적합성 피드백을 이용한 자동 음차표기의 성능향상 기법

오종훈°, 최기선
한국과학기술원 전자전산학과/전문용어언어공학연구센터/언어자원은행
{rovellia°, kschoi}@world.kaist.ac.kr

Improving English-to-Korean Transliteration through Automatic Relevance Feedback

Jong-Hoon Oh°, Key-Sun Choi
Department of EECS
Korea Advanced Institute of Science and Technology/KORTERM/BOLA

요 약

음차표기란 외국어의 단어를 글자나 발음을 이용하여 자국어로 표기하는 것으로 정의된다. 자동음차표기는 기계번역과 정보검색 등의 자연언어처리 응용에서 사용된다. 기계번역에서는 번역사전에 등재되어 있지 않은 고유명사나 전문용어를 번역하는 방법으로 사용되며, 정보검색에서는 단어불일치 문제의 해결과 질의확장 등에 사용된다. 하지만 지금까지의 영-한 자동 음차표기 연구들은 대부분 주어진 원어에 대하여 가장 적합한 음차표기를 생성하는 연구에 초점을 맞추었다. 또한, 원어로부터 가능한 음차표기를 파악하는 연구에서도 해당 음차표기에 대한 적합성을 파악하지 않고 단순 리스트 형태로 음차표기를 생성함으로써, 음차표기생성결과에 대한 품질이 낮았다. 본 논문에서는 이러한 문제점을 해결하기 위하여, 주어진 원어로부터 가능한 음차표기를 생성하고 이들에 대한 적합성을 자동으로 파악하는 음차표기 모델을 제안한다. 본 논문의 기법은 약 72%의 단어 정확률을 나타내었으며 기존의 기법보다 최고 31%의 성능향상을 나타내었다.

1. 서론

음차표기란 외국어의 단어를 글자나 발음을 이용하여 자국어로 표기하는 것으로 정의된다[1]. 한국어에서 음차표기된 단어들은 주로 영어에 기원을 두고 있기 때문에, 한국어에서 음차표기는 주로 영-한 음차표기를 지칭한다. 전문분야문서와 같이 외국어 원의 용어가 많이 포함된 문서를 처리할 때, 음차표기를 올바르게 파악하고 이를 효과적으로 처리하는 것은 매우 중요하다. 이는 전문용어의 많은 부분이 외국어에 기원을 두고 있고, 음차표기되거나 원어 그대로 사용되는 경우가 많기 때문이다. 전문분야 문서에서 음차표기와 원어의 혼재는 정보검색과 같은 자연언어응용에서 단어불일치 문제[1]를 야기한다. 음차표기로 인한 단어 불일치 문제를 해결하기 위한 방법으로 영-한 음차표기 대역쌍을 포함하는 사전을 이용하는 방법이 있다. 하지만 음차표기되는 용어들이 대부분 사전에 등재되지 않는 경우가 많기 때문에 이를 처리하기 위한 방법으로 자동음차표기에 대한 연구가 활발히 진행되어 왔다[1,2,3,4,5,6,7].

자동음차표기는 기계번역과 정보검색 등의 자연언어처리 응용에서 주로 사용된다. 기계번역에서는 번역사전에 등재되어 있지 않은 고유명사나 전문용어를 번역하는 방법으로 사용되며, 정보검색에서는 단어불일치문제의 해결과 질의확장 등에 사용된다. 따라서 자동음차표기시스템이 해결해야 할 문제는 1) 원어에 대응되는 정확한 음차표기를 찾는 것과 2) 원어로 생성가능한 다양한 음차표기의 변이를 찾는 것이다. 예를 들어 기계번역에서는 *data*에 대한 가장 적합한 음차표기인 '데이터'를 생성하는 자동음차표기 기법이 필요하며, 정보검색에서는 *data*의 가능한 음차표기인 '데이터', '데이타', '데타' 등을 생성할 수 있는 음차표기 기법이 필요하다.

하지만 지금까지의 영-한 자동 음차표기 연구들은 대부분 주어진 원어에 대하여 가장 적합한 음차표기를 생성하는 연구에 초점을 맞추었다. 또한, 원어로부터 가능한 음차표기를 파악하는 연구에서도 해당 음차표기에 대한 적합성을 파악하지 않고 단순 리스트 형태로 음차표기를 생성함으로써, 음차표기생성결과에 대한 품질이 낮았다.

본 논문에서는 이러한 문제점을 해결하기 위하여, 주어진 원어로부터 가능한 음차표기를 생성하고 이들에 대한 적합성을 자동으로 파악하는 음차표기 모델을 제안한다. 본 논문의 기법은 여러가지 음차표기 모델을 이용하여 다양한 음차표기를 생성한 후, 생성된 음차표기가 나타나는 웹문서의 정보로부터 음차표기의 적합성을 자동으로 파악한다. 이를 통하여 신뢰성 있는 음차표기 결과를 생성할 수 있을 뿐만 아니라 적합성에 의해 순위화된 음차표기 리스트를 통하여 가장 적합한 음차표기를 생성할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 관련연구에 대하여 기술하고 3장에서는 본 논문의 기법에 대하여 설명한다. 4장에서는 실험 및 결과에 대하여 기술하고 5장에서 결론을 맺는다.

2. 관련연구

기존의 영-한 음차표기 기법은 **글자기반방법**과 **음소기반방법** 그리고 **글자 및 음소기반 방법**으로 분류된다. 글자기반 방법은 영어단어를 한국어 음차표기로 직접 변환하는 방법으로, 발음지식이 필요한 다른 방법에 비하여 비교적 간단하게 음차표기를 생성할 수 있다는 장점이 있다. 즉 영-한 자소 변환 규칙의 학습만으로 음차표기를 생성할 수 있다. 이러한 특성으로 인하여 기존의 영-한 음차표기 연구[1,2,3,4,5]에서는 글자기반방법에 기반한 연구들이 주로 이루어져 왔다. 이들 연구들로는 확률기반[1,3,4], 결정트리기반[2], 음차표기 네트워크 기반 방법[5]이 있다. 하지만, 한국어 음차표기의 많은 부분이 음소에 기반하여 생성되기 때문에 글자기반방법은 올바른 음차표기를 생성하는 데 한계가 있다.

음소기반방법[1]은 영어단어의 발음을 이용하여 한국어 음차표기를 생성하는 방법으로 주어진 영어단어에 대한 발음을 파악하는 과정과 발음을 한국어 음차표기로 변환하는 과정으로 이루어진다. 발음에서 한국어음차표기로의 변환은 음소에 기반하여 이루어진다. 음소기반방법은 글자기반 방법에 비하여 발음지식이 필요하다는 점과 두 단계 과정을 거치기 때문에 원어의 발음을 파악하는 과정의 오류가 발음에서 음차표기로 변환하는 과정에 파급된다는 단점이 있다.

[1] 단어 불일치 문제란 같은 의미의 단어가 다른 형태로 나타날 경우 다른 용어로 취급하는 문제를 지칭한다.

글자 및 음소 기반 방법[7]은 글자와 글자에 대응되는 음소 정보를 이용하여 음차표기를 생성하는 기법이다. 글자 및 음소 기반 방법은 음소기반 방법과 마찬가지로 발음지식이 필요하며, 주어진 영어단어에 대한 발음을 파악하는 과정과 발음을 한국어 음차표기로 변환하는 과정으로 이루어진다. 음소기반방법과 다른 점은 발음을 한국어 음차표기로 변환하는 과정에서 음소뿐만 아니라 글자를 이용하는 점이다. 글자 및 음소기반 기법은 글자와 음소 정보를 모두 사용하기 때문에 글자기반방법과 음소기반방법보다 효과적으로 음차표기를 생성하는 장점이 있다.

본 논문에서는 이들 세 가지 음차표기 기법을 이용하여 다양한 음차표기를 생성한 후 이들 결과들에 대한 적합성 피드백을 이용하여 음차표기를 생성한다.

3. 적합성 피드백을 이용한 자동음차표기 모델

3.1 시스템 구조도

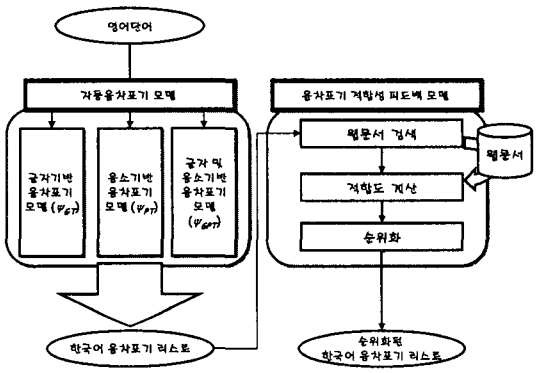
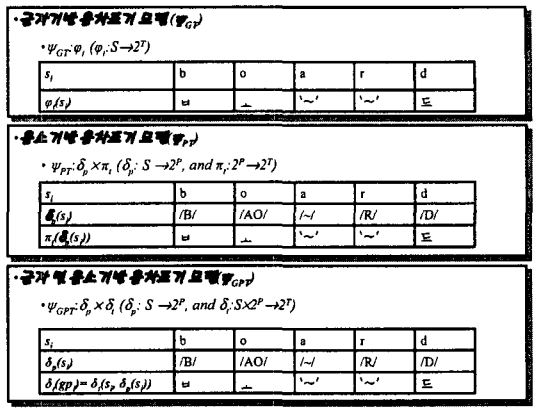


그림 1. 시스템 구조도

그림 1은 적합성 피드백을 이용한 자동음차표기 시스템의 구조도를 나타낸다. 제안하는 시스템은 자동음차표기 과정과 음차표기 적합성 피드백 과정으로 구성된다. 자동음차표기 과정에서는 세가지 음차표기기법을 이용하여 영어단어에 대한 가능한 한국어 음차표기를 생성한다. 음차표기 적합성 피드백 과정에서는 자동음차표기 모델에서 생성된 음차표기 리스트에 대하여 웹문서정보를 이용하여 해당 음차표기의 적합도를 계산한 후 순위화된 한국어 음차표기리스트를 생성한다.

3.2. 한국어 음차표기 리스트의 생성

그림 2는 본 논문에서 사용하는 글자기반 음차표기 모델 (ψ_{GT}), 음소기반 음차표기 모델 (ψ_{PT}), 글자 및 음소기반 음차표기 모델 (ψ_{GPT})을 나타낸다. ψ_{GT} 는 함수 φ_t 로 구성되며, ψ_{PT} 는 함수 δ_p 와 π_t 로 구성된다. 또한, ψ_{GPT} 는 함수 δ_p 와 δ_t 로 구성된다. φ_t 는 주어진 원어의 글자를 대상어의 글자로 변환하는 함수로서 글자에 기반한 음차표기를 수행하는 함수이다. 예를 들어 $\varphi_t(b)$='ㅂ'로 표현된다. δ_p는 주어진 원어의 글자를 음소로 변환하는 함수로서 ψ_{PT} 와 ψ_{GPT} 에서 원어의 발음을 생성하기 위해 사용된다. π_t 는 음소정보를 이용하여 음차표기의 글자를 생성하는 함수이다. 예를 들어 π_t(/B/[2])='ㅂ'로 표현된다. δ_t 는 π_t 와 달리 글자와 음소를 함께 이용하여 음차표기의 글자를 생성하는 함수이다. 예를 들어 $\delta_t(b, /B/)$='ㅂ'로 표현된다. 본 논문에서 사용된 함수들은 현재 글자(C)와 좌우문맥 (L3~L1, R3~R1)을 이용하여 글자 및 음소를 생성하며, 메모리기반 학습방법인 TiMBL[8]을 이용하여 구현되었다. 표 1 은 구현된 δ_p 에 의해 *board* 의 발음이 생성되는 과정을 나타낸다.



• ψ_{GT}: φ_t (φ_t: $S \rightarrow 2^T$)

s_i	b	o	a	r	d
$\varphi_t(s_i)$	ㅂ	ㅗ	'~'	'~'	ㄷ

• ψ_{PT}: $\delta_p \times \pi_t$ (δ_p: $S \rightarrow 2^P$, and π_t: $2^P \rightarrow 2^T$)

s_i	b	o	a	r	d
$\delta_p(s_i)$	/B/	/AO/	/~/	/R/	/D/
$\pi_t(\delta_p(s_i))$	ㅂ	ㅗ	'~'	'~'	ㄷ

• ψ_{GPT}: $\delta_p \times \delta_t$ (δ_p: $S \rightarrow 2^P$, and δ_t: $S \times 2^P \rightarrow 2^T$)

s_i	b	o	a	r	d
$\delta_p(s_i)$	/B/	/AO/	/~/	/R/	/D/
$\delta_t(gp_i) = \delta_t(s_i, \delta_p(s_i))$	ㅂ	ㅗ	'~'	'~'	ㄷ

그림 2 세가지 음차표기 모델: /~/는 묵음을, '~'는 공문자를 나타낸다.

L3	L2	L1	C	R1	R2	R3	δ_p(C)
$	$	$	**b**	o	a	r	/B/
$	$	b	**o**	a	r	d	/AO/
$	b	o	**a**	r	d	$	/~/
b	o	a	**r**	d	$	$	/R/
o	a	r	**d**	$	$	$	/D/

표 1 *board*에 대한 발음 추정의 예: $는 단어의 시작과 끝을 나타내며, /~/는 묵음을 나타낸다.

3.3. 음차표기 적합성 피드백을 통한 순위화된 음차표기 리스트의 생성

음차표기 적합성 피드백 과정에서는 자동음차표기 과정에서 생성된 음차표기리스트에 대한 적합성을 자동으로 파악하고 순위화된 음차표기리스트를 생성하는 것을 목표로 한다. 적합성 파악의 "실제 문서에서 자주 사용되는 음차표기일수록 해당 원어에 대한 적합한 음차표기이다."라는 가정에 기반한다. 이러한 가정에 기반하여 적합성은 원어와 음차표기가 함께 사용된 웹문서의 개수로서 파악된다. 이를 위하여 검색엔진을 사용하였으며, 검색되는 웹문서의 적용률을 높이기 위하여 4가지 검색엔진을 사용하였다. 검색된 웹문서에 의한 적합성(Relevance)은 식 (1)과 같이 표현된다. 여기에서 t_i는 e가 생성하는 i번째 음차표기를 나타내며, $hit_j(x,y)$는 질의어 x,y에 의해 j번째 검색엔진으로 검색되는 웹문서의 개수를 나타낸다. 그리고 Z_j는 j번째 검색엔진에서 검색되는 웹문서에 대한 정규화 인수이며, T는 e로부터 자동음차표기모델에 의해 생성되는 음차표기의 집합을 나타낸다. 적합성값은 각 검색엔진에서 검색한 웹문서의 상대빈도($hit_j(e,t_i)/Z_j$)의 합으로 표현되므로 적합성값의 범위는 0~n으로 표현된다. 여기에서 n은 검색엔진의 개수를 나타낸다. 적합성값이 높은 음차표기일수록 적합한 음차표기로 판단된다. 만약 생성된 모든 음차표기에 대한 적합성 값이 0일 경우 세 가지 음차표기 모델에 의한 투표방식에 의해 적합성이 파악된다.

$$\mathrm{Relevance}(e,t_i) = \sum \frac{1}{Z_j} hit_j(e,t_i); Z_j = \sum_{t_k \in T} hit_j(e,t_k) \quad (1)$$

[2] 본 논문에서는 발음 및 음소를 표현하기 위하여 ARPAbet 기호를 사용한다. ARPAbet 기호는 음소를 ASCII 기호로 표현하기 위한 방법이다.

표 2는 식 (1)에 의해 계산된 data의 음차표기들에 대한 적합성값을 나타낸다. 표에서는 각 음차표기에 대하여 $hit_j(x,y)$와 $hit_j(x,y)/Z_j$을 표현하였으며, 식 (1)의 의해 계산된 적합성 값을 Relevance로 표현하였다. 예를 들어 검색엔진 A에 의해 $hit_j(data_i.\text{데이터})=94100$이며, $hit_j(data_i.\text{데이터})/Z_j=0.581$이다. 각 음차표기의 적합성은 각 검색엔진에 대한 $hit_j(x,y)/Z_j$의 합으로 표현되며, 이를 통하여 Relevance($data_i$.데이터)=2.177이 된다. 적합성에 의해 t_1, t_2, t_3의 순위화된 음차표기 리스트를 생성할 수 있으며, 이 중 t_1이 가장 적합한 음차표기로 판단된다.

검색엔진	t_1=데이터	t_2=데이타	t_3=데타
A	94100 (0.581)	67800 (0.4186)	54 (0.0003)
B	633 (0.4922)	633 (0.4922)	20 (0.0156)
C	101834 (0.796)	26132 (0.2042)	11 (0.0001)
D	1358 (0.3080)	3028 (0.6868)	23 (0.0052)
Relevance	2.1770	1.8018	0.0212

표 2. data의 음차표기 '데이터', '데이타', '데타'에 대한 음차표기 적합성 피드백 결과

4. 실험

4.1 실험환경

본 논문에서는 실험 및 평가를 위하여 7,185개 영어-한국어 음차표기 쌍으로 구성된 실험집합을 사용하였다. 이 중 6,185쌍은 학습데이터로 1,000쌍은 시험데이터로 사용하였다. 사용한 실험집합은 [2,5,7]에서 사용한 실험집합으로 본 논문의 기법과 [2,5,6,7]의 성능평가를 위하여 사용한다. 평가 방법은 음차표기 평가 방법으로 널리 사용되는 단어 정확도(W.A.)와 글자 정확도(C.A.)를 이용한다. 단어 정확도와 글자 정확도는 식 (2)와 같이 표현된다.

$$W.A.=\frac{\#of\ correct\ words}{\#of\ generated\ words}, C.A.=\frac{L-(i+d+s)}{L} \quad (2)$$

여기에서 L은 원문자열의 길이를 나타내며, i,d,s는 각각 원문자열에서 목표문자열로 변환하기 위해 필요한 삽입, 삭제, 치환의 개수를 나타낸다. 만약 $L < (i+d+s)$이면 $C.A.$는 0으로 판단한다 [10].

4.2 실험결과

Method	C.A	W.A	성능향상
[2]	81.80%	51.40%	31.65%
[6]	85.36%	52.40%	30.10%
[5]	88.05%	55.10%	25.83%
ψ_{GT}	85.37%	57.60%	21.89%
ψ_{PT}	86.41%	54.30%	27.09%
ψ_{GPT}	90.45%	63.50%	12.60%
Top1	92.35%	71.5%	
Top2	92.52%	72.7%	
Top3	92.54%	72.8%	

표 3. 실험결과

표 3은 실험결과를 나타낸다. 표에서는 기존의 연구[2,5,6]에 대한 성능과 본 논문에서 사용한 자동음차표기모델(ψ_{GT}, ψ_{PT}, ψ_{GPT})에 대한 성능, 그리고 음차표기 적합성 피드백에 의해 순위화된 결과 중 Top1~Top3에 대한 성능을 각각 나타낸다. 실험결과 기존연구와 ψ_{GT}, ψ_{PT}, ψ_{GPT}보다 최고 31.6%의 성능향상을 나타냄을 알 수 있으며, 이를 통하여 본 논문에서 사용한 음차표기 적합성 피드백 모델이 유용함을 알 수 있었다. 음차표기 피드백 모델이 본 논문에서 사용한 ψ_{GT}, ψ_{PT}, ψ_{GPT}보다 높은 성능을 나타내는 이유는 세 가지 기법의 결과를 음차표기 피드백을 통하여 효과적으로 통합할 수 있었기 때문으로 분석된다. 본 논문의 기법은 Top3에서 수렴되는데 이는 순위화된 결과에서 Top3개의 결과만으로 적합한 음차표기를 파악할 수 있음을 나타낸다. 또한 Top1과 Top3의 성능차이가 크지 않는데 이는 본 논문에서 사용한 적합성 피드백 기법이 매우 효과적임을 나타낸다.

5. 결론

본 논문에서는 적합성 피드백을 이용한 자동음차표기 모델을 제안하였다. 본 논문의 기법은 글자기반, 음소기반, 글자 및 음소기반 자동음차표기 모델을 통하여 생성된 음차표기에 대하여 원어와 음차표기가 나타나는 웹문서의 개수를 통하여 적합성을 파악하였다. 실험결과 본 논문의 기법은 기존의 기법보다 최고 31%의 성능향상을 나타내었으며, 본 논문에서 사용한 각각의 자동음차표기 기법보다 최고 27%의 성능향상을 나타내었다.

향후 다양한 한국어 음차표기를 생성할 수 있는 기법에 대한 추가적인 연구가 필요하며 정보검색 등과 같은 응용시스템에 적용하여 제안한 기법의 효용성을 검증하는 것이 필요하다.

감사의 글

이 논문은 과학기술부, 과학재단의 지원에 의하여 이루어졌습니다.

참고문헌

[1] 이재성, (1999), 다국어 정보검색을 위한 영-한 음차표기 및 복원 모델, 박사학위논문, 한국과학기술원 전산학과

[2] 강병주, (2001), 한국어 정보검색에서 외래어와 영어로 인한 단어불일치문제의 해결, 박사학위논문, 한국과학기술원 전산학과

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

[4] Jeong, Kil Soon, Sung Hyun Myaeng, Jae Sung Lee and Key-Sun Choi, (1999) "Automatic identification and back-transliteration of foreign words for information retrieval," Information Processing and Management, No.35, pp. 523-540.

[5] Kang I.H. and G.C. Kim, (2000), "English-to-Korean Transliteration using Multiple Unbounded Overlapping Phoneme Chunks", In Proceedings of the 18th International Conference on Computational Linguistics.

[6] Goto I., N. Kato, N. Uratani and T. Ehara (2003) Transliteration Considering Context Information Based on the Maximum Entropy Method, In Proceedings of MT-Summit IX

[7] 오종훈, 배선미, 최기선 (2004) 글자 및 발음 기반 영-한 음차표기 모델, 정보과학회춘계학술대회.

[8] Daelemans W., Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, (2002), "TiMBL: Tilburg Memory Based Learner, version 4.3 Reference Guide", ILK Technical Report 02-10, 2002.

[9] 남영신, (1997), 최신 외래어 사전, 국어사전 별책부록, 서울: 성안당 출판사

[10] Hall, P., and G. Dowling, (1980), "Approximate string matching", Computing Surveys, 12(4), 381-402.