어원 정보를 이용한 외래어의 자동 원어 복원*

포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터
이상율† · 강인수 · 나승훈 · 이종혁

Automatic Back-Transliteration with Word Origin Information

Sang-Yool Lee, In-Su Kang, Seung-Hoon Na, Jong-Hyeok Lee
Department of Computer Science and Engineering, Division of Electrical and Computer Engineering, Pohang University of Science and Technology, and Advanced Information Technology Research Center(AITrc), Pohang, Korea

요 약

음차 표기된 외래어로부터 원어를 복원하는 문제는 원어의 발음정보를 이용한 통계적인 방법을 많이 사용한다. 하지만 지금까지의 연구들은 대부분 영어단어만을 그 대상으로 하였기 때문에 '도쿄(Tokyo)', '하인리히(Hinrich)'와 같이 어원이 영어가 아닌 단어들의 복원에는 좋은 결과를 보여주지 못했다. 이러한 문제를 해결하기 위하여 한글로 표기된 외래어의 어원을 판단할 수 있는 방법을 찾아내고, 이 방법을 통해 외래어를 어원별로 분리하여 학습모델을 구축함으로써 다양한 어원을 가진 외래어들의 복원 정확률을 높이고자 하였다. 위의 방식으로 구현된 시스템은 영어, 일본어, 중국어, 프랑스어의, 서로 다른 4개의 어원을 가진 데이터의 복원 실험에서 기존의 방식에 비해 13%의 성능 향상을 보였다.

서 론

교차언어 문서검색(Cross-Language Text Retrieval)은 검색대상이 되는 문서집합과 질의어가 서로 다른 언어로 구성된 경우의 문서검색을 말한다. 이 때, 서로 다른 두 언어를 하나의 언어로 일치시켜 검색이 가능하게 해 주어야 하는데, 이 경우 문서를 번역하는 방법보다는 질의어를 변환시키는 쪽이 더 간편하기 때문에 많이 사용된다. 또 질의어의 변환은 일반적으로 사전정보에 기반하여 이루어지는데 이는 사전의 확장을 통한 어휘의 추가가 편리하다는 장점이 있다. 하지만 이런 방식들도 단일어 검색에 비하면 60%정도의 성능만을 보여주게 되는데(Hull, 1996), 그 대표적인 원인으로 어의중의성(Word Sense Ambiguity) 문제와 미등록어(Out of Vocabulary) 문제를 들 수 있다.

특히 미등록어 문제는 일반명사 보다는 고유명사에서 대부분 발생하고 있으며(Tompson, 1997)에 따르면 신문기사를 대상으로 한 검색에서 고유명사가 검색시 질의어로 사용되는 빈도가 전체 질의문 중에서 63%를 차지할 만큼 그 비중이 크기 때문에 외래어 고유명사에 대한 미등록어 문제 해결이 매우 중요하다.

따라서 미등록어에 대한 해결책으로 거론되고 있는 것이 미등록어의 원어를 자동으로 찾아내는 방법에 대한 연구이다. 현재 외래어 고유명사는 해당 언어의 원어 발음에 최대한 가깝게 한글로 표기하도록 되어있다(문화관광부, 2000). 따라서 각 언어로부터 한국어 단어 생성시에 언어별로 서로 다른 변환규칙을 사용하며, 이러한 내용을 한글로 표기된 외래어의 원어 복원에도 고려해 주어야 한다.

관련연구

1. 외래어 복원

한글로 표기된 외래어의 원어복원에 대한 기존 연구는 크게 규칙기반 방식과 통계기반 방식으로 나뉜다. 먼저 규

*본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.
†E-mail : gilbert@postech.ac.kr
E-mail : dbaisk@postech.ac.kr
E-mail : nsh@postech.ac.kr
E-mail : jhlee@postech.ac.kr

칙기반 방식에는 가장 대표적인 것으로 국어의 로마자 표기법을 들 수 있다. 하지만 이 방법은 영어에서 음차 표기된 한글 단어를 대상으로 하는 것이 아니라, 한글로 이루어진 우리 고유의 지명, 인명 등을 대상으로 가장 한글 발음을 유사하게 유지하는 방식으로 로마자 표기가 이루어지기 때문에 외래어의 복원 문제에서는 적합하지 않다.

다음으로 김병혜(1991)의 연구에서는 주어진 영어철자를 한국어로 변환함에 있어서 미리 작성된 규칙을 기반으로 영어 철자를 발음기호로 변환하고 이 발음 기호를 외래어 표기법을 이용하여 한국어를 생성해 내는 방식을 사용하였다. 이는 일본의 Yuichi(1990)에서 영어 철자를 규칙에 기반하여 발음기호로 변환한 것과 유사하다. 그러나 기존의 규칙 기반 방법은 미리 정해진 규칙에 의존하여 발음을 생성하기 때문에 동일 철자의 여러 변화에 대한 내용을 반영하기 힘들다. 따라서 이를 보완하기 위한 연구로 SERI(1995)에서는 한글 모음에 해당하는 자소에 대해서 변화가 가능한 조합들을 미리 정의해 두고 이를 통해 다양한 변이형을 생성하는 방법을 제안하기도 하였으나 불필요하게 많은 변이체를 생성하여 실제 응용에서는 그리 효과적인 방법이 되지 못하였다.

통계기반 방식의 기존 연구 중에서는 신경망을 이용한 방법인 정길순(1998), 김정재(1999)의 연구와HMM(Hidden Markov Model)을 이용한 이재성(1998)의 방법이 있었으며, 또한 정길순(1998)에서는 사전 매칭을 통한 후처리 과정으로 성능의 향상을 꾀할 수 있음을 함께 보였다. 그리고, 이재성(1998)에서는 직접 영어철자로 변환하는 방식, 영어철자의 표준 발음을 추출하고 이를 이용한 변환 방식 등 2가지 방식으로 변환을 진행하여, 철자 위주의 영어단어표기(눈말표기)와 발음 위주의 영어단어표기(입말표기)를 둘 다 고려하였다.

하지만 위에 언급된 여러 실험들은 외래어 복원문제에서 모든 데이터들의 철자들이 영어발음을 충실히 반영한다는 가정하에 Fig. 1과 같이 하나의 확률모델을 가지고 시스템을 제작하게 되는데, 이로 인해 흔히 접할 수 있는 영어가 아닌 언어로부터 발생된 단어인 '하인리히(Hinrich)' 나 '도쿄(Tokyo)'와 같은 것들에 대해서는 영어 발음 위주의 확률모델에서 정확한 결과를 얻기가 힘들다.

따라서 이것을 고려해 주는 연구가 필요한데, 이상율(2003)에서 이것과 관련하여, 영어 발음위주의 데이터와 그렇지 않은 데이터를 규칙기반으로 구분해 주고 이를 통해 각각의 확률모델에 대한 학습데이터를 분리해서 사용하는 방법을 제안하였다. 그리고 이를 통해서 어느 정도의 성능 향상이 있었지만, 기본적으로 영어와 영어이외의 단어,

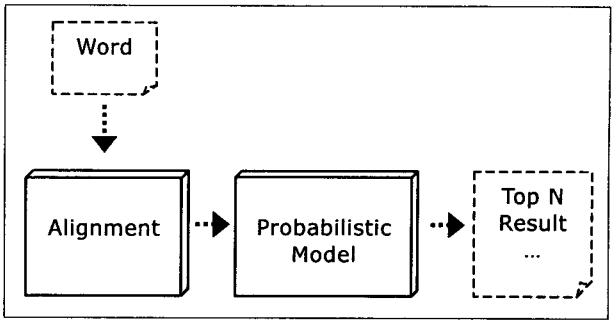

Fig. 1. 전통적인 통계기반의 외래어 복원모델.

이렇게 이진 분류를 사용함으로써, 영어 이외의 단어를 전부 뭉쳐서 하나로 만들기 때문에 발생하는 오류에 대해서는 역시 해결이 힘들었다.

본 논문에서는 위에 언급된 문제들을 해결하기 위해 철자에 나타난 정보를 통해서 각각의 학습 데이터들의 어원을 구분하여 주고, 이를 이용해서 각 언어별로 독립적인 확률모델을 만든 후에, 복원의 대상이 되는 단어가 어떤 언어에 속하는지를 사전에 판단하여 해당 단어에 가장 적합한 확률모델에서 처리하는 방식을 제안한다.

2. 외래어의 어원구분

한글로 표기된 외래어를 대상으로 한 어원 구분에 관한 연구는 현재까지 제대로 소개된 바가 없었다. 하지만 영어권에서는 음성인식 분야에서 LID(Language IDentification)라는 주제로 연구가 활발히 진행되고 있으며, 이는 실제의 음성데이터 분석을 통해 사용된 언어를 알아내는 연구분야이다.

이 연구에서는 각 언어별로 발음상의 차이점을 이용하여 음소레벨의 n-gram 정보를 가지고 확률값을 구하는 방법을 주로 사용한다. 이 방법을 이용한 Corredor-Ardoy(1997)에서는 영어, 프랑스어, 독일어, 스페인어의 4개 국어를 대상으로 한 시스템에서 약 90~91%의 정확률을 보였다.

또 다른 연구로 Combrinck(1995)에서는 텍스트 기반의 언어 판단에 관한 실험을 진행하였는데, 이탈리아어, 영어, 독일어, 프랑스어, 스페인어, 포르투칼어, 그리고 아프리카 6개 언어, 이렇게 총 12개의 언어를 대상으로 각 언어들로 이루어진 문서내의 모든 철자를 ASCII 코드에 정의된 알파벳으로 정규화시키고, 이 철자 정보를 바탕으로 문서에 씌여진 언어를 판단하는 문제로 100%의 정확도를 보였다. 이는 각 언어마다 철자로부터 독립적인 특징들을 추출해 낼 수 있으며, 이런 특징들을 바탕으로 사용된 언어를 유추하는 것이 충분히 가능하다는 것을 보여준다. 하

지만, 언어의 유추에 사용된 정보의 양이 각 언어당 평균 5,000 글자 정도 였으며, 실제 판단에 상당히 많은 정보를 사용하였기 때문에 이를 한국어 단어(3~4글자)수준에서 바로 적용하기는 현실적으로 문제가 많다.

제안모델

1. SVM을 이용한 외래어의 어원구분

앞서 살펴본 바와 같이, 한글 외래어를 어원별로 분류하기 위해서는 각 언어마다 가지는 고유한 발음 특성을 이용해야 한다. 다행히 한글은 자소 하나 하나마다 각각의 고유한 음이 정해져 있으며, 따라서 철자를 통해서 음의 연속된 패턴이나 특정한 음의 발생유무를 추측할 수 있다. 그리고 이런 정보들을 이용하여 언어별 발음특성을 규정할 수 있다. Table 1은 각 언어별로 높은 빈도수를 보인 n-gram들의 목록이다. 여기서 굵게 * 표시된 n-gram들은 다른 언어에서는 많이 등장하지 않는 발음들이며, 이것들이 분류과정에서 각 언어별 특징을 나타내는 역할을 하리라 예상된다. Table 1에서 괄호( )로 묶인 글자들은 한글의 초성과 종성을 구분하기 위해서 사용되었으며, 종성을 나타낸다.

Table 1를 토대로 본 연구에서는 한글 자소의 n-gram을 통해 발음상의 특성들을 추출해내고, 추출된 n-gram들을 SVM을 이용해서 학습 및 분류를 시도하였다. n-gram 정보로는 1~3-gram을 함께 사용하였으며, 하나의 단어에서도 각각의 n-gram 들의 발생횟수에 가중치를 부여하여 SVM의 학습과정에 사용하였다. Table 2는 각 언어별로 n-gram을 추출한 예를 몇 개 보여준다.

이렇게 추출된 n-gram을 가지고 SVM의 학습 데이터로 이용하게 되는데, SVM은 잘 알려진 바와 같이 +1에 해당하는 긍정적 데이터와, -1에 해당하는 부정적 데이터를 학습하여 이 둘을 분리하는 결정면을 찾는 이원 패턴 분리를 위한 알고리즘이므로 각 학습 데이터들을 이에 맞게 구성하여 학습을 하면 된다. 즉, 영어 단어로부터 추출된 n-gram 정보들은 영어 패턴 분리를 위한 모델에서는 긍정적 학습데이터로, 프랑스어나 일본어, 중국어 패턴 분리 모델에서는 부정적 데이터로 적용하게 된다. 그리고 학습되어진 개별 언어 판단 모델로부터 테스트 데이터를 적용하여, 최대값을 내는 모델로 결과를 내면 된다. 이렇게 만들어진 SVM의 구성도는 Fig. 2에서 볼 수 있다..

2. 규칙기반의 정렬기법

주어진 한글 외래어로부터 원어를 복원하는 문제를 통계기반으로 처리하기 위해서는 한글 자모 및 알파벳으로부터 하나의 발음을 낼 수 있는 단위로 쪼개고, 각각의 한글 자모들이 어떤 알파벳으로 표현될 수 있는지에 대한 확률값을 학습을 통해서 정하게 된다. 이 때, 모든 한글 자모가 한

Table 1. 각 언어별 최다 빈도 n-gram 목록

Top 10	영어	일본어	중국어	프랑스어
1	ㅣ	ㅏ	(ㅇ)	ㄹ
2	ㄹ	ㅣ	ㅏ	ㅏ
3	ㅏ	ㅗ	ㅣ	ㅔ
4	(ㄴ)	ㅋ*	(ㄴ)	ㅗ
5	ㅓ	ㅅ	ㅜ	ㅣ
6	ㅅ	ㅜ	ㅓ	(ㅇ)
7	ㅗ	ㅁ	ㅅ	(ㄹ)
8	(ㄹ)	ㄹ	ㅈ*	ㅅ
9	ㅔ	ㅅㅣ*	ㅏ(ㄴ)*	ㅜ
10	ㅌ	ㅔ	ㅣ(ㅇ)*	ㅌ

Table 2. 각 단어별 n-gram 추출 예

영어	제임스	ㅈ, ㅔ, ㅇ, ㅣ, (ㅁ), ㅅ, ㅡ, ㅈㅔ, ㅔㅇ, …, (ㅁ)ㅅ, ㅅㅡ, ㅈㅔㅇ, ㅔㅇㅣ, …, (ㅁ)ㅅㅡ
	워싱턴	ㅇ, ㅝ, ㅅ, ㅣ, …, ㅓ, (ㄴ), ㅇㅝ, ㅝㅅ, …, ㅌㅓ, ㅓ(ㄴ), ㅇㅝㅅ, ㅝㅅㅣ, ㅅㅣ(ㅇ), …ㅌㅓ(ㄴ)
일본어	도쿄	ㄷ, ㅗ, ㅋ, ㅛ, ㄷㅗ, ㅗㅋ, ㅋㅛ ㄷㅗㅋ, ㅗㅋㅛ
	오사카	ㅇ, ㅗ, ㅅ, ㅏ, ㅋ, ㅏ, ㅇㅗ, ㅗㅅ, ㅅㅏ, ㅏㅋ, ㅋㅏ, ㅇㅗㅅ, ㅗㅅㅏ, ㅅㅏㅋ, ㅏㅋㅏ
중국어	베이징	ㅂ, ㅔ, ㅇ, ㅣ, ㅈ, ㅣ, (ㅇ), ㅂㅔ, ㅔㅇ, …, ㅈㅣ, ㅣ(ㅇ), ㅂㅔㅇ, ㅔㅇㅣ, …, ㅈㅣ(ㅇ)
	상하이	ㅅ, ㅏ, (ㅇ), ㅎ, ㅏ, ㅇ, ㅣ, ㅅㅏ, ㅏ(ㅇ), (ㅇ)ㅎ, …, ㅇㅣ, ㅅㅏ(ㅇ), ㅏ(ㅇ)ㅎ, …, ㅏㅇㅣ
프랑스어	파리	ㅍ, ㅏ, ㄹ, ㅣ, ㅍㅏ, ㅏㄹ, ㄹㅣ, ㅍㅏㄹ, ㅏㄹㅣ
	몽마르뜨	ㅁ, ㅗ, (ㅇ), …, ㅡ, ㄸ, ㅡ, ㅁㅗ, ㅗ(ㅇ), (ㅇ)ㅁ, …, ㄸㅡ, ㅁㅗ(ㅇ), ㅗ(ㅇ)ㅁ, …, ㅡㄸㅡ

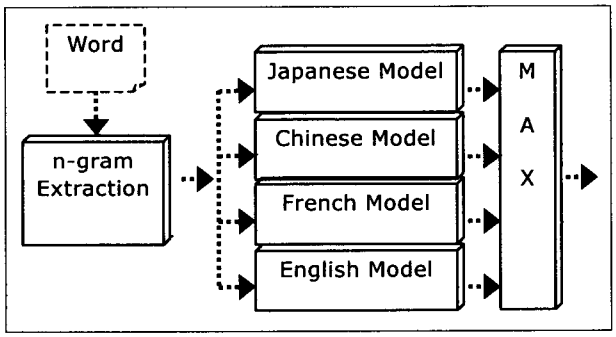

Fig. 2. SVM을 이용한 다중 언어 분류기.

단위가 되지는 않고, 또한 모든 단일 알파벳이 한 단위가 되지 않는다. 한글과 영어 알파벳과의 관계가 1 : 1 대응이 아니기 때문이다. 따라서 발음을 기준으로 그 범위를 정하게 되는데, 이때 정해진 한 단위를 발음단위(Pronunciation Unit)이라는 용어로 부른다(이재성, 1999).

따라서 학습 데이터로부터 확률값을 학습하기 위해서는 학습 데이터의 한글-영어 쌍을 발음단위별로 구분하여줄 필요가 있는데, 이 과정을 정렬(alignment)이라 한다. 이 정렬 과정은 단순히 자모단위로 이루어질 수가 있고, 복잡한 연산과정을 통해서 통계기반으로 이루어질 수가 있지만, 본 논문에서는 몇 가지의 간단한 규칙을 기반으로 정렬을 시도한다.

일단 한-영 철자의 정렬을 위해서 간단히 한-영 철자와 발음과의 관계를 따져볼 필요가 있다. 보통의 경우 'a' - '아', 'b' - 'ㅂ', 'c' - 'ㅅ'와 같이 한글 자모 하나에 영어 알파벳 하나씩 대응이 되고 있지만, 'a' - '에이', 'o' - '오우'와 같이 영어 알파벳 하나에 한글 모음이 2개가 대응되는 경우나 'ng' - 'ㅇ', 'th' - 'ㅅ'처럼 영어 알파벳 두개에 한글 자모 하나가 대응되는 경우가 있다. 또, 드물기는 하지만 '퀸'의 'uee' - 'ㄲ'처럼 영어 알파벳 3개가 한글 자모 하나에 대응되기도 한다. 이처럼 다양한 길이로 이루어지는 정렬과정도 몇 가지 규칙을 가지고 있는데, 발음을 기반으로 하는 단어정렬에서 모음이 자음과 대응되거나, 자음이 모음과 대응되지는 않는다는 것이 그 중에서 가장 중요한 규칙이다. 즉, 'a'가 한글 자음과 대응될 수 없고, 'b'는 한글 모음과는 대응될 수 없다.이를 이용하여 간단한 정렬규칙을 설정해 보았다.

아래의 알고리즘에서 영어 단어의 알파벳 순서는 E_i, 한글 단어의 자모 순서는 K_j로 하고, 영어 알파벳의 길이는 m, 한글 자모의 길이는 n로 한다. 또, 모음은 V, 자음은 C로 표시한다.

```
Loop (from i = j = 0 to i == m or j == n)
{
  if(Ei == V and Kj == V)
  {
    match(Ei , Kj)
    i++ and j++
  }
  else if(Ei == C and Kj == C)
  {
    match(Ei, Kj)
    i++ and j++
  }
  else
  {
    if(Ei == Ei-1) {
      append_to_array_E(Ei)
      i++
    }
    else if(Kj == Kj-1) {
      append_to_array_K(Kj)
      j++
    }
  }
}

if(i == m)
  assign_NULL(from Kj to Kn)
else if(j == n)
  assign_NULL(from Ei to Em)
```

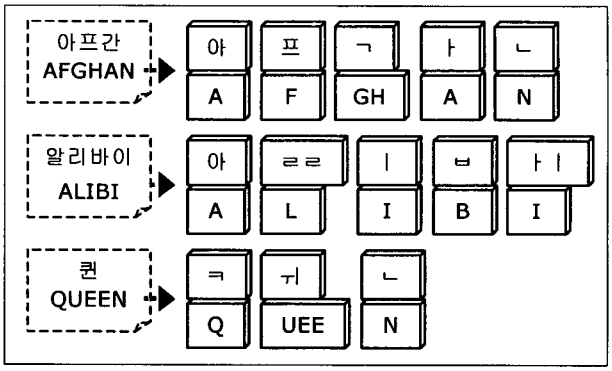

Fig. 3. 규칙기반의 단어정렬의 결과 예시.

위의 알고리즘에서 일반적으로 영어 알파벳 'y'와 같이 자음이지만 모음의 소리를 내는 것은 모음으로 취급한다. 그리고, 'er', 'or', 등과 같이 영어 알파벳의 자음과 모음이 하나로 뭉쳐서 '어', '오' 등의 한글 모음을 내는 경우가 있는데, 주로 '모음+r'의 경우가 많다. 이런 경우에 대해서는 하나의 모음묶음으로 처리를 해 주었다.

Fig. 3은 위의 규칙을 통한 정렬 결과의 몇 가지 예를 보여준다.

이와 같은 단순 규칙에 의한 정렬만으로도 사람이 직접 정렬한 결과의 97~98%에 달하는 정확률을 얻을 수 있었다.

3. HMM 기반의 외래어 복원 모델

한글 외래어의 철자 정보로부터 해당되는 영어단어를 생성해 내기 위해서 확률모델을 적용할 수 있다. 이때, 주어진 한글 외래어를 정렬한 결과가 K_1~K_n이라 하고, 올바른 영어 철자의 정렬결과가 E_1~E_m이라 할 때, $P(K)=P(K_1, K_2, \cdots, K_n)$, $P(E)=P(E_1, E_2, \cdots, E_m)$가 된다. 따라서 이것을 수식으로 정리하면 아래와 같다.

$$\arg\max_E P(E \mid K) = \arg\max_E P(E)P(K \mid E)$$

이는 정확한 영어 철자 순서를 찾기 위한 최적의 한글 정렬 결과를 구하는 원리이며, 이를 HMM에 적용하면 관찰열(한글정렬결과)을 최대로 나타내는 상태변화(정확한 영어철자)를 구하는 문제로 생각할 수 있다. 따라서, HMM에서 최적상태경로를 구하는 비터비(Viterbi) 알고리즘을 적용해서 주어진 문제를 해결할 수 있다. 이 문제를 도식화하면 다음의 Fig. 4가 된다.

4. 전체 시스템

위의 어원구분 모듈과, 단어정렬, 그리고 HMM 기반의 확률모델을 적용한 전체 시스템의 구조는 Fig. 5와 같다.

실험 및 결과분석

본 논문에서 외래어 어원 분류에 사용된 데이터는 세종전자사전의 2002년 까지의 구축 완료분에서 추출한 중국어, 일본어, 영어, 프랑스어 고유명사들을 대상으로 하였다. 본 사전에서 추출한 각 고유명사의 개수가 서로 일정하지 않기 때문에 각각 500개씩을 무작위 추출하여 사용하였다. 단, 사전에서 추출된 일본어 고유명사의 개수는 500개에 못 미치기 때문에, 웹에서 수작업을 통해 추출한 고유명사를 추가하였다.

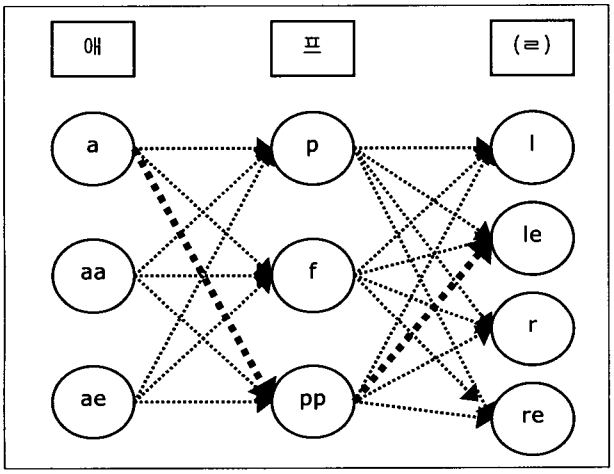

Fig. 4. HMM의 최적상태경로 문제에 적용한 외래어 복원 문제.

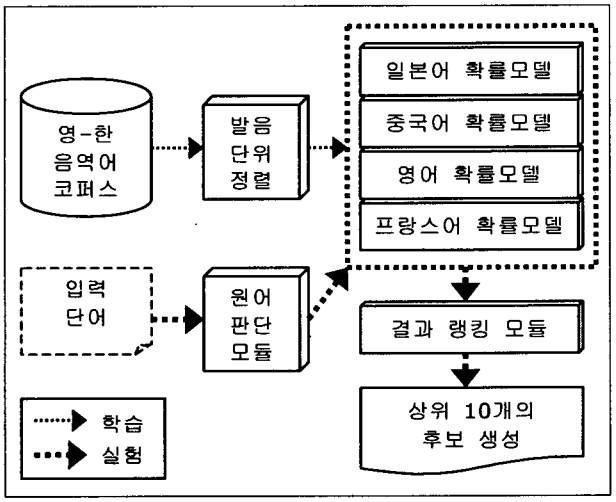

Fig. 5. 어원분류 모듈을 포함한 전체 시스템의 구조.

Table 3은 추출된 고유명사들을 대상으로 원어 판단 실험을 한 결과이다. 각 언어별로 450 단어를 학습에, 나머지 50 단어를 테스트에 사용하였다.

실험 결과에 따르면, 일본어의 경우 학습데이터와 실험데이터 모두에서 100%의 정확률을 보였다. 이는 한글 철자상의 일본어가 다른 언어와 구별되는 매우 뚜렷한 특징이 존재한다는 점을 보여주는데, 한글로 변환된 일본어에서 사용되는 자모가 다른 언어들에 비하여 매우 제한적인 점이 그 원인일 수 있다. 또한 프랑스어도 영어/중국어/일본어와 구별되는 나름의 특징이 있었다. 하지만 중국어와 영어의 경우 비교적 유사한 한글자모의 분포를 가지고 있기 때문에 이로 인해 분류성능이 낮아진 것으로 보이며, 이는 Table 4에서 볼 수 있는 오류의 예를 통해 확인할 수 있다.

오류들을 일단 살펴보면 단어내에서 받침이 존재하지 않거나, 'ㄴ'만 사용되었을 경우, 또한 모음이 복잡하지 않고, 'ㅏ', 'ㅣ' 등으로 구성되었을 경우, 거의 예외 없이 일본어로 판단하면서 문제가 발생되었다. 하지만, 중국어의 경우 '레이', '베이' 등의 철자들이 영어에서도 동일하게 발견되고, 이로 인해 영어와의 구분에서 문제가 발생함을 볼 수 있다. 따라서 성능 개선을 위해서는 중국어와 영어 사이에서 구별 가능한 새로운 철자상의 특징들을 찾아낼 필요가 있다.

위의 실험을 통해서 4개국 언어의 어원 판단에서 91.5%의 성능을 얻을 수 있었다. 하지만 현재의 모듈을 그대로 적용하여서 원어를 복원한다면, '미시간', '조지아', '고비노' 등은 일본어로 판단이 이루어져서, 일본어 모델로 변

Table 3. 어원 분류 실험결과

	영어	일본어	중국어	프랑스어	평균
학습 데이터	90	100	90	94	93.5
실험 데이터	86	100	86	94	91.5

Table 4. 어원 분류 실험의 오류 예

	영어	중국어	프랑스어
실험 데이터	미시간(J) 조지아(J)	마오레이(E) 허베이(E) 위린스(E)	고비노(J)

Table 5. 오류 단어에 대한 각 언어 모듈별 결과값

	영어	일본어	중국어	프랑스어
미시간	-0.84	-0.06	-0.53	-2.55
조지아	-0.66	-0.23	-1.609	-0.99
마오레이	-0.41	-1.08	-0.74	-1.69
허베이	-0.2	-1.08	-0.83	-2.09
위린스	0.527	-3.34	0.221	-1.79
고비노	-1.53	-0.06	-1.59	-0.23

Table 6. 외래어의 원어 복원 실험 결과

	학습 데이터	실험 데이터
어원판단 모듈 미포함	62.0 %	56.5 %
어원 판단 모듈 포함	73.0 %	69.0 %

환이 이루어지므로 영어나 혹은 프랑스어의 원어를 제대로 복원하는 것이 불가능해 진다. 따라서 위의 모델을 바로 적용하는 것보다, 일종의 가중치를 통해서 "일본어일 가능성이 높기는 높지만, 영어의 가능성도 배제할 수 없는 경우", 또는 "영어일 가능성이 높지만, 중국어일 가능성도 상당한 경우" 등에 대해서도 고려를 해 줄 수 있는 방법이 필요한데, 실제 실험에서도 최대값을 가지는 하나의 부류를 택하는 경우에 91.5% 이던 성능이 top1과 top2를 함께 고려함으로써 99%의 결과를 낼 수 있었다. 그리고, top1과 top2에 해당하는 결과를 가중치로 사용함으로써 좀 더 정확한 어원 판단이 이루어지게 하였다. Table 5는 오류가 난 6 단어들에 대한 각 언어 모듈들의 결과값을 정리하였다. 검게 처리된 칸은 각 데이터에서 top1, top2를 차지한 모듈을 나타낸다.

위의 결과를 바탕으로 최대값을 가지는 하나의 모듈로부터 단어를 생성하지 않고, top1과 top2의 확률차가 크지 않을 경우에는 top2 까지 고려한다. 그리고 가중치를 두는데 top1=1, top2=top2/top1 값을 가중치로 사용한다. 그리고 두 가지 언어 모델을 통하여 원어복원 후보를 각각 10개씩 만들고, 위의 가중치를 적용하여 20개의 후보들 중에서 10개를 선택하는 방식을 취한다.

결과후보에 대한 랭킹 시스템이 포함된 원어 복원 모델의 실험은 아래와 같이 진행하였다. 먼저 복원 대상이 될 단어는 원어 판단 시스템에서 이용된 세종사전에서 추출한 2,000 단어를 대상으로 하였고, 역시 각 언어별로 HMM 모델의 학습에 450 단어, 실험에 50 단어를 사용했다. 총 10개의 후보를 생성하는 실험에서 결과는 아래와 같았으며, 결과의 비교를 위해서 어원 판단 모듈을 포함하지 않는 단일 확률모델을 사용한 시스템을 함께 테스트 하였다.

Table 7. 각 언어별 복원 성능

	학습 데이터	실험 데이터
일본어	96%	96%
중국어	66%	60%
프랑스어	56%	52%
영어	74%	68%

Table 8. 단어정렬의 오류와 데이터 희귀성의 예

단 어	정렬 결과
나페트	ㄴ ㅏ ㅍ ㅔ ㅌ
Nappetit	n a pp e tit
듀	ㄷ ㅠ ?
Deux	d eu x

Table 6, 7에 따르면 언어분류를 하지 않았을 때 보다 약 13% 정도의 높은 성능 향상이 있었다. 특히 언어별 원어복원 결과를 살펴보면, 일본어의 경우에는 거의 100%에 가까운 복원 정확율을 보이는데, 이는 일본어의 발음을 영어 철자로 변환시에 사용되는 발음들과, 한글로 변환시에 사용되는 발음등에 큰 차이가 없기 때문일 것이다. 하지만, 프랑스어, 영어, 중국어 등에서는 한글에서 사용되지 않는 고유의 발음들이 매우 많고, 이러한 발음들을 한글로 변환하는 과정에서 변이형이 대량으로 발생할 수 있기 때문에 좋은 결과를 보여주지 못했다.

하나의 예로, '메사추세츠' - 'Messachusetts' 와 같은 단어들은 한글 각 자모에 해당되는 변이형이 매우 많은 관계로 제대로 영어로 판단하더라도 확률모델이 결정한 top 10 후보에 나타날 확률이 거의 없다. 또한 기존의 연구들에서 사용된 '데이터' - 'data', '컴퓨터' - 'computer' 와 같은 일반명사보다는 고유 명사들이 더 특정한 발음으로 많이 사용되는 경향이 있기 때문에 이로 인하여 전체 복원의 성능이 저하되었다.

본 실험을 통해 발견한 또 다른 문제는 한-영 단어의 정렬과정에서 정한 규칙들이 모든 언어에 일반적인 규칙이 될 수 없고, 영어에 적합한 규칙으로 이루어져 있기 때문에, 중국어나 프랑스어에서는 단어 정렬 과정에서 잘못 처리되거나, 확률모델에서 학습 데이터의 희귀성 문제 때문에 복원 성능을 떨어뜨리기도 하였다. Table 8은 그 예를 보여준다.

결 론

주어진 한글 외래어의 어원별 분류를 통해 원어를 복원하는 시스템은, 어원을 구분하지 않고, 하나의 확률모델로 처리할 때 보다 약 13% 정도의 성능향상이 있었다. 이는 기

존의 외래어 복원 연구에서 고려되지 않았던 부분이다. 물론, 기존의 연구들에서는 영어단어 위주의 변환이 대부분이었기 때문에 실험데이터의 특성이 다르므로 직접적인 실험결과의 비교는 타당하지 않다고 할 수 있다. 하지만, 영어 이외의 외국어 고유명사의 사용이 꾸준히 증가하고 있는 점을 감안해 본다면, 영어 이외의 외래어 복원에 관한 연구도 반드시 필요하다고 볼 수 있다.

REFERENCES

이재성(1998) : "다국어 정보검색을 위한 영-한 음차 표기 및 복원모델", 박사학위논문, 한국과학기술원

정길순, 맹성현(1998) : "외래어의 자동음역을 통한 영어단어 생성", *1998*년 한국정보과학회 춘계학술발표논문집(*B*), *pp429-431*

김병혜(1991) : "영 단어의 한글로의 자동변환", 석사학위논문, 서강대학교

김정재(1999) : "신경망을 이용한 발음단위 기반 자동 영-한 음차 표기 모델", 1999년 한국인지과학회 춘계 학술대회 발표논문집, *pp247-252*

이주호, 최기선, 이재성(2000) : "자동정렬을 통한 영한 복합어의 역어 추출", 제 12회 한글 및 한국어 정보처리 학술발표 논문집, *pp309-314*

이상율, 강인수, 나승훈, 이종혁(2003) : "음차표기된 외래어의 발음특성을 이용한 자동 영어단어 복원", *2003*년 한국정보과학회 춘계학술발표논문집(*B*), *pp525-527*

문화관광부(2000) : "국어의 로마자 표기법", 문화관광부 고시 제2000-8호

Jeong KS, Myaeng SH, Lee JS, Choi KS(1999) : *"Automatic identification and back-transliteration of foreign words for information retrieval", Information Processing and Management 35th, pp523-540*

Jung SY, Hong SL, Eunok Paek(2000) : *"An English to Korean Transliteration Model of Extended Markov Window", Coling 2000 Volume 1 : The 18th International Conference on Computational Linguistics, pp383-389*

Corredor-Ardoy C, Gauvain JL, Adda-Decker M, Lamel L(1997) : *"Language Identification with Language-independent Acoustic Models", Proceedings of Eurospeech '97*

Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Chung Tsai (1998) : *"Proper Name Translation in Cross-Language Information Retrieval", Coling-Acl '98 Volume 1 : The 17th International Conference Computational Linguistics, pp232-236*

Tompson P, Dozier C(1997) : *"Name Searching and Information Retrieval.", Proceedings of Second Conference on Empirical Methods in Natural Language Processing, Providence, Phode Island*

Hull DA, Grefenstette G(1996) : *"A dictionary-based approach to multilingual information retrieval", in Proc. of the 19th ACM SIGIR Conference, pp49-57*

Thorsten Joachims(1998) : *"Text Categorization with Support Vector Machines : Learning with Many Relevant Features", Proceedings of ECML '98*

Kevin Knight, Jonathan Graehl(1998) : *"Machine Transliteration", Computational Linguistics Dec. 1998, pp599-612*

Combrinck HP, Botha EC(1995) : *"Text-Based Automatic Language Identification", Proceedings of the sixth annual South Africa Workshop on Pattern Recognition '95*

Thorsten Joachims, "SVM^{light} : http://svmlight.joac hims.org/"