동적 윈도우와 토크나이저를 이용한 영-중 음차표기 대역쌍 자동 추출

김성국[O1] 나승훈[2] 김동일[3] 이종혁[2]
포항공과대학교 정보통신대학원 정보처리학과[1] 포항공과대학교 컴퓨터공학과[2]
중국연변과학기술대학 언어공학연구소[3] 첨단정보기술 연구센터
{chengguo[1], nsh1979[2], jhlee[2]}@postech.ac.kr, dongil[3]@ybust.edu.cn

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

Cheng-Guo Jin[O1], Seung-Hoon Na[2], Dong-Il Kim[3], Jong-Hyeok Lee[2]
Dept. of Graduate School for Information Technology, POSTECH, Korea[1]
Dept. of Computer Science & Engineering POSTECH, Korea[2]
Language Engineering Institute, YUST, China[3]
Advanced Information Technology Research Center(AITrc)

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 교류상의 문제를 해결하기 위하여 중국에서는 1993 년 65 만개 외국인 인명을 수록한 세계인명대사전을 출판하였다. 이 사전은 약 40 년 동안에 거쳐 편집한 것으로서 외국인 이름뿐만 아니라 각 나라 이름에 대하여 어떤 발음을 사용하고 그 발음에 대하여 어떤 한자를 사용할 지에 대해서도 규정지었다. 일반적인 중국어 단어 사전에 10 만~20 만개 정도의 단어를 포함하고 있는 것에 비하여 60 만개 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.

음차표기에 대한 기존 연구는 크게 두가지로 나눌 수 있다. 하나는 주어진 영어 단어에 대하여 자동으로 음차표기를 생성해주는 것이고 다른 하나는 이중언어 문서에서 자동으로 음차표기 대역쌍을 추출하는 것이다. 중국어에서는 지금 신화사(“新华社”)의 음차표기 표준을 많이 따르고 있지만 한 개 발음에 대하여 여러 개 한자가 존재하고 어떤 때에는 한자의 뜻도 고려하여 음차표기하기 때문에 사람들이 자주 사용하는 음차표기를 자동으로 생성하기는 쉽지 않다. 그러므로 본 논문에서는 대역쌍 추출에 초점을 맞춘다. 음차표기 대역쌍을 추출하는 방법에는 두 언어 문서에서 각각 음차표기 후보를 찾고 그 후보들 사이의 음성적 유사도를 계산하여 대역쌍을 추출하는 방법과 한 언어에서만 음차표기 후보를 찾고 그 후보를 기준으로 다른 언어에서 음차표기 대역쌍을 추출하는 방법이 있다. 중국어는 영어와 한국어처럼 띄어쓰기를 하지 않고 일본어와 같이 외래어를 특수표기법(가타카나)으로 구분지어 표기하지도 않으므로 중국어 문장에서 음차표기 후보를 찾기는 아주 어렵다. 그러므로 영어와 중국어에서 모두 음차표기 후보를 찾고 그것들 사이 유사도를 계산하는 방법을 사용하면 높은 성능을 기대하기가 어렵다. Lee 의 논문[1]에서는 영어 문서에서만 음차표기 후보를 찾고 그 후보를 기준으로 통계기반 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 기법과 토크나이저 기법을 제안한다.

본 논문의 영-중 음차표기 자동 추출 모델은 먼저 영-중 병렬 말뭉치의 영어 문장에 대하여 고유명사 인식 모듈을 적용하여 고유명사를 추출한 후 그 중에서 음차표기될 영어 단어만 선택하여 대응되는 중국어 문장에서 음차표기 단어를 추출하였다. 중국어 음차표기 대역쌍 추출에서는 일반적으로 중국어 한자의 로마표기법, 즉 병음을 사용하여 영어와 비교한다. 예를 들면 “克林顿”(클린턴) 이란 중국어 단어는 먼저 KeLinDun 이란 병음으로 변환 한 후 영어 단어 Clinton 과 병음 KeLinDun 의 음성적 유사도를 계산하여 비교한다. 본 논문에서는 E 는 영어, C 는 중국어, TU(Transliteration Unit) 는 음차표기 단위로 가정한다. 그러면 조건확률 P(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 TU 로 사용하였다. 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.

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 있다면 그것을 윈도우 크기로 설정하여 음차표기를 찾으면 아주 높은 성능을 낼 수 있다. 왜냐하면 정확한 음차표기일수록 영어 TU 와 중국어 TU 사이에 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 정렬한 확률값을 구한다. 이런 방식으로 윈도우 크기를 점차적으로 증가시키면서 가장 높은 확률값을 갖는 중국어 문자열을 찾고 그 문자열을 역추적하여 음차표기를 추출한다. 각 윈도우 사이의 점수를 비교할 때에는 윈도우 크기가 커짐에 따라 전반적으로 점수가 낮아지므로 윈도우 크기로 정규화하여 비교한다. 동적 윈도우 기법을 적용하면 통계기반 음역 모델을 적용했을 때 생기는 대부분 오류들을 해결할 수 있다. 그러나 일부 정확한 윈도우를 적용했을 때에도 여전히 오류가 생기는 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.

토크나이저 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 조사나 사용빈도수가 매우 높은 문자로서 고유명사 주위에 자주 나타나므로 이런 문자들과 정확한 음차표기가 붙어서 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. [1]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 추출된 음차표기에 정확한 음차표기를 포함하고 있는 경우에만 적용될 수 있다. 만약 추출된 음차표기에 정확한 음차표기를 포함하고 있지 않다면 후처리 기법을 통하여 이런 조사들을 제거 한다 할 지라도 정확한 음차표기를 추출할 수 없다. 본 논문의 토크나이저 기법에서는 음차표기에 전혀 사용하지 않는 문자는 사전에 제거하여 이런 문자를 기준으로 한 문장을 여러 부분으로 나누어서 처리한다. 이처럼 오류를 유발하는 요소를 사전에 제거함으로 후처리 기법을 통하지 않고도 더 높은 성능을 낼 수 있다. 뿐만 아니라 토크나이저 기법을 적용하여 전체 문장을 여러 부분으로 나누면 시간 복잡도도 크게 줄어들게 된다. 이와 같이 동적 윈도우 기법과 토크나이저 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.

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 평가한다. 본 논문의 알고리즘은 한번에 하나의 영어 단어에 대해서만 고려하기 때문에 단어 정확률과 재현율은 같게 된다. 동적 윈도우 기법과 토크나이저 기법의 타당성을 증명하기 위하여 아래와 같은 실험을 수행하였다. 첫번째 실험은 본 논문의 베이스라인인 통계기반 음차표기 모델만 적용한 실험이다. 두번째 실험은 이런 통계기반 음차표기 모델에 동적 윈도우를 적용한 실험이다. 세번째 실험은 토크나이저를 적용한 실험이고 네번째 실험은 동적 윈도우와 토크나이저를 동시에 적용한 실험이다. 마지막으로 [1]방법과의 성능 비교를 위하여 통계기반 음차표기 모델에 [1]에서 제안한 후처리 기법을 적용한 실험을 수행하였다.

통계기반 음차표기 모델만 적용하여 대역쌍을 추출하면 약 75%정도의 성능을 낼 수 있다. 실제로 통계기반 음차표기 모델은 짧은 문장에서는 비교적 좋은 성능을 낼 수 있으나 문장 길이가 길어짐에 따라 노이즈가 많아지고 따라서 성능도 많이 떨어진다. 반면에 통계기반 음차표기 모델에 동적 윈도우를 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 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. 이런 단점은 토크나이저 기법으로 극복할 수 있다.

토크나이저 기법을 적용했을 때 통계기반 음차표기 모델에 비하여 3%정도 성능이 향상된다. 비록 많은 성능향상을 가져 오지는 못했지만 3%에는 대부분 동적 윈도우 기법에서 해결하지 못한 문제들을 포함한다. 그러므로 동적 윈도우와 토크나이저를 동시에 적용한 결과 동적 윈도우만 적용했을 때 보다 3%정도 성능이 더 향상되었다. 이는 윈도우 기법으로 낼 수 있는 최고 성능보다도 1% 더 높은 성능이다. 동적 윈도우만 적용했을 때 시간 복잡도는 통계기반 음차표기 모델만 적용했을 때의 약 27배임을 알 수 있다. 그러나 동적 윈도우와 토크나이저를 동시에 적용하면 시간 복잡도가 원래의 1/5로 크게 줄어든다. 즉 토크나이저 기법은 동적 윈도우 기법으로 해결하지 못하는 일부 문제를 해결함으로써 성능을 조금 더 향상 시키고 동시에 시간 복잡도도 크게 줄여주는 역할을 한다.

기존 방법과의 성능 비교를 위하여 통계기반 음차표기 모델에 [1]에서 제안한 후처리 기법을 적용한 실험을 수행하였다. [1]의 후처리 기법을 적용했을 때 통계기반 확률모델만 적용했을 때에 비하여 약 12%정도 성능이 향상되었으나 본 논문에서 제안하는 방법인 동적 윈도우와 토크나이저를 동시에 적용하면 베이스라인에 비하여 약 23%정도 성능이 향상된다. 이는 정확한 길이를 예측하는 동적 윈도우 기법과 오류를 유발하는 요소를 사전에 제거하는 토크나이저 기법이 [1]의 규칙기반 후처리 기법보다 약 11%정도 성능이 더 향상됨을 보여준다.

본 논문에서는 통계기반 음차표기 모델에 기초하여 동적 윈도우와 토크나이저를 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 토크나이저 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 일본어처럼 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. 본 논문은 영어 문서에서 추출한 고유명사 중에서 음차표기 될 영어 단어에 대해서만 고려하였다. 앞으로 대량의 비교 가능한 말뭉치(comparable corpus)에서 음차표기 대역쌍을 자동으로 추출하는 연구를 하려고 한다. 비교 가능한 말뭉치에 적용할 때에는 추출한 영어 단어가 음차표기가 되지 않을 수 있고, 중국어 문서에 대응되는 음차표기가 없을 수도 있다. 즉 비교 가능한 말뭉치에서의 대역쌍 추출은 병렬 말뭉치에서의 대역쌍 추출보다 더 많은 노이즈를 포함하므로 처리하기가 더 어렵다. 그러나 대량의 비교 가능한 말뭉치에서 대역쌍을 추출하면 같은 단어가 여러번 나타날 수 있으므로 단어 빈도수, 혹은 추출된 중국어 음차표기의 앤트로피(Entropy)정보 등 여러 가지 정보를 이용하여 정확한 대역쌍을 추출할 수 있을 것이다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통한 과학재단 및 2007년도 두뇌한국21사업의 지원을 받았습니다.

참고문헌

[1] 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)

[2] 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)