자동 음차표기를 이용한 영-한 음차표기 대역쌍의 자동 추출*

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

An Algorithm for extracting English-Korean Transliteration pairs using Automatic E-K Transliteration

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

요 약

지금까지 기계번역과 교차언어 정보검색 등과 같은 자연언어응용에서 사용되는 번역지식을 자동으로 구축하는 연구가 활발히 진행되어 왔다. 번역지식을 자동으로 구축하는 연구는 대역사전에 등재되어 있지 않은 미등록어에 대한 대역정보를 문서에서 자동으로 획득하는 것을 목표로 한다. 최근에는 이러한 미등록어 중 음차표기 번역지식에 대한 연구가 활발히 진행되고 있다. 음차표기는 주로 영어 단어를 발음에 기반하여 비영어권의 언어로 표기하는 것을 의미한다. 음차표기된 단어들은 새로운 개념을 나타내는 신조어가 많기 때문에 사전에 등재되어 있지 않은 경우가 많다. 따라서 효과적인 번역지식 구축을 위해서는 이러한 음차표기 번역지식을 자동으로 획득하는 것은 매우 중요하다. 본 논문에서는 영-한 음차표기 대역쌍을 문서에서 자동으로 추출하는 알고리즘을 제안한다. 본 논문의 기법은 한국어 음차표기의 인식, 영-한 자동음차표기, 한국어 음차표기와 자동음차표기된 영어단어간의 음성적 유사도 비교를 통하여 음차표기 대역쌍을 추출한다. 본 논문의 기법은 약 83%의 정확률과 68%의 재현율을 나타내었다.

1. 서론

지금까지 기계번역과 교차언어 정보검색 등과 같은 자연언어응용에서 사용되는 번역지식을 자동으로 구축하는 연구가 활발히 진행되어 왔다[1,2,3]. 번역지식을 자동으로 구축하는 연구는 대역사전에 등재되어 있지 않은 미등록어에 대한 대역정보를 문서에서 자동으로 획득하는 것을 목표로 한다. 최근에는 이러한 미등록어 중 음차표기 번역지식에 대한 연구가 활발히 진행되고 있다. 음차표기는 주로 영어 단어를 발음에 기반하여 비영어권의 언어로 표기하는 것을 의미한다. 음차표기된 단어들은 새로운 개념을 나타내는 신조어가 많기 때문에 사전에 등재되어 있지 않은 경우가 많다. 따라서 효과적인 번역지식 구축을 위해서는 이러한 음차표기 번역지식을 자동으로 획득하는 것은 매우 중요하다.

주어진 영어단어에 대한 음차표기 대역어를 획득하는 연구로는 자동 음차표기와 음차표기 대역쌍 추출 등이 있다. 자동 음차표기는 주어진 영어 단어를 비영어권의 언어의 단어로 음차표기하는 기법이다[4,5,6,7]. [8,9]과 같은 기존의 교차언어 정보검색 연구에서 자동 음차표기 방법은 번역사전에 존재하지 않는 영어단어의 음차표기를 자동으로 생성함으로써 교차언어 정보검색의 성능을 향상시키는 방법으로 사용되었다. 음차표기 대역쌍 추출은 이중언어 문서 (bilingual corpora)에서 영어와 영어에 대응되는 음차표기된 단어를 자동으로 추출하는 기법이다[10,11,12]. 이들 연구는 번역사전의 적용범위를 높이기 위하여 단어의 번역지식을 이중언어 문서로부터 자동으로 추출하는 연구로서, 번역지식은 음차표기 대역쌍으로 한정하였다.

자동 음차표기와 음차표기 대역쌍 추출은 음차표기된 단어를 처리하기 위한 방법으로 연구가 활발히 진행되고 있지만, 이들 두 방법을 통합적으로 사용하는 연구는 미흡한 상태이다. 음차표기는 영어 단어의 발음에 기반하여 자국어의 단어로 표기하는 방법이다. 따라서 음차표기 대역쌍을 파악하는 것은 음성적으로 유사한 영어 단어와 자국어의 단어를 찾는 작업으로 정의된다. 그런데 영어단어와 자국어의 단어간의 글자체계가 다르기 때문에 음성적 유사도를 비교하기 위해서는 음성적 변환 과정 (phonetic convert process)이 필요하다. 음성적 변환 과정은 자동 음차표기와 같이 한 언어의 단어를 음성적으로 동등한 다른 언어의 단어로 변환하는 작업이다. 따라서 자동 음차표기는 영어 단어에 대한 음성적 변환과정을 자동으로 수행해 주는 모듈로서 음차표기 대역쌍 추출에서 사용될 수 있다.

본 논문에서는 자동 음차표기 방법을 이용한 음차표기 대역쌍의 자동 추출 모델을 제안하다. 제안하는 모델은 이중언어 문서에서 음차표기 대역쌍을 자동으로 추출하는 것을 목표로 한다. 하지만 음차표기가 포함된 대량의 이중언어문서를 획득하는 것은 매우 어려운 작업이므로 본 논문에서는 영-한 이중언어 사전을 이용하여 음차표기 대역쌍을 추출한다. 영-한 이중언어 사전에서 영어 표제어와 대응되는 한국어 표제어를 정렬된 가상의 영-한 문장으로 구성하고 이를 이용하여 가상의 이중언어문서를 구성한다. 그리고 가상의 이중언어 문서를 이용하여 음차표기 대역쌍을 추출한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대하여 기술하고, 3장에서는 제안하는 음차표기 대역쌍의 자동추출 모델에 대하여 기술한다. 4장에서는 실험에 대하여 기술하고 5장에서 결론을 맺는다.

2. 관련연구

기존의 음차표기 대역쌍 추출은 주로 영어-일어[10,11,12]를 대상으로 하였다. Collier 등[11]은 고유명사에 대한 영어-일본어 음차표기 대역쌍을 추출하기 위한 모델을 제안하였다. 그의 연구에서는 일본어 음차표기를 나타내는 '카타카나'를 파악하여 음차표기 대역쌍의 후보를 추출한 후 영어와 일본어 후보에 대하여 대역관계를 파악하였다. 음차표기 대역관계는 NPT (Nearest Phonetic Transliteration)라고 불리는 음성코드를 이용하여 일본어 후보를 영어 단어로 변환한 후 영어단어와의 유사도를 비교하여 파악하였다. Tsuji[12]는 Collier의 모델을 일반 단어까지 확장하여 음차표기 대역쌍을 추출하였다. 또한 다이스 계수에 기반한 문자열 비교 알고리즘을 통하여 유사도를 비교하였다. Brill[10]은 영어-일본어 음차표기 대역쌍 추출을 위하여 잡

*본 논문은 과학기술부, 과학재단, 한국과학기술원 BK21 정보기술사업단의 지원에 의해 이루어짐.

음채널 오류 모델 (noisy-channel error model)과 학습 가능한 편집거리함수(trainable edit distance function)를 이용하였다.

기존의 연구와 본 논문에서 제안하는 기법은 다음과 같은 두 가지 차이점이 있다. 첫째, 일본어에서는 음차표기된 단어가 일반적으로 특정 문자 집합인 '카타카나'로 표기되기 때문에 음차표기 인식 과정이 필요하지 않지만 한국어는 음차표기를 나타내는 문자세트가 없기 때문에 추가적으로 음차표기를 파악하는 알고리즘이 필요하다. 둘째, 기존의 연구들에서는 일본어를 영어의 철자로 변환한 방법인데 비해, 본 논문의 기법은 영어를 한국어로 변환하는 방법이다. Knight[13]의 정의에 따르면 전자를 음차복원 (back-transliteration), 후자를 음차표기(transliteration)이라고 한다. 일반적으로 원어에서 목적어로 음차표기 하는 것은 정보의 손실이 있기 때문에 음차복원의 과정은 손실된 정보를 복원해야 한다는 점에서 음차표기보다는 어려운 작업이다. 이로 인해 음차표기의 성능이 음차복원의 성능보다 높게 나타나며, 음차표기를 이용한 경우가 원어의 단어를 음성적으로 같은 다른 목적어의 단어로 변환시키는 방법으로는 효과적이라 할 수 있다. 본 논문에서는 음차표기 기법을 이용함으로써 보다 정확하게 음성적 변환 과정을 수행한다

3. 글자 및 발음 기반 영-한 음차표기 모델

본 논문에서 제안하는 음차표기 대역쌍 추출 시스템은 '음차표기된 한국어 후보 추출', '음차표기 대역쌍 후보의 추출', '영-한 자동 음차표기', '음성적 유사도 비교를 이용한 음차표기 대역쌍 추출'의 네 단계로 구성되어 있다. 본 장에서는 각 단계에 대하여 3.1절에서 3.4절에 걸쳐 자세히 기술한다.

3.1 음차표기 인식을 통한 한국어 후보의 추출

첫 번째 단계인 한국어 후보 추출 단계에서는 한국어 표제어로 구성된 문서 (한국어 문서)로부터 한국어 음차표기를 인식하여 영어-한국어 음차표기 대역쌍의 한국어 후보를 추출한다. 영-한 병렬 문서에서 모든 영어단어가 한국어로 음차표기되는 것은 아니기 때문에, 음차표기 대역쌍을 추출하기 위해서는 한국어 문서에서 한국어 음차표기 단어를 추출해야 한다. 본 논문에서는 이를 위하여 은닉 마르코프 모델에 기반한 한국어 음차표기 인식 알고리즘을 이용한다[14]. 한국어 음차표기 인식 기법에서는 식 (1)에 기반한 음절태깅 기법을 이용하여 음차표기를 인식한다. 음절태깅이란 주어진 어절에 나타나는 음절이 순수 한국어를 구성하는 음절일 경우에는 *K*라는 태그를, 외래어를 구성하는 음절일 경우에는 *F*라는 태그를 할당하는 문제로 정의된다. 예를 들어, 어절 '오페라(opera)+는'과 '미테랑(Mitterant)'은 각각 '오/F+페/F+라/F+는/K'과 '미/F+테/F+랑/F'으로 음절태깅될 수 있다.

$$\phi(S) = \arg\max_T = P(T \mid S) = \arg\max_T P(S \mid T)P(T) \quad (1)$$

여기에서 *S*는 어절을 나타내며, *T*는 어절 *S*에 대한 음절태그 열을 나타낸다.

음절 태깅을 통한 음차표기 추출은 태깅된 결과에서 *F*태그의 연속을 추출하는 작업으로 생각할 수 있다. 따라서 위의 예제에서 '오/F + 페/F + 라/F'와 '미/F + 테/F + 랑/F'이 음차표기 단어로 추출된다. 추출된 음차표기된 단어는 음차표기 대역쌍 추출을 위한 한국어 후보로 사용된다. 본 논문에서는 첫 번째 단계에서 추출된 한국어 후보의 집합을 $CK=\{ck_1,...,ck_n\}$이라 정의하며, '*CK*={보드, 시스템, 데이터...}'와 같이 표현된다.

3.2 음차표기 대역쌍 후보의 추출

두 번째 단계에서는 첫 번째 단계에서 추출한 한국어 후보를 이용하여 음차표기 대역쌍 후보를 추출한다. 이를 위하여 영어단어의 첫 알파벳과 한국어 후보 단어에 나타나는 첫 음절의 초성간의 음성적 매핑 정보를 사용한다. 예를 들어, 영어 단어의 *board*의 첫 알파벳 *b*는 음성적으로 한국어 단어 '보드'의 첫 자소 'ㅂ'와 대응된다. 본 논문에서는 이를 첫 자소의 음성적 매핑 또는 첫 자소의 음성적 일치라고 정의한다. 영어단어와 음차표기된 한국어에서 첫 자소의 음성적 매핑관계는 매우 정형화된 형태로 나타나기 때문에, 주어진 음차표기된 한국어에 대응되는 영어단어를 추출하는데 도움이 된다. 첫 자소의 음성적 일치 정보는 영-한 음차표기 대역사전[15]을 이용하여 획득하였다. *MR*을 첫 자소의 음성적 매핑 정보의 집합이라고 정의하면 $MR=\{mr(a),mr(b),...,mr(z)\}$라고 표현할 수 있다. 여기에서 $mr(e_i)$는 알파벳 e_i에 대한 첫 자소의 음성적 매핑규칙으로 $mr(e_i)=\{k_1,k_2,...,k_m\}$ 와 같이 표현된다. 예를 들어 영어 알파벳 *b*에 대한 규칙은 $mr(b)=\{ㅂ, ㅃ\}$이다.

본 논문에서는 주어진 한국어후보와 첫 자소의 음성적 매핑 관계를 가지는 영어단어는 음차표기 대역쌍 후보로 정의한다. 음차표기 대역쌍 후보에서 하나의 한국어 후보 ck_i에 대하여 대역 가능한 *k*개의 영어대역후보가 생성된다. 한국어후보 ck_i에 대한 영어대역후보의 집합을 CE_i라고 정의하면 $CE_i = \{ce_{i1},ce_{i2},...,ce_{ik}\}$로 표현된다. 두 번째 단계의 결과로 나타나는 음차표기 대역쌍 후보집합 *CP*는 $CP=\{cp_1,cp_2,...,cp_n\}$과 같이 정의된다. 여기에서 $cp_i=<ck_i, CE_i>$이며, cp_i=<보드, {board, bootstrap, bool,...}>과 같이 표현된다.

3.3 영-한 자동 음차표기를 통한 영어단어의 음성적 변환

세 번째 단계에서는 영-한 자동 음차표기를 통하여 음차표기 대역쌍 후보에서 영어단어의 음성적 변환을 수행한다. 한국어와 영어는 글자 체계가 다르기 때문에 그 자체로 음성적 유사도를 비교하기 어렵다. 따라서 한 언어의 단어를 음성적으로 동일한 다른 언어로 변환하는 음성적 변환 작업이 필요하다. 본 논문에서는 영-한 자동음차표기 기법을 이용하여 영어단어의 음성적 변환을 수행한다. 영-한 음차표기를 통하여 음차표기 대역쌍 후보의 한국어 단어에 대응되는 모든 영어단어를 한국어 음차표기로 자동 생성한다.

본 논문에서는 자소 및 음소 정보에 기반한 자동 음차표기 기법을 이용하여 음차표기를 생성한다. 자소 및 음소 정보에 기반한 자동 음차표기 기법은 영어단어에 대한 발음을 생성한 후 영어단어의 글자와 발음 정보를 이용하여 한국어 음차표기를 생성하는 기법이다. 발음생성은 발음사전을 이용하는 방법과 발음사전에 등재되지 않은 단어에 대한 발음 추정 방법을 사용한다. 자동 음차표기에서 발음 추정과 한국어 음차표기 생성을 위하여 결정트리[16]와 메모리기반 학습[17]을 이용하였다.

세 번째 단계의 결과는 영어단어가 음차표기된 영-한 음차표기 대역쌍 후보, $CPT=\{cpt_1,cpt_2,...,cpt_n\}$로 표현한다. 여기에서 $cpt_i=<ck_i, CE_i, TCE_i>$, $TCE_i=\{tce_{i1},tce_{i2},...,tce_{ik}\}$이며, TCE_i는 CE_i를 영-한 음차표기를 통하여 변환한 집합을 나타낸다. 예를 들어 한국어 음차표기 단어 '보드'에 대하여 cpt_i=<보드, {board, bootstrap, bool,...}, {보드, 부트스트랩, 불,....}>와 같이 표현된다.

3.4 음성적 유사도 비교를 통한 음차표기 대역쌍 추출

네 번째 단계에서는 음성적 유사도 비교를 통하여 음차표기 대역쌍 후보 *CPT*에서 음차표기 대역쌍을 추출한다. 즉, $cpt_i=<ck_i, CE_i, TCE_i>$에서 ck_i와 TCE_i의 원소들 간의 음성적 유사도를 비교하여 음성적으로 가장 유사한 영어후보의 음차표기 형태 tce_{ij}와 tce_{ij}에 대응되는 ce_{ij}를 파악한다. 본 논문에서는 편집거리에 기반한 음성적 유사도를 이용하여 음차표기 대역쌍을 추출한다. "편집거리" 는 주어진 문자열 *s*를 문자열 *t*로 변환하는데 필요한 삭제, 삽입, 치환의 개수로 정의된다[18]. $LD(ck_i,tce_{ij})$를 ck_i와 TCE_i의 원소 tce_{ij} 간의 편집거리로 정의하면, 편집거리에 기반한 음성적 유사도는 식 (2)로 정의된다.

$$sim_p(ck_i,tce_{ij}) = \frac{length(ck_i) - LD(ck_i,tce_{ij})}{length(ck_i)} \quad (2)$$

여기에서, *length(s)* 는 문자열 *s*의 문자개수를 나타낸다.

식 (2)의 음성적 유사도를 이용하여 TCE_i의 원소 중 ck_i와 가장 높은 음성적 유사도를 가지면서 임계치 δ보다 큰 tce_{ij}를 추출한다. 본 논문에서는 임계치를 0.7로 사용하였다. 임계치는 실험적으로 결정되었다. tce_{ij}는 ce_{ij}와 대응되므로 한국어 음차표기 단어 ck_i에 대한 음차표기 대역쌍 tp_i=<ck_i,ce_{ij}>를 식 (3)과 같이 추출할 수 있다.

$$TP_{EK} = \{tp_1, tp_2, ..., tp_n; tp_i = < ck_i, ce_{ij} >\} \quad (3)$$

$$tce_{ij} = \arg\max_{tce_{ik} \in TCE} sim_p(ck_i, tce_{ik})$$

여기에서 CE_i와 TCE_i의 j번째 원소를 각각 ce_{ij}, tce_{ij}라고 나타낸다.

4. 실험

4.1 실험환경

실험을 위하여 전문분야 영-한 대역사전을 사용하였다. 전문분야 대역사전은 물리, 화학, 생물, 기계 등의 20여 분야의 사전이며 총 1,400,000개의 영-한 대역쌍을 포함한다. 실험을 위하여 영어표제어로 영어문서를 구성하고 대응되는 한국어 표제어로 한국어문서를 구성한다. 그리고 영어 문서와 한국어 문서에서 음차표기 대역쌍을 추출한다.

평가를 위하여 본 논문에서 제안하는 기법과 기존연구와의 비교 실험을 수행한다. 그런데 기존의 연구들에서는 음차표기 인식과정과 영어의 음성적 변환 과정이 포함되어 있지 않을 뿐만 아니라 음성적 유사도를 계산하는 방법에 연구의 초점을 맞추고 있기 때문에 직접적인 비교가 어렵다. 따라서 본 논문에서는 음차표기 인식과 영-한 음차표기 과정은 본 논문에서 제시한 기법을 그대로 사용하고 음성적 유사도를 이용한 음차표기 대역쌍 추출 부분에 대해서만 비교실험을 수행한다.

평가는 정확률 (precision), 재현율 (recall), F-값(F-value)으로 평가한다. 정확률은 추출된 음차표기 대역쌍 중 올바른 음차표기 대역쌍의 비율을 나타내고, 재현율은 문서에 나타난 음차표기 대역쌍 중 올바르게 추출한 음차표기 대역쌍의 비율을 나타낸다. F-값은 정확률과 재현율을 통합한 평가 기준이다[19].

4.2 실험결과

표 1은 실험결과를 나타낸다. 표 1에서 A, B, C는 각각 다이스 계수[12], KODEX 알고리즘[5][1], 본 논문의 기법을 나타낸다. 실험결과 본 논문의 기법은 다이스 계수 기법에 비해 약 3.7%의 성능향상을 나타내며, KODEX방법에 비하여 약 21.2%의 성능향상을 나타낸다. 또한 정확률에서는 최고 27.7%의 성능향상을 나타내며, 재현율에서는 최고 13.4%의 성능향상을 나타내었으며, 정확률과 재현율 모두에서 기존의 기법보다 좋은 성능을 나타내었다.

표 1. 실험결과

	정확률	재현율	F-값
A	79.64%	65.55%	71.91%
B	73.07%	53.11%	61.51%
C	82.86%	67.81%	74.58%

5. 결론

본 논문에서는 영-한 음차표기 대역쌍을 자동으로 추출하는 알고리즘에 대하여 기술하였다. 본 논문에서는 자동음차표기를 통하여 영어를 한국어 음차표기로 변환한 후 변환된 음차표기와 한국어 음차표기의 음성적 유사도를 이용하여 음차표기 대역쌍을 추출하였다. 실험 결과, 본 논문의 기법은 약 83%의 정확률과 68%의 재현율을 나타내었으며, 기존연구에 비해 최고 21%의 성능향상을 나타내었다.

향후 본 논문의 기법을 이중언어 코퍼스에 적용하는 연구가 추가로 수행될 예정이다. 본 논문의 기법은 이중언어 코퍼스에 대해서도 제안한 알고리즘의 큰 변화 없이 음차표기 대역쌍을 추출할 수 있을 것으로 기대된다.

[1] KODEX알고리즘은 한국어 자음에 기반한 음차표기된 한국어간의 음성적 유사도 계산 알고리즘이다.

참고문헌

[1] Dagan, I., Church, K.W., and Gale, W.A. (1993), Robust bilingual word alignment for machine aided translation. In Proceedings of the workshop on Very Large Corpora, pp. 1-8

[2] Wu, D., and X. Xia (1994) Learning An English-Chinese Lexicon From A Parallel Corpus, in Proceedings of AMTA 94, pp. 206-213

[3] Smadja F., K. R. McKeown, and V. Hatzivassiloglou, (1995), Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics, 22(1):1- 38

[4] Lee, J. S. and K. S. Choi, (1998), English to Korean Statistical transliteration for information retrieval. Journal of Computer Processing of Oriental Languages, 12(1):17-37..

[5] Kang B.J. and K-S. Choi, (2000), Automatic Transliteration and Back-transliteration by Decision Tree Learning, In Proceedings of LREC'2000.

[6] Oh J.H., and Key-Sun Choi, (2002), An English-Korean Transliteration Model using Pronunciation and Contextual rules, In proceedings of COLING 2002

[7] 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

[8] Qu Yan, Gregory Grefenstette, David A. Evans, (2003), Automatic transliteration for Japanese-to-English text retrieval, In Proceedings of ACM SIGIR'2003, pp. 353-360

[9] Virga Paola and Khudanpur (2003), Transliteration of Proper Names in Cross-Lingual Information Retrieval, in ACL 2003's Workshop on Multilingual and Mixed-language Named Entity Recognition

[10] Brill E., Gary Kacmarcik, Chris Brockett, (2001): Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs. NLPRS 2001: 393-399

[11] Collier, N., Kumano, A. and Hirakawa, (1997), H. Acquisition of English-Japanese proper nouns from noisy-parallel newswire articles using KATAKANA matching. In Proceedings of NLPRS'97.

[12] Tsuji, K. (2002), Automatic Extraction of Translational Japanese-KATAKANA and English Word Pairs from Bilingual Corpora, Journal of Computer Processing of Oriental Languages, vol.15, no.3, p.261-279

[13] Knight, K. and J. Graehl, (1997). "Machine Transliteration". In Proceedings. of ACL'97, Madrid, Spain.

[14] Oh, J.H, and Key-Sun Choi, (2003), A statistical model for Automatic Extraction of Korean Transliterated Foreign words, Journal of Computer Processing of Oriental Languages, 16(1).

[15] 남영신, (1997), 표준외래어사전 성안당출판사

[16] Quinlan, J.R., (1993), "C4.5: Programs for Machine Learning", Morgan Kauffman.

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

[18] Levenshtein V. I. (1965) Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR 163(4) p845-848,

[19] Salton, G. and McGill, M. (1983), Introduction to Modern Information Retrieval, New-York: McGraw-Hill