복수 음운 정보를 이용한 영·한 음차 표기

강인호° 김길창
한국과학기술원 전산학과
ihkang@csone.kaist.ac.kr, gckim@cs.kaist.ac.kr

English-to-Korean Transliteration using Multiple Unbounded Overlapping Phonemes

In-Ho Kang° Gil Chang Kim
Dept. of Computer Science
Korea Advanced Institute of Science and Technology

요 약

본 연구에서는 기존의 한정된 길이의 영어 또는 한글의 발음 단위를 이용하던 자동 음차 표기 방식과 달리, 학습 데이터에서 추출한 임의 길이의 음운 패턴을 사용하는 방법을 제안한다. 통계적 정보에 기반하여 추출한 음차 표기 패턴과 외래어 표기 규칙에 기반하여 기술한 음차 표기 패턴을 위치와 길이에 관계없이 사용하여 주어진 영어 단어의 한글 음차 표기를 얻어낸다. 제안하는 방법은 먼저 주어진 영어 단어의 가능한 모든 발음 단위를 기준으로 한글 표기 네트워크를 만든 후, 학습 데이터에서 추출한 음운 패턴을 교차 적용시켜 네트워크 각 노드의 가중치를 결정한다. 가중치가 결정된 네트워크에서 가중치의 합이 가장 좋은 경로를 찾아냄으로 해서 음차 표기를 수행한다. 본 연구에서 제안하는 방법으로 실험을 한 결과 자소 단위 86.5%, 단어 단위 55.3%의 정확률을 얻을 수 있었다.

1 서론

외국과의 문화와 정보의 활발한 교류에 의해 외래어, 외국어 사용이 빈번하게 발생한다. 영어와 한글을 고려할 경우, 한글이 영어 문서에서 나타나는 경우보다는 영어가 한글 문서에 나타나는 경우가 많다. 한글 문서에 나타나는 영어들은 많은 경우 한글로 표기된 단어, 즉 외래어 표기 형태로도 나타난다. 현재 우리 나라의 외래어 표기법은 발음을 중심으로 표기하고 있으나, 예외를 인정하고 있어서(문화체육부, 1995)[1], 실생활에 쓰이는 외래어는 여러 가지 방식에 의해 표기되고 있다. "데이타"의 경우, 영어 *data*의 마지막 알파벳 'a'의 대표음을 "아"로 보아 "데이터" 대신 "데이타"로 표기했다고 볼 수 있다. 이와 같이 영어 철자에 근거하여 표기하는 것을 **눈말 표기**라고 하며, 영어 발음에 근거하여 표기하는 것을 **입말 표기**라고 한다.

표 1: 입말 표기와 눈말 표기의 예

	입말 표기	눈말 표기
data	데이터	데이타
digital	디지틀	디지탈
radio	레이디오	라디오

[1]제 5항: 이미 굳어진 외래어는 관용을 존중하되, 그 범위와 용례는 따로 정한다.

영어와 그에 대한 다양한 외래어 표기 방식으로 인해 원어와 외래어 표기간의 관계를 설정하는 일이 필요하다. 다국어 정보 검색에서 질의어 번역 방식이나 문서 번역 방식의 경우, 영어 단어와 그 단어의 외래어 표기를 제대로 찾아내지 못하면 원하는 작업을 수행할 수 없게 된다. 특히 외래어로 표기된 단어들은 고유 명사이거나 전문 용어 등으로써 문서를 특징짓는 중요한 키워드일 가능성이 매우 높기 때문에 문제 해결 필요성이 더욱 크다. 그런데 대부분의 전문 용어나 고유 명사는 그 크기가 한정된 것이 아니라 계속적으로 만들어지며 그 양 또한 매우 크기 때문에, 모든 영어 단어에 대해 한글 표기를 가지는 것이 힘들다. 따라서 영어 질의어에 대한 외래어 표기와 한글 문서에서 영어에 대한 외래어 표기를 자동으로 추출하는 방법이 필요하다. 본 연구에서는 학습 데이터에서 자동으로 추출한 음운 패턴을 이용하여 영·한 음차 표기(transliteration)를 수행하는 방법을 제안한다.

2 관련 연구

영어 단어에 대한 한글 음차 표기를 구하는 방법에는 영어 발음 표기를 찾아낸 다음 그 발음을 한글로 표기하는 피벗 방식(김병혜, 1991a)과 그 중간 과정을 거치지 않는 직접 방식(김정재, 이재성, 최기선, 1999)이 있다. 일반적으로 피벗 방

식은 입말 표기를 잘 다루며, 직접 방식은 눈말 표기에 효과적이다(이재성, 1999c).

2.1 영어 발음 추출

영어 철자로부터 발음을 추출하는 기술은 다양한 방법으로 많은 연구가 진행되어 왔음에도 불구하고, 아직도 완벽한 단계에 이르지 못하고 있어 이를 향상시키기 위한 연구가 지속되고 있다. 이러한 발음 추출에 사용된 방법의 예로는 기계학습 방법인 결정 트리(G., Hild, and Bakiri, 1995), 신경망(J. and R, 1986), 확률 정보 기반의 HMM(Hidden Markov Model)을 이용한 방법들이 있다. 이러한 방법들이 아직도 완벽하게 영어 발음을 생성하지 못하는 이유는 영어가 순수한 표음문자가 아니고, 어원이나 단어에서의 위치 및 주변 음절에 따라 제각기 다른 발음을 내기 때문이다. 영어 철자에서 자동으로 외래어 표기를 할 경우에도 이러한 문제점을 그대로 가진다.

2.2 영·한 음차 표기

자동적인 영·한 음차 표기에 사용되는 방법은 크게 HMM으로 대표되는 확률 정보 기반 방식과 신경망 및 결정 트리를 이용한 방법이 있는데, 영어 발음 추출에 사용하는 방법과 흡사하다.

2.2.1 확률 정보 기반

영어 단어 E는 발음 단위로 $e_1, e_2, \ldots, e_n$으로 나뉘어지며 거기에 대응되는 한글 표기 K는 영어 발음 단위에 맞추어서 $k_1, k_2, \ldots, k_n$으로 나타난다. 영어 단어 *dressing*에 대해서 발음 단위로 나타내면 그림 1[2]과 같다.

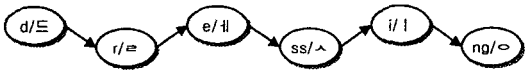

그림 1: *dressing*에 대한 한글 음차 표기 예

확률 정보 기반의 영·한 음차 표기는 주어진 영어 단어 E에 대해서 가장 좋은 확률값을 얻는 한글 음차 표기 K를 찾아내는 것으로 정의할 수 있다.

$$arg\ \max_K p(K|E) = arg\ \max_K p(K)p(E|K) \qquad (1)$$

$p(K)$와 $p(E|K)$는 한글 표기(k_i)는 주위 몇 개의 한글 표기($k_{i-1}, k_{i-2}, ...$)에 의해서 결정된다고 가정하는 Markov Independent Assumption에 의해 간략화 된다. 바로 전 발음 단위에 의해서만 결정된다고 가정하면 식 2와 같이 근사화 시킬 수 있다.

$$p(K) \cong p(k_1)\prod_{i=2}^{n} p(k_i|k_{i-1}),\quad p(E|K) \cong \prod_{i=1}^{n} p(e_i|k_i) \qquad (2)$$

[2]편의상 영어 발음 단위와 한글 표기를 '/'를 이용해서 품사 태깅의 형식으로 나타낸다.

일반적으로 주위 문맥으로 사용하는 발음 단위의 개수를 늘림에 따라 정확률이 올라가지만 자료 희귀성(data sparseness) 문제가 생긴다. 이를 극복하기 위해서 여러 종류의 확률 정보를 함께 사용할 수 있는 MEM(Maximum Entropy Model), Back-off, 선형 결합(Linear Interpolation)의 방법으로 확률 분포를 구하기도 한다(강인호, 1999b). 직접 방식을 사용할 경우에는 특정한 하나의 영어 발음열만 고려하는 것이 아니라 발생 가능한 모든 영어 발음 경우를 고려한다(이재성, 1999c). 즉 data의 경우 $d-a-t-a$, $d-at-a$, $da-ta$ 등으로 발음될 가능성이 있다고 보고 세 경우에 대해서 확률값을 구한 뒤 제일 좋은 표기열을 선택한다[3].

2.2.2 신경망 기반

신경망과 결정 트리 방법은 영어 발음이나 음절 단위에 해당하는 한글 표기를 결정적으로 찾을 수 있게 해준다. 현재 고려하고 있는 영어 발음 단위의 좌우 2-3개의 발음 단위를 입력층으로 한글 표기를 출력층으로 해서 학습을 시킨다. 그러나 입력층으로 고려하는 범위를 벗어나는 영어 음절에 의해서 그 값이 결정될 경우에는 올바른 값을 결정할 수 없다는 단점이 있으며, 영어 단어에 대한 발음 단위가 결정되지 않은 경우에서 직접 방식으로 한글 표기를 추출할 경우, 확률 정보 기반 방법과 같이 발음 단위 개수에 따른 정규화(normalization) 문제를 가진다.

2.2.3 후처리

(정길순, 맹성현, 1998)에서는 사전을 이용한 후처리 방법을 제안했다. 후처리는 모델을 통하여 나온 결과와 사전에서 가장 유사한 단어를 결과로 내보내는 방법이다. 사전과의 매칭 방법은 사전에 나타나 있는 단어와 후보 영어와의 트라이그램(trigram) 단위로 유사도를 측정하는 방법을 사용하였다. 후처리 방법이 한글에서 영어로의 복원에는 유용할 지 모르나 한글 음차 표기에는 한글 표기 정답이 없는 경우가 대부분이기 때문에 사용하기 힘들다. 그러나 생성한 한글 표기에 대해 연음법칙이나 두음 법칙을 적용하여 자연스러운 연결이 되게 하는 작업은 필요하다.

3 음운 패턴을 이용한 영·한 음차 표기

새로운 영어 단어를 보고 영어 발음이나 한글 표기를 유추하는 것은 일정한 규칙에 기반하기도 하지만 경험에 근거하는 경우도 많다. 기존에 알고 있던 단어의 한글 표기를 적용시킴으로써 새로운 단어의 한글 표기를 찾아내는 것이다. *scalar*라는 단어를 예로 들면 유사한 음운을 가지는 영어 단어로써 *scale*, *casino*, *koala*, *car* 등을 들 수 있다. 이러한 영어 단어들의 표기 형태를 통해서 그림 2와 같이 *scalar*의 표기 형

[3]형태소 품사 태깅과 같이 중복된 연산을 줄이기 위해 하나의 구조로 표현하여 처리한다.

태를 추측할 수 있다. 이와 같이 본 연구에서는 특정한 음운 패턴이나 유사한 단어를 이용하여 새로운 단어의 한글 표기를 유추할 수 있다고 가정한다. 본 연구에서는 표기나 발음에 있어서 규칙을 나타내는 음운 패턴과 학습 말뭉치에서 통계적으로 의미 있는 음운 패턴을 이용하여 영·한 음차 표기를 수행한다.

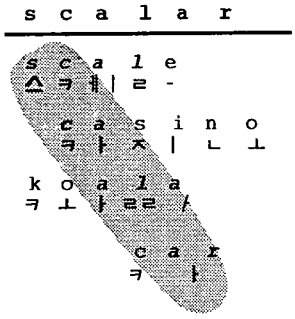

그림 2: *scalar*의 한글 음차 표기 유추

3.1 음운 패턴 정보의 추출

본 연구에서 사용하는 음운 패턴은 한글과 영어의 발음 단위로 정렬된 말뭉치에서 뽑아낸다. 발음 단위로 정렬된 학습 데이터를 사용하는 이유는 영어의 일부분과 한글의 일부분을 관계 짓기 위함이다. 주어진 영어 단어에 대해서 나열할 수 있는 모든 형태의 발음 단위와 그에 따른 한글 표기를 결합하여 음운 패턴을 뽑아낸다. 즉 음운 패턴이란 영어의 발음열과 그에 따른 한글 표기열이라고 할 수 있다. 음운 패턴을 추출할 때 영어 단어의 처음(@sow@)과 끝부분(@eow@)에 표시를 하여 처음과 끝이 들어가는 음운 패턴은 표시를 한 음운 패턴과 그렇지 않은 음운 패턴 두 개를 추출한다. 이는 단어 중간에서 나타나는 경우와 그렇지 않은 경우에 표기가 다르기 때문이다. 그림 1에서 본 *dressing*의 경우 표 2와 같이 음운 패턴을 추출할 수 있다.

표 2: *dressing*에서 추출할 수 있는 음운 패턴의 일부

적용 문맥	해당 표기
@sow@d	d/드
d	d/드
@sow@dr	d/드+r/ㄹ
r	r/ㄹ
@sow@dre	d/드+r/ㄹ+e/ㅖ
⋮	

이렇게 추출한 음운 패턴은 발생 빈도를 이용하여 걸러낸다. 이는 학습 데이터에 존재하는 오류를 제거하기 위함이다. 추출한 음운 패턴 중에서 하나의 발음 단위만으로 구성된 음운 패턴을 제외하고 높은 빈도수를 가지는 음운 패턴을 보이면 표 3과 같다. 표 3에서 '-'는 묵음을 나타낸다.

표 3: 높은 발생 빈도를 가지는 음운 패턴들

적용 문맥	해당 표기	적용 문맥	해당 표기
e@eow@	e/-	@sow@c	c/ㅋ
n@eow@	n/ㄴ	@sow@b	b/ㅂ
er@eow	er/ㅓ	ne	n/ㄴ+e/-
in	i/ㅣ+n/ㄴ	le	l/ㄹ+e/-

편의상 음운 패턴에서 영어 발음 단위에 해당하는 부분을 적용 문맥이라고 하겠다. 추출한 음운 패턴은 처음과 끝을 나타내는 심벌을 제외한 적용 문맥 길이와 가능한 해당 표기의 개수에 따라서 가중치가 결정된다. 식 3에서 $C(x)$는 학습 말뭉치에서 x의 발생 빈도수를 뜻한다.

$$weight(\text{적용문맥} : \text{해당표기}) = \frac{C(\text{해당표기})}{C(\text{적용문맥})} \qquad (3)$$

두 음절 이상의 영어 발음 단위는 다른 두 영어 발음 단위로 나눌 수 있다. 이때 나눠진 영어 발음 단위 두개가 학습 데이터에서 발견할 수 있는 연결이면 전체 음운 패턴의 가중치에 각각의 가중치를 더해주고, 그렇지 않은 경우에는 두 발음 단위의 결합은 있을 수 없다고 표시한다. 예를 들어 *pp*의 경우 영어 발음 단위 p와 p의 결합으로 나눌 수 있다. 그러나 학습 말뭉치에서 p와 p가 연속해서 나타나는 경우가 없기 때문에 p와 p의 연속은 있을 수 없다. 반면 $ar : ar/ㅏ$의 경우는 $a : a/ㅏ$의 가중치와 $r : r/-$의 가중치는 $ar : ar/ㅏ$ 가중치에 더해진다. 후자의 경우는 자동 정렬에서 나타나는 일관되지 못한 오류를 보정하기 위함이다. 음운 패턴의 가중치는 각각의 발음 단위에 할당이 된다. 이때 발음 단위의 길이에 비례해서 가중치를 할당한다. 이는 문맥이 긴 정보를 우선하겠다는 뜻이다.

추출한 음운 패턴은 결합 정보로도 사용된다. 결합 정보는 두 한글 표기가 연속해서 나타난 경우가 있는지를 나타내는 것으로 부자연스러운 한글 생성을 막기 위해서다. 일종의 연음법칙과 두음 법칙을 해결하기 위한 방법이다. 결합 정보는 두 음절 이상 길이의 발음 단위에 의한 접속 불가 정보와 한글 표기의 결합 여부를 포함한다.

3.2 한글 표기 네트워크

주어진 영어 단어에 대해서 있을 수 있는 모든 발음 단위를 이용해서 한글 표기 네트워크를 만든다. 한글 표기 네트워크는 하나의 영어 발음 단위와 그에 대한 한글 표기 그리고 가중치를 가지는 노드와 노드 간의 연결로 표현된다. *scalar*의 경우 s, c, a, l, a, r 외에 sc, al, ar이 하나의 발음 단위가 될 수 있기 때문에 각각의 경우를 다 고려해서 노드가 만들어지며, 원 단어 순서에 맞춰 노드가 연결된다(그림 3). 한글 표기 네트워크 구성 시 결합 정보에서 연결이 허용되지 않는 두 개의 노드는 연결하지 않는다. 그림 3에서 점선 화살표는 결합 정보에 나타나지 않아서 연결되지 않는 것을 나타낸다.

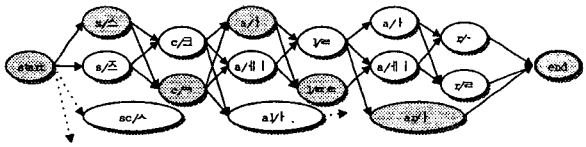

그림 3: scalar에 대한 한글 표기 네트워크의 일부

한글 표기 네트워크가 만들어진 뒤, 한글 표기 네트워크에 적용할 수 있는 모든 음운 패턴을 사용한다. 이때 각 노드의 가중치는 음운 패턴에서 해당 발음 단위가 가지고 있는 가중치 만큼 더해진다. 그림 4에서는 (scal:s/스+c/ㅋ+a/ㅐㅣ+l/ㄹ)의 각 발음 단위 가중치인 α는 한글 표기 네트워크에서 해당되는 s/스, c/ㅋ, a/ㅐㅣ, l/ㄹ, 노드에 각각 더해진다. 적용할 수 있는 모든 음운 패턴이 사용되고 난 후에는 viterbi 알고리즘을 이용해서 가중치의 합이 제일 좋은 경로를 찾아내고 그 경로의 해당 표기를 나열함으로 해서 원하는 한글 표기를 얻어낸다.

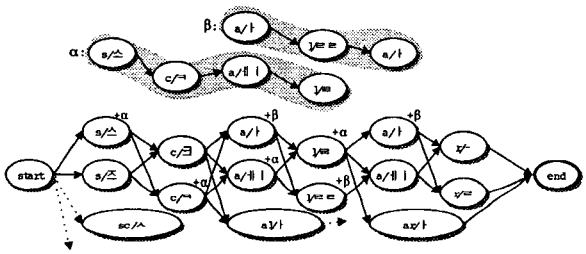

그림 4: 가중치 적용 예

4 실험 및 결과 분석

학습 및 평가 데이터는 이희승(1994)의 저서에서 발췌한 표준 외래어를 사용하였다. 학습 데이터와 시험용 데이터의 크기는 비교를 위해서 (이재성, 1999c)와 같이 1500개를 학습 데이터로 150개를 시험용 데이터로 사용하였다. 본 실험에서 사용한 발음 단위 정렬대역데이터에는 영어의 발음단위가 총 156개, 한글 발음 단위는 총 85개인데, 영어 발음단위는 평균 1.166개의 알파벳으로 이뤄졌다. 그리고 한글과 영어의 발음 단위 자동 정렬은 (이재성, 1999c)에서 제안한 방법에 의해 이루어져 있다. 실험을 통해서 총 49,000개의 가능한 음운 패턴 정보를 추출할 수 있었다.

정확도는 크게 단어 단위 정확도(word accuracy)와 자소 단위 정확도(character accuracy)를 사용한다. 자소 단위 정확도는 정확한 한글 음차 표기와 실험 결과로 나온 음차 표기를 자동 정렬하여 최대로 매칭되는 비율을 정확도로 계산한다. 식 5에서 L은 정답 표기의 길이를 나타내며, i, d, s는 시스템 출력 결과를 정답 표기와 일치 시키기 위해서 수행한 삽입, 삭제, 교체 횟수를 나타낸다. 결과가 음수일 경우는 0으로 처리한다.

$$word\ accuracy = \frac{number\ of\ correct\ words}{number\ of\ generated\ words} \quad (4)$$

$$character\ accuracy = \frac{L-(i+d+s)}{L} \quad (5)$$

표 4와 표 6은 확률 정보 기반의 (이재성, 1999c)의 모델과 좌우 한 발음 단위씩 고려한 신경망 모델 (김정재, 이재성, 최기선, 1999)과 본 연구에서 제안한 모델의 정확도를 비교한 것이다. 다른 모델과 달리 (이재성, 1999c)의 단어 단위 정확률은 시스템이 출력한 20개의 결과 값 중에서 정답이 있을 경우, 정답이라고 간주했을 때의 수치이다.

표 4: 학습 데이터 실험 결과

모델	단어 단위 정확률	자소 단위 정확률
확률 기반	72.7%*	82.3 %
신경망	59.6%	86.2 %
패턴 기반	99.6%	99.6 %

학습 데이터에 대해서 총 6개의 오류가 있었는데 표 5와 같다.

표 5: 학습 데이터 실험 결과 오류

영어 단어	정 답	출력값
set	세트	셋
lot	로트	롯
net	네트	넷
tar	타르	타
piment	피망	피만트
grand	그랑	그랜드

여기서 *grand*는 '그랜드'와 '그랑'이라는 영어, 불어 두 가지 표현이 학습 데이터에 있었기 때문이며, *set*, *lot*, *net*, *tar*는 단어의 앞부분에서 쓰일 경우 축약하여 쓰여지기 때문에 가중치에 의해서 축약된 형태가 나타났다. 이는 규칙 형태로 기술하여 높은 가중치를 부여한다면 해결 할 수 있다. *piment*는 불어 표현으로써 영어 표현에 치중된 가중치에 영향을 받았다.

표 6: 미학습 데이터 실험 결과

모 델	단어 단위 정확률	자소 단위 정확률
확률 기반	40.7 %*	69.3 %
신경망	35.1 %	79.0 %
패턴 기반	55.3 %	86.5 %

미학습 데이터의 오류 중 모음에 의한 오류가 많았는데, 이는 자음에 비해 모음이 주변 상황에 따라 가능한 표기가 많기 때문이다. 또한 학습 데이터에서 추출한 결합 정보에 나타나지 않은 연결에 의한 오류도 있었다.

5 결론 및 향후 연구

본 연구에서는 기존의 한정된 길이의 영어 또는 한글의 발음 단위를 이용하던 자동 음차 표기 방식과 달리, 영어 단어를 인식할 때 사용하는 음운 패턴을 통계적 기법을 이용하여 자동으로 추출하는 방법과 추출한 음운 패턴을 이용하는 음차 표기 방법을 제안했다. 본 연구에서 제안하는 방법으로 실험을 한 결과 단어 단위로 55.3%, 자소 단위로 86.5%의 정확률을 얻을 수 있었다. 본 연구에서 제안한 모델은 특별한 학습 과정이 필요 없다는 장점과 사용하는 정보의 길이에 영향을 덜 받는다는 장점을 가진다. 또한 모델에서 발생하는 오류 분석을 통한 새로운 규칙(음운 패턴의 형태)및 임의의 규칙을 추가하기 쉽다는 장점도 가진다. 다른 모델보다 좋은 정확률을 얻게 된 이유는 먼저 다양한 길이의 문맥 정보를 쉽게 사용할 수 있었다는 것과, 자동 추출한 결합 정보에 의해 두음 법칙과 연음 법칙을 보완할 수 있었기 때문이다.

앞으로 본 논문에서 사용한 간단한 가중치 설정 외에 다양한 형태로 길이와 빈도를 고려할 수 있는 가중치 설정 방법이 연구되어야 한다. 또한 단어 구성 정보를 이용하는 연구가 필요하다. *cameraman*의 경우 *camera*와 *man*을 독립적으로 입력하면 올바른 결과가 나오지만 두 개를 합치면 '카메라만'이라고 출력하는데 이렇게 합성어인 경우에는 분리해서 시작할 때의 정보 즉 중간에 끊어져서 새로 발음이 시작된다는 정보를 이용할 수 있다. 단어적으로 분리할 수 있는 정보는 문제의 크기도 줄이고 애매성도 줄일 수 있다. 이처럼 문제 크기를 줄이는 정보나 애매성을 줄이는 정보에 대한 연구가 필요하다. 그리고 본 연구에서 제시한 방법으로 영·한 음차 복원에 적용하여 음차 표기와 음차 복원의 관련성과 차이점을 알아볼 필요가 있다. 마지막으로 grand, piment 같이 어원이 영어가 아닌 불어에 대해서는 좋은 결과를 보이지 않았는데 어원에 따른 표기 규칙의 구분에 대한 연구가 필요하다.

참고문헌

김정재, 이재성, 최기선. 1999. 신경망을 이용한 발음단위 기반 자동 영·한 음차 표기 모델. 한국 인지과학회 춘계 학술대회, pages 147–252.

정길순, 맹성현. 1998. 외래어의 자동음역을 통한 영어단어 생성. 한국정보과학회 춘계학술대회, pages 429–431.

문화체육부. 1995. 외래어 표기법. 문화체육부 고시 제 *1995-8호(1995.3.16)*.

문교부. 1984. 국어의 로마자 표기법. 문교부 고시 제 *84-1호(1984.1.13)*.

김병혜. 1991a. 영단어의 한글로의 자동변환. 석사학위논문, 서강대학교.

강인호. 1999b. 최대엔트로피 모델을 이용한 한국어 품사 태깅. 석사학위논문, 한국과학기술원.

이재성. 1999c. 다국어 정보검색을 위한 영·한 음차 표기 및 복원 모델. 박사학위논문, 한국과학기술원.

Bertjan Busser, Walter Daelemans, Antal van den Bosch. 1999. Machine learning of word pronunciation: the case against abstraction. In *Eurospeech99*, pages 2123–2126.

G., Dietterich Thomas, Hermann Hild, and Ghulum Bakiri. 1995. A comparison of id3 and backpropagation for english text-to-speech mapping. In *Machine Learning*, pages 51–80.

Lee, Jae Sung and Key-Sun Choi. 1998. English to korean statistical transliteration for information retrieval. *Computer Processing of Oriental Languages.*

J., Sejnowski T. and Rosenburg C. R. 1986. Nettalk: a parallell network that learns to read aloud. In *JHU/EECS-86.*

R., Quinlan J. 1986. Induction of decision trees. In *Machine learning*, pages 81–106.