한-영 자동 음차 복원

강병주, 최기선
한국과학기술원 전산학과
전문용어언어공학연구센터
대전시 유성구 구성동 373-1, 우:305-701
{bjkang, kschoi}@world.kaist.ac.kr

Automatic Korean-English Back-Transliteration

Byung-Ju Kang, Key-Sun Choi
Department of Computer Science
Korea Terminology Center for Language and Knowledge Engineering
Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu, Taejon, 305-701

요약

최근 다국어 정보검색, 기계번역 등과 관련하여 자동 음차 표기 및 복원에 대한 필요성이 증대되고 있다. 특히 영어와 한국어 같이 그 음운구조의 차이가 큰 언어 쌍인 경우에는 간단한 문제가 아니다. 더구나 외래어를 영어로 복원하는 것은 표기의 경우보다 훨씬 어렵다. 본 논문에서는 결정트리 학습을 통한 한/영 자동 음차 복원 방법을 제안하고 기존의 방법 및 로마자 표기법에 기반한 방법에 비교하여 매우 정확하게 복원이 가능함을 보인다.

1 서론

음차 표기 (transliteration)는 외국어의 발음을 한국어로 표기하는 것을 말하며 음차 복원 (back-transliteration)이란 그 역으로 음차 표기된 외국어로부터 원어의 철자를 복원하는 것을 의미한다. 예를 들어 영어 "internet"의 음차표기는 "인터넷"이고, "인터넷"의 올바른 음차복원은 "internet"이다 (그림 1).

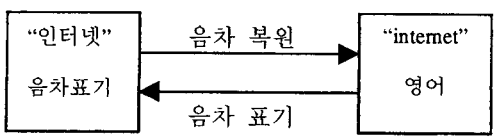

그림 1. 음차 표기와 복원

음차 표기 및 복원에 대한 연구는 주로 다국어 정보검색[4], 기계번역[7], 그리고 다양한 음차표기 및 영어의 혼용으로 인한 단어불일치 문제해결[6] 등과 관련하여 주로 이루어져 왔다.

다국어 정보검색이나 기계번역에서 특히 음차복원은 매우 중요한데 한-영 번역의 경우 최근에 수입된 외래어나 음차 표기된 이름의 경우 사전에 없는 경우가 많아 이를 영어로 번역하기 위해서는 자동 음차 복원이 필수적이다. 또한 한국어 정보검색에서도 음차 복원이 유용하게 사용될 수 있다. 한국어에서는 같은 영어에 대해서 다양한 음차표기가 사용되고 있고 심지어 영어 자체가 그대로 사용되기도 한다. 이렇게 같은 개념에 대해 다양한 표층표현이 존재함에 따라 정보검색에서 심각한 단어불일치 문제를 야기하게 된다. 따라서 이들 다양한 표기형태를 대표하는 표준형태로 색인하는 것이 필요하다. 다양한 음차표기들과 영어의 혼용일 경우 영어를 표준형태로 하는 것이 바람직하고 이를 위해서는 음차표기를 영어로 복원하는 것이 필요하다[6].

자동으로 음차 표기하는 문제도 쉬운 일이 아니지만 음차 복원은 더욱더 어려운 문제이다. 음차 표기 과정에서 원시어 (영어)의 음운정보가 상실되기 때문에 이 잃어버린 정보 없이 완벽한 복원은 원리적으로 불가능하다. 따라서 일반적으로 자동 음차 복원에는 사전[4, 6] 및 언어모델 (language model)[7]과 같은 정보들이 이용된다.

음차 복원의 또 다른 어려움 중의 하나는 다중단어 문제이다. 영어의 다중단어가 한국어로 음차 표기될 경우 일반적으로 각각의 음차표기를 함께 붙여 쓰는 경향이 있다. 예를 들어 "Web server"는 "웹서버"라고 음차 표기된다. 따라서 음차 표기된 외국어를 복원할 경우 다중단어일 가능성을 고려하여야 할 것이다. 일본어의 경우도 마찬가지 어려움이 있고 Knight와 Graehl [7]은 영어 다중단어를 고려하여 일-영 확률적 복원모델을 만들었다. 하지만 이 때문에 정확도의 저하를 감수해야만 하였다. 본 논문에서는 다중단어는 고려하지 않는다.

한국어 음차표기에는 입말표기와 눈말표기가 혼재되어 있다[1, 4]. 입말표기는 원어의 발음으로 표기하는 것을, 그리고 눈말표기는 원어의 철자를 보고 표기하는 것을 의미한다. 예를 들어 영어 "radio"의 음차표기 "레이디오"는 입말표기이고 "라디오"는 눈말표기이다[8]. 따라서 음차표기와 영어 쌍에서 직접적으로 규칙을 학습하는 방법이 효과적일 수 있다라는 근거가 될 수 있다[4, 8]. 본 논문에서도 중간적인 음성적 표현을 거치지 않고 음차표기에서 영어로 직접 대응시키는 접근 방법을 사용한다.

본 논문에서는 영어 단어와 그 대응 한국어 음차표기 쌍들에서 한글 자소의 영어 스트링으로의 대응규칙을 자동으로 학습하는 방법을 제안한다. 개략적인 방법은 다음과 같다. 먼저 학습데이터로 각 한글 자소와 영어 스트링의 대응 실례가 많이 필요하다. 이를 위해서는 음소 단위로 정렬된 한국어-영어 단어쌍들이 필요한데 이는 일반적으로 쉽게 구할 수 없다. 따라서 우리는 자동 정렬 방법을 고안하였다. 자동 정렬 알고리즘에 대해서는 다음 장에서 설명한다. 일단 음소수준에서 정렬된 한국어-영어 단어쌍으로부터 충분한 학습데이터가 구축되면 각 자소별로 결정트리 (decision tree)를 만든다. 자음의 경우 초성과 종성을 구분하였기 때문에 총 46 (초성:18[1], 종성: 7[2], 중성: 21) 개의 결정트리가 만들어 진다 (표 1). 새로운 한글 음차표기의 복원 과정은 매우 간단하다. 각 자소별로 대응 영어 스트링을 결정트리로부터 구하고 이들을 하나의 스트링으로 연결하면 된다. 이렇게 구해진 영어 알파벳 스트링이 올바른 영어 단어를 구성할 가능성은 적다. 따라서 후처리로 영어 사전을 탐색하여 가장 비슷한 철자의 영어 단어를 해답으로 결정한다 (그림 2). 하지만 본 논문에서는 후처리 부분은 고려하지 않는다.

[1] 초성은 19개이지만 'ㅇ'은 음가가 없기 때문에 대응 영어 음소가 존재하지 않는다.

[2] 외래어표기법 (문화체육부 고시 제1995-8호, 1995년 3월 16일)에 따라 음차표기의 받침으로 7개의 자음만을 허용한다.

표 1. 결정트리를 구축할 46개 자소

초성 자음 (18)	ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ (ㅇ) ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ
종성 자음 (7)	ㄱ* ㄴ* ㄹ* ㅁ* ㅂ* ㅅ*
모음 (21)	ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ

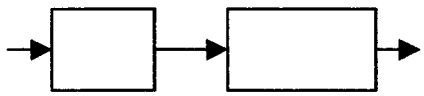

그림 2. 음차 복원 과정

2 단어 정렬

임의의 한국어 음차표기와 영어가 주어졌을 때 약간의 두 언어에 대한 음운적 지식을 가진 사람이라면 거의 완벽하게 영어 알파벳과 한글 자소 사이의 대응관계를 결정할 수 있다. 이 것은 대부분의 경우 정확한 대응에 필요한 충분한 음운적 단서가 존재하기 때문이다. 본 정렬 알고리즘은 다음과 같은 휴리스틱을 이용한다:

(1) 각 한글 자소에는 일반적으로 대응되기 쉬운 영어 알파벳들이 존재한다. 예를 들어 'ㅂ'은 'b', 'p', 'v'로 대응된다.
(2) 자음은 자음, 모음은 모음과 대응되는 경향이 있다.

이 두 가지 휴리스틱은 임의의 한글 자소와 영어 알파벳 사이의 벌점 테이블로 실현된다. 모든 가능한 정렬을 구하고, 각 정렬에 대해 총벌점을 계산한 후, 가장 적은 벌점을 가지는 정렬을 선택한다. 다음은 이렇게 구해진 "보드"와 "board"의 정렬 예이다.

한국어	ㅂ	ㅗ	-	-	ㄷ	ㅡ
	|	|	|	|	|	|
영어	b	o	a	r	d	-

하지만 이렇게 구해진 일대일 정렬은 한국어 쪽에 null이 있어 결정트리 학습 방법의 적용을 곤란하게 만든다. 따라서 다음과 같이 한국어 쪽에 null을 허용하지 않고 일대일 대응을 일대다 대응으로 변환한다.

한국어	ㅂ	ㅗ	ㄷ	ㅡ
	|	|	|	|
영어	b	oar	d	-

이렇게 되면 각 한글 자소에 대해 결정트리를 만들면 되므로 모두 46개의 결정트리가 만들어지게 된다.

1,688개의 영어/외래어 쌍에 대해 평가한 결과 10개만 틀리게 정렬되어 약 99.4%의 매우 높은 정확도를 보였다.

3 결정트리 학습

일단 정렬된 한국어 음차표기와 영어 쌍이 만들어지면 여기서 결정트리 학습에 필요한 학습데이터를 만드는 일은 매우 간단하다. 각 음차표기-영어 대응 쌍에서 한글 자소 수 만큼의 대응 예 (example)가 만들어 진다. "보드"와 "board"의 예를 다시 들면 다음과 같은 4개의 학습데이터가 만들어 진다.

L3	L2	L1	(K)	R1	R2	R3		E
<	<	<	(ㅂ)	ㅗ	ㄷ	ㅡ	→	b
<	<	ㅂ	(ㅗ)	ㄷ	ㅡ	>	→	oar
<	ㅂ	ㅗ	(ㄷ)	ㅡ	>	>	→	d
ㅂ	ㅗ	ㄷ	(ㅡ)	>	>	>	→	-

여기서 '<'과 '>'는 단어의 처음과 끝을 나타내는 특수 문자이다. 각 예는 학습 한글 자소의 왼쪽, 오른쪽 3개씩의 자소를 자질 (attribute)로 하는 자질 벡터로 구성된다. 그리고 각 예에 대한 클래스 레이블은 대응 영어 스트링이 된다.

결정트리 학습에는 Quinlan의 ID3 알고리즘[13]과 유사한 방법이 사용된다. ID3는 단순한 greedy 하향식 (top-down)의 결정트리 구축 알고리즘으로 루트 (root)에서 시작하여 재귀적 (recursively)으로 트리가 구축된다. 각 노드에서 분기할 자질이 선택되고 그 노드의 예들은 자질의 값에 따라 분할된다. 모든 예가 같은 클래스 값을 가지거나 더 이상 테스트할 자질이 없을 때 까지 분기과정이 반복된다. 분기할 자질의 선택은 클래스 변수를 C, 자질 변수를 A라고 할 때 상호정보 (mutual information) $I(A;C)$가 최대로 되는 자질을 선택한다. 이 상호정보를 *정보이득* (*information gain*)이라고 부른다. 우리는 정보이득 대신에 보다 효과적인 도메인 지식에 기반한 자질 선택 방법을 사용한다. 자세한 내용은 다음 절에서 설명한다.

이렇게 각 한글 자소에 대해 독립적으로 결정트리가 만들어지면 학습이 완료된다. 새로운 한국어 음차표기의 복원은 매우 간단하다. 먼저 각 자소 단위로 풀어쓰고 각 자소에 대해 대응 영어 스트링을 결정트리를 이용해 구하고 이들 영어 스트링을 단순히 연결하면 된다.

4 실험

4.1 학습 및 평가 데이터

실험 데이터로 7,000개의 외래어-영어 쌍을 준비하였다. 이 데이터는 남영신의 외래어사전 [2]에서 다중단어, 약어, 특수기호를 포함한 단어를 제외하고 추출한 것이다. 여기에는 독어, 불어, 이탈리아어 등, 여러 어원의 알파벳 단어가 다수 포함되어 있다. 이들 단어들은 노이즈로 작용하여 학습의 효과를 떨어뜨릴 가능성이 높다. 이 7,000개의 단어 쌍에서 1,000개를 무작위로 추출하여 평가데이터로 사용하고 나머지 6,000개를 학습데이터로 사용하였다.

4.2 평가 방법

복원 정확도는 전체 단어수에 대한 정확하게 복원된 영어 단어수의 백분률로 측정된다. 이를 *단어정확도*라고 부르는데 부분적으로 일치하는 경우를 고려하지 않기 때문에 정밀한 평가에는 문제가 있다. 따라서 출력된 영어 스트링과 올바른 영어 단어사이의 편집거리 (edit distance)를 계산하는 *글자정확도*를 도입한다[4]. 따라서 글자정확도는 복원되어야 할 정확한 영어단어에 얼마나 가까운지 근사도를 평가한다. 또한 각 자소별로 결정트리가 얼마나 잘 학습되었는지 평가하기 위하여 *자소정확도*를 평가한다.

$$\text{단어정확도} = \frac{\text{정확하게 복원된 단어의 수}}{\text{총 단어의 수}}$$

$$\text{글자정확도} = \frac{L-(i+d+s)}{L}$$

$$\text{자소정확도} = \frac{\text{정확하게 복원된 자소의 수}}{\text{총 자소의 수}}$$

여기서 L은 정확한 영어 단어의 길이이고, i, d, s는 각각 삽입, 삭제, 치환의 수이다. 분자가 음수이면 ($L < (i + d + s)$) 전체 글자정확도를 0으로 간주한다.

단어정확도는 단어를 구성하는 각 자소들의 자소정확도를 곱한 값과 비슷한 값을 가져야 하지만 실제 값은 자소의 빈도 분포에 따라 조금씩 달라질 수 있다. 그리고 글자정확도와 자소정확도의 상관관계는 분명하지 않지만 실험결과를 보면 대충 글자정확도는 자소정확도보다 3~4%정도 약간 낮은 값을 보이고 있다.

표 2. 자질 선택 방법의 비교

자질선택 방법	단어 정확도	글자 정확도	자소 정확도	트리 크기
정보이득	34.2	78.2	81.5	7767
근접도	34.7	78.6	82.1	7738

2.3 자질 선택

ID3의 induction bias는 큰 트리보다는 작은 트리를 선호한다는 것이다[11]. 이는 직관적으로 학습데이터에 일치하는 가능한 단순한 가설이 복잡한 가설에 비해 효과적이라는 것을 의미한다. 결정트리 학습 알고리즘에서 트리의 크기를 결정하는 가장 중요한 요소들 중의 하나가 자질의 선택 순서이다. 좋은 자질을 선택하기 위한 평가 방법이 많이 제안되었지만 도메인에 상관없이 절대적으로 우수한 평가 방법은 아직 존재하지 않는 것처럼 보인다[9]. 이들 대부분의 평가 방법은 정보이론 (information theory)에 기반한 방법이며 제한된 양의 학습데이터의 통계분석에서 좋은 자질을 결정하는 일은 근본적인 한계가 있을 수 밖에 없다.

하지만 우리의 한-영 음차복원 도메인에서는 좋은 자질을 선택하는데 단서가 될 수 있는 도메인 지식이 존재한다. 먼저 복원할 자소에 인접해 있는 자소가 해당 자소의 복원에 미치는 영향이 크다. 그리고 왼쪽에 있는 자소를 먼저 선택할 지 아니면 오른쪽을 먼저 선택할 지에 대한 방향의 결정은 복원할 자소에 따라 달라질 수 있다라는 점이다. 따라서 우리가 제안하는 자질 선택 순서는 왼쪽 우선일 경우는 L1 > R1 > L2 > R2 > L3 > R3 또는 오른쪽 우선의 경우는 R1 > L1 > R2 > L2 > R3 > L3 와 같이 되고 방향은 별도의 검증데이터를 사용하여 결정된다. 즉, 두 가지 순서 중 검증데이터에서 성능이 더 좋은 방향을 선택한다.

본 실험에서는 전체 학습데이터 중에서 3000단어를 학습데이터로, 나머지 3000단어를 검증데이터로 사용하였다. 표 2는 본 논문에서 제안하는 근접도 방법을 ID3의 정보이득 (information gain) 방법과 비교한 결과이다. 실험결과 정보이득 방법에 비해 많이 좋아지지는 않았지만 자소 정확도가 0.6% 정도 개선되었다. 이후의 본 논문의 모든 실험에서는 근접도 자질 선택 방법이 사용된다.

2.4 가지치기 (Pruning)

ID3 알고리즘은 모든 학습 예들이 완벽하게 분류될 때까지 트리를 확장한다. 하지만 데이터에 노이즈가 있을 경우 문제가 발생한다. 이것은 ID3 알고리즘이 데이터에 내재되어 있는 개념뿐만아니라 특수한 노이즈까지 학습하려고 들기 때문이다. 이러한 경우 학습데이터에 대해서는 좋은 성능을 보일지 몰라도 새로운 데이터에 대해서는 좋지 않은 성능을 보이게 된다. 이러한 현상을 과도학습 (overfitting)이라고 한다[11].

우리의 경우 노이즈의 원인이 될 수 있는 것이 단어 정렬 오류로 인해 잘못된 학습데이터가 만들어 지는 경우와 영어 이외의 다른 어원들 때문에 발생하는 노이즈의 경우를 생각해 볼 수 있다. 하지만 우리의 정렬 알고리즘은 매우 정확하여 정렬로 인한 오류는 무시할 수 있는 정도라고 생각되기 때문에 여러 다른 어원의 문제가 더 가능성이 있다.

과도학습을 완화하는 한 가지 방법은 트리를 만들때 완벽하게 학습데이터가 분류되도록 하지 않는 것이다. 이를 위해 트리 가지치기 (pruning)가 사용되는데 미리-가지치기 (pre-pruning)와 나중-가지치기 (post-pruning)의 두 가지 가지치기 방식이 있다. 미리-가지치기 방법은 학습 예들이 완벽하게 분류되기 전에 트리 확장을 멈추는 방법이고, 나중-가지치기는 완벽한 트리를 우선 만들어 놓고 나중에 가지치기하는 방법이다. 미리-가지치기는 어느 시점에서 트리 확장을 멈출 것인가 하는 멈추기 기준 (stopping criterion)이 중요하다. 많은 방법이 제안되었지만 우리는 정보이득 (information gain) 방법만을 사용하여 실험하였다. 나중-가지치기에도 많은 방법이 존재하지만[13] 일반적으로 도메인에 상관없이 대체적으로 꾸준한 성능을 보인다고 하는 reduced error pruning [12] 방식만을 사용하여 실험하였다. 표 3은 이 두 가지 가지치기 방식을 3000단어 학습집합과 1000단어 평가집합을 사용하여 실험한 결과이다. 실험결과를 보면 미리-가지치기나 나중-가지치기 모두 글자 및 자소정확도를 약 1% 정도 개선하는데 그쳤다. 따라서 생각보다는 과도학습이 심하지 않다고 볼 수 있다. 하지만 트리의 크기는 미리-가지치기와 나중-가지치기 모두에서 각각 28%와 46%의 큰 폭으로 줄어들었다.

표 3. 가지치기 실험 결과

	단어 정확도	글자 정확도	자소 정확도	트리 크기
가지치기 이전	34.7	78.6	82.1	7738
미리-가지치기	34.7	79.1	82.4	5513
나중-가지치기	32.3	79.3	82.3	4123

2.5 학습 데이터의 크기

지금까지는 3000 단어 학습데이터만을 사용하였는데 학습데이터의 크기에 따라 복원 정확도가 얼마나 차이 나는지 그리고 더 많은 학습데이터를 사용하면 정확도가 올라가는지 실험하였다. 1000단어부터 6000 단어까지 1000개씩 증가시키면서 실험한 결과는 표 4와 같다. 6000 단어일 때 단어정확도가 37.2%로 1000 단어일 때 27.3% 비해 정확도가 크게 향상되었다. 하지만 3000단어 이후부터는 향상의 폭이 현저히 줄어들고 있다.

표 4. 학습 데이터의 크기에 따른 복원 정확도 비교

데이터 크기	단어 정확도	글자 정확도	자소 정확도	트리 크기
1000	27.3	75.2	79.2	3104
2000	31.7	77.6	81.2	5565
3000	34.7	78.6	82.1	7738
4000	36.4	79.2	82.6	10646
5000	35.9	79.5	83.0	13167
6000	37.2	80.0	83.3	15423

2.6 로마자 표기법과의 비교

로마자 표기법[3]을 음차 복원에 사용하는 것이 가능하다. 하지만 로마자 표기법은 그 목적이 한국어의 음운체계를 외국인에게 전달하는데 있으므로 영어의 음운체계를 고려하여야 하는 음차 복원에 사용하는 데는 무리가 있다. 우리는 본 논문에서 제안하는 자동 규칙 학습에 의한 방법이 수동으로 만들어진 규칙기반의 복원 방법에 비해 우수함을 보이기 위하여 로마자 표기법과 성능을 비교한다.

표 5. 로마자 표기법과 성능 비교

	단어 정확도	글자 정확도	자소 정확도	트리 크기
제안된 방법	34.7	78.6	82.1	7738
로마자 표기법	9.7	59.4	NA	NA

자음				
ㄱ	ㄲ	ㄴ	ㄷ	ㄸ
k/g	k	n	t/d	t
ㄹ	ㅁ	ㅂ	ㅃ	ㅅ
r/l	m	p/b	p	s
ㅆ	ㅇ	ㅈ	ㅉ	ㅊ
s	ng	j/ch	tch	ch
ㅋ	ㅌ	ㅍ	ㅎ	
k	t	p	h	
모음				
ㅏ	ㅐ	ㅑ	ㅒ	ㅓ
a	ae	ya	yae	eo
ㅔ	ㅕ	ㅖ	ㅗ	ㅘ
e	yeo	ye	o	wa
ㅙ	ㅚ	ㅛ	ㅜ	ㅝ
wae	oe	yo	u	weo
ㅞ	ㅟ	ㅠ	ㅡ	ㅢ
we	wi	yu		eui
ㅣ				
i				

(1) 'ㄱ, ㄷ, ㅂ, ㅈ'이 모음과 모음 사이, 또는 'ㄴ, ㄹ, ㅁ, ㅇ'과 모음 사이에서 유성음으로 소리날 때에는 각각 'g, d, b, j'로 적고, 이 밖에는 각각 'k, t, p, ch'로 적는다.
(2) 'ㄹ'은 모음 앞에서는 'r'로 적고, 자음 앞이나 낱말의 끝에서는 'l'로, 'ㄹㄹ'은 'll'로 적는다.

그림 3. 변형된 로마자 표기법

표준 로마자 표기법을 그대로 사용하는 것은 매우 비효과적이므로 다음과 같이 약간 수정하여 사용하였다. 특수 기호를 포함한 로마자를 영어로 바꾸었다 (그림 3). 경음 ('ㄲ', 'ㄸ', 'ㅃ')은 영어에

없는 음소이므로 격음 ('ㅋ', 'ㅌ', 'ㅍ')과 같이 취급한다. 모음 'ㅡ'는 영어에 없는 음소이므로 null로 대응시켰다.

실험결과는 표 5와 같다. 예측했던 대로 변형된 로마자 표기법은 본 논문에서 제안하는 방법에 비해 현저히 낮은 성능을 보였다.

2.7 기존의 방법과의 비교

한국어에서 자동 음차 복원에 대한 연구는 아직 많지 않고 정길순, 맹성현, 이재성, 최기선[6], 이재성[4], 정길순과 맹성현[5]의 연구가 대표적이다. [4]와 [6]은 HMM에 기반한 방법이고 [5]는 신경망을 이용한 방법이다.

[6]에서는 후처리 전 단계의 복원정확도에 대한 평가 데이터를 제시하지 않고 있기 때문에 이재성[4]과만의 비교를 하였다. [4]에서는 자신의 데이터를 사용해 실험한 복원 정확도를 제시하고 있다. 하지만 방법의 특성상 가능한 후보 영어스트링를 확률순으로 원하는 만큼 출력할 수 있는데 단어정확도는 평가하지 않았고 글자정확도는 상위 20개 결과의 평균 글자정확도를 취하였다. 따라서 그의 실험결과와 직접적으로 비교하기는 곤란하고 간접적인 비교를 보인다.

[4]는 1500개 학습데이터와 100개의 평가데이터를 사용하고 있는데 이 데이터는 우리가 사용하고 있는 7000개 실험데이터의 부분집합이다. 우리는 그의 1600개 데이터를 입수하였다. 보다 정확한 성능 평가를 위하여 이 1600개 데이터를 무작위로 1500개와 100개의 학습 및 평가 집합으로 나뉘고 이를 5회 반복하여 평균값을 취하였다. 표6에서 글자정확도를 비교할 때 본 논문에서 제안한 방법이 훨씬 우수하다는 것을 볼 수 있다. 자소정확도가 글자정확도에 비해 3~4% 높은 값을 가지므로 글자정확도에서 자소정확도를 추정해볼 수 있다. 따라서 이재성 방법의 자소정확도가 65% 정도라고 보면 평균 단어길이가 6 자소라고 가정할 경우 단어정확도를 추정해보면 약 7.5% (= 0.65^6) 정도 밖에 되지 않는다.

하지만 이재성[4]과의 직접적인 비교는 적절하지 않을 수도 있다. 우리의 경우는 좌우 3개씩의 자소를 보고 있는데 반하여 이재성[4]은 좌우문맥을 거의 보지 않고 있다. 따라서 더 많은 좌우문맥을 볼 경우에 성능이 좋아질 가능성이 매우 높다. 하지만 trigram 이상으로 문맥을 넓힐 경우 통계적모델의 약점인 심각한 데이터부족 (data sparseness) 문제를 일으킬 수 있다.

표 6. 이재성[4] 방법과의 성능 비교

	단어 정확도	글자 정확도	자소 정확도	트리 크기
제안된 방법	31.0	77.1	80.5	4309
이재성*	-	60.5	N/A	N/A

3 결과 분석 및 향후 연구

최대 37% 이상의 단어정확도를 얻었다. 이것은 후처리 전단계의 정확도라고 생각할 때 매우 실용적인 정도의 성능이라고 보여진다. 이러한 높은 정확도는 매우 정확한 단어음소정렬에서 오는 것으로 믿어진다. 사실 어떠한 감독하 학습 (supervised learning) 알고리즘을 사용하던지 정확하게 정렬된 영어-음차표기 데이터가 필수적이다.

대부분의 오류는 '프라이머'의 음차 복원인 'frimer' ('primer')의 'f'와 'p'와 같이 한국어에서는 구분이 되지않는 영어 음소들 때문이다. 이러한 애매성은 좌우 문맥으로도 일반적인 규칙을 유도하기 힘든 매우 불규칙한 현상이기 때문에 근본적인 해결은 쉽지 않을 것 같다. 향후연구에서는 이러한 애매성을 보다 효과적으로 다룰 수 있는 방법에 대한 연구를 수행할 예정이다.

또한 다중단어를 고려한 보다 정교한 복원 방법 및 복원의 성능을 더욱 높일 수 있는 보다 효과적인 사전매칭에 의한 후처리에 대해서도 연구가 필요하다.

4 결론

본 논문에서 우리는 결정트리 학습을 통한 한-영 음차복원 방법을 제안하였다. 또한 각 자소별 결정트리 학습을 위해 정확한 영어-음차표기 정렬 방법을 고안하였다. 특히 영어-음차표기 정렬방법은 어떠한 감독하 학습 알고리즘을 적용하더라도 필수적인 부분이기 때문에 그 의의가 매우 크다고 생각된다. 실험을 통해 기존의 음차 표기 및 복원 모델인 이재성 방법에 비해서 월등한 성능을 보임을 보일 수 있었다. 또한 기존의 로마자 표기법은 매우 단순하여 한-영 음차복원에 사용되기에는 매우 부적절함을 실험을 통해 입증하였다.

참고문헌

[1] 김병혜, "영어단어의 알파벳표기로부터 한글표기로의 자동변환,". 석사학위논문, 서강대학교 공공정책대학원, 1991.

[2] 남영신, 최신외래어사전, 국어사전 별책부록, 성안당, 1997.

[3] 문교부, "국어의 로마자 표기법," 문교부 고시 제 84-1 호 (1984.1.13).

[4] 이재성, "다국어 정보검색을 위한 영-한 음차표기 및 복원 모델,". 박사학위논문, 한국과학기술원 전산학과, 1999.

[5] 정길순, 맹성현, "외래어의 자동음역을 통한 영어단어 생성," 한국정보과학회 춘계학술대회, 1998.

[6] Jeong K., Myaeng, H., Lee, J. S., and Choi, K. S., "Automatic identification and back-transliteration of foreign words for information retrieval,". *Information Processing and Management*, 35(4), 1999, pp. 523-540.

[7] Knight K. and Graehl, J., "Machine Transliteration," in *Proceedings of the 35th Annual Meeting of ACL*, 1997, pp. 17-22.

[8] Lee, J. S., and Choi, K. S., "English to Korean statistical transliteration for information retrieval," *Computer Processing and Oriental Languages*, 12(1), 1999, pp. 17-37.

[9] Mingers, J., "An empirical comparison of selection measures for decision tree induction," *Machine Learning*, 3, 1989, pp. 319-342.

[10] Mingers, J., "An empirical comparison of pruning methods for decision tree induction," *Machine Learning*, 4, 1989, pp. 227-243.

[11] Mitchell, T. M., "Machine Learning," The McGraw-Hill Companies, Inc., 1997.

[12] Quinlan, J. R., "Rule induction with statistical data – a comparison with multiple regression," *Journal of the Operational Research Society*, 38, 1987, pp. 347-352.

[13] Quinlan, J. R., "Induction of decision trees," *Machine Learning*, 1, 1986, pp. 81-106.