 |
|---|---|---|
| 8,699 | 91,146 | 21,145 |

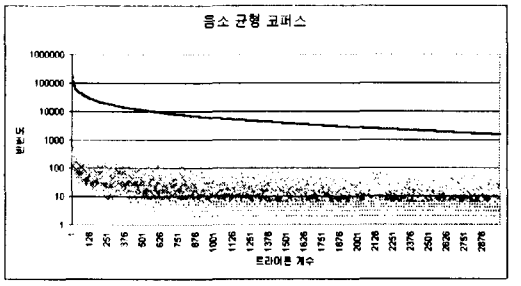
<표. 2> 음소균형 코퍼스
<Table 2> Phonetically Balanced Corpus

<표 3>은 음소 균형 코퍼스 상호명의 예를 나타내었다. 3.1 절에서 기술한 알고리즘에 따라 선정된 상호명은 잘 사용되고 있지 않는 트라이폰을 포함하고 있는 단어를 먼저 선택하였기 때문에 외래어로 이루어지는 상호명이 우선적으로 선택될 확률이 많다. 특히 트라이폰 특성상 좌, 우측이 묵음으로 되어 있는 한 개의 트라이폰으로구성되어 있는 상호 명( 예: "우", "어", "위" 등)일 경우도 우선적으로 선정되도록 알고리즘이 구성되어 있다.

| |
|---|
| 우, 어, 위, 휘, 비, 게파, 퓨쳐위즈, 시쿼이어, 벨뷔, 디쌤, 앙뛰, 트웽, 희우 |
| 카웨, 앵뷔떼, 삐우 ,췹, 겝 ,쉬어, 써꺼써꺼, 서오, 괴외잡 ,꼬께뜨, 여의, 나아, 뉘아주 |
| 코웬, 쩡, 버쯔, 뚜삐, 카아, 서아, 서우, 핫윙, 내꺼에요, 후띠에, 위위즈, 이두웨어 |
| 비위치, 바에자수, 위아, 데 |

<표. 3> 음소균형 코퍼스 내용

<Table 2> Contents of Phonetically Balanced Corpus




[그림. 3] 상호명 코퍼스 분포도(실선) 및 음소 균형 코퍼스 분포도(점선)

[Fig. 3] Company Names' Corpus (Line) and PBc(Dotted Line) Characteristics

[그림 3]은157만개의 상호명 코퍼스와 음소 균형 코퍼스의 분포도를 비교하였다. 상호명 코퍼스를 이루고 있는 21,145개의 트라이폰을 빈번도 순으로 정렬하여 실선으로 그렸으며 가장 빈번도가 많은 트라이폰은 20만 회가 넘고 있다. 반면 음소 균형코퍼스의 트라이폰 분포도는 점선으로 표시하였으며 상호명 트라이폰을 기준으로 음소균형 코퍼스내의 빈번도를 표시하였다. 상호명 코퍼스에서 10만개 이상의 빈번도를 나타내어 주는 첫번째 트라이폰의 경우 음소 균형 코퍼스에서는 800개 정도의 빈번도를 나타내고 있다. 전반적으로 음소 균형 코퍼스내의 트라이폰 빈번도는 상호명 코퍼스내의 트라이폰 빈번도와 어느 정도 유사하게 표현되고 있다. 그러나 음소균형 코퍼스는 상호명 코퍼스 내에 존재하는 모든 트라이폰이 존재하는 최소 코퍼스를 구성하였기 때문에 상호명 코퍼스내의 트라이폰 빈번도와 많은 차이점을 알 수 있다. [그림 3]에서 가로 축은 선형적으로 표현되었으며 세로 축은 로그 함수로 표현되었다.

### 4.1 음소분포상호명코퍼스(PDC)

음소 분포 코퍼스란 4.1절에서 설명한 음소 균형 코퍼스를 기준으로 하여 상호명 코퍼스 내에 존재하는 트라이폰의 빈번도와 유사한 기능을 갖는 최소 단어로 구성되어 있는 코퍼스를 말한다. <표 4>에는 상호명 코퍼스의 트라이폰 빈번도와 음소 분포 코퍼스내의 트라이폰 빈번도와의 유사도(R)를 0.90, 0.92로 하였을 경우 3.2절에서 기술한 알고리즘에 따라 구한 음소 분포 상호명 코퍼스내의 상호명 수와 트라이폰의 정보를 표현하였다. 유사도가 0.90일 경우 음소 분포 상호명 코퍼스내의 상호명 수는 157만개의 상호명 코퍼스의 0.69%인 10,860개의 상호명으로 형성이 되며 유사도가 0.92일 경우에는 1.07%인 16,783개의 상호명으로 구성되어 있다. 또한 사용되는 트라이폰 수는 157만개의 상호명 코퍼스에 존재하고 있는 트라이폰 수의 각각 0.47%(R=0.9), 0.83%(R=0.92) 정도 줄어들었다. 그러나 트라이폰 종류는 유사도와 관계없이 21,145개로 구성되어 있다.

| 유사도 (R) | 상호명 수 | 트라이폰 개수 | 트라이폰 종류 |
|---|---|---|---|
| 0.90 | 10,860 | 118,615 | 21,145 |
| 0.92 | 16,783 | 207,398 | 21,145 |

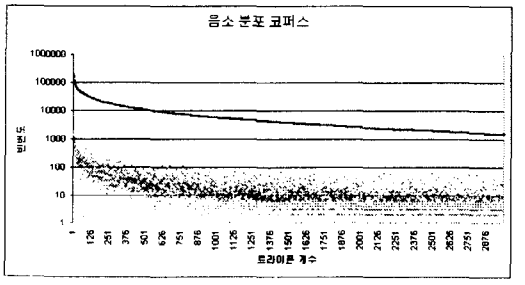
<표. 4> 음소 분포 코퍼스

<Table 4> Phonetically Distributed Corpus

| |
|---|
| 영점, 원점, 서점, 영원, 우원지점, 영주점, 주점, 영서점, 안경점, 영경사, 영외매점,영원서점 |
| 원점주점, 빙점, 시경산점, 영명사, 성영사, 원서점, 영매점, 우영사, 영영상사, 영상사 |
| 영영, 영, 애영사, 의령상점, 영영반점, 초점, 오영사, 영반점, 명원, 종점, 명안경점 |
| 영산원, 영애원, 서영사, 매점, 명대리점, 영전, 노점 |

<표. 5> 음소 분포 코퍼스 내용
<Table 2> Contents of Phonetically Distributed Corpus

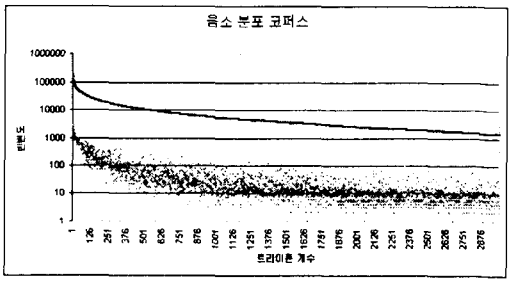
<표 5>에는 음소 균형 코퍼스 상호명의 예를 나타내었다. 3.2 절에서 기술한 알고리즘에따라 추가로 선정된 상호명은 많이 사용되고 있는 트라이폰을 포함하고 있는 단어를 먼저 선택하였기 때문에 일반적으로 많이 사용되는 상호명이 우선적으로 선택될 확률이 많다. 예를 들면 "영점", "원점", "서점" 등과 같이 많이 사용되고 있는 트라이폰의 빈번도가 높아야 하므로 동일한 음절 등이 반복되어 있는 단어가 선정되는 경향이 있다.




[그림4] 상호명 코퍼스 분포도(실선) 및 음소 분포 코퍼스 분포도(점선,R=0.90)
[Fig. 4] Company Names' Corpus (Line) and PDC(Dotted Line, R=0.92) Characteristics

[그림 4]에는 157만개의 상호명 코퍼스와 유사도 0.90 값을 갖는 음소 분포코퍼스와의 관계를 157만개의 상호명 코퍼스에 존재하는 트라이폰을 기준으로 분포도를 비교하였다. 실선은 157만개의 상호명 코퍼스 내에 존재하는 트라이폰의 빈번도를 기준으로 정열 한 것을 나타내며 점선은 유사도 0.92 값을 갖도록 구한 음소 분포코퍼스 내에 존재하는 트라이폰을 157만개의 상호명 코퍼스 내에 존재하는 트라이폰을 기준으로 빈번도를계산하여 표시하였다. 가장 높은 빈번도를 나타내는 트라이폰은 약 2000회를 기록하고 있으며 이것은 [그림 3]에서 그려진 음소 균형 상호명 코퍼스 내의 동일한 트라이폰의 800회 보다 많은 것이다. 또한 [그림 3]의 음소 균형 코퍼스 와 [그림 4]의 음소 분포 코퍼스를 비교하면 [그림 4]의 점선이 실선에 유사하게 그려져 있는 것을 알 수 있다. 이 것은 음소 균형 코퍼스 보다 음소 분포 코퍼스의 트라이폰의 분포도가 157만개의 상호명 코퍼스내의 트라이폰 분포도와 유사하다는 것을 나타내어 준다.

[그림 5]는 유사도를 0.92로 하였을 경우에 음소 분포 코퍼스의 트라이폰 빈번도를 점선으로표시한 것이다. [그림 5]에서 가장 많은 빈번도를 나타내어 주는 첫번째 트라이폰의 경우 빈번도는 약 5000회를 기록하고 있으며 이것은 유사도 0.90일 경우 2000회 보다 훨씬 많으며 또한 157만개의 상호명 내에 존재하는 트라이폰의 빈번도에 더욱 가깝다는 것을 알 수 있다. 즉 [그림 5]에서는 [그림 4]보다 점선의 폭이 상대적으로 작으며 실선에 보다 충실하게 가깝다는 것을 알 수 있다.




[그림. 5] 상호명 코퍼스 분포도(실선) 및 음소 분포 코퍼스 분포도(점선,R=0.92)

[Fig. 5] Company Names' Corpus (Line) and PDC(Dotted Line, R=0.92) Characteristics

## 5. 결론

본 논문에서는 114 안내 시스템에 구축 되어 있는 상호명 중 서울을 중심으로 저장되어 있는 157만개의 상호명을 가장 잘 음운적으로 표현이 가능한 최적의 코퍼스를 구하였으며 이를 분석하였다. 이러한 최적 코퍼스를 구하기 위하여 157만개에 존재하는 트라이폰의 종류가 모두 포함되는 최적의 코퍼스를 구하는 음소 균형 상호명 코퍼스를 일차적으로 구했으며 이 코퍼스를 기준으로 하여 트라이폰의 빈번도를 157만개의 상호명에 존재하는 트라이폰의 빈번도와 유사하게 되도록 하는 음소 분포 상호명 코퍼스도 구하였다. 그 결과 157만개의 상호명 코퍼스 내에 존재하는 트라이폰의 수는 21,145개 였으며 음소 균형 상호명 코퍼스로 8,699개의 상호명이 선정되었다. 음소 분포 상호명 코퍼스는 유사도가 0.90일 경우에는 10,860의 상호명으로 구성되었으며 유사도가 0.92 일 경우에는 16,783개의 상호명으로 증가 하였다. 그러나 유사도가 0.92일 경우 트라이폰의 빈번도에 대한 통계가 157만개의 상호명에 존재하는 트라이폰의 빈번도 통계와 매우 유사함을 알 수 있었다.

향후에는 음소 균형 상호명 코퍼스와 음소 분포 상호명 코퍼스를 이용하여 실제 음성인식 및 합성에 적용하는 연구가 진행되어야 할 것이다.

## 참고문헌

[1] R. Sproat, Multi-lingual TTS System, Bell Lab's approach, Kluwer Academic Publishers, 1998.

[2] Jia-Lin Shen, et. Al, "Automatic selection of phonetically distributed sentence sets of speaker adaptation with application to large vocabulary Mandarin speech recognition," Computer Speech and Language,"1999. pp. 79-97

[3] W. Zhu, et. al., "Corpus building for data-driven TTS systems," IEEE 2002, TTS Workshop

[4] H. Li. et. al., "Generating script using statistical information of the content variation unit vector," Proc. of ICSLP 2002.

[5] H. Kawai et. al., "A design method of speech corpus for text-to-speech synthesis taking account of prosody," Proc. of ICSLP 2000.

[6] 이선정 ,"한국어 음운환경을 고려한 최적 구문 코퍼스 설계에 관한 연구", 컴퓨터산업교육학회 논문지, 제 3권 제 11호, pp. 1615-1620, 2002.

이선정

1980.2 숭실대학교 전자계산학과(학사)
1985. 2. 서울대학교 대학원 전산과학과(이학석사)
1994. 2. 서울대학교 대학원 전산과학과(이학박사)
1985~1994. 2. 한국통신 멀티미디어 연구소 선임연구원
1994.4~현재 시립인천전문대학 전자계산학과 부교수
관심분야 : 자연어 처리, 한국어정보처리