음차표기된 외래어의 발음특성을 이용한 자동 영어단어 복원

이상율[0] 강인수 나승훈 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터
(gilbert[0], dbaisk, nsh, jhlee)@postech.ac.kr

Automatic Back-Transliteration from Foreign Word to English Word

Sang-Yool Lee[0] 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)

요 약

음차 표기된 외래어의 원어 복원 문제에 있어서 확률모델을 이용한 방법들이 기존에 많이 사용되었다. 이는 '발음단위'개념(이재성 1998)을 이용하여 서로 대응될 수 있는 한글발음단위와 영어발음단위의 쌍들을 대역어 집합으로부터 추출하고 이를 확률모델에 적용하는 방법이다. 하지만 영어 철자를 영어 발음단위로 변환하는 과정에서 그 단어의 어원에 따라 서로 다른 발음상의 특징을 보이게 되는데, 이것이 기존의 연구에서 성능을 떨어뜨리는 원인이 되었다. 따라서 본 논문에서는 학습 데이터(대역어 집합)들을 발음 특성에 따라 분류하고, 분류된 각 데이터 집합을 학습과정에서 따로 적용함으로써 서로 다른 특성을 가지는 여러 개의 복원 모델을 얻을 수 있고, 이를 이용하여 원어 복원에 대한 성능을 높일 수 있음을 보여준다.

1. 서론

대량의 문서집합에서 원하는 정보가 포함된 문서들을 찾아내기 위해서는 각 문서에 특정 색인어가 포함되어 있는지의 여부를 확인하는 방법을 사용한다. 이 때 색인어와 문서내의 핵심어는 동일한 단어로 이루어져 있어야 한다. 그렇지만, 최근의 정보검색에 사용되는 문서들은 비록 한국어로 쓰여진 문서라고 하더라도 핵심 색인어는 원어를 그대로 사용하는 경우가 많이 있고[1], 그러한 경우에는 단순한 문자열 매칭 만으로는 원하는 결과를 얻을 수 없는 문제가 발생하게 된다.

이런 문제를 해결하기 위해서는 한글로 표기된 외래어들의 원어 표기를 알아내야 할 필요가 있다. 특히 원어로의 복원에서도 외래어를 영어 철자로 복원을 해 주는 문제가 가장 중요한 부분을 차지 하는데, 그 이유는 한글문서에서 영어가 핵심 색인어로 사용되는 비율이 다른 언어에 비해 압도적으로 많기 때문이다.

외래어의 영어표기 복원에 대해서 살펴보기 전에 먼저 영어단어가 어떤 방식을 거쳐서 한글로 표기되는 지를 알아둘 필요가 있다. 영어 단어의 한글 표기는 크게 2가지 방식으로 나뉘는데, 그 중 한가지는 단어의 철자를 보고 발음을 추측하여 표기하는 '눈말표기방식'이며 'digital'을 '디지탈'로 표현하는 것이 그 한 예가 된다. 다른 한가지 방법은 '입말표기방식'인데 이는 단어의 실제 발음을 기준으로 한글을 대응하는 방식이다. 'data'를 '데이타'로 표현하는 경우가 여기에 해당된다.

따라서 이러한 내용을 바탕으로 주어진 한글 외래어를 원래의 영어표현으로 자동 복원하는 시스템을 만들어 정보검색의 질의어 확장에 사용한다면 기존의 문서 검색 시스템의 성능에 많은 향상이 있을 것으로 기대된다.

2. 기존연구

한글로 표기된 외래어로부터 영어단어를 복원하는 방법에 대한 기존 연구는 크게 규칙기반 방식과 통계기반 방식으로 나뉜다.

먼저 규칙기반 방식에는 국어의 로마자 표기법을 들 수 있다. 하지만 이 방법은 영어에서 온 외래어를 그 대상으로 하기보다 우리말로 되어 있는 고유의 지명, 이름 등을 소리 그대로 로마자로 표현하는 방법이기 때문에 외래어의 복원 문제에는 적합하지 않다.

다음으로 김병혁(1991)의 연구에서는 주어진 영어철자를 한국어로 변환함에 있어서 미리 작성된 규칙을 기반으로 영어철자를 발음기호로 변환하고 이 발음 기호를 외래어 표기법을 이용하여 한국어를 생성해 내는 방식을 사용하였다. 이는 일본의 Yuichi(1990)에서 영어철자를 규칙에 기반하여 발음기호로 변환한 것과 유사하다. 그러나 기존의 규칙기반 방법은 미리 정해진 규칙에 의존하여 발음을 생성하기 때문에 발음의 여러 변이들에 대한 효과적인 표현이 힘들다는 단점이 존재한다. 따라서 이를 보완하기 위한 연구로 SERI(1995)에서는 한글모음에 해당하는 자소를 다양하게 변화시키는 방법을 제안하기도 하였으나 불필요하게 많은 변이체를 생성하여 그리 효과적인 방법이 되지 못했다.

통계기반 방식의 기존연구중에서 정길순(1998)은 외래어의 복원을 위해서 확률모델과 신경망에 기반한 방법을 사용했다. 신경망의 입력으로는 외래어의 한글 음소가 되고, 출력은 변환된 영어 철자들이 된다. 그리고 신경망을 통해서 만들어진 결과를 사전을 이용하여 후처리 함으로써 성능이 향상될 수 있음을 보였다. 이와 유사하게 신경망을 이용하지만 입출력에 사용되는 단위로 발음단위 개념을 이용한 김정재(1999)의 연구가 있다. 김정재(1999)에서는 한-영 발음단위를 보다 실제 활용에 가깝게 재현하기 위해서 수동으로 직접 묶어주는 방법을 사용했다. 그 외 확률모델과 발음단위의 자동정렬을 이용한 이재성(1998)의 방법이 있었는데, 정렬된 발음 단위를 가지고 외래어에서 직접 영어철자로 변환하는 방식과 영어철자의 표준 발음을 추출하고 이를 외래어 표기법에 적용하는 2가지 방식을 사용하여 눈말표기와 입말표기 둘 다를 고려했다.

기존의 연구결과를 살펴보면 규칙에 기반한 방법보다는 확률기반의 방법이 좀 더 나은 성능을 보이며, 철자 중심의 눈말표기와 발음 중심의 입말표기를 함께 고려할 때 더 좋은 결과를 얻을 수 있음을 알 수 있다.

또한 실제 한글 문서에 쓰이는 외래어 중에는 'doeuvre(되브르)', 'huguenot(위그노)' 등과 같은 영어가 아닌 다른 외국어에서 변환된 단어들이 다수 존재하는데, 기존의 영어발음에 기반한 모델에서 성능을 떨어뜨리는 원인이 되어왔다.

따라서 본 논문에서는 어원을 구분하는 효과를 얻기 위해서 학습 데이터들을 발음 특성에 따라 분류하고, 분류된 각각의 데이터들을 학습과정에 개별적으로 이용하여 서로 다른 특성을 갖는 분류모델을 만드는 방법을 제안한다.

[1] 권윤형(1996)에 따르면 KTSET에서 임의 추출한 문서내의 색인어중 26% 가 외래어 및 외국어로 이루어져 있다.

3. 제안모델

3.1 규칙 기반의 발음단위 자동정렬

'발음단위'개념은 통계기반 방법을 이용한 연구에서 자주 사용되었는데, 정길순(1998)에서는 한글 자소에 대응하는 영어 알파벳의 집합을 사용해서 발음단위를 비교적 간단하게 구성하였다. 이 방식은 아래와 같이 '바나나(banana)' 의 경우에는 아무 문제가 없지만, '데이타(data)'와 같은 경우에는 적절하게 대응되지 못하는 한글 자소가 발생하게 된다.

b- ㅂ, a- ㅏ, n- ㄴ, a- ㅏ, n- ㄴ, a- ㅏ (o)
d- ㄷ, a- ㅔ, ?- 이, t- ㅌ, a- ㅏ (x)

반면에 김정재(1999)에서는 사람이 직관적으로 판단하여 -tion(션), non-(넌) 과 같이 최대한 큰 묶음으로 발음단위를 구성한다. 이 방법은 보다 정확한 발음 분석을 하고 있기는 하지만, 자동화하기가 어렵다는 단점을 가진다.

이재성(1998)에서는 자동정렬 방법을 사용하는데, 이는 실제 매우 유용하지만 정렬을 위하여 거리계한 등의 복잡한 정렬알고리즘을 사용한다. 따라서 본 논문에서는 보다 간단한 방법을 이용하여 자동 정렬을 수행하기 위하여 [그림 1]과 같은 규칙을 사용한다.

- 규칙 1 : 한글 자음은 영어 자음과 대응된다.
- 규칙 2 : 한글 모음은 영어 모음과 대응된다.
- 규칙 3 : 'y', 'w' 등 실제 자음이지만 한글에서 모음 소리를 내는 철자는 모음과 대응된다.
- 규칙 4 : 단어의 마지막에 남는 자음이나 모음은 NULL에 대응한다.
- 규칙 5 : 대응이 안되는 자음은 바로 앞의 자음과 결합하고, 대응이 안되는 모음은 바로 앞의 모음과 결합한다.
- 규칙 6 : 서로 대응되는 자음이라 하더라도, 연속하여 나타날 경우 중간에 적절한 분리를 해야 할 필요성이 있는지를 검사한다. 이때 대응 가능한 관계들을 미리 정의해 둔다.

[그림 1] 자동정렬 규칙

위의 규칙을 기반으로 각 단어에 대한 정렬을 시도한 예는 다음과 같다.

시작 -> d a t a
시작 -> ㄷ ㅔ 이 ㅌ ㅏ

a. 먼저 'd' 와 'ㄷ'는 모두 자음이므로 대응될 수 있다.
b. 다음으로 'a'와 'ㅔ'는 모두 모음이므로 대응될 수 있다.
c. 그 다음 't'와 '이'는 자음과 모음의 관계이므로 대응이 불가능하다. 따라서 '이' 를 바로 앞의 'ㅔ'와 합쳐서 'ㅔ이' 라는 발음 단위를 만든다. 이때, 't'는 앞의 단어가 모음이므로 결합할 수 없다.
d. 그 다음은 't'와 'ㅌ'를 대응시키고, 'a'와 'ㅏ'를 대응시키면 된다.
e. 정렬된 결과는 아래와 같이 나타난다.

d a t a
ㄷ ㅔ이 ㅌ ㅏ

다른 하나의 예로 'cream(크림)'을 정렬시키면 다음과 같다.

c r ea m
ㅋ ㄹ ㅣ ㅁ

실제 위의 자동정렬 규칙을 적용하여 한-영 대역어 코퍼스로부터 정렬을 시도하고 개별 발음단위에 대해서 정확률을 측정하였더니 사람이 수동 정렬한 결과의 97% - 98%에 근접하는 좋은 결과를 얻었다.

3.2 HMM을 이용한 변환모델

실험에 사용할 변환 모델로 신경망, HMM 확률모델 등의 여러 가지가 있을 수 있는데 본 논문에서는 학습 데이터의 구분에 따라 하나의 확률모델을 다양하게 학습을 시도하여 동일한 모델이 여러 특성을 갖도록 만드는 것이 목적이기 때문에 특정 확률모델에 구애 받지 않지만, 이번 실험에서는 확률모델로 HMM을 사용한다.

3.3 발음 특성에 따른 학습 데이터의 분류

기존의 연구들에서 알려진 바와 같이 외래어의 영어복원(back-transliteration) 혹은 외래어로의 변환(transliteration)에서 실제 어원이 영어가 아닌 외래어 데이터들은 영어의 발음과는 다른 특성을 보이며, 이로 인해서 전체적인 복원 성능에 나쁜 영향을 끼쳐왔다.

이러한 문제를 해결하기 위해서는 해당 언어의 어원에 따른 발음 특성을 감안하여 서로 다른 처리를 해 주는 것이 더 좋은 결과를 낸다고 알려져 있다. 하지만 현실적으로 외래어의 단어를 바탕으로 어원을 찾아내는 것이 매우 어렵고, 우리가 알고 싶은 것은 외래어의 어원 정보 라기보다 어원에 따른 발음 특성이므로, 한글-영어 철자 쌍으로 이루어진 학습데이터를 통하여 비교적 유사한 발음 특성을 가지는 것들을 분류하여 활용함으로써 어원에 기반한 외래어 복원과 유사한 환경을 만들고자 하였다. 이 작업을 위해서 일단 정렬된 데이터로부터 얻어낸 영어와 한글 발음단위들을 발음 특성에 따라 구분해야 하는데 [그림 2]와 같은 분류규칙을 사용한다.

- 규칙 1 : 눈말표기로 이루어진 데이터를 찾아내는 규칙
  주어진 영어단어의 모든 발음단위가 하나의 철자로 이루어져있고 각 철자에 대응되는 모든 한글 발음들이 눈말표기에서 사용되는 음으로만 이루어져 있는지를 확인한다. 아래의 예에서 '아다지오'의 경우는 눈말표기로 이루어진 데이터로 분류되고, '베이스'의 경우는 a(ㅔ이) 조합이 눈말표기라 보기 힘들기 때문에 눈말표기가 아닌 데이터로 분류될 수 있다.
  예) a(아) - d(ㄷ) - a(ㅏ) - g(ㅈ) - i(ㅣ) - o(오) (o)
  b(ㅂ) - a(ㅔ이) - se(스) (x)

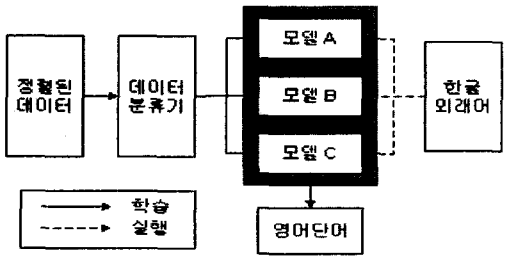
- 규칙 2 : 규칙 1에서 분류되지 않은 데이터들 중에서 영어의 표준 발음을 따르는 경우와 영어의 발음 특성을 가지지 않는, 즉 어원이 영어가 아닐것으로 예측되는 단어를 구분해내기 위해서 각 단어의 정렬된 데이터로부터 영어발음이 아닐것으로 예상되는 음이 하나라도 존재하는지를 확인한다. 아래의 예시 중에서 '에이스'는 입말표기 데이터로 분류되지만, '되브르'의 경우는 eou(ㅢ) 조합이 영어에서 왔다고 보기 힘들기 때문에 영어가 아닌 외래어로 분류된다.
  예) a(에이) - ce(스) (o)
  d(ㄷ) - eou(ㅢ) - v(브) - re(르) (x)

[그림 2] 발음 특성에 따른 분류 규칙

위의 규칙 1을 통해서 눈말표기라 예상되는 데이터를 따로 분류하고, 규칙 2를 통해서 나머지 데이터들 중 외래어라 예상 되는 것들을 분류할 수 있다. 그리고 각 분류된 데이터를 이용한 학습으로 서로 다른 특성을 갖는 3개의 변환 모델을 생성할 수 있는데, 그 각각은 다음과 같다.

- 1 : 눈말표기된 단어집합을 학습데이터로 사용한 모델
- 2 : 입말표기된 단어집합을 학습데이터로 사용한 모델
- 3 : 영어의 발음특성을 가지지 않는 단어집합을 학습데이터로 사용한 모델

전체적인 시스템은 아래의 [그림 3]에 잘 나타나 있다.



[그림 3] 전체적인 시스템

4. 실험 및 결과분석

실험을 위해서 5만 단어의 국어 사전에서 외래어 음차표기된 2,800 단어를 추출하고 여기에 적절한 영어단어를 수동으로 붙여준 데이터를 사용했다. 이 단어들 중에서 'art nouveau - 아르 누보', 'dal segno - 달 세뇨' 등과 같이 프랑스어가 어원인 단어들 중에 한글, 영어가 두 단어가 묶여서 표현되는 경우는 각각 분리하여 'art - 아르', 'nouveau - 누보' 와 같은 형태로 변형시켜서 학습데이터로 활용하였는데, 이는 비록 한글표기가 '아르누보'처럼 하나의 단어로 표현되더라도 실제 원어가 두개의 단어로 이루어져 있기 때문에 하나로 붙일 경우 실제의 발음과 전혀 다른 형태의 '철자-발음' 쌍을 가질 가능성이 크기 때문이다.

또한 각각의 학습데이터에 대한 HMM모델의 구성을 간편하게 하기 위해서 미리 정렬된 학습데이터로부터 모든 표현 가능한 영어발음단위와 한글발음단위의 수를 파악하여 이를 은닉상태집합과 관찰가능상태집합의 고정값으로 사용하였다. 이 작업을 통하여 구해진 영어발음단위의 총 개수는 142개 였으며, 한글발음단위의 총 개수는 65개 였다.

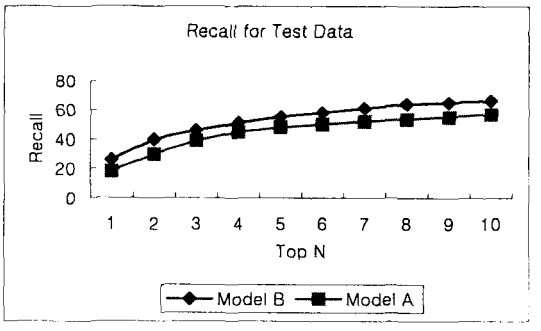
실험은 주어진 한글 외래어를 입력값으로 하여 학습된 3개의 모델에 각각 적용하고 확률값이 가장 높은 10개의 결과들로부터 원하는 영어단어를 얻을 수 있는지를 알아보았다. 이를 위해 전체 2,800개의 대역어 데이터들 중 2,500 단어를 학습에 사용하고 나머지 300 단어를 실험에 사용하였다. 또 성능의 비교를 위해 전체 학습데이터를 발음특성에 따른 구분을 하지 않고 학습시켜서 얻은 모델을 함께 테스트하였다. 실험의 결과는 아래 [표 1]에 나타난다.

[표 1] HMM을 이용한 외래어 복원 실험 결과

	학습 데이터	미학습 데이터
데이터를 분류하지 않은 모델 Model A	56.8 %	51.1 %
데이터를 분류한 모델 Model B	66.2 %	60.3 %

[표 1]에서 나타난 바와 같이 외래어를 정렬과정에서 발음특성에 따라 분류하였을 경우에 그렇지 않은 경우보다 원어복원의 성능이 뛰어남을 알 수 있다.

아래의 [그림 4]에서는 미학습데이터에 대하여 복원 실험을 진행하고 확률값이 가장 좋은 1개만 택했을 경우부터 순차적으로 1씩 증가시키면서 10개를 택했을 경우까지의 성능의 향상을 알아보았다. 실제 실험은 top 20 까지도 조금씩 꾸준히 증가하였으나 값의 큰 변화를 보여주지는 않았기 때문에 그 결과는 아래의 그림에서 제외하였다.



[그림 4] 실험 결과

[표 2]는 전체 데이터를 발음 특성에 따라 3가지로 분류 하였을 때 각각의 비율과 개별 변환모델의 성능을 보여준다. 여기서 확인할 수 있는 것은 눈말표기, 즉 철자위주의 표기일 경우 영어 철자와 한국어 철자 사이의 많은 변화가 없기 때문에 확률모델에서 좋은 결과를 얻을 수 있으며, 입말표기의 경우에는 눈말표기의 경우와 비교하여 한글 철자로부터 생성될 수 있는 영어 발음 및 철자의 종류가 매우 다양하고 심지어는 '알콜' - 'alcohol'과 같이 영어 'h'의 복원을 위해서 필요한 한글 철자상의 정보가 없는 경우도 발생한다. 따라서 눈말표기 방식의 데이터에 비해서 성능이 낮게 나타나는 이유가 된다. 또 하나 [표 2]에서 알 수 있는 것은 영어발음이 아닌 외래어의 표기일 경우에 다른 모델과 비교하여 복원성능이 좋지 못한데, 이는 프랑스, 러시아, 독일어 등 서로 발음상의 연관성이 없는 단어들을 하나의 모델로 사용했기 때문에 나타난 결과라 보여지며, 따라서 발음규칙에 따라 학습 데이터를 더욱 세분화 하고, 복원 모델들을 각각 만든다면 더 좋은 성능을 낼 것이라 생각된다.

[표 2] 개별 변환모델의 성능 비교

	데이터의 수	복원 정확률 (%)
영어 철자 위주의 데이터	41.6 %	81.6 %
영어 발음 위주의 데이터	53.0 %	47.8 %
영어가 아닌 발음의 데이터	5.3 %	18.8 %

5. 결론 및 향후연구

통계기반에서 음차표기된 외래어의 복원 모델을 만들 때, 각 단어들의 발음 특성의 차이로 인하여 발생하는 오류를 줄이기 위해 학습 데이터를 발음 특성에 따라 분류하고 분류된 데이터들을 개별적으로 학습모델에 적용함으로써 서로 다른 특성을 갖는 여러 개의 복원모델을 만들 수 있었고, 이런 방법으로 전체 외래어 복원의 성능을 향상시킬 수 있었다.

특히 음악용어 'andante(안단테)', moderato(모데라토)' 등의 이탈리아어를 어원으로 하는 단어들의 경우 매우 좋은 복원성능을 보였는데, 이것은 눈말표기에 가까운 방식으로 단어가 구성되어 있기 때문으로 보여진다.

그러나 'huguenot(위그노)', fohn(푄)' 등의 영어의 발음 특성과 큰 차이를 보이는 단어들은 비록 영어와 분리하여 개별적인 모델로 처리를 하고자 하였으나 기대한 만큼의 성능을 보이지 못하고 전체적인 성능을 많이 떨어뜨리는 요인으로 작용하였다. 이는 분류된 데이터들이 프랑스를 어원으로 갖는 말들만 존재하는 것이 아니라 '불세비키'와 같은 러시아 혹은 기타의 다양한 언어들을 세밀하게 분류하지 못함으로써 발생하는 문제이다. 따라서 이러한 문제를 개선하기 위해서는 보다 정교하게 발음특성별 분류를 시도해야 할 것으로 보인다. 향후에는 언어별 발음특성을 좀 더 깊이 연구하여 보다 정확하고 세밀한 분류규칙을 찾아내는 연구가 필요하다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

참고문헌

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

[2] 정길순, 맹성현 (1998), "외래어의 자동음역을 통한 영어단어 생성", 1998년 한국정보과학회 춘계학술발표논문집(B) pp. 429-431

[3] 김병혜 (1991), "영 단어의 한글로의 자동변환", 석사학위논문, 서강대학교

[4] 김정재 (1999), "신경망을 이용한 발음단위 기반 자동 영-한 음차 표기 모델", 1999년 한국인지과학회 춘계 학술대회 발표논문집 pp.247-252

[5] 이주호, 최기선, 이재성 (2000), "자동정렬을 통한 영한 복합어의 역어 추출", 제 12회 한글 및 한국어 정보처리 학술발표 논문집 pp. 309-314

[6] 문화관광부 (2000), "국어의 로마자 표기법", 문화관광부 고시 제2000-8호

[7] 문화체육부(1995), "외래어 표기법", 문화체육부 고시 제1995-8호

[8] K.S. Jeong, S.H. Myaeng, J.S. Lee, K.S. Choi (1999), "Automatic identification and back-transliteration of foreign words for information retrieval", Information Processing and Management 35th pp. 523-540

[9] S.Y. Jung, S.L. Hong, Eunok Paek (2000), "An English to Korean Transliteration Model of Extended Markov Window", Coling 2000 Volume 1: The 18th International Conference on Computational Linguistics pp. 383-389