영-한 병렬 코퍼스로부터 외래어 표기 사전의 자동 구축

이재성†

요 약

본 논문에서는 한국어 번역문과 영어 원문으로 구성된 병렬 코퍼스로부터 자동으로 외래어 표기 사전을 구축하는 시스템을 제안한다. 구축 시스템은 첫 단계로 한국어 문서에서 명사를 추출하고, 두 번째 단계에서 추출된 명사 중 언어 모델에 근거하여 외래어만을 추출한 후, 마지막 세 번째 단계에서 확률적 정렬 방법을 이용하여 외래어에 대응되는 영어를 추출한다. 특히, 외래어는 한국어 어미나 조사가 붙어서 같이 쓰이기 때문에, 한국어 어절 내에서 정확하게 외래어 부분만을 분리하기 위해, 병렬 코퍼스 내에 존재하는 대응 영어 단어 정보를 활용하였다. 또, 문자체계가 다른 두 단어를 같은 문자로 변환하지 않고 직접 옮운 유사도를 비교할 수 있도록 했다. 실험 결과, 성능은 전처리 단계인 한국어 미등록어 및 외래어 추정에 영향을 많이 받았고, 수작업으로 전처리를 한 모델 중 가장 성능이 높은 것은 재현률 85.4%, 정확률 91.0%를 보였고, 전 과정을 자동으로 한 모델 중에서는 재현률 68.3%, 정확률 89.2%를 보였다.

Automatic Construction of Foreign Word Transliteration Dictionary from English-Korean Parallel Corpus

Jae Sung Lee†

ABSTRACT

This paper proposes an automatic construction system for transliteration dictionary from English-Korean parallel corpus. The system works in 3 steps: it extracts all nouns from Korean documents as the first step, filters transliterated foreign word nouns out of them with the language identification method as the second step, and extracts the corresponding English words by using a probabilistic alignment method as the final step. Specially, the fact that there is a corresponding English word in most cases, is utilized to extract the purely transliterated part from a Korean word phrase, which is usually used in combined forms with Korean endings(Eomi) or particles(Josa). Moreover, the direct phonetic comparison is done to the words in two different alphabet systems without converting them to the same alphabet system. The experiment showed that the performance was influenced by the first and the second preprocessing steps; the most efficient model among manually preprocessed ones showed 85.4% recall, 91.0% precision and the most efficient model among fully automated ones got 68.3% recall, 89.2% precision.

1. 서론

서로 다른 언어 문화권 사이의 정보교류가 활발해지면서, 많은 정보들이 번역되어 유통되고 있다. 특히, 새로운 용어나 전문용어 등은 그 언어를 창조해 낸 언어권의 원어를 대개 그대로 사용하거나 음차(transliteration)하여 사용하는 경우가 많다. 우

† 정회원:충북대 컴퓨터정보통신연구소/컴퓨터교육과 교수
논문접수: 2002년 12월 20일, 심사완료: 2003년 4월 7일
* 이 논문은 2000년 충북대학교 발전기금재단연구비에 의하여 연구되었음

리 나라에서도 많은 용어가 영어를 중심으로 많은 서구 언어들로부터 음차되어 외래어로 사용되고 있다. 외래어 표기법은 이런 음차 표기의 기준을 정하기 위해 정해졌지만, 실제적으로는 표준 외래어의 정의가 애매하여 다양한 표기를 허용하고 있다. 즉, 외래어 표기는 기본적으로 그 단어의 발음을 근거로 이루어지지만, 원어 발음 자체가 여러 가지인 경우, 단어의 철자는 같으나 그 언어가 영어, 불어, 독일어 등의 다른 언어 발음으로 표기될 경우 등과 같이 다양한 음차 표기가 가능하기 때문에 같은 철자의 한 단어가 여러 가지 외래어로 표기될 수 있다. 또한, 외래어 표기법에서는 "외래어의 1음운은 원칙적으로 1기호로 적는다"로 규정하고 있으나, 그 음운의 위치에 따라 여러 가지로 다양하게 표기하고 있다. 예를 들어 발음 [t]의 경우, 단어의 끝에서 "트" 나 받침 "ㅅ"으로 표기되기도 하고, 초성의 경우 "ㅌ"으로 표기되기도 한다. 또한, "이미 굳어진 외래어는 관용을 존중하되, 그 범위와 용례는 따로 정한다"라는 규정은 사실상 애매하여, 어느 단어를 표준어로 해야 할 지 일반 사람들이 정할 수 없어서, 표준단체에 의해 표준언어로 정해지기 전까지는 서로 다르게 표기된 문서의 유통이 불가피하다[1,2].

외래어로 표현된 단어들은 대개 전문용어나 고유명사 등으로 정보검색이나 번역에서 주요한 키워드 역할을 하는 단어들이 많다. 따라서, 이러한 외래어의 표기가 통일되지 않거나 다양한 표기를 같은 단어로 인식하지 못할 경우, 기계번역이나 정보검색, 교차 언어(cross lingual) 정보검색 등의 분야에서 그 성능이나 품질이 저하될 수 있다. 이러한 문제를 해결하기 위한 한 방법으로는 새로운 용어가 등장하는 즉시 표준어 심의를 하여 표준 용어로 확정하고 이를 기준으로 번역이 이루어져야만 한다. 하지만, 현실적으로 모든 분야의 용어에 대해 외래어 표기 표준을 만들기는 불가능하다. 보다 실질적인 방법으로는 이미 번역된 용어를 참조하여 필요한 번역을 하거나, 여러 가지 표현을 모두 찾아서 정보검색에 활용하는 방법이다. 이를 위해서는 기존 많은 번역물에서 외래어 표기의 용례를 추출하여 대역어 사전을 구축해야하는 방대한 작업이 필요하다.

대역어 사전을 자동으로 구축하는 방법으로는 대량의 병렬 코퍼스(parallel corpus: 원문과 대응되는 번역문을 모아 놓은 문서집합)로부터 단어의 공기(co-occurrence) 정보를 이용하여 구축하는 방법이 있다[3-8]. 그러나, 순수하게 통계적 정보만 사용할 경우, 방대한 양의 병렬 코퍼스가 필요하고, 그 계산량도 많이 요구되며, 특히, 언어 구조가 다른 번역물들 사이의 정렬은 더욱 더 많은 코퍼스와 계산량을 필요로 한다. 이러한 문제점을 해결하기 위해, 기존의 사전 정보를 이용하거나, 같은 어족의 특징을 이용함으로써, 비교적 적은 양의 코퍼스를 이용하면서도 정확도 높은 정렬을 할 수 있는 방법들이 제시되고 있다[5,6,9].

본 논문에서는 영어 및 대응되는 한국어 번역문으로 구성된 병렬 코퍼스로부터 음운 유사도를 이용하여 외래어 표기 사전을 자동 구축하는 방법을 제안한다. 이 방법은 적은 양의 병렬 코퍼스에 대해 적용할 수 있으며, 상대적으로 적은 계산으로도 추출이 가능하다. 이 결과로 만들어진 사전은 다시 한국어와 영어 병렬 코퍼스 정렬에 사용될 수 있고, 일관적인 번역을 확인하는 수단으로 사용될 수도 있으며, 다국어 정보검색 및 기계번역의 대역어 사전으로 활용될 수 있다.

음운 유사도를 이용하여 외래어와 원어를 비교하기 위해서는 우선 한국어 문서에서 사용된 외래어를 정확하게 추출해야 한다. 외래어 추출은 주로 한국어와 영어를 혼용한 단일 문서내에서 추출하기 위한 방법으로 연구되었지만[10,11,12], 본 연구에서는 병렬 코퍼스의 특징을 이용하여, 외래어와 그에 대응되는 영어 단어의 음운 정보를 활용하여 보다 정확하게 외래어 부분만을 추출하는 방법을 제안한다. 또, 분리된 외래어와 영어를 비교하기 위해서 같은 문자체계로 변환하여 비교하는 기존의 방법과는 다르게, 통계적 정보를 이용하여 음운상으로 비슷한 단어들을 직접 비교할 수 있는 방법을 제안하였다.

논문의 순서는 우선 2장에서 관련 연구로서, 한국어 및 일본어에 대해 진행된 외래어 추출 방법과 그에 관련된 방법들을 간단히 소개하고, 3장에서는 본 논문에서 제안하는 외래어 사전 자동 구축 방법과 처리 모델들을 단계적으로 설명한다. 이어 4장에서는 제안된 모델들을 실험한 결과를 소개하고, 5장에서 그 결과에 대한 분석 및 토의를 하며, 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

2.1 외래어 판별 및 추출

병렬 코퍼스에서 음차 표기된 외래어와 대응 영어를 추출하기 위해서는, 일반적으로 우선 음차 표기된 단어를 찾고, 그 다음 단계로 음차 표기된 단어에 대응되는 영어를 찾는다.

일본어의 경우, 음차 표기된 단어는 항상 가다카나로 표기하므로, 문자 코드가 다른 히라카나로 표시한 순수 일본어로부터 쉽게 구분해 낼 수 있다. 즉, 프로그램으로 자동 추출을 하기 위해서는 단순하게 문자 범위 값을 비교하여 가다카나 문자만을 추출할 수 있다[13,14,15].

한국어의 경우는 외래어와 순수 한국어가 모두 같은 문자 코드를 사용하므로, 외래어 부분을 판단할 수 있는 방법이 필요하다. 일반적으로 외래어 어절을 분별하는 방법은 문자들이 연속되어 나타나는 확률을 기존의 언어 데이터로부터 학습시킨 후, 이를 언어 모델로 정의하고, 새로운 문자열이 어느 언어 모델에 더 확률적으로 높게 생성될 수 있는가를 비교하여 분별한다[16].

예를 들어 외래어 "리듬", "니켈", "케이크" 등은 한국어초성에서는 잘 사용하지 않는 "ㄹ", "ㄴ", "ㅋ" 등이 초성으로 사용되어 순수 한국어와 쉽게 구분된다. 이러한 정보들을 통계적으로 자동 학습한 후, 각 단어를 분석함으로써 순수 한국어와 외래어를 구분해 낸다.

그러나 다음과 같이 조사 혹은 접사들과 함께 사용된 경우, 다시 순수한 외래어만을 추출해야 하는 문제가 발생된다.

오페라는: 오페라(명사) + 는(조사) 혹은 오페(명사) + 라는(조사)
스캔들이: 스캔들(명사) + 이(조사) 혹은 스캔(명사) + 들(접미사) + 이(조사)

첫 번째 경우, "오페라"가 명사이고, 그 뒤에 조사 "는"이 붙은 것으로 파악할 수 있고, "오페"라는 명사에 다시 조사 "라는"이 붙은 것으로 파악할 수도 있다. 마찬가지로 두 번째 경우도 "스캔들"이라는 명사에 조사 "이"가 붙은 것으로 보거나 "스캔"이라는 명사에 접미사 "들"과 조사 "이"가 붙은 것으로 경우에 따라 파악할 수 있다.

Jeong[11]의 연구에서는 어절이 외래어일 확률과 한국어일 확률을 언어 모델에 근거하여 계산하고 외래어일 확률이 높은 어절을 외래어로 구분했다. 다음 단계로 외래어의 전후에 있는 한국어로 된 조사나 어미, 접사 등을 분리하였다. 이를 위해, 단어의 임의 위치를 분리하여 그 부분에 대한 각각의 언어 확률을 계산하고 각 언어의 확률값을 최대로 하는 부분을 외래어의 분리 위치로 하였다.

오종훈[12]의 연구에서는 외래어 구분 및 분리 문제를 태깅문제로 변환하였다. 즉, 각 음절들을 사람이 판단하여 한국어 또는 외래어로 태깅하고, 이를 은닉 마코프 모델로 학습한 후, 새로운 문자열 데이터에 대해 프로그램이 학습된 확률정보에 근거하여 태깅하도록 하였다. 태깅된 결과에 따라, 외래어 음절로 태깅된 연속의 문자열들을 하나의 외래어로 인식하고 분리하여 추출하였다. 따라서, 외래어의 인식과 순수 외래어 부분의 추출을 동시에 하였다.

위에서 언급한 두가지 통계적 방법의 문제는 똑같은 어절이라고 하더라도 문맥에 따라 다르게 해석되는 경우, 이를 판별할 방법이 없다는 것이다. 즉, 앞의 예에서 "스캔들이"를 확률적으로 계산하여 "스캔들"+"이"로 분리되면, 이를 "스캔(scan)"+"들이" 라고 해석해야 할 경우에도 계속 "스캔들" + "이"로 분리하게 된다.

2.2 외래어와 영어의 음운 비교

한글로 표현된 외래어와 알파벳으로 표현된 영어는 다른 문자체계로 인해 단순하게 문자 비교로는 유사도 계산이 불가능하다. 따라서, 원어를 일단 자동으로 음차하여 표기한 후 이를 외래어와 직접 비교하는 방법이나 공통의 기호체계로 일단 바꾼 후에 둘을 비교하는 방법을 기존 연구에서는 사용하였다.

일본어-영어간의 외래어 추출방법에서는 일본어를 수동으로 작성한 규칙에 따라 일단 로마자로 표기한 후, 영어와 비교하기도 하고, 일본어와 영어를 모두 발음기호로 바꾼 후, 비교하는 방법을 이용하기도 했다. 이러한 연구 중에는 일본어를 NPT(Nearest Phonetic Transliteration)인 영어표기 형태로 바꾼 후, 일본어-영어간의 병렬 코퍼스

에서 외래어(일본어)와 그 원어(영어)를 추출한 것으로 Collier의 연구[14]와 Hepburn 표기 규칙으로 일본어를 영어로 바꾼 후 비교하여 추출한 Keita의 연구[15]가 있다.

한국어 문서에 나타난 외래어를 영어와 비교하는 연구는 주로 다국어 정보검색이나 한국어 정보검색에서 다양하게 사용된 외래어 색인어들을 찾기 위한 방법으로 연구되었다. 이 방법은 모두 일단 영어를 외래어, 혹은 외래어를 영어로 변환한 후, 같은 문자체계로 비교를 하였다.

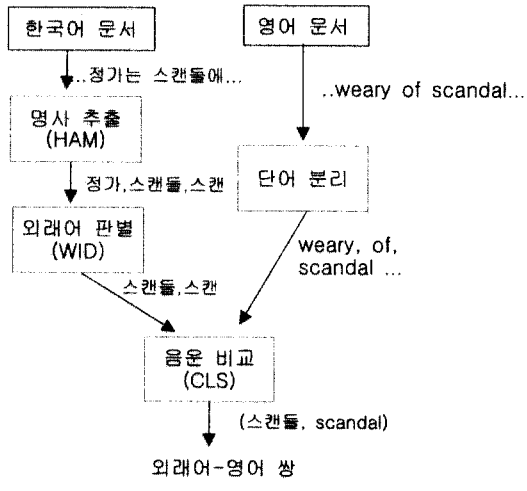
외래어의 표기 즉, 영어를 외래어로 표기하기 위해서는 영어에서 발음기호로 표기한 후, 이를 다시 규칙에 따라 한국어인 외래어로 표기하는 방법이 있고, 영어 알파벳에서 직접 문맥에 따라서 외래어로 표기하는 방법이 있다. 전자를 피봇방식, 후자를 직접방식으로 부른다[10]. 직접방식과 피봇방식의 성능 및 이 둘을 혼합한 혼합방식을 객관적으로 비교한 결과, 정확도면에서 혼합방식이 더 우수한 결과를 보였다(자소 단위의 정확도로 직접방식 71.7%, 피봇방식 70.6%, 혼합방식 75.8%)[10]. 직접방식이 다른 방식들에 비해 처리가 쉬우면서도 성능면에서 그다지 차이가 많지 않고, 특히, 역으로 변환하면 외래어 복원에 쉽게 사용될 수 있으므로 많이 사용된다. 또한, 직접방식은 은닉 마코프 모델(HMM)로 주로 구축되었고, 모델을 직접방식에 맞도록 좀더 조절하여 성능을 향상시키고 있다. 또, 직접방식은 결정 트리[17]나 신경망[18] 등으로 구축되어 성능을 향상시키기도 한다.

Jeong[11]이나 이재성[10]의 경우, 은닉 마코프 모델로 직접방식에 근거한 복원 혹은 표기를 하여 한국어와 영어 혼용문의 정보검색에 적용하였다. Jeong[11]의 경우, 외래어로 된 색인어가 영어 색인과 같은지를 비교하기 위해 우선 직접방식으로 외래어 복원(back-transliteration)을 한 후, 복원된 영어 단어를 원래의 영어 단어들과 비교하였다. 복원이 쉽지 않기 때문에 정확도를 높이기 위해 근사 일치를 쓰거나, 여러 가지의 후보 단어들을 복원하여 그 중에서 일치되는 것을 골랐다. 이재성[10]의 경우, 영어 질의어를 한국어로 변환하기 위해, 직접방식, 간접방식, 혼합방식 모두를 사용하여, 영어를 여러 가지 가능한 외래어로 음차 표기한 후, 이들을 확장된 색인어로 검색을 할 수 있는 방법을 제안하였다. 결국, 다양한 표기를 통해 실제 사용된 표기를 찾을 수 있도록 했다.

두 방법 모두 한국어와 영어가 혼용된 혼용문에서 색인어가 일치하는지를 찾는 문제이지만, 정보검색의 관점에서 처리되었기 때문에, 외래어 부분의 추출시에 대응되는 영어에서 제공되는 정보를 활용하지는 않았고, 외래어와 영어를 비교하기 위해 같은 문자(영어나 한국어)로 변환하여 비교하는 방식을 취했다.

3. 외래어 사전의 자동 구축 단계

영어 문서와 그에 대응되는 한국어 문서가 주어졌을 경우, 외래어 사전을 자동으로 구축하기 위한 방법으로 본 논문에서는 (그림 1)과 같이 3단계 처리 방법을 제안한다. 즉, 한국어 문서에서 명사 부분의 추출, 외래어 판별, 음운 비교를 이용한 대응 영어의 검색 부분으로 나누어 처리한다.



(그림 1) 외래어 사전 자동 구축 과정

명사 부분의 추출은 형태소 분석기를 이용하여 처리하고, 앞에서 설명한 바와 같이, 추정 명사가 여러 개로 나올 경우, 이중에 적당한 것을 다음 단계인 외래어 판별에서 하거나, 그 다음 단계인 대응 영어 검색 부분에서 선택하도록 한다.

외래어 판단 부분은 추출된 명사에 한해 처리하므로 간단한 언어 모델을 이용하여 구현하였다. 대응영어 검색 부분은 주어진 외래어(한국어)를 직접

영어와 비교하여 유사도를 계산하도록 했다. 각 단계의 구체적인 처리 방법은 다음과 같다.

3.1 명사 부분의 추출(HAM)과 단어 분리

한국어 문서에서 외래어는 주로 명사로서 사용된다. 그러나 문장 내에서 사용될 경우, 다른 명사와 마찬가지로 조사나 접사 등이 붙어 주어, 목적어, 부사어, 서술어 등으로 사용되므로 순수한 외래어 부분만을 추출하기 위해서는 우선 명사만을 추출할 필요가 있다.

명사 추출은 형태소 분석 기능 중 미등록어 추정 기능을 이용하여 이루어진다. 미등록어의 추정은 조사나 어미 등의 접사를 수집하여 그 어휘부분을 단어에서 제거하고 나머지 부분을 미등록어로 추정하거나, 미등록어의 패턴을 파악하여 이를 추정하는 방법 등이 있다[19]. 본 논문에서는 이런 방법이 적용된 HAM을 사용하였다[20]. 따라서, 이 과정에서 잘못된 분리가 일어날 가능성이 있으므로, 본 논문에서는 가능한 명사 후보를 모두 출력한다.

예를 들어 (그림 1)에서도 "정가는 스캔들에..."의 문장에서 "정가"라는 명사와 "스캔들"이라는 명사만을 추출해 낼 수 있다. 물론 여기에서 "정가"는 외래어가 아니지만, "스캔들"이라는 외래어를 추출하기 위해서는 이런 추출과정이 필요하다. 특히, "스캔들"과 "스캔"이라는 두가지 단어가 나온 이유는 2.1절에서 설명한 바와 같이 어절 "스캔들에"로부터 미등록어 추정을 통해 분석한 것으로 정확히 어느 부분까지 명사인지를 형태소 분석기가 판단할 수 없으므로 두 가지 후보를 모두 출력한 것이다. 다음 단계인 외래어 판별 단계나 대응 영어 검색(교차 언어 검색) 단계에서는 추출된 명사 중에서 정확하게 명사 부분만 분리한 외래어 단어만을 선택한다.

외래어와 비교하기 위한 영어 단어는 특별히 품사에 제한을 두지 않고 모든 단어를 대상으로 하였다. 대부분의 경우, 띄어 쓴 상태 그대로 외래어와 비교할 수 있으므로 빈칸과, 알파벳 이외의 문자들을 분리기준으로 하여 각 단어를 분리하고 비교할 수 있도록 했다.

3.2 외래어 판별(WID: Word IDentification)

미등록어의 추정이 끝난 단어는 한국어일 수도 있고, 외래어일 수도 있다. 이를 판별하기 위한 방법으로는 기존의 언어 구분(language identification) 방법을 사용했다. 언어 구분 방법은 각 언어에 대해 모델을 만들고, 각 단어에 대해 엔트로피를 계산하여 가장 유사한 언어 모델을 선택하는 방법이다[16].

언어 모델은 여러 가지로 만들어 질 수 있는데, 본 논문에서는 단어를, 외래어 표기에서 하나의 단위로 표기되는 발음단위, 혹은 자주 연속되어 사용되는 1개 또는 그 이상의 자소 단위로 분리한 후, 이들 사이의 연결관계를 언어 모델로 표기한 것을 사용했다. 단어를 W, 발음단위를 U_i라고 하면, 언어 모델은 수식 (1)과 같다. 예를 들어 "케이크"가 외래어일 확률은 발음단위(음차 표기시 한 단위로 자주 나타나는 단위)로 분리되어 "ㅋ", "ㅔㅣ", "ㅋ"로 분리되고 이는 P(ㅋ|$)*P(ㅔㅣ|ㅋ)*P(ㅋ|ㅔㅣ)*P($|ㅋ)의 확률로 계산된다(여기에서 $는 단어 앞뒤의 끝을 의미하는 특수 기호이고, 초성 'ㅇ'은 음가가 없으므로 무시된다.)

$$M(W)=\prod P(U_i|U_{i-1}) \quad \cdots\cdots\cdots (1)$$

위 모델을 순수 한국어 단어 집합과 외래어 단어 집합에 대해 학습시킨 모델을 각각 M_k, M_f라 표기하자. 단어 W에 대한 언어 확률은 각각 $M_k(W)$, $M_f(W)$로 표기되며, $M_k(W) > M_f(W)$일 경우는 한국어, $M_k(W) < M_f(W)$일 경우는 외래어로 판정한다. 즉 식(2)에서 FW(W)값이 1 보다 크면 외래어로 판별한다.

$$FW(W)=\frac{M_f(W)}{M_k(W)} \quad \cdots\cdots\cdots (2)$$

3.3 교차 언어 검색(CLS: Cross Lingual Search)

2.2절에 소개한 기존 연구 방식에서는 외래어와 영어를 비교하기 위해서 하나의 문자체계로 변환하는 방법을 사용했지만, 이 논문에서는 다른 문자체계를 직접 비교하는 방법을 제안한다. 즉, 통계적 음차 표기 방법을 변형하여 음운 정렬의 확률 계산에 직접 이용한다.

원언어에서 음운구조가 다른 대상언어로 음차 표기가 될 경우, 대상언어의 음운구조에 맞게 원언

어의 음운이 변형되어 표기된다[21]. 이러한 과정 중에 새로운 음운이 추가되거나 생략되므로, 대상 언어나 원언어에서 몇 개의 음운이 하나의 단위처럼 음차 표기되기도 한다. 예를 들어 영어-한국어의 경우, "pitcher", "hanger"의 단어 끝에 있는 "er"은 하나의 단위처럼 "ㅓ"로 표기된다. 또 "ch"나 "sh" 등의 문자열도 문맥에 따라 "ㅊ", "치" 나 "ㅅ", "시" 등으로 표기되기도 한다. 이재성[10]에서는 이를 처리하기 위해 통계적으로 자주 함께 사용되는 문자열을 하나의 발음단위로 취급하여 통계적 방법으로 음차 표기하였다.

직접방식에 의한 발음단위의 음차 표기의 확률은 다음과 같이 표기될 수 있다. 주어진 외래어 표기 K에 대해 원어인 영어 E가 대응될 확률은 P(E|K)로 표시된다. 이것은 다시 베이스 규칙으로 식(3)과 같이 쓸 수 있으며, P(K)는 이미 완성된 한국어 단어에 대한 확률이므로 1로 볼 수 있어 식(4)와 같이 된다[4]. 식(4)는 외래어로부터 대응되는 영어를 복원하기 위한 식으로 사용될 수 있으며, P(E)는 복원된 영어 단어가 올바른 영어 단어일 확률을 나타낸다. 그러나 이 식이 외래어와 주어진 영어 단어에 대한 비교식으로 사용될 때, P(E)는 이미 완전한 영어 단어에 대한 확률로 보아 1로 처리될 수 있다. 이를 식으로 쓰면 식(5)와 같다. 결국, 주어진 한국어(외래어)에 대한 영어 단어의 음차 확률 P(E|K)는 P(K|E)와 같다. 다시 말해, 주어진 외래어에 대해 어떤 영어 단어가 음차될 확률을 구하기 위해서는, 영어 단어에서 한국어(외래어)로 음차 표기되는 확률을 계산하면 된다. 이런 확률은 외래어 표기 용례를 통해 계산이 가능할 것이다. 다시 수식으로 돌아와서 식(5)는 단어에 대한 값이므로 이를 좀더 일반적인 단어들에게도 적용하기 위해 좀더 작은 단위인 한국어 발음단위 KU 및 영어 발음 단위 EU의 식으로 바꾸고, 발음단위로 분리되는 확률을 무시하여 식을 간단하게 나타내면 (6)과 같게 된다.

$$P(E|K) = \frac{P(K|E)\times P(E)}{P(K)} \quad \cdots\cdots(3)$$

$$= P(K|E)\times P(E) \quad \cdots\cdots(4)$$

$$= P(K|E) \quad \cdots\cdots(5)$$

$$\cong \prod_{i=1}^{n} P(KU_i|EU_i) \quad \cdots\cdots(6)$$

이 식으로 각 단어들의 음운 정렬의 확률을 계산할 수 있다. 또, 이 식에서 사용되는 각 발음단위들의 정렬 확률은 자동 혹은 수동으로 미리 발음단위를 분리한 영-한 외래어 단어 쌍으로부터 자동 학습할 수 있다.

위 수식은 한글 단어와 영어 단어가 1:1로 대응되는 경우이지만, 여러 가지 대응에 대해 확장하여 사용할 수 있다. 실제 한국어와 영어 단어의 대응은 1:1, 1:2, 2:1, 2:2 등으로 다양하게 발생될 수 있으며, 1:1이 가장 많고, 다음으로 1:2인 경우가 많다. 복합어의 경우, 한국어 명사 2개가 하나로 붙여져서 복합어로 사용되는 경우가 많다고 볼 수 있다. 본 연구에서는 식(6)를 기본으로 하여 단일어 정렬 모델과 복합어를 처리할 수 있는 다중어 정렬모델을 유도했다. 다중어 정렬모델은 다양한 대응 정렬 중 가장 많이 나타나는 1:2 대응(한국어 1단어: 영어 2단어)만을 처리할 수 있도록 우선 제한하였다. 각각의 확률 모델은 다음과 같다.

단일어 모델 CLSS (Simple word)

수식 (6)를 다시 대응된 발음단위 갯수에 따라 정규화하여 표시한 것이 (7)이다.

$$CLSS(E|K) = P(E|K) \cong \sqrt[n]{\left(\prod_{i=1}^{n} P(KU_i|EU_i)\right)} \quad \cdots\cdots (7)$$

다중어 모델 CLSM (Multiword)

하나의 한국어 다중어에 대해 연속한 두 영어 단어를 직접 연결하여 한 단어처럼 취급한 후, CLSS 모델의 방법을 적용한 것으로 식(8)과 같다. 예를 들어 CLSM(color, printer | 컬러프린터) 는 CLSS(colorprinter | 컬러프린터)로 변환하여 계산한 것이다.

$$CLSM(E_w, E_{w+1}|K)$$
$$= P(E_w, E_{w+1}|K)$$
$$\cong P(E_w E_{w+1}|K)$$
$$= CLSS(E_w E_{w+1}|K) \cdots\cdots (8)$$

3.4 명사 후보의 선택 방법

형태소 분석기는 사전을 중심으로 한 어절을 분석하여 문법적으로 가능한 명사를 추출해 낸다. 그러나, 문장 내의 쓰임에 따라 해석이 여러 가지로 되는 경우, 그 가능한 해석을 모두 출력할 수 있다. 특히, 미등록어의 경우는 그 단어가 어느 부분까지 명사이고, 접사 내지는 조사인지를 구분하기가 힘들다. 따라서, 후보 명사를 여러 가지로 출력하게 된다.

특히, 외래어는 미등록어가 많으므로 가능한 후보를 모두 고려하여 외래어 추출을 할 경우, 정확한 범위의 외래어로 된 한국어 명사를 분리하지 못하는 경우가 발생하여, 결과적으로 잘못된 외래어 표기를 추출하게 된다. 이 문제를 해결하는 방법으로서, 2.1절에서 통계적 방법으로 한국어 정보만을 이용하여 처리하는 두 가지의 기존 연구를 소개하였다[11, 12]. 본 논문에서는 2개 이상의 후보 명사 중 정확한 외래어 후보를 선택하기 위한 방법으로 3가지 방법을 제안한다. 즉, 특별한 제한이 없는 방법인 CLSX-HAM과 통계적 정보를 이용하되 한국어 언어 정보만을 사용하는 방법인 CLSX-WID와, 원어의 정보를 활용하는 방법인 CLSX-CLS를 제안한다. 여기에서 CLSX로 표시된 부분은 앞절에서 설명한 단일어 모델(CLSS)과 다중어 모델(CLSM)로 대체될 수 있으며, 각 방법에 대한 좀더 구체적인 설명은 다음과 같다.

1. CLSX-HAM: HAM단계에서 후보 선택을 하는 모델로 HAM의 출력으로 나온 후보 명사를 모두 명사로 취급하여 추출
2. CLSX-WID: WID단계에서 후보 선택을 하는 모델로 각 후보 명사를 WID로 계산한 후, 외래어로 판별된 후보 명사들 중에서 외래어 확률값이 가장 높은 것 1개를 명사로 인정
3. CLSX-CLS: CLS단계에서 후보 선택을 하는 모델로 각 후보 명사를 WID로 일단 계산하여 외래어인 명사를 추출하고(한 어절에 여러 후보가 나올 수 있음), 다시 그 명사들 중에서 영어와 음운 비교(CLS)를 하여 가장 유사도가 높은 명사 1개를 선택

CLSX-WID는 한국어 정보만을 이용하여 외래어 부분을 정확하게 추출하기 위한 방법이다. 예를 들어, "멜로드라마" 와 "멜로드라마적"이라는 단어에 대해 외래어일 확률을 계산하면, 순수한 외래어만 포함된 "멜로드라마"가 더 확률이 높을 것이라는 가정에 근거한 방법이다. 이 방법은 기존의 연구인 Jeong[11]과 오종훈[12]의 연구와 근본적으로는 유사점이 많으나, 기존 연구는 단어에서 임의의 부분을 분리하고 각각에 대한 외래어 확률 혹은 태깅 확률을 구하여 비교한데 반해, 이 방법은 형태소 분석 결과로 나온 각각의 후보 명사들에 대해서만 외래어일 확률을 구하여 그 중 가장 확률이 높은 후보 명사를 선택한 것이 차이가 있다.

CLSX-CLS는 한국어 정보 뿐만아니라, 병렬 코퍼스의 장점을 살려, 영어에서 제공된 정보를 활용한 방법이다. 즉, 후보 명사 "멜로드라마"와 "멜로드라마적"이 나왔을 경우, 영어 단어 "melodrama"가 대개 병렬 코퍼스에 존재하기 때문에, 이에 대한 음운 유사도를 계산할 경우, 올바른 답이 되는 "멜로드라마"가 "멜로드라마적"보다 유사도가 높을 것이라는 가정에 근거한 방법이다.

4. 실험 및 결과

4.1 학습 코퍼스

실험에 필요한 모델 중 학습이 필요한 것은 외래어 판별을 위한 언어 모델과 교차 언어 검색을 위한 단일어 모델이다. 다중어 모델은 단일어 모델에서 자동 생성되므로 직접적인 학습은 필요 없다.

단일어 모델의 학습을 위해서는 논문[10]에서 모델 학습에 사용된 1,500 단어 쌍의 외래어-영어 단어 집합을 사용하였다. 한국어 언어 모델(M_k)을 학습하기 위해, 형태소 분석기용으로 구축한 명사 사전에 있는 순수 한국어 단어 1,559개를 사용하였고, 외래어 언어 모델(M_f)은 마찬가지로 논문[10]에서 사용된 1,691개의 외래어 단어를 사용했다.

4.2 테스트 코퍼스

사전 추출에 필요한 실험용 코퍼스는 2가지를 사용했는데, 하나는 정보검색용으로 구축된 코퍼스인 KTSET 2.0 [22]에 있는 이중언어 코퍼스와 다른 하나는 뉴스위크 코리아의 웹페이지[23]에서 추출한 영-한 대역문이다. 각 코퍼스의 특성은 다음과 같다.

KTSET 코퍼스

원래의 4,404문서 중에서 한-영 문서가 있는 처음의 100문서를 선택했다. 이 부분은 정보과학회지의 논문 요약으로 한글 요약과 그에 대한 영어번역을 포함한 문서이고, 그 중에서 문서 설명이나 제목을 제외한 요약만을 대상으로 했다. 문서당 영어의 평균 단어수는 88개이고 한국어의 평균 어절수는 57개이다.

NWK 코퍼스 (뉴스위크 한글판)

뉴스위크 코리아의 웹사이트에 게재되어 있는 잡지의 영-한 대역 컬럼에서 추출했다. 추출된 문서쌍은 모두 97개이며 353호부터 364호까지 실린 것으로 주로 시사적인 내용을 담고 있고, 많은 고유명사와 외래어 표기가 포함되어 있다. 문서당 영어의 평균 단어수는 92개이고 한국어의 평균 어절수는 68개이다.

4.3 단일어 모델과 다중어 모델의 비교

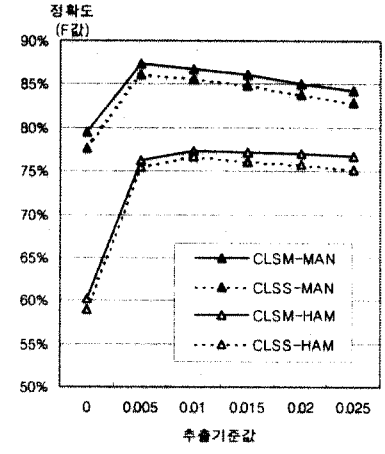
단일어 모델(CLSS)과 다중어 모델(CLSM)은 주어진 외래어에 대해 영어 단어가 음차되어질 확률을 나타내므로, 일정한 값 이하일 경우, 음차되지 않은 것으로 판단할 수 있다. 이 판단의 기준값을 추출 기준값이라고 정의하고 이를 실험에 사용한다. 또, 실험의 결과를 측정하기 위해 재현률(전체 정답에 대해 추출된 정답 비율)과 정확률(전체 추출된 단어쌍 중에 정답이 포함된 비율)로 측정하고 이를 한 단위로 표현한 F-값을 사용한다. F-값은 재현률과 정확률의 비율을 조절할 수 있는데, 여기에서는 1:1로 고정하였으며 그 식은 (9)와 같다.

$$F-\text{값} = \frac{2\times\text{재현률}\times\text{정확률}}{(\text{재현률}+\text{정확률})} \quad \cdots\cdots (9)$$

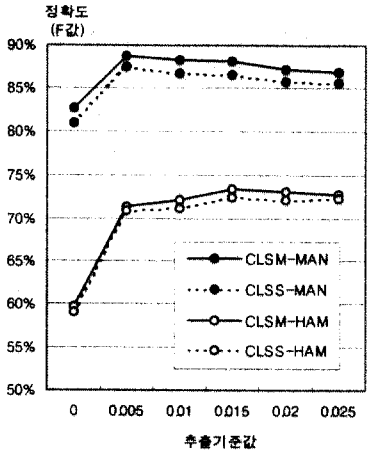
비교를 위해 실험에서 하나의 모델을 더 추가한다. 즉, 전처리 단계를 수작업으로 하여 외래어를 추출한 후, 이를 음운 비교(CLS) 방법으로 대응되는 영어 단어를 추출하는 모델로 CLSX-MAN이다. 이 모델도 단일어 모델과 다중어 모델로 나누어, CLSS-MAN과 CLSM-MAN으로 실험한다. 이 모델은 전처리 단계에서 형태소 분석기와 언어 구분 프로그램을 사용한 CLSX-HAM, CLSX-WID, CLSX-CLS와의 비교 모델로 사용된다.

단일어 모델과 다중어 모델을 비교하기 위해 전처리 과정을 자동으로 한 것 중 가장 기본적인 모델인 CLSX-HAM을 실험하였고, 전처리 과정에 관계없이 순수하게 음운비교(CLS)의 효과를 측정하기 위해 CLSX-MAN 모델을 실험하였다. 추출 기준값을 0.0에서 0.005씩 증가시켜가며 0.025까지의 성능을 F-값으로 나타낸 것이 (그림 2)와 (그림 3)이다. (그림 2)는 KTSET 데이터에 대해, (그림 3)은 NWK 데이터에 대한 결과이다.

실험 결과에서 보듯이 모든 모델이 추출 기준값을 0에서 점차 증가시킴에 따라, 모델의 성능이 증가하다가 특정 기준값부터 다시 점차 감소하였다. 최고점이 되는 부분은 각 모델에서 약간씩 차이가 있다. 전처리를 수동으로 처리한 CLSS-MAN이나 CLSM-MAN은 모두 테스트 데이터에 관계없이 0.005점에서 최고점을 나타냈고, 전처리를 자동으로 한 CLSS-HAM과 CLSM-HAM은 KTSET 테스트 데이터에서는 0.01점에서 최고, NWK 테스트 데이터에서는 0.015점에서 최고를 나타냈다. 전처리 방법이나 테스트 데이터, 추출 기준점에 관계없이 일관적으로 CLSM모델이 CLSS모델보다 좀 더 나은 성능을 보였다.



(그림 2) CLSS와 CLSM의 비교 (KTSET)



(그림 3) CLSS과 CLSM의 비교 (NWK)

4.4 전처리 단계의 정확도 비교

외래어 추출의 전처리 과정은 3가지로, 한국어 문서에 대한 형태소 분석과 외래어 추출, 영어 문서에 대한 단어 분리 과정이다. 이중 영어 문서의 단어 분리 과정은 특수 문자 등을 기준으로 분리하면 비교적 간단하게 처리되므로 정확도에 그리 큰 영향을 주지 않는다. 그러나 한국어 문서에서 외래어 추출은 문장 내에서의 쓰임이나 미등록어 등의 문제로 인해 좀더 성능에 영향을 준다. 즉, 일단 명사를 형태소 분석기가 분리해 내고, 그 중에서 외래어로 추정되는 단어들을 선별하는 과정에서 잘못된 외래어를 추출하는 경우가 많다.

그 과정의 문제점을 파악하기 위해 외래어 추정 명사를 얼마나 정확하게 추출하는가를 분석하였다. 즉, HAM과 WID의 출력 결과를 최종 답에 나타나는 외래어 단어와 비교하여 얼마만큼 일치하는가를 재현률, 정확률, F-값으로 계산하였다. 그 결과는 <표 1>에 나타나 있고, 이 결과는 KTSET 테스트 데이터에 대해 추출 기준값 0.01로 CLSM을 처리했을 때 추출되는 외래어 명사의 정확도를 각 단계별로 나타낸 것이다. HAM 단계에서는 추출된 후보 명사가 3,559개로 그 중에는 정답 326개가 포함되었다. HAM의 출력에 대해 다시 WID를 수행한 결과 457개의 후보 명사를 추출했고, 그 중에 정답인 명사가 326개 포함되어, HAM단계에서 추출된 정답은 모두 그대로 추출했다. 이 결과를 다시 CLS에 적용한 결과, 321개 후보 명사를 추출했고, 그 중 269개의 정답 명사만을 선택해 냈다. CLS 과정에서는 외래어와 영어의 비교시 정렬 확률이 낮게 계산되어 제거된 정답 단어가 57(=326-269)개이었고, 잘못된 단어와 정렬되어 정답으로 나온 단어가 52(=321-269)개이었다. (여기에서 주의해야 할 것은 이 성능은 외래어-영어 단어 쌍의 추출이 아닌 외래어 명사의 추출만을 고려한 것이므로, WID의 수행 결과의 F-값이 CLS의 수행 결과보다 높을 수 있다는 것이다.)

<표 1> 명사만의 추출 성능

단계	답	추출	일치	재현률	정확률	F값
HAM	409	3559	326	79.7%	9.2%	16.4%
WID	409	457	326	79.7%	71.3%	75.3%
CLS	409	321	269	65.8%	83.8%	73.7%

4.5 명사 후보의 선택 방법 비교

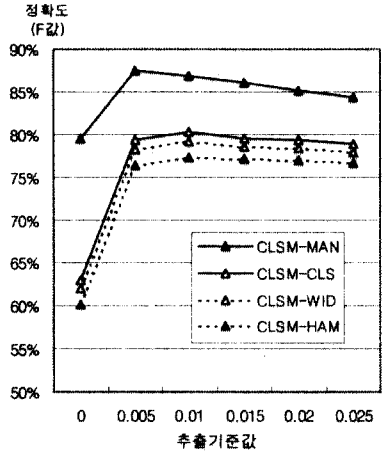
명사 후보 선택 방법의 3가지 실험은 앞의 모델 비교에서 우수한 성능을 보였던 다중어 모델을 이용하였으며, 두가지 테스트 데이터에 적용한 결과는 (그림 4)와 (그림 5)에 나타냈다. 비교를 위해 앞 절에서 실험한 모델인 CLSM-MAN 모델의 성능도 함께 나타냈다.

CLSM-CLS 모델이 항상 CLSM-WID나 CLSM-HAM보다 우수한 결과를 보였다. 그러나, CLSM-WID는 CLSM-HAM에 비해 항상 좋은 결과만을 내놓지는 않았다. 즉, KTSET에 대해서는 CLSM-WID가 좀더 우수했지만, NWK에 대해서는 오히려 CLSM-HAM보다도 성능이 나빴다. 그 이유는 WID에서 후보 명사의 선택이 잘못되어, 올바른 정답을 초기에 제거시켰기 때문으로 해석된다. (KTSET의 경우, WID는 HAM의 출력 중 정답에 해당되는 명사를 대부분 추출했지만, NWK는 일부를 제거하여 성능이 떨어졌다.)

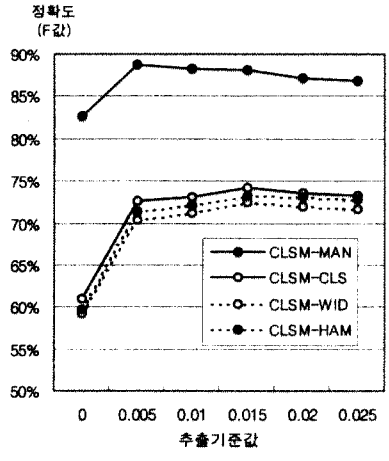
전체적으로 수작업에 의한 모델인 CLSM-MAN이 가장 높은 성능(재현률 85.4%, 정확률 91.0%)을 보였고, 전 과정을 자동으로 한 모델 중에서는 CLSM-CLS 모델이 가장 높은 성능(재현률 68.3%, 정확률 89.2%)을 보였다. <표 2>는 두 모델의 최고점에서의 성능을 재현률 및 정확률로 표시한 것이다.

<표 2> CLSM-MAN과 CLSM-CLS의 최고성능

모델	테스트 데이터	추출 기준값	재현률	정확률	F값
CLSM-MAN	KTSET	0.005	85.2%	89.8%	87.4%
	NWK	0.005	85.5%	92.2%	88.7%
	평균		**85.4%**	**91.0%**	**88.1%**
CLSM-CLS	KTSET	0.01	73.5%	88.3%	80.2%
	NWK	0.015	63.2%	90.0%	74.3%
	평균		**68.3%**	**89.2%**	**77.2%**



(그림 4) 명사 후보 선택 방식 비교(KTSET)



(그림 5) 명사 후보 선택 방식 비교(NWK)

4.6 구축된 사전 예

본 실험에서는 4.1절에서 설명한 태스트 코퍼스로부터 사전을 자동 구축하여, 중복된 단어 쌍을 모두 제거하여 KTSET에서는 163개의 단어 쌍, NWK는 359개의 단어 쌍을 생성했다. (본 논문에서는 각 방법의 성능을 평가하기 위해 적은 양을 구축하였지만, 더 많은 이중언어 코퍼스에 대해 처리하면, 많은 단어 쌍을 구축할 수 있을 것이다.) <표 3>은 KTSET에서 CLSM-CLS 방법으로 구축된 사전의 앞부분 일부를 보여준다. 이 예에는 올바른 단어 쌍으로 구성된 부분도 있지만, "데이타-that" 처럼 잘못된 부분도 있다. 이런 오류는 대개, 순수 한국어를 외래어로 잘못 판별하거나, 외래어는 존재하지만 대응되는 영어가 존재하지 않는 경우에, 가장 비슷한 영어 단어를 선택해 내기 때문에 발생된다. 오류의 다양한 원인은 5장 토의에서 자세히 분석한다.

<표 3> 외래어 표기 사전 구축 예

그래프	graph
그래픽	graphic
네트	net
네트워크	network
네트워크	networks
노드	node
노드	nodes
데드라인	deadline
데이타	data
데이타	that
데이타베이스	data_base
데이타베이스	database
데이타베이스	databases
데이터	data
데이터	that
도큐멘트	document
:	

5. 토의

영어-한국어 번역문에서 직접 외래어 사전을 추출하는 연구가 아직 없었으며, 단지, 관련 연구에서 설명했듯이 한영 혼용문서에서 색인어를 비교하기 위한 방법으로는 사용되었었다. 따라서, 본 연구 결과와 직접 비교할 실험 결과는 없다. 간접적으로 결과를 비교하기 위해 영어-일본어 번역문에서 외래어를 추출한 Kang[13]의 결과를 보면,

일본어와 영어를 모두 음성기호로 바꾼 후, 두 문자열을 비교하여 일본어로 표기된 외래어를 찾아낸 비율이 80%이었다. 일본어에서 외래어는 문자 코드가 다른 가타카나를 사용하므로 특별한 작업 없이도 분리되어 찾아 낼 수 있기 때문에 그 결과는 본 실험의 CLSS-MAN 방법이나 CLSM-MAN 방법과 비교해 볼 수 있다. 즉, CLSM-MAN 방법의 재현률이 NWK, KTSET 모두에 대해 약 85%이므로, 매우 우수함을 알 수 있다. 또한, NPT방법을 사용한 Collier[14]의 연구 결과(재현률 75%, 정확률 82%)와 비교해 보아도 <표 2>에서 나타낸 바와 같이 본 실험의 CLSM-MAN 방법의 결과(재현률 85.4%, 정확률 91%)가 훨씬 정확함을 알 수 있다. 또 최근의 Keita[15]의 연구 결과도 75%의 재현률 상황에서 코퍼스 종류에 따라 83%-100%의 정확률을 보였다. 대개 재현률이 조금 증가함에 따라 정확률이 급격히 떨어지는 경향을 고려해 볼 때, CLSM-MAN 방법이 우수함을 알 수 있다. 즉, 언어의 특성을 고려하지 않는다면 일본어의 경우에 비해서 본 논문에서 제시된 방법이 매우 뛰어남을 알 수 있다.

실험 결과에 나타난 오류의 원인을 분석해 보면 다음과 같다.

1) 잘못된 한국어의 띄어쓰기 문제

한국어 복합명사는 띄어쓰기와 붙여쓰기를 모두 인정한다. 띄어 쓴 명사는 단일어 모델로 처리가 가능하고, 붙여 쓴 경우에 대해서는 다중어 모델을 이용하여 처리할 수 있다. 그러나, 영어 복합어에 대해 외래어 표기시 잘못 띄어쓰기를 하여 표기를 하는 경우가 있다. 예를 들어 "database" (데이터베이스)를 잘못 판단하여 "데이터_베이스"로 잘못 띄어 쓰는 경우가 있었다. (여기에서 밑줄표시 "_"는 빈칸을 나타낸다.) 이런 경우는 원칙적으로 번역문에서의 띄어쓰기 오류로 판단해야 할 것이다.

2) 외래어와 한국어가 혼합되어 사용된 경우

이 경우는 "부시스템" (subsystem), "다중프로세서" (multiprocessor) 등과 같이 번역된 부분과 음차된 부분이 합쳐져 있는 경우이다. 이를 처리하기 위해서는 어느 부분까지 음운 정렬을 해야 하는지를 판단해야 하고, 또 그에 따라 적절한 부분만을 분리해 내야 한다.

3) 변형된 영어 단어에 대한 표기 문제

영어의 단어가 복수형이나 형용사, 동사 등으로 사용되는 경우라도, 외래어로 표기할 경우에는 대개 명사형으로 표기된다. 예를 들어 "net"와 "nets"는 모두 "네트"로 표기되고, "relation", "relations", "relational" 등도 모두 "릴레이션"으로 표기된다. 이 경우는 음운 유사도가 좀 낮더라도, 관련된 단어로 판단해야 한다.

이러한 문제 외에도 단순한 단어수준의 파악만으로는 불가능하고 일부 혹은 전체적인 문맥을 파악하여 처리해야 하는 경우도 있다.

4) 영어의 분리 문제

영어의 경우에도 경우에 따라 분리할 것인가 한 단위로 할 것인가를 판단해야 한다. 예를 들어 "Denny's"의 경우 음식점을 나타내는 고유명사이어서 한 단위로 "데니스"와 정렬되어야 한다. 하지만, "Cliton's wife" (클린턴의 부인)의 예에서는 "Cliton"을 분리하여 "클린턴"과 정렬해야 한다. 본 실험에서는 단순하게 모든 특수문자들을 분리시켰지만, 보다 정확한 정렬을 위해서는 문맥을 파악해서 이러한 문제를 처리할 필요가 있다.

5) 단어 단위의 외래어 판별 방법의 한계

외래어 판별은 현재 한 단어에 한해서 판단하고 있다. 이 방법으로는 그 문맥에 따라 여러 가지로 다르게 쓰이는 단어가 있으므로 판단의 한계가 있다. 예를 들어, "톱", "집"은 외래어로 볼 경우, "top", "zip"으로 판단될 수 있으나, 한국어로 "톱"(연장), "집"(주택)으로 판단할 수도 있다. 따라서 문맥 내에서의 의미에 따라 외래어와 한국어로 판단해야 한다.

6) 대응되는 원어의 부재

번역문의 경우, 독자에게 그 의미를 정확하게 전달하기 위해 의역을 하는 경우가 많다. 외래어 표기에 관련되어서도 이러한 의역을 고려해야 한다. 예를 들어 영어 문장에서 "Bill Gates"를 의미하는 단어로 "Bill"이라는 단어가 쓰이지만, 번역문에서는 보다 친숙한 "게이츠"로 번역되어 나타났다. 또, "Washington to Manila"라는 구를 "미국에서 필리핀까지"라고 번역하고 있다. 이런 경우, 외래어가 사용되었지만, 대응되는 원어를 음운 비교로 찾아낼 수 없어, 잘못된 대응어를 내 놓을 수도 있다.

6. 결론

본 논문에서는 영어 및 대응되는 한국어 번역문으로 구성된 병렬 코퍼스로부터 외래어 표기 사전을 자동 구축하는 방법을 제안하였다. 이 방법은 명사 추출, 한국어 명사의 외래어 판별, 한국어 외래어와 영어의 음운 비교를 하는 3단계로 수행된다. 특히, 명사 추출 단계에서는 병렬 코퍼스에 있는 원어 정보를 이용하여, 어미나 조사와 함께 사용된 외래어를 보다 정확하게 추출하는 방법을 제안했다. 또, 음운 비교 단계에서는, 서로 다른 문자체계의 외래어와 영어를 비교하기 위해서 일단 같은 문자체계로 변환하여 비교하는 기존의 방법과는 다르게, 통계적 정보를 이용하여 음운상으로 비슷한 단어들을 직접 비교할 수 있는 방법을 제안하였다.

실험 결과, 외래어 추출 전 과정을 자동으로 한 모델 중에서 가장 높은 성능을 보인 모델은 다중어를 처리하고, 명사 추출시 원어 정보를 음운 비교시에 활용한 모델(CLSM-CLS)로서 재현률 68.3%, 정확률 89.2%를 보였다. 각 단계별 성능은 전처리 단계인 한국어 미등록 명사 추출 및 외래어 판별에 영향을 많이 받았으며, 직접 수작업으로 전처리를 했을 경우, 가장 성능이 높은 모델은 재현률 85.4%, 정확률 91.0%를 보였다. 이 결과를 언어 특성상 명사 추출과 외래어 판별이 100%로 정확한 일본어 외래어 추출 연구인 Collier[14](재현률 75%, 정확률 82%)와 Keita[15](재현률 75%, 정확률 83%-100%)와 간접적으로 비교해 볼 때 우수한 결과이다.

본 연구를 좀더 많은 병렬 코퍼스에 적용할 경우, 다국어 정보검색을 위한 대역어 사전의 구축, 기계번역용 사전 등에 활용할 수 있을 것이며, 외래어 추출 방법은 번역어의 용어 일치 점검 등에 활용될 수 있을 것이다. 또, 음운 유사도에 의한 비교 방법을 다른 언어에 적용할 경우, 영어에서 온 외래어 처리뿐만아니라, 불어나 독일어 등으로 확대할 수 있고, 일본어와 영어, 중국어와 영어 등의 다른 언어 병렬 코퍼스에도 적용이 가능할 것이다.

참고문헌

[1] 문화체육부 고시 제1995-8호, 1995, 외래어 표기법, 1995년 3월 16일.

[2] 임동훈, 외래어 표기법의 원리와 실제, 1996, 새국어생활 제6권 제4호, pp. 41-61.

[3] P. F. Brown, J. C. Lai, and R. L. Mercer, 1991, Aligning sentences in parallel corpora, In *Proceedings 29th annual meeting of the ACL*, Berkeley, CA, pp. 169-176.

[4] P. F. Brown and et al, 1993, The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics*, Vol. 19, No. 2, pp. 263-311.

[5] I. Dagan, K. Church, and W. Gale, 1993, Robust bilingual word alignment for machine aided translation, In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pp. 1-8.

[6] K. Church, 1993, Char_align: A program for aligning parallel texts at the character level, in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Ohio, pp. 1-8.

[7] M. Kay and M. Roscheisen, 1994, Text-translation alignment, in Using large corpora, edited by Susan Armstrong, The MIT Press, pp. 121-142.

[8] J. Kupiec, 1993, An algorithm for finding noun phrase correspondences in bilingual corpora, in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Ohio, pp. 17-22.

[9] P. Fung and K. W. Church, 1994, K-vec: A new approach for aligning parallel texts, in *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto Japan, pp. 1096-1102.

[10] 이재성, 1999, 다국어 정보검색을 위한 영-한 음차 표기 및 복원 모델, 박사학위논문, 한국과학기술원.

[11]Kil Soon Jeong, 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.

[12] 오종훈, 최기선, 1999, 은닉 마르코프 모델을 이용한 과학기술문서에서의 외래어 자동 추출 모델, 제11회 한글 및 한국어 정보처리 학술대회, pp. 137-141.

[13] Y. Kang and A. A. Maciejewski, 1996, An algorithm for generating a dictionary of Japanese scientific terms, *Literary and Linguistic Computing*, Vol. 11, No. 2, 1996, pp. 77-85.

[14]N. Collier, A. Kumano and H. Hirakawa, 1997, Acquisition of English-Japanese proper nouns from noisy-parallel newswire articles using Katakana matching, in *Proceedings of Natural Language Processing Pacific Rim Symposium*, Phuket, Thailand, pp. 309-314.

[15] Keita Tsuji, 2001, Automatic extraction of translational Japanese-KATAKANA and English word pairs from bilingual corpora, in *Proceedings of 19th International Conference on Computer Processing of Oriental Languages*, pp. 245-250.

[16]E. Charniak, 1993, Statistical Language Learning, The MIT Press, pp. 21-38.

[17]Kang, B. J. and Choi, K. S., 2000, Automatic Transliteration and Back-transliteration by Decision Tree Learning, in *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athnes, Greece.

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

[19]강승식, 1995, 한국어 자동 색인을 위한 형태소 분석 기능, 제 22회 한국정보과학회 봄 학술발표논문집, 22권 1호, pp. 929-932.

[20] http://nlp.kookmin.ac.kr/ (HAM 5.0), 2002.

[21]D. Silverman, 1992, Multiple scansions in loanword phonology: evidence from Cantonese, *Phonology* 9, pp. 289-328.

[22]박영찬, 최기선, 김재군, 김영환, 1996, 한국어 정보 검색 연구를 위한 시험용 데이터 모음 2.0 (KTSET 2.0) 개발, 한국정보과학회 인공지능연구회 춘계학술발표대회 논문집, 서울, pp. 59-65.

[23]뉴스위크(한국어판), 1999, http://nwk.joongang.co.kr/.

이 재 성

1983년 서울대학교 컴퓨터공학과 (학사)
1985년 한국과학기술원 전산학과 (석사)
1999년 한국과학기술원 전산학과 (박사)
1985년~1988년 큐닉스컴퓨터 개발부 과장
1988년~1989년 미국 마이크로소프트 S/W 설계자
1988년~1993년 마이크로소프트 개발부 차장
1999년~2000년 한국전자통신연구원 선임연구원/팀장
2000년~현재: 충북대 컴퓨터교육과 조교수
관심분야 : 컴퓨터교육, 정보검색, 자연언어 처리
e-mail: jasonl@cbu.ac.kr