

동적 윈도우와 토큰라이저를 이용한 영-중 음차표기 대역쌍 자동 추출

(Automatic Extraction of
English-Chinese Transliteration
Pairs using Dynamic Window and
Tokenizer)

김 성 국 [†] 나 승 훈 ^{**}
(Cheng-Guo Jin) (Seung-Hoon Na)

김 동 일 ^{***} 이 종 혁 ^{****}
(Dong-Il Kim) (Jong-Hyeok Lee)

요 약 인터넷의 발달로 대량의 이중언어 문서 구축이 가능해짐에 따라 이런 언어자원을 이용하여 음차표기 대역쌍을 추출하는 연구가 활발히 진행되고 있다. 이런 연구들은 대부분 통계기반 음차표기 모델을 기반으로 하고 있다. 본 논문에서는 기존의 통계기반 음차표기 모델의 문제점에 대하여 분석하고 동적 윈도우와 토큰라이저 기법을 제안하여 약 99%의 단어 정확률을 나타냈으며 그 결과 기존의 통계기반 음차표기 모델에 비하여 약 23%정도의 성능 향상을 보였다.

· 본 연구는 첨단정보기술 연구센터를 통한 과학재단 및 2007년도 두뇌한국21사업의 지원을 받았고 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

· 이 논문은 2007 한국컴퓨터종합학술대회에서 '동적 윈도우와 토큰라이저를 이용한 영-중 음차표기 대역쌍 자동 추출'의 제목으로 발표된 논문을 확장한 것임

[†] 학생회원 : 포항공과대학교 정보통신연구소
chengguo@postech.ac.kr

^{**} 학생회원 : 포항공과대학교 컴퓨터공학과
nsh1979@postech.ac.kr

^{***} 비회원 : 중국연변과학기술대학 컴퓨터공학과 교수
dongil@ybust.edu.cn

^{****} 종신회원 : 포항공과대학교 컴퓨터공학과 교수
jhlee@postech.ac.kr

논문접수 : 2007년 9월 27일

심사완료 : 2007년 10월 29일

: 개인 목적이거나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회는문지: 컴퓨팅의 실제 및 레터 제13권 제6호(2007.11)

Copyright©2007 한국정보과학회

키워드 : 음차표기, 동적윈도우 기법, 토큰라이저 기법, 통계기반 음차표기 모델

Abstract Recently, many studies have focused on extracting transliteration pairs from bilingual texts. Most of these studies are based on the statistical transliteration model. The paper discusses the limitations of previous approaches and proposes novel approaches called dynamic window and tokenizer to overcome these limitations. Experimental results show that the average rates of word and character precision are 99.0% and 99.78%, respectively.

Key words : transliteration, dynamic window, tokenizer, statistical transliteration model

1. 서론

음차표기는 원어의 발음에 근거하여 번역하는 일종의 번역방식이다. 중국어에서는 보통 인명, 지명, 회사명 등 고유명사에 대하여 음차표기를 한다. 그러나 일반적으로 원어의 한 발음에 대하여 여러 개의 비슷한 중국어 발음이 존재하고 또 한가지 중국어 발음에 대해서도 여러 개의 한자가 존재하기 때문에 한 외래어에 대하여 다양한 음차표기 결과가 있을 수 있다. 이런 다양성으로 인하여 생기는 문제를 해결하기 위하여 중국에서는 1993년 65만개 외국인 인명을 수록한 세계인명대사전[1]을 출판하였다. 이 사전은 약 40년 동안에 걸쳐 편집한 것으로서 외국인 이름뿐만 아니라 각 나라 이름에 대하여 어떤 발음을 사용하고 그 발음에 대하여 어떤 한자를 사용할 지에 대해서도 규정되었다. 일반적인 중국어 단어 사전에 10만~20만개 정도의 단어를 포함하고 있는 것에 비하여 60만개 단어는 매우 큰 수임에도 불과하고 새로운 외래어들은 포함하고 있지 않다. 이런 번역 사전 구축은 지속적인 노력과 시간이 필요하므로 본 논문에서는 영-중 병렬 말뭉치에서 자동으로 이런 대역쌍들을 추출하는 방법을 제안한다.

음차표기 대역쌍을 추출하는 방법에는 두 언어 문서에서 각각 음차표기 후보를 찾고 그 후보들 사이의 음성적 유사도를 계산하여 대역쌍을 추출하는 방법[2,3]과 한 언어에서만 음차표기 후보를 찾고 그 후보를 기준으로 다른 언어에서 음차표기 대역쌍을 추출하는 방법[4]이 있다. 중국어는 영어와 한국어처럼 띄어쓰기를 하지 않고 일본어와 같이 외래어를 특수 표기법(가타카나)으로 구분지어 표기하지도 않으므로 중국어 문장에서 음차표기 후보를 찾기는 아주 어렵다. 그러므로 [2]와 같이 영어와 중국어에서 모두 음차표기 후보를 찾고 그것들 사이 유사도를 계산하는 방법을 사용하면 높은 성능을 기대하기가 어렵다. Lee의 논문[4]에서는 영어 문서

에서만 음차표기 후보를 찾고 그 후보를 기준으로 통계 기반 음차표기 모델을 사용하여 중국어 문서에서 대응되는 음차표기를 추출하는 방법을 제안하였다. 그리고 규칙에 기반한 후처리 기법을 통하여 성능을 조금 더 향상시켰다.본 논문에서는 통계기반 음차표기 모델을 기반으로 후처리 과정을 거치지 않고도 높은 성능을 낼 수 있는 동적 윈도우(Dynamic Window) 기법과 토크나이저(Tokenizer) 기법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 동적 윈도우 와 토크나이저에 기반한 음차표기 대역쌍 추출 과정에 대하여 기술하고, 3장에서는 실험에 대하여 기술하고 4장에서는 결론을 맺고 향후 계획에 대하여 기술한다.

2. 영-중 음차표기 자동 추출 모델

본 논문의 영-중 음차표기 자동 추출 모델은 먼저 영-중 병렬 말뭉치의 영어 문장에 고유명사 인식 모듈을 적용하여 고유명사를 추출한 후 그 중에서 음차표기될 영어 단어만 선택하여 대응되는 중국어 문장에서 음차표기 단어를 추출하였다.

2.1 통계기반 음차표기 모델

중국어 음차표기 대역쌍 추출에서는 일반적으로 중국어 한자의 로마표기법, 즉 병음을 사용하여 영어와 비교한다. 예를 들면 “克林顿(클린턴)”이란 중국어 단어는 먼저 KeLinDun이란 병음으로 변환한 후 영어 단어 Clinton과 병음 KeLinDun의 음성적 유사도를 계산하여 비교한다. 본 논문에서는 E는 영어, C는 중국어, TU (Transliteration Unit)는 음차 표기 단위로 가정한다. 그러면 조건확률P(C|E)는 P(克林顿Clinton)로 치환되어 P(KeLinDun|Clinton)확률을 구하는 문제로 전환할 수 있다. 본 논문에서는 영어는 유니그램(Unigram), 바이그램(Bigram), 트라이그램(Trigram)을, 중국어는 병음의 첫 음절, 마지막 음절 혹은 병음 전체를 TU로 사용하였다. 그러면 TU의 정의에 근거하여 P(克林顿Clinton)는 아래 식처럼 여러 개 TU로 분할하여 그 근사치를 구할 수 있다.

$$P(\text{克林顿} | \text{Clinton}) \cong P(\text{kelindun} | \text{Clinton}) \cong P(\text{ke} | \text{C})P(\text{l} | \text{l})P(\text{in} | \text{in})P(\text{t} | \text{d})P(\text{un} | \text{on}) \quad (1)$$

[4]에서는 식(1)에 매치타입이라는 정보를 더 추가하였다. 매치타입은 영어 TU의 크기와 중국어 TU의 크기에 의하여 정의 된다. 예를 들면 위의 식에서 P(ke|C)의 매치타입은 영어 TU C의 크기가 1, 중국어 TU ke의 크기가 2이므로 2-1이다. 매치타입은 파라미터 추정 과정에서 학습이 안된 새로운 파라미터에 대하여 보완 작용을 하므로 [4]에서처럼 발음사전 없이 통계기반 음

차표기 모델을 적용할 때 더 좋은 성능을 기대할 수 있다. 본 논문에서도 발음사전 없이 EM 알고리즘을 적용하여 파라미터를 자동으로 추정하는 방법을 사용하였으며 음차표기 모델에 매치타입 정보를 추가하였다. 식 (1)에 매치타입(M) 정보를 추가하면 식 (2)와 같게 된다. 식 (3)에서 u, v는 각각 영어 TU와 중국어 TU를 의미하며 m은 u와 v의 매치타입을 의미한다.

$$P(C|E) \approx \max_M P(C|M,E)P(M|E) \approx \max_M P(C|M,E)P(M) \quad (2)$$

$$\log P(C|E) \approx \max_M \sum_{i=1}^N ((\log P(v_i | u_i) + \log P(m_i)) \quad (3)$$

영어단어: Clinton			
음차표기: 克林顿(KeLinDun)(클린턴)			
중국어 문장: 明天克林顿将访问韩国.			
(내일 클린턴이 한국을 방문하게 될 것이다.)			
영어	병음	매치타입	한자
	mi	[0,2]	明
	n	[0,1]	
	g	[0,1]	
	t	[0,1]	天
	l	[0,1]	
	an	[0,2]	
C	ke	[1,2]	克
li	li	[2,2]	林
n	n	[1,1]	
t	d	[1,1]	顿
on	un	[2,2]	
	ji	[0,2]	将
	an	[0,2]	
	g	[0,1]	
	f	[0,1]	访
	an	[0,2]	
	g	[0,1]	
	we	[0,2]	问
	n	[0,1]	
	h	[0,1]	韩
	an	[0,2]	
	g	[0,1]	国
	uo	[0,2]	

그림 1 중국어 문장에서 음차표기를 찾아주는 예

그림 1은 영어 단어 Clinton에 대하여 대응되는 중국어 문장에서 정확한 음차표기 “克林顿”(KeLinDun)을 찾아주는 예이다.

2.2 제안 방법

통계기반 음차표기 모델을 적용하여 음차표기 대역쌍을 추출할 때 만약 한 문장에 주어진 영어 단어와 발음상 비슷한 중국어 문자열이 여러 개 존재할 경우 오류가 자주 발생한다. 본 논문에서는 이런 오류를 해결하고자 동적 윈도우 기법과 토크나이저 기법을 제안한다.

2.2.1 동적 윈도우 기법

동적 윈도우 기법은 중국어 문장에 대하여 한번에 최적화된 경로를 찾는 것이 아니라 주어진 영어 단어에 근거하여 가능한 중국어 음차표기 단어크기의 범위를

설정하고 그 범위 내의 윈도우를 각각 앞으로 이동하면서 음차표기를 찾는 기법이다.

만약 중국어 음차표기 단어의 실제 길이를 알 수 있고 그것을 윈도우 크기로 설정하여 음차표기를 찾으려면 아주 높은 성능을 낼 수 있다. 왜냐하면 일반적으로 정확한 음차표기일수록 영어 TU와 중국어 TU사이의 정렬이 더 잘 되기 때문이다. 이런 특성은 중국어뿐만 아니라 다른 언어에도 공통적으로 나타나는 특성이다.

그러나 정확한 중국어 음차표기의 크기를 예측하기 어려우므로 본 논문에서는 학습데이터에서 영어 단어 길이와 중국어 단어 길이 사이의 분포에 대한 분석을 통하여 음차표기 단어 크기의 가능한 범위를 예측하였다. 동적 윈도우를 적용하는 과정은 다음과 같다. 먼저 주어진 영어 단어에 근거하여 윈도우 범위를 예측한 후 예측한 범위 내의 윈도우를 각각 앞으로 이동하면서 주어진 영어 단어와 현재 윈도우가 포함하는 중국어 문자열에 대하여 식 (3)을 이용하여 확률값을 구한다. 이런 방식으로 윈도우 크기를 점차적으로 증가시키면서 가장 높은 확률값을 갖는 중국어 문자열을 찾고 그 문자열을 역추적하여 음차표기를 추출한다. 그림 2는 동적 윈도우를 적용하여 정확한 음차표기를 찾아주는 과정을 나타낸 것이다. 각 윈도우 사이의 점수를 비교할 때에는 윈도우 크기가 커짐에 따라 전반적으로 점수가 낮아지므

로 윈도우 크기로 정규화하여 비교한다.

동적 윈도우 기법을 적용하면 그림 2와 같이 통계기반 음역 모델을 적용했을 때 생기는 대부분 오류들을 해결할 수 있다. 그러나 정확한 윈도우를 적용했을 때에도 여전히 오류가 생기는 일부 문장에 대해서는 해결할 수 없다. 또한 동적 윈도우만 적용할 경우 전체 문장에 대해 다양한 크기의 윈도우로 여러번 정렬 과정을 거치야 하므로 시간 복잡도가 너무 커지게 되는 단점이 있다. 아래에 시간 복잡도를 크게 줄여주고 성능도 조금 더 향상 시켜주는 토큰라이저 기법에 대하여 기술한다.

2.2.2 토큰라이저 기법

토큰라이저 기법은 중국어 음차표기에 전혀 사용되지 않는 문자를 기준으로 중국어 문장을 먼저 여러 부분으로 나누고 각 부분에 대하여 통계기반 음차표기 모델을 적용하여 음차표기를 추출하는 기법이다.

중국어에는 “施(shi), 德(de), 勒(le), 赫(he) ...”와 같은 음차표기에 자주 사용하는 문자 집합이 있는 반면에 발음은 비슷하나 “是(shi), 的(de), 了(le), 和(he), ...” 등 음차표기에는 전혀 사용하지 않는 문자 집합이 있다. 이런 문자들은 보통 조사로서 고유명사 주위에 자주 나타나므로 이런 문자들과 정확한 음차표기가 결합하여 오류를 낼 때가 많다. 예를 들면 “David”의 음차표기는 마지막 “d” 발음을 생략하여 “大卫”(DaWei)로 음차표기한다. 여기서 만약 이런 명사 뒤에 생략한 문자 “d”와 비슷한 발음을 내는 “的”(De)와 같은 조사가 붙으면 “大卫的”(DaWeiDe)로 잘못 인식될 수 있다. [4]에서는 규칙에 기반한 후처리 과정을 거쳐 추출한 음차표기 양 끝에 자주 사용하지 않는 문자가 있으면 제거해주는 방식으로 어느 정도 이런 문제를 해결하였다. 그러나 이런 후처리를 통한 기법은 그림 3과 같은 오류에 대해서는 여전히 해결할 수 없다. 왜냐하면 조사 “是”(Shi)가 다른 문자와 결합하여 음성적으로 영어 단어 Jacey와 비슷한 “者是”(ZheShi)가 정확한 음차표기 “杰西”(JieXi) 대신 인식되었기 때문이다. 이때 후처리 기법을 통하여 “是”를 제거한다 하더라도 나머지 부분 “者”는 정확한 음차표기가 아니다. 본 논문의 토큰라이저 기법에서는 “是”(Shi)와 같은 문자는 음차표기에 전혀 사용하지 않는 문자 집합에 속하기 때문에 이런 문자들을 사전에 제거한다. 그러면 남은 문자 “者”(Zhe)와 Jacey의 음성적 유사도는 “杰西”(JieXi)와 Jacey의 음성적 유사도보다 훨씬 낮아짐으로 정확한 음차표기를 추출할 수 있다. 뿐만 아니라 토큰라이저 기법을 적용하여 전체 문장을 여러 부분으로 나누면 시간 복잡도도 크게 줄어들게 된다. 그림 3은 토큰라이저를 적용하여 정확한 음차표기를 추출하는 과정이다.

영어 단어	Ziegler	
중국어 문장	齐格勒与意大利化学家共同获得了1963年诺贝尔化学奖。(齐格勒와 이태리 화학자 줄리오의 공동으로 1963년 노벨 화학상을 수상하였다.)	
실제 찾은 음차표기	家居里奥 (JiaJuLiAo)	
정확한 음차표기	齐格勒 (QiGeLe)	
절 차		
1. 먼저 영어 단어 Ziegler 의 길이에 근거하여 중국어 음차표기의 가능한 범위를 2 부터 7 사이로 설정한다. (학습데이터에서 영어 단어 길이와 중국어 음차표기 길이 분포를 분석한 결과 영어 단어 길이가 L 일 때 중국어 단어 길이는 모두 L/3 부터 L 사이에 있었다.)		
2. 각 윈도우를 앞으로 이동하면서 최고값을 구한다.		
3. 각 윈도우들에서의 최고값 중에서 다시 최대치를 구하여 그에 대응되는 문자열을 역추적하여 음차표기를 추출한다.		
윈도우 크기	각 윈도우의 최대치에 대응되는 문자열. (역추적하여 얻은 결과는 밑줄로 표시)	점수(윈도우 크기로 정규화)
2	齐格 (QiGe)	-9.327
4	齐格勒与 (QiGeLeYu)	-8.433
5	齐格勒与意 (QiGeLeYuYi)	-9.719
6	家居里奥共同 (JiaJuLiAoGongTong)	-10.458
7	齐格勒与意大利 (QiGeLeYuYiDaLi)	-10.721

그림 2 동적 윈도우를 적용하여 정확한 음차표기를 찾는 과정

영어 단어	Jacey	
중국어 문장	这本书的作者佩尼娜汤姆森和杰西格雷厄姆。(이책의 저자는 페니아 톰슨과 제이시 그레이엄이다.)	
실제 추출된 음차표기	者是(ZheShi)	
정확한 음차표기	杰西(JieXi)	
절 차		
1. 먼저 자주 안 쓰는 문자 “这, 的, 是, 和”(은정부호, 숫자, 영어 자모 등도 포함)을 기준으로 위의 문장을 아래와 같이 여러 부분으로 분할 한다. 本书 作者 佩尼娜汤姆森 杰西格雷厄姆		
2. 그리고 각 부분에 대하여 통계기반 음차표기 모델을 적용하여 점수를 계산하고 최고값을 갖는 문자열을 역추적하여 음차표기를 추출 한다.		
번호	각 부분에 대응되는 문자열 (역추적 하여 얻은 결과를 밑줄로 표시)	점수(각 부분의 문자열 길이로 정규화)
1	本书 (BenShu)	-24.79
2	作者 (ZuoZhe)	-15.83
3	佩尼娜汤姆森(PeiNiNaTangMuShen)	-16.32

그림 3 토큰라이저를 적용하여 정확한 음차표기를 찾는 과정

이와 같이 동적 윈도우 기법과 토큰라이저 기법은 서로 다른 문제를 해결하므로 두가지 방법을 함께 적용하면 더 높은 성능을 낼 수 있으며 동시에 동적 윈도우만 적용했을 때 생기는 시간 복잡도도 크게 줄여줄 수 있다.

3. 실험 및 평가

3.1 실험환경

실험을 위하여 영-중 병렬 말뭉치에서 지명, 인명, 제품명을 등 각종 음차표기 대역쌍을 포함한 300개 문장을 선택하였다. 학습 데이터는 860개 영-중 음차표기 단어쌍을 사용하였다. 성능 평가를 위하여 본 논문에서 제안하는 기법과 기존연구와의 비교 실험을 수행하였다. 성능은 단어 정확률, 글자 정확률, 글자 재현율로 평가

동적 윈도우 기법과 토큰라이저 기법의 타당성을 증명하기 위하여 아래와 같은 실험을 수행하였다. 첫번째는 본 논문의 베이스라인인 통계기반 음차표기 모델 (STM)만 적용한 실험이고 두번째는 통계기반 음차표기 모델(STM)에 동적 윈도우(DW)와 토큰라이저(TOK)를 각각 적용한 실험이고 세번째는 동적 윈도우와 토큰라이저를 동시에 적용한 실험이다. 마지막으로 기존방법과의 성능 비교를 위하여 통계기반 음차표기 모델에 [4]에서 제안한 후처리 기법을 적용한 실험을 수행하였다.

3.2 동적 윈도우 기법과 토큰라이저 기법의 성능 평가

표 1에서 보여주듯이 통계기반 음차표기 모델(STM)만 적용하여 대역쌍을 추출하면 약 75%정도의 성능을 낼 수 있다. 실제로 통계기반 음차표기 모델은 짧은 문

장에서는 비교적 좋은 성능을 낼 수 있으나 문장 길이가 길어짐에 따라 노이즈도 많아지고 따라서 성능도 많이 떨어진다. 반면에 통계기반 음차표기 모델에 동적 윈도우(STM+DW)를 적용하면 현저한 성능향상(약 21%)을 보일 뿐만 아니라 문장 길이가 길어져도 성능은 크게 떨어지지 않는다. 동적 윈도우 기법으로 도달할 수 있는 최고 성능을 측정하기 위하여 실험데이터에서 미리 중국어 음차표기 단어의 길이를 측정하고 그 길이를 윈도우 크기로 설정하여 성능을 측정하였다. 그 결과 (STM+CW) 베이스라인에 비해 약 23%정도 성능이 향상되었다. 이는 동적 윈도우를 적용했을 때와 거의 비슷한 성능 향상이다. 즉 동적 윈도우를 적용하면 윈도우 기법으로 도달할 수 있는 최고 성능에 가까운 성능을 낼 수 있다. 그러나 동적 윈도우만 적용하면 시간적인 복잡도가 지나치게 커지는 단점이 있다.

표 1에서 토큰라이저 기법(STM+TOK)을 적용했을 때 통계기반 음차표기 모델에 비하여 3%정도 성능이 향상되었음을 알 수 있다. 비록 많은 성능향상을 가져 오지는 못했지만 3%에는 대부분 동적 윈도우 기법에서 해결하지 못한 문제들을 포함한다. 그러므로 동적 윈도우와 토큰라이저를 동시에 적용한 결과(STM+DW+TOK) 동적 윈도우만 적용했을 때 보다 3%정도 성능이 더 향상되었다. 이는 윈도우 기법으로 낼 수 있는 최고 성능보다도 1% 더 높은 성능이다.

표 2는 각 방법의 시간 복잡도에 대한 비교 실험이다. 표 2에서 동적 윈도우(STM+DW)만 적용했을 때 시간 복잡도는 통계기반 음차표기 모델(STM)만 적용했을 때의 약 27배임을 알 수 있다. 그러나 동적 윈도우와 토큰라이저(STM+DW+TOK)를 동시에 적용하면 시간 복잡도가 원래의 1/5로 크게 줄어든다. 즉 토큰라이저 기법

표 1 동적 윈도우 기법과 토큰라이저 기법을 적용한 실험 결과

방법	단어 정확률	문자 정확률	문자 재현율
STM (baseline)	75.33%	86.65%	91.11%
STM+DW	96.00%	98.51%	99.05%
STM+TOK	78.66%	85.24%	86.94%
STM+DW+TOK	99.00%	98.78%	99.72%
STM+CW	98.00%	98.81%	98.69%
STM+CW+TOK	99.00%	99.89%	99.61%

표 2 본 논문에서 제안하는 각 방법의 시간 복잡도

방법	실행시간
STM (baseline)	5초 (5751 밀리초)
STM+DW	2분 34초 (154893 밀리초)
STM+TOK	4초 (4574 밀리초)
STM+DW+TOK	32초 (32751 밀리초)

은 동적 윈도우 기법으로 해결하지 못하는 일부 문제를 해결함으로써 성능을 조금 더 향상 시키고 동시에 시간 복잡도도 크게 줄여주는 역할을 한다.

3.3 기존 방법과의 성능 비교

기존 방법과의 성능 비교를 위하여 통계기반 음차표기 모델에 [4]에서 제안한 후처리 기법을 적용한 실험을 수행하였다. 표 3에서 [4]의 후처리 기법을 적용했을 때 통계기반 확률모델만 적용했을 때에 비하여 약 12%정도 성능이 향상되었으나 본 논문에서 제안하는 방법인 동적 윈도우와 토큰나이저를 동시에 적용하면 베이스라인에 비하여 약 23%정도 성능이 향상된다. 이는 정확한 길이를 예측하는 동적 윈도우 기법과 오류를 유발하는 요소를 사전에 제거하는 토큰나이저 기법이 [4]의 규칙기반 후처리 기법보다 약 11%정도 성능이 더 향상됨을 보여준다.

표 3 본 논문의 방법과 기존 방법과의 성능 비교

방법	단어 정확률	문자 정확률	문자 재현율
STM (baseline)	75.33%	86.65%	91.11%
STM+[4]의 후처리	87.99%	90.17%	91.11%

4. 결론

본 논문에서는 통계기반 음차표기 모델에 기초하여 동적 윈도우와 토큰나이저를 사용하는 두가지 방법을 제안하였다. 이 두 방법을 사용하면 별도의 후처리 과정을 거치지 않고도 높은 성능을 낼 수 있다. 동적 윈도우 기법은 기본적으로 정확한 음차표기일수록 대역쌍 사이에 서로 매치가 잘 되는 특성을 이용하였다. 이런 특성은 영-중 언어쌍에만 존재하는 것이 아니라 다른 언어쌍에도 공통적으로 존재하는 특성이다. 본 논문의 토큰나이저 기법은 중국어 음차표기에 쓰이지 않는 문자들을 먼저 제거하는 방식으로 한 문장을 여러 부분으로 나누어서 처리하는 기법이다. 이 방법도 다른 언어에 쉽게 적용할 수 있다. 예를 들면 띄어쓰기를 하는 한국어에서는 띄어쓰기, 문장부호, 숫자 등을 기준으로 문장으로 여러 부분으로 미리 분할하여 처리할 수 있다. 통계기반 음차표기 모델의 파라미터는 발음사전 없이 음차표기 대역쌍 리스트에서 자동으로 추정하여 얻은 것이다. 그러므로 본 논문의 기법은 영-중 언어쌍뿐만 아니라 다른 언어쌍에도 쉽게 적용할 수 있다. 만약 일본어 처럼 외래어를 쉽게 인식할 수 있는 언어에서는 고유명사 인식 모듈과 같은 도구도 이용할 수도 있다. 그러나 앞에서 설명했듯이 아직까지 중국어는 형태소 분석이 어려우므로 고유명사 인식 모듈과 같은 도구를 사용하여 더 좋은 성능을 기대하기 어렵다.

참고 문헌

[1] Xinhua Agency, Names of the world's peoples: a comprehensive dictionary of names in Roman-Chinese (世界人名翻译大辞典), (1993)

[2] Richard Sproat, Tao Tao, ChengXiang Zhai, Named Entity Tranliteration with Comparable Corpora, in: Proceedings of the 21st International Conference on Computational Linguistics. (2006).

[3] Jong-Hoon Oh, Sun-Mee Bae, Key-Sun Choi, An Algorithm for extracting English-Korean Transliterationpairs using Automatic E-K Transliteration In Proceedings of Korean Information Science Society (Spring). (In Korean), (2004).

[4] C.-J. Lee, J.S. Chang, J.-S.R. Jang, Extraction of transliteration pairs from parallel corpora using a statistical transliteration model, in: Information Sciences 176, 67-90 (2006).

[5] J.S. Lee and K.S. Choi, "English to Korean statistical transliteration for information retrieval," International Journal of Computer Processing of Oriental Languages, pp. 17-37, (1998).

[6] K. Knight, J. Graehl, Machine transliteration, Computational Linguistics 24 (4), 599-612, (1998).

[7] W.-H. Lin, H.-H. Chen, Backward transliteration by learning phonetic similarity, in: CoNLL-2002, Sixth Conference on Natural Language Learning, Taipei, Taiwan, (2002).

[8] J.-H. Oh, K.-S. Choi, An English-Korean transliteration model using pronunciation and contextual rules, in: Proceedings of the 19th International Conference on Computational Linguistics (COLING), Taipei, Taiwan, pp. 758-764, (2002).

[9] C.-J. Lee, J.S. Chang, J.-S.R. Jang, A statistical approach to Chinese-to-English Backtransliteration, in: Proceedings of the 17th Pacific Asia Conference on Language, Information, and Computation (PACLIC), Singapore, pp. 310-318, (2003).

[10] Jong-Hoon Oh, Jin-Xia Huang, Key-Sun Choi, An Alignment Model for Extracting English-Korean Translations of Term Constituents, Journal of Korean Information Science Society, SA, 32(4), (2005).

[11] Chun-Jen Lee, Jason S. Chang, Jyh-Shing Roger Jang: Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. ACM Trans. Asian Lang. Inf. Process. 5(2): 121-145 (2006).

[12] Lee, C. J. and Chang, J. S., Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts Using a Statistical Machine Transliteration Model, In. Proceedings of HLT-NAACL, Edmonton, Canada, pp. 96-103, (2003).