# 중국어 음성합성을 위한 지지 벡터 기반 실시간 미등록어 처리*

포항공과대학교 컴퓨터공학과
하주홍† · 정 옥 · 이근배

# Real-time Unknown Word Identification Using Support Vector Machine For Chinese Text-to-Speech

Ju-Hong Ha, Yu Zheng, Gary G. Lee
*Department of Computer Science Engineering, POSTECH, Pohang, Korea*

**요 약**

음성 합성 시스템 구축에 있어서 입력 텍스트를 정확한 발음 표기로 변환하는 것은 매우 중요하다. 중국어에는 하나의 한자가 의미나 사용에 따라 다르게 발음되는 다음자(polyphony)들이 존재한다. 다음자의 처리는 상당히 복잡한 문제이기 때문에 본 논문에서는 그 중 가장 발음에 영향을 미치는 요소인 인명과 지명에 대한 미등록어 처리를 수행했다. 무엇보다 실시간 음성 합성 시스템을 위해서는 처리 속도의 향상이 요구된다. 따라서 본 연구에서는 미등록어 후보 구간 선정을 선행하고, 선정된 후보에 대해 추정하는 두 단계로 진행하였다. 후보 구간 선정은 단일 한자 단어(monosyllable word)의 확률과 간단한 패턴들을 이용한다. 최종 선정된 후보의 미등록어 추정은 SVM(Support Vector Machine)을 기반으로 실시하였다.

## 서 론

음성 합성 시스템(Text-to-Speech)에서 텍스트 문자를 정확한 발음 표기로 변환하는 것은 매우 중요한 문제이다. 변환된 발음 표기는 합성기가 음성 DB에서 해당 단위 합성음을 선택하는데 중요한 정보를 제공하기 때문에 음성 합성 시스템 전체 성능에 상당한 영향을 미치는 요소이다. 특히 자동 발음 변환 모듈(grapheme-to-phoneme conversion)에서 가장 중요한 문제는 미등록어에 대한 발음 기호 생성이다. 중국어를 포함한 여러 언어의 단어들 중에는 그 단어가 가지는 품사에 따라 발음이 변하는 경우가 있기 때문이다. 텍스트 분석의 첫 단계인 단어 분할과 품사 태깅 과정을 수행하기 위해서는 어휘 사전이 필요하다. 하지만 아무리 많은 말뭉치로부터 사전을 생성한다고 하더라도 모든 단어를 포함할 수는 없다. 따라서 보다 자연스러운 합성음을 발화하기 위해서는 미등록어에 대한 정확한 발음 처리가 필요하다.

지금까지 영어를 비롯한 알파벳 언어를 위한 여러 가지 발음 변환 방법이 제안되었다. 하지만 이와 달리 중국어 발음 변환의 어려움은 다음자 한자에 대한 처리에 있다. 중국어의 다음자 한자의 발음 처리는 상당히 복잡한 문제이다. 일반적으로 통일된 명확한 패턴이 없다.[8] Xiu shirong의 "중국어 다음자 한자 사전"에서는 중국어의 다음자를 크게 '뜻의 구별'과 '사용의 구별'의 두 가지로 분류하고 있다.[9] 이 두 가지는 다시 12가지로 상세하게 구별하고 있다. 본 논문에서는 위 세부 분류 중 발음 변환에 가장 많이 영향을 미치는 '고유명사의 특수성'을 고려하며 고유명사 중에서도 중국인명, 외국인명, 지명 등에 대한 미등록어 처리를 수행한다. 기관명에 대한 처리는 복합 명사 처리 문제로 그 성질을 달리 하고 자연스러운 발화를 위한 휴지 구간(Pause) 추정과 연관되어 있기 때문에 본 논문에서는 처리 대상에서 제외시켰다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 최근 연구되고 있는 중국어 미등록어 처리에 대한 방법들을 살펴

*본 연구는 2003년 한국과학재단의 특정기초연구 사업인 "무제한 단어 중국어 TTS(Text-to-Speech) 시스템을 위한 자연어 처리" 과제의 일환으로 수행되었습니다.
†E-mail : miracle@postech.ac.kr
E-mail : zhengyu@postech.ac.kr
E-mail : gblee@postech.ac.kr

보고, 3장에서는 본 연구실에서 개발한 대용량 말뭉치를 사용한 통계적 방법 기반 중국어 문장의 단어 분할 및 품사 태깅 시스템인 POSTAG/C[1, 1]를 간략하게 살펴보도록 한다. 4장에서는 실시간 음성 합성 시스템에서의 미등록어 처리를 위해서 문장에서 미등록어 후보 구간을 빠르게 선정하는 방법에 대해 설명한 후, 5장에서는 SVM 기반 미등록어 분류 추정 방법을 새롭게 제안한다. 6장에서는 인명과 지명에 대한 실험 및 분석을 하며, 마지막 7장에서는 결론 및 고찰을 기술하도록 한다.

## 관련연구

중국어는 한국어나 영어와는 달리 문장 내에서 단어들의 구분을 하지 않는다. 따라서 중국어 텍스트 처리를 위해서는 가장 먼저 단어 분할(word segmentation)이 수행되어야 한다. 하지만 단어 분할 단계에서는 사전에 없는 단어들의 경우 개별 한자들로 분할하여 출력하게 되고, 이 결과는 태깅 단계에서 전혀 엉뚱한 품사를 할당되게 한다. 이러한 사전에 없는 미등록어 문제를 해결하기 위한 연구들이 중국어 처리에 있어서 활발하게 진행되어 왔다.

Chen Ken-Jian과 Ma Wei-Yun[7]는 통계적 방법을 제안했는데 대용량의 Sinica 말뭉치[4]에서 형태소 룰과 통계적 룰을 생성하여 미등록어를 추정하는 방식으로 인명과 복합어에 대해 89%의 정확율(precision)과 68%의 재현률(recall)의 실험 결과를 나타내고 있다.

Kenvin Zhang 등[5]은 Markov 모델 기반 방식으로 품사 태거 시스템과 유사하게 비터비(Viterbi) 알고리즘으로 미등록어를 추정했다. 말뭉치는 1998년 북경 인민일보[12] 2개월 분을 사용하였다. 실험 결과는 인명에 대해서 69.88%의 정확율과 91.65%의 재현율을 보여주고 있다.

Goh Choori Ling 등[3]은 Markov 모델 기반 품사 태거와 글자 자질(character features)을 사용한 SVM 기반 chunker로 미등록어를 추정했다. 실험은 1998년 북경 인민일보[12] 1개월 분으로 학습과 추정을 실시하였다. 성능은 인명과 기관명에 대해서는 84.31%의 정확율과 73.85%의 재현율을 나타내고 있다.

본 논문에서는 Goh Choori Ling 등[3]의 방법과 유사하게 SVM 기반으로 미등록어를 추정한다. 하지만 실시간 음성 합성 시스템을 위한 미등록어 추정을 위해서는 문장에서 미등록어 구간만을 추출해서 미등록 단어의 정확한 범위와 품사를 고속으로 추정하는 것이 필요하다. 따라서 본 논문에서는 먼저 미등록어 후보 구간을 선정하고, 선정된 후보 구간에 대해 미등록어 추정을 수행하는 2단계 미등록어 처리를 제안한다.

1 POStech TAGger Chinese version

## 단어 분할 및 품사 태깅

앞에서 언급했듯이 중국어 텍스트 처리에 있어서 가장 먼저 수행되어야 하는 것이 단어 분할과 품사 태깅 과정이다. 본 연구에서는 POSTAG/C[1]의 간체 버전을 사용하여 단어 분할과 품사 태깅을 수행하였다. POSTAG/C는 대용량 말뭉치를 사용하여 규칙과 사전 기반의 단어 분할 모듈과 HMM(Hidden Markov Model) 기반 품사 태깅 모듈이 결합된 시스템이다. 간체 버전의 성능은 정확율, 재현율 모두 95%를 나타내고 있다.

## 미등록어 후보 선정

음성 합성 시스템을 구성하는 각 모듈들은 실시간으로 처리 되어야 한다. 하지만 입력되는 텍스트의 처음부터 끝까지 미등록어 여부를 확인한다면 속도는 매우 느려질 것이다. 또한 본 연구에서 사용하고자 하는 SVM의 느린 속도를 감안하여 본다면 보다 효율적인 방법이 필요하다. SVM은 기계 학습 방법들 중 가장 좋은 성능을 보이는 방법 중 하나이다. 그러나 수행 시간이 오래 걸리는 단점이 있다. 본 논문에서는 이러한 문제를 해결하기 위해서 문장 전체를 검사하는 대신 문장 내 미등록어가 발생 가능한 구간을 미리 선정하여, 그 구간에 대해서만 미등록어 추정을 실시한다. 따라서 보다 정확한 미등록어 후보 구간을 선정하는 것이 중요하다.

일반적인 중국어 단어 분할 시스템과 마찬가지로 POSTAG/C도 사전에 없는 미등록어에 대해서는 단어의 최소 단위인 단일 한자의 연속으로 출력한다. 따라서 태거의 결과에서 단일 한자가 연속적으로 발생한 구간을 미등록어 후보 구간의 출발 점으로 하면 된다. 중국어 처리에서 단일 한자가 연속적으로 발생하는 구간에 미등록어가 포함될 확률이 90% 이상이라는 연구 결과[10]가 이를 뒷받침해준다. 하지만 중국어는 단일 한자가 하나의 단어로 문장 내에 포함될 경우가 빈번하기 때문에 단순히 연속되는 단일 한자들을 전부 미등록어 후보 구간으로 사용할 수는 없다. 실제로 본 논문에서 테스트로 사용한 텍스트들에서 인명의 경우 단일 한자들의 연속(2개 이상)인 구간은 총 83,314개였으나, 인명 미등록어를 포함한 구간은 14,901개 뿐이었다. 결과적으로 미등록어 구간보다 그렇지 않은 구간이 약 5.6배 많기 때문에 미등록어를 제대로 처리하지 못할

뿐만 아니라, 오히려 올바른 결과를 잘못 처리하여 오류를 더욱 증가시킬 가능성이 많다.

따라서 본 논문에서는 다음과 같이 두 단계를 통해서 전체 단일 한자들의 연속 구간들 중 미등록어 추정에 사용할 후보 구간들을 선정하였다. 우선 말뭉치 내 존재하는 단일 한자 단어(monosyllable word)들에 대해 1)의 확률을 계산한다.

$$\frac{\text{단일 한자 단어의 출현 빈도}}{\text{단일 한자 단어로 쓰이는 한자의 출현 빈도}} \geq \lambda \quad 1)$$

계산된 $\lambda$를 단일 한자 단어 확률이라고 하고, 실험을 통해 연속되는 단일 한자들 중 $\lambda$가 0.7 이상인 한자를 후보 구간에서 제외한다.

두 번째 단계는 학습 말뭉치에 존재하는 인명과 지명들로부터 간단한 바이그램(bigram) 패턴을 구축하여 미등록어 후보 구간들 중 미등록어를 포함하는 구간들을 추출한다. 사용된 패턴은 인명과 지명들에 나타나는 연속된 두 한자를 묶어 하나의 패턴으로 사용하였다. 학습 말뭉치로부터 생성된 패턴은 인명이 34,727개, 지명이 15,982개이다.

## SVM 기반 미등록어 추정

미등록어 후보 구간 선정 단계를 거쳐 나온 후보구간들에 대해 보다 정밀한 처리를 위해서 본 논문에서는 SVM 기반으로 미등록어 후보에 정확한 품사를 부여한다. 본 연구에서는 Chang Chih-Chung의 LIBSVM[2, 2)]의 2.0 버전을 사용했다. 커널 함수는 최적 파라미터를 구할 수 있는 RBF를 사용하였다. LIBSVM에 대한 자세한 내용은 해당 사이트를 참고하길 바란다.

### 1. 자질들

SVM 학습을 위해 사용된 자질들은 Fig 1에서 보는 바와 같이 10개의 자질들을 사용하였다.

후보 구간 내 각 한자들은 자신을 포함한 좌우 두 개의 한자들에 대한 어휘 정보와 태그 정보를 이용하여 자신의 태그를 추정한다. 각 한자에 할당된 태그 정보는 그 한자가 속한 단어의 태그 정보와 단어에서의 위치 정보를 결합하여 나타낸다. 이렇게 개별 한자들을 자질 단위로 선정한 이유는 POSTAG/C가 사전에 없는 미등록어에 대해서 단일 한자들의 연속으로 분할하여 출력하기 때문이다. 또 하나의 이유는 중국어 형태소의 최소 단위가 단일 한자여서

2 http://www.csie.ntu.edu.tw/~cjlin/libsvm

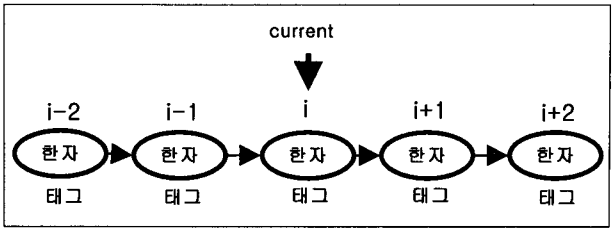


**Fig 1.** 지지 벡터 머신을 위한 자질들.

**Table 1.** 위치 태그 정보

| 위치 태그 | 설 명 |
|---|---|
| S | 단일 한자 단어 |
| B | 다한자 단어(두자 이상)의 첫 번째 한자 |
| I | 다한자 단어의 중간(세자 이상) 한자들 |
| E | 다한자 단어(두자 이상)의 마지막 한자 |

미등록어 처리에 더 효율적이기 때문이다. 예를 들어 미등록어 후보 구간인

唐 / 國 / 强 /还

가 본 연구에서 구현한 미등록어 추정 모듈을 거치게 되면

唐 國 强 / 还

로 미등록어 구간 내에서도 미등록어와 단일 한자 단어를 정확하게 구분해낸다.

### 2. 위치 태그 정보(Position tag)

미등록어 처리를 위해 POSTAG/C를 통해 나온 결과들 중 미등록어 후보 구간이라고 판단된 구간의 경우 단어 내 각 한자들에게 위치 태그 정보를 부여한다. 예로 단어가 단일 한자 단어일 경우는 <품사_S>의 형태로 단어의 품사와 위치 태그를 결합한 태그 정보를 부여하고, 두 개의 한자로 구성된 단어는 각각 <품사_B>, <품사_E>를 부여하며, 세 개 이상일 경우에는 단어의 처음과 끝에 각각 <품사_B>, <품사_E>를 중간 한자들에 대해서는 <품사_I>를 부여한다. 태그 정보는 Table 1에서 정리하고 있다.

### 3. 미등록어 추정

미등록어 추정 과정은 선정된 후보 구간들 내 각 한자들에 대해 위치 태그를 부여하여 SVM에 사용될 자질을 생성한다. 이렇게 생성된 자질들은 LIBSVM에서 사용할 수 있는 입력 형태로 변형시켜 학습과 추정을 실시한다. 본 논문에서는 음성 합성 시스템의 성능에 가장 영향을 미칠 것으로 판단되는 중국인 인명, 외국인 인명, 지명에 대해서 미등록어 처리를 수행하였다.

미등록어 추정은 인명과 지명 미등록어 후보 구간에 대해 앞 절에서 설명한 정확한 위치 태그를 추정하고, 추정된 위치 태그를 결합시켜 최종 단어를 추정한다. 예를 들

어 唐国强还에 대한 태거 결과는 唐/Tg_S 國/n_S 强/Ng_S 还/d_S 인데, 이 결과가 미등록어 추정과정을 거치게 되면 唐/nr_B 國/nr_I 强/nr_E 还/O로 바뀌게 되고, 최종적으로 唐國强가 인명이라는 정확한 결과를 얻게 된다. 여기서 O가 의미하는 것은 인명이나 지명이 아닌 다른 품사라는 것을 나타내며, O의 경우 그 한자의 품사는 태거 결과를 그대로 따르게 된다. 참고로 본 연구에서는 단일 한자 인명과 지명에 대해서는 <품사_S> 대신 <품사_B> 태그를 사용하였다. 그 이유는 1. 말뭉치 구성 및 전처리에서 설명할 전처리 과정을 거친 말뭉치에서는 인명과 지명이 단일 한자로 사용된 경우가 거의 없어서 <품사_S>로 나타낼 경우 성능의 저하를 가져오게 된다.

## 실험 및 결과

본 논문에서는 중국어 간체 텍스트를 대상으로 음성 합성 시스템을 위한 미등록어 처리를 실시하였으며, 중국인명, 외국인명, 지명의 세 부분에 대해 SVM 기반 미등록어 학습 및 추정 실험을 진행하였다.

### 1. 말뭉치 구성 및 전처리

본 연구에 사용된 말뭉치는 중국 본토에서 사용되는 간체로 북경 인민일보 1998년 1월부터 6월까지 6개월 분의 기사[12]를 사용하였다. 본 말뭉치는 약 650만개의 단어로 구성되어 있으며 수작업으로 단어 형태소 분석과 태깅이 미리 되어 있다. 여기에서 사용된 품사들은 총 43개로 이루어져 있으며, 이 말뭉치의 특징의 하나는 고유명사인 인명, 지명, 기관명에 대해 각각 nr, ns, nt의 품사들로 표기하고 있어 본 연구와 같은 개체명(named entity)에 대한 연구에 유용하다.

효율적인 실험을 수행하기 위해서 실험 전 말뭉치에 대해 전처리를 실시하였다. 원시 말뭉치에서는 江/nr 泽民/nr 과 같이 성과 이름을 분리하여 인명을 표기하고 있다. 또한 복합 단어 기관명을 [中國/ns 政府/n]nt로 분리 표기하고 중괄호로 묶어 하나의 기관명으로 표기하고 있다. 전처리 과정에서는 말뭉치 내의 위와 같은 표기들을 모두 하나의 단어로 통합해서 재구성하고, 단어 분할 및 품사 태거 시스템의 어휘 사전을 구성하였다.

실험을 위해서 단어 분할 및 품사 태거 시스템의 사전들을 1월부터 5월까지의 5개월 분 기사들만 학습해여 생성하였으며, 미등록어 추정 효과를 높이기 위해 사전 생성 과정에서 인명과 지명 단어들을 모두 제외시켰다. 그 다음 추정 실험을 위해 마지막 6월 분 기사에 대해 단어 분할과 품사 태깅을 실시하였다.

**Table 2.** 미등록어 후보 구간 선정(인명)

| | 후보구간 추정 전 | 후보구간 추정 후 | 감소율(%) |
|---|---|---|---|
| 총 후보 구간 | 83,314 | 14,901 | 82.11 |
| 인명 미등록어 포함 구간 | 14,028 | 12,533 | 10.66 |

**Table 3.** 미등록어 후보 구간 선정(지명)

| | 후보구간 추정 전 | 후보구간 추정 후 | 감소율(%) |
|---|---|---|---|
| 총 후보 구간 | 88,649 | 19,735 | 77.74 |
| 지명 미등록어 포함 구간 | 15,693 | 14,803 | 5.67 |

### 2. 실험 결과

본 논문에서의 실험은 크게 두 가지로 나눌 수 있다. 첫 번째 실험은 얼마나 정확하게 미등록어 후보 구간을 선정하는지에 대한 것이다. 미등록어가 발생한 구간을 정확하게 추출하여, 올바른 품사를 추정하는 것은 정확성을 유지하면서도 발음 변환 속도를 향상시킬 수 있기 때문에 실시간 음성 합성 시스템에 있어서 매우 중요한 요소이다. 미등록어 후보 구간 선정의 정밀성을 높이기 위해서 본 논문에서는 4장의 설명대로 2단계 과정을 수행하였다. Table 2, 3은 인명과 지명 각각에 대한 후보 구간 선정 적용 전, 후의 미등록어 후보 구간의 변화량을 나타내고 있다.

위 Table들에서 보는 바와 같이 2단계 방법 적용 후 실제 미등록어 포함 구간이 인명, 지명 각각 10.66%, 5.67% 감소하긴 했지만, SVM 기반 추정을 위한 총 후보 구간은 각각 82.11%, 77.74%로 상당한 감소율을 보임을 알 수 있다. 이와 같이 후보 선정 단계들을 거친 결과는 그렇지 않은 결과에 비해 SVM 기반 미등록어 추정의 성능 향상에 상당한 영향을 준다는 것을 실험을 통해 확인하였다. 일부분의 미등록어 포함 구간이 후보 구간에서 제외되는 문제는 본 연구에 사용된 말뭉치가 미등록어 처리를 위해 구축된 말뭉치가 아니기 때문에 패턴 구축의 한계로 인해 발생하는 것으로 해석할 수 있다. 미등록어 처리를 위한 대량의 개체명으로 패턴을 구축한다면 미등록어 포함 구간의 감소율(구간 선정 에러율)은 상당히 줄어들 것이다.

2단계 미등록어 후보 구간 선정 방법의 적용 유무에 따른 수행 속도 변화를 Table 4에서 확인할 수 있다. 결과는 인명에 대한 수행시간이며, 말뭉치 중 6월분 기사의 첫 160문장과 300문장을 대상으로 수행 속도를 측정하였다. 표에서 보는 바와 같이 2단계 미등록어 후보 구간 선정 방

**Table 4.** 후보 선정 방법 적용에 따른 처리 속도(인명)

| | | 후보선정시간(ms) | 추정시간(ms) | 총시간(ms) |
|---|---|---|---|---|
| 160문장 | 적용 전 | | 126,118 | 126,118 |
| | 적용 후 | 250 | 2,116 | 2,366 |
| 300문장 | 적용 전 | | 261,600 | 261,600 |
| | 적용 후 | 520 | 5,040 | 5,560 |

**Table 5.** 인명에 대한 미등록어 추정 성능

| | 정확율(%) | 재현율(%) | F-measure(%) |
|---|---|---|---|
| 기본 10개 자질 | 79.33 | 88.64 | 83.73 |
| 성씨 자질 추가 | 82.46 | 89.44 | 85.81 |
| 성씨 자질 및 외국인명 한자 자질 추가 | 84.59 | 91.57 | 87.94 |

**Table 6.** 지명에 대한 미등록어 추정 성능

| | 정확율(%) | 재현율(%) | F-measure(%) |
|---|---|---|---|
| 기본 10개 자질 | 72.84 | 84.95 | 78.43 |

법을 적용한 경우 약 50배 정도 속도가 향상되는 것을 확인할 수 있다.

두 번째 실험은 후보 선정 단계를 거친 결과들로 SVM을 학습시키고, 학습된 모델로 미등록어 추정 성능을 확인하는 실험이다. 실험은 후보 선정 단계를 거친 총 후보 구간을 5등분하여 4/5를 학습 데이터로 사용하고, 나머지 1/5을 추정 데이터로 사용하였다. 평가 방법은 정확율(precision), 재현율(recall), F-measure을 사용하였다. 인명에 대한 실험 결과는 Table 5와 같다.

인명을 대상으로는 Table 5에서와 같이 실험을 실시했는데, 처음에는 앞 1. 자질들에서 기술한 자질들만으로 실험을 하였고, 이후 성씨 자질과 외국인명 한자 자질들을 추가하면서 실험을 실시하였다. 성씨 자질로는 중국인 성씨들 중 사용 빈도가 가장 높은 200개를 자질로 사용하였다. 참고로 상위 200개의 성씨는 174,900명의 조사 대상자 중 96% 이상을 차지하는 것으로 조사되었다.[3] 외국인명 한자 자질은 외국인명을 중국어로 표기할 때 쓰이는 한자들로써 중국에서는 외국인명을 표기할 때 정해진 특정 범위의 한자들로 표기하는 것이 일반적이다. 본 논문에서는 [6]과 [11]에서 획득한 520개의 한자들을 외국인명 한자 자질로 사용하였다. 인명에 대해서는 위에서 기술한 모든 자질은 사용했을 때 F-measure 값이 87.94%로 좋은 성능을 보여주었다.

Table 6에서는 지명에 대한 실험 결과를 나타내고 있는데 지명에 대해서는 기본 10개의 자질만 실험에 사용하였다. 지명에 대한 미등록어 추정 성능은 78.43%로 인명의 성능에 비해 약 5% 정도 낮은 성능을 보였다.

지명에 대한 추정 성능이 인명의 결과보다 낮은 이유는 일반적으로 인명이 지명보다 고유성이 높은데 있다고 생각할 수 있다. 실험적으로는 미등록어 후보 구간 선정 단계를 거친 결과를 살펴보면 지명의 경우 실험 말뭉치 내에 존재하는 총 개체명 수에 대한 선정된 후보 구간에 포함된 개체명 수의 차이가 인명보다 상대적으로 크다는 사실에서 성능 저하가 생김을 추정할 수 있다. 실제로 태거의 어휘(lexical) 사전을 살펴본 결과 지명의 단어 중 두 개 이상의 품사를 가지는 단어가 807개 존재함을 확인하였다. 주로 일반 명사(n)나 기타 고유 명사(nz)와 함께 쓰이는 경우가 많았다. 이는 말뭉치 오류로 말뭉치 생성 과정이 일관되게 진행되지 못했음을 보여준다.

마지막으로 전체 실험 결과를 살펴보면 재현율보다 정확율이 상대적으로 낮음을 볼 수 있다. 그 이유는 최종 단어 추정 전 SVM을 통한 위치 태그 추정 결과를 살펴보면 알 수가 있다. 결과에서 인명이나 지명 단어의 한자들의 위치 태그의 추정 정확율이 평균 약 93%으로 높은 반면, 그 외의 한자들의 정확율은 80% 이하로 낮음을 확인할 수 있었다. 즉, 인명이나 지명에 속하지 않는 한자들의 오류가 정확율을 낮추는 요인이라 할 수 있겠다.

## 결론 및 고찰

본 연구에서는 중국어 음성 합성 시스템 구축에 있어 중요한 요소인 텍스트의 정확한 발음 변환을 위해 사전에 존재하지 않는 미등록어 처리를 수행하였다. 처리 대상은 발음에 가장 영향을 미친다고 판단된 중국인명, 외국인명, 지명을 대상으로 실시하였다. 미등록어 처리 과정은 고속 처리를 위해 두 단계로 처리하였다. 먼저 미등록어 후보 구간을 자동으로 선정하고, 선정된 구간에 대해서만 SVM 기반 미등록어의 품사를 추정하는 과정을 실시하였다. 후보 구간 대상은 태거 결과 중 단일 한자 연속 구간으로 한정했다. 전체 후보 구간은 미등록어를 포함한 구간에 비해 상당히 많기 때문에 단일 한자 단어의 확률과 한자 패턴으로 미등록어 후보 구간을 선정하였다. 미등록어 추정 과정은 1. 자질들에서 기술한 기본 10개의 자질들로 SVM 기반 추정을 실시하였다. 성능 향상을 위해 인명에 대해서는 중국인 성씨 자질, 외국인명 한자 자질들을 차례로 첨가해서 실험을 진행하였다. 실험 결과 인명에 대해서는 F-measure 값이 약 88%로 과거 연구들에 비해서 상당히 좋은 결과를 보였고, 지명의 결과는 약 78%의 성능을 나타냄을 확인할 수 있었다.

3 http://zhongwen.com/x/xingshi.htm

실험 결과 전체적으로 재현율이 정확율보다 높게 나타났으며, 이는 SVM 학습에 있어서 인명이나 지명이 아닌 다른 한자들의 오류로 인한 것으로 해석할 수 있다. 또한 말뭉치 내 단어에 대한 품사 할당의 일관성 부족이 후보구간 성능이나 전체 성능에 악영향을 미치고 있음을 확인할 수 있었다.

향후 코퍼스 정제를 통해 추정 성능을 향상 시킬 수 있을 것으로 기대된다. 또한 SVM 자체의 처리 속도를 좀 더 향상시킨다면(예, 점진적 SVM) 보다 신속하고 정확한 미등록어 처리를 수행할 수 있을 것이다.

## REFERENCES


1) 하주홍, 정 옥, 이근배(2002) : "품사 사전 자동 학습을 통한 중국어 단어 분할 및 품사 태깅", 제 *14* 회 한글 및 한국어 정보처리 학술대회
2) CC Chang and CJ Lin(2003) : *"LIBSVM : a Library for Support Vector Machines", a guide of beginners*
3) CL Goh, Msasayuki Asahara, Yuji Matsumoto(2003) : *"Chinese Unknown Word Identification Using Character-based Tagging and Chunking", ACL-2003 Interactive Poster*
4) KJ Chen, CR Huang, LP Chang and HL Hsu(1996) : *"SINICA CORPUS : Design Methodology for Balanced Corpora", Proceedings of PACLIC 11th Conference, pp167-176*
5) KJ Chen, WY Ma(2002) : *"Unknown Word Extraction for Chinese Documents", COLING-2002, Institute of Information science, Academia Sinica*
6) Kevin Zhang, Q Liu, H Zhang and XQ Cheng(2002) : *"Automatic Recognition of Chinese Unknown Words Based on Roles Tagging, SIGHAN Workshop, ACL*
7) L Chen(1995) : *"A Chinese Text Display supported by the Chinese Segmentation Algorithm", Master Thesis, Department of Computer Science, New Mexico State University*
8) ZR Zhang and M Chu(2002) : *"A Statistical Approach for Grapheme-to-Phoneme Conversion in Chinese", ISCSLP, Microsoft Research Asia*
9) 陈亚川, 郑懿德 "多音字全解词典", 陕西人民出版社 *(1998)*
10) 吕雅娟, 赵铁军, 杨沐的, 于活, 李生 "基于分解与动态规划策略的汉语未登录识别", 中文信息学报*(2000)*
11) "英语姓名译名手册"、新华社、商务出版社 *(1997)*
12) 俞士汶，现代汉语语料库加工——词语切分与词性标注规范与手册，北京大学计算语言学研究所*(1999)*