이중언어 코퍼스로부터 외래어 표기 사전의 자동구축

이재성 한국전자통신연구원 지식정보연구부
jasonl@etri.re.kr

Automatic Construction of a Transliteration Dictionary from Bilingual Corpus

Jae Sung Lee, Knowledge Information Department,
Electronics and Telecommunications Research Institute

요약

외국문명의 영향으로 많은 외래어가 한국어 문서 내에서 사용되고 있으며, 이러한 단어는 주로 전문용어, 고유명사, 신조어 등으로 사전에 등록되지 않는 것이 많다. 본 논문에서는 이중언어 코퍼스로부터 자동으로 외래어 사전을 추출해 내는 확률적 정렬 방법과 실험결과를 소개한다. 확률적 정렬 방법은 통계적 음차 표기 모델에서 사용된 방법을 변형하여 적용한 것이며, 문서단위로 정렬된 두 종류의 영-한 이중언어 코퍼스에 대해 실험하여 재현률과 정확률을 측정하였다. 성능은 전처리단계인 한국어 미등록어 추정에 영향을 많이 받았는데, 미등록어 추정을 대략하였을 경우, 재현률은 평균 58%였고, 정확률은 평균 74%이었으며, 수동으로 미등록어 명사를 분리했을 경우, 재현률 평균 86%, 정확률 평균 91%로 외래어와 대응되는 원어를 추출해 냈다.

1. 서론

기계번역이나 다국어 정보검색 및 교차언어 정보검색에서 매우 중요한 자원으로 사용되고 있는 이중언어사전(bilingual dictionary)은 주로 기존의 종이사전으로부터 추출해 내거나, 이중언어 코퍼스로부터 추출해 낸다(Brown 1991, Brown 1993, Dagan 1993, Church 1993, Kay 1993, Kupiec 1993). 종이사전으로부터 추출된 이중언어사전은 쉽게 정보를 얻을 수 있는 반면, 그 용례가 정형화되어 있고, 쓰임에 따른 다양한 형태의 번역을 모두 반영하지는 못하고 있다. 특히, 전문용어나 고유명사 등은 사전에 등재되어 있지 않아 처리에 한계를 주고 있다. 이러한 문제점을 해결하기 위한 방법으로 이중 코퍼스로부터 사전을 구축하는 연구가 진행되고 있으며, 특히, 방대한 양의 코퍼스를 처리해야 하므로 계산량 축소와 정확도 향상을 위한 연구가 관심을 받고 있다 (Brown 1993, Dagan 1993, Fung 1994).

코퍼스에서 이중언어사전 혹은 대역사전을 효과적으로 구축하기 위해서는 기존의 사전을 이용하여 정렬의 효과를 높이기도 하고, 언어의 공통적인 특성을 이용하기도 한다(Church 1992). 한국어와 영어에 대한 정렬 연구의 경우, 언어의 이질성 때문에 계산량 문제가 영어-불어의 경우보다 더 심각하다(신중호 1996). 영어와 한글의 경우 일반적으로 언어의 유사성이 없지만, 외래어의 경우는 음운상 그 유사도가 높으므로 이를 이용하면 정렬의 효율을 높일 수 있으며, 외래어 사전을 자동으로 추출할 수 있다. 특히, 영어 문서를 번역한 것에는 전문용어나 고유명사 등 많은 단어들이 음차표기되어 사용되고 있어서 효과적이다. 실제 조사에 의하면, 많은 양의 외래어가 전문서적, 웹 등에서 사용되고 있으며, 이는 정부에서 정한 표준 외래어 표기가 나오

기 전에 사용되는 것들이 많아서, 이러한 추세는 더 심해 질 것을 보인다.

이 논문에서는 음운상으로 비슷한 단어들의 비교 방법을 제안하고 이 방법을 이용하여 이중언어 코퍼스로부터 외래어 표기 사전을 자동구축하는 방법을 소개한다[1]. 이 방법은 외래어 표기된 단어로부터 음운상 유사도(정렬 확률)를 학습한 다음, 이 유사도를 다시 코퍼스에 적용하여 음운유사도가 높은 단어를 추출한다. 이렇게 추출된 단어쌍들이 바로 외래어 표기 사전이 되며, 이것은 기계번역용 사전이나 다국어 정보검색, 교차언어 정보검색 등에 이용될 수 있다.

2. 관련연구

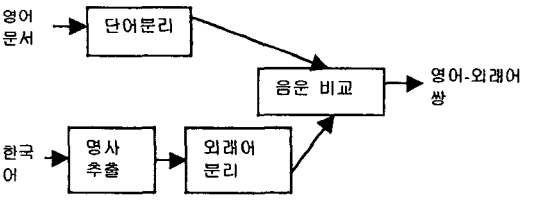
코퍼스에서 자동으로 외래어 사전을 추출하는 방법에 대한 연구로는 일본어-영어간에 이루어 진 것이 있다(Kang 1996). 이 연구에서는 일본어를 로마자로 수동으로 작성한 규칙에 따라 표기한 후, 일정한 규칙으로 변형하여, 영어와 비교하기도 하고, 일본어와 영어를 모두 발음기호로 바꾼 후, 비교하는 방법을 이용하기도 했다. 일본어는 외래어 표기를 순수 일본어와는 다른 문자체계인 가타가나로 표기하여 외래어의 추출에는 별문제가 없다. 이렇게 추출된 외래어에 대해, 영어 문자체계나 발음기호체계로 변경하여 비교함으로써 유사한 단어를 추출해 내었다. 또 이와 유사하게 일본어-영어간의 이중언어 코퍼스에서 외래어를 추출한 것으로 Collier (1997)가 있다.

한국어 문서에 나타난 외래어를 영어와 비교하는 연구는 주로 외래어 생성이나 혹은 복원의 관점에서 연구되었다. 외래어의 생성 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 부르며, 통계적 방식의 모델에 적용하여 실험한 결과, 정확도면에서 후자가 약간 더 우수한 결과를 보였다 (이재성 1999). 또, 두가지 방법을 적절히 혼합하여 좀더 좋은 결과를 내놓는 방법인 혼합방식도 있으나, 본 논문에서는 구현이 편리하고 성능도 비교적 우수한 직접방식을 사용한다.

이외에도 김병혜(1991)는 피봇방식으로 규칙에 기반하여 외래어를 생성하였으며, 성능은 그다지 높지 않았다. 또, 직접방식을 신경망 및 통계적 방식으로 구현한 연구로 김정재 (1999), Jeong (1997) 등이 있다.

3. 구축 단계

외래어를 한국어 문서에서 추출하고, 다시 이 단어를 영어 문서내의 단어와 비교하기 위해서는 3 단계를 거쳐서 한다. 즉, 명사부분의 추출, 외래어의 판별, 대응 영어의 검색으로 나누어 처리할 수 있다.



3.1명사부분의 추출

한국어문서에서 외래어의 쓰임은 주로 명사적으로 사용된다. 형용사적이거나 동사적으로 쓰일 경우는 대개 "하다"와 같은 파생접미사와 연결되어 사용된다. 이에 대응되는 영어 단어는 명사, 형용사, 동사 등이 있고, 이들은 대개 띄어쓰기가 명확하여, 특별한 경우를 제외하고는 띄어 쓴 상태 그대로 외래어와 비교할 수 있다.

이와는 다르게 한국어의 경우, 외래어는 조사나 접두사/접미사 등과 함께 쓰여서 순수한 외래어 추출을 위한 별도의 작업을 해야 한다. 예를 들어 다음과 같이 3가지 경우를 들 수 있다.

---

[1] 외래어표기는 모든 단어를 사전에 넣을 수 없기 때문에 자동 음차표기와 보완적으로 사용될 수 있다. 즉, 외래어 표기사전은 이미 알려진 단어에 대해 정확도 높은 대역어(음차어)를 제공해 주며, 자동 음차 표기는 미등록어에 대한 생성 또는 검색에 이용될 수 있을 것이다.

(1) 테스트를 한다
(2) 미마이크로소프트사는
(3) 섹시하다.

(1)의 경우, 외래어 "테스트"에 조사 "를"이 붙은 형태이고, (2)의 경우, 외래어 고유명사 "마이크로소트"에 미국을 의미하는 접두사 "미"와 회사를 의미하는 접미사 "사"가 붙여진 형태이다. 또한 (3)의 경우 "섹시"라는 영어 형용사에 형용사 파생 접미사 "하"가 붙여져 형용사로 쓰인 형태이다. 외래어 부분을 이미 알고 있다면, 쉽게 외래어 부분을 추출해 낼 것이다. 하지만, 사전 구축의 주된 목적으로 볼 때, 아직 사전에 등록되지 않은 외래어를 추출하여 구축하는 것이 중요하므로 미등록어 부분을 정확히 추정하는 일이 중요하다.

미등록어의 추정은 조사나 접사를 수집하여 그 어휘부분을 단어에서 제거하고 나머지 부분을 미등록어로 추정하거나, 미등록어의 패턴을 파악하여 이를 추정하는 방법 등이 있다(강승식 1995). 이 과정에서 잘못된 분리가 일어날 가능성이 있고, 이러한 문제를 통계적으로 해결하기 위한 시도도 있었다(Jeong 1997). 본 논문에서는 문제점은 있으나 단순한 처리를 위해, 조사나 접사를 최장일치 방법으로 인식하고 나머지 부분을 미등록어로 처리하는 방식을 사용한 KAIST의 형태소 해석기를 그대로 사용했다.

3.2외래어 판별

미등록어의 추정이 끝난 단어는 한국어일 수도 있고, 외래어일 수도 있다. 이를 판별하기 위한 방법으로는 언어구분(language identification) 방법을 사용했다. 언어구분 방법은 각 언어에 대해 모델을 만들고, 각 단어에 대해 엔트로피를 계산하여 가장 유사한 언어모델을 선택하는 방법이다(Charniak 1993).

언어모델은 여러가지로 만들어 질 수 있는데, 본 논문에서는 자소를 외래어표기에서 하나의 단위로 표기되는 발음단위로 분리한 후 이들 사이의 연결관계를 언어모델로 표기한 것을 사용했다. 이를 수식으로 표현하면 다음과 같다.

$$M(W) = \sum P(U_i \mid U_{i-1})$$

위 모델을 순수 한국어 단어 집합과 외래어 단어집합에 대해 학습시킨 모델을 M_K, M_F라 표기하자. 단어 W에 대한 언어확률은 $M_K(W)$, $M_F(W)$로 표기되며, $M_K(W) > M_F(W)$일 경우, W는 한국어, $M_K(W) < M_F(W)$일 경우는 외래어로 판정한다.

3.3확률적 정렬

원언어에서 음운구조가 다른 대상언어로 음차표기가 될 경우, 대상언어의 음운구조에 맞게 원언어의 음운이 변형되어 표기된다 (Silverman 1992). 이러한 과정중에 새로운 음운이 추가되거나 생략되므로, 대상언어나 원언어에서 몇 개의 음운이 하나의 단위처럼 음차표기되기도 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(1999)에서는 이를 처리하기위해 통계적으로 자주 함께 사용되는 문자열을 하나의 발음단위로 취급하여 통계적 방법으로 음차표기하였다.

발음단위의 음차표기는 다음과 같이 표기될 수 있다. 주어진 외래어표기 K에 대해 원어인 영어 E가 대응될 확률은 P(E|K)로 표시된다. 이것은 다시 (1)와 같이 쓸 수 있다. 이 수식은 주어진 두 언어 사이의 정렬이므로 올바른 언어로 구성될 확률을 나타내는 P(E)는 항상 1이라고 처리할 수 있다. 따라서 단순한 정렬수식이 (2)와 같이 정해지고 이것을 발음단위로 나타내면 (3)과 같게 된다.

$$P(E \mid K)$$
$$= P(K \mid E) \times P(E) \quad (1)$$
$$= P(K \mid E) \quad (2)$$
$$= \prod P(KU_i \mid EU_i) \quad (3)$$

수식 (3)을 길이에 따라 정규화하여 표시한 것이 (4)이다.

$$P(E \mid K) = \left(\prod_i^n P(KU_i \mid EU_i) \right)^{\frac{1}{n}} \quad (4)$$

이것은 발음단위로 분리된 영-한 외래어 단어쌍으로부터 자동학습할 수 있다.

4. 실험

실험용 코퍼스는 2가지를 사용했는데, 하나는 정보검색용으로 구축된 코퍼스인 KTSET2.0 (박영찬 1996)에 있는 이중언어 코퍼스와 다른 하나는 뉴스위크 코리아의 웹페이지(뉴스위크 1999)에서 추출한 영-한 대역문이다. 각 코퍼스의 특성은 다음과 같다.

KTSET

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

NWK(뉴스위크 한글판)

뉴스위크 코리아의 웹사이트에 게재되어 있는 잡지의 영-한대역 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.

4.1 실험방법

실험은 각각 다음의 모델에 대해서 했다.

PA-0

주어진 외래어에 최대한의 확률로 정렬된 영어단어를 추출한다.

PA-0.001

주어진 외래어에 대해 확률이 0.001이상으로 정렬된 중 최대 확률을 갖는 영어단어를 추출한다.

PA-0.01

주어진 외래어에 대해 확률이 0.01이상으로 정렬된 중 최대 확률을 갖는 영어단어를 추출한다.

PA-D

주어진 외래어에 대해 확률이 0.01이상으로 정렬된 중 최대 확률을 갖는 영어단어를 추출한다. 이때 단독어 이외에도 인접한 단어들을 2개씩 붙여서 복합어로 가정하고 정렬확률을 계산하여, 확률이 0.01이상일 경우, 결과로 추출한다.

PA-MAN

외래어를 수동으로 추출해내어 이에 대한 영어단어를 정렬하여 확률인 0.01이상인 것 중 최대인 영어단어를 추출한다.

4.2 실험 결과

표 1. 각 모델에 대한 재현률 및 정확률

모델 \ 데이터	NWK		KTSET	
	재현률	정확률	재현률	정확률
PA-0	0.4618	0.4701	0.7579	0.4983
PA-0.001	0.4618	0.4927	0.7579	0.5255
PA-0.01	0.4540	0.7095	0.7105	0.7759
PA-D (복합어처리)	0.4521	0.6525	0.7000	0.7000
PA-MAN (수동분리)	0.8591	0.9052	0.8579	0.9106

결과는 코퍼스에 있는 외래어-영어 쌍으로부터 얼마나 많은 쌍을 추출해 냈는가의 비율을 측정하는 재현률과 추출된 모든 쌍 중에서 정확하게 추출된 쌍의 비율을 나타내는 정확률로 측정했다. 그 결과는 표 1에 나타나 있으며, 이를 정리하면 다음과 같다.

1) 모델에서 최저 확률값을 적절히 적용하여 정렬된 쌍을 조절할 경우, 재현률은 약간 떨어지지만 정확률이 급격히 상승함을 알 수 있다. PA-0에서 PA-0.001로 할 경우, 재현률의 변화는 없고, 정확률이 약간씩 높아 졌다. 또, PA-0.001에서 PA-0.01로 할 경우, KTSET에서 정확률은 0.5255에서 0.7759로 약 0.25정도 상승했으나, 재현률은 0.7579에서 0.7105로 0.05정도만 하락했다. NWK에 대해서도 비슷하다.
2) 복합어처리는 단독어 처리에 비해 비효율적이다. PA-0.01과 PA-D를 비교해 보면, PA-D가 재현률 및 정확률 면에서 모두 약간씩 나빴다. 복합어처리를

할 경우, 단독어에서 발견되지 않은 복합어 쌍을 더 발견해 낼 수 있으리라 생각했지만, 실제로는 임의의 복합어 결합으로 인해 확률계산이 잘못되어 더 나쁜 결과를 낸 것으로 추측된다.

3) 외래어부분을 수동으로 정확하게 추출한 PA-MAN은 재현률 및 정확률 면에서 모두 월등한 성능을 보였다. 즉, KTSET에 대해서, PA-0.01보다 재현률은 0.16, 정확률은 0.21 향상되었으며, NWK에 대해서는 재현률이 0.41, 정확률이 0.20정도 향상되었다. 결과적으로 두개의 데이터 집합에 대한 평균 재현률이 0.86정도이고, 정확률이 0.91정도이어서 매우 우수한 결과를 보였다.

4) PA-MAN을 제외한 모델에 대해 NWK가 KTSET에 비해 재현률이 현저하게 낮게 나타났다. 즉, 자동으로 외래어를 추출해 낼 경우, 성능이 나쁘게 나왔다. 이는 NWK가 KTSET보다 다양한 외래어를 포함하고 있어서 순수외래어를 자동으로 판별하여 추출하기 어렵기 때문인 것으로 보인다.

표 2는 KTSET에서 수동, PA-0.01, PA-MAN의 방법으로 구축된 외래어 표기사전의 앞부분 일부를 보여준다. 3번은 PA-0.01에서 '너비'를 외래어로 잘못 판별하여 잘못된 결과를 내 놓은 예이다. 이와 반대로 24번은 '도큐멘트"를 순수 한국어로 잘못 판단하거나 정렬확률이 낮아 결과를 생성하지 못한 예이다. 7, 10, 17번은 실제 외래어가 사용되었지만, 영어문서에는 그 단어가 없어서 다른 단어를 찾아낸 예이다. 12, 13, 14는 한국어와 외래어가 함께 붙여서 쓰임으로써, 외래어로 판단되지 않아 제대로 처리되지 않은 예이다. 이렇게 순수한국어와 영어가 혼합된 형태에 대한 처리는 한국어와 영어 모두의 분리가 필요하고, 음운적인 정렬 외에 의미상의 비교를 통해 분리하여 정렬할 필요가 있다. 11번이나, 22번은 원어가 "diagram", "database"이어서 외래어 표기가 "다이어그램", "데이터베이스"로 되어야 하지만, 실제문서에서 잘못 띄어쓰기를 하여 "다이어"와 "데이터"만을 가지고 음운정렬을 시도했기 때문에 엉뚱한 단어를 추출한 예이다. 이러한 문제는 수동일 경우 미리 파악하여 23번에서처럼 한단위로 할 수 있지만, 자동일 경우는 PA-D와 같은 방법으로 처리해야 한다. 하지만, PA-D를 사용해도 전체적인 성능면에서는 떨어졌기 때문에 좀더 효과적인 방법이 필요하다.

5. 토의

오류의 원인을 분석해 보면 다음과 같다.

1) 한국어 미등록어에 대한 분리문제

이미 앞에서도 언급했듯이 형태소해석기에서 미등록어를 추정할 경우, 오류를 생성하는 문제이다. 예를 들어 "컴파일"에서 "일"을 접미사로 취급하여 "컴파"와 "일"로 분리하였으며, "팔레스타인"을 "팔레스"와 "타인"으로 분리하기도 했다.

2) 외래어와 한국어가 혼합되어 사용된 경우

이 경우는 어느 부분까지 음운정렬을 해야 하는지를 판단해야 하고, 또 그에 따라 적절한 부분만을 분리해 내야 한다. "부시스템 (subsystem)", "다중프로세서 (multiprocessor)" 등이 이러한 예이며, 이를 처리하기 위한 방법이 연구되어야 한다.

3) 영어의 분리문제

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 특수 문자들을 분리시켰지만, 보다 정확한 정렬을 위해서는 이러한 문제를 고려하여야 한다.

4) 한국어의 띄어쓰기 문제

한국어 문장에서 복합명사에 대한 띄어쓰기는 대개 두가지를 모두 인정한다. 이러한 쓰기 방법의 영향으로 "database (데이터베이스)"를 잘못 판단하여 "데이터_베이스"로 잘못 띄어 쓰는 경우가 있고, 반대로 "head_literal"과 같은 경우에 "헤드리터럴"로 붙여 쓰기도 한다. (여기에서 밑줄표시 "_"는 빈칸을 나타낸다.) 띄어쓰기 규칙상 붙여 쓴 것은 문제가 없기는 하지만, 음운정렬의 관점에서는 모든 경우를 고려하여 처

표 2. 각각의 방법으로 구축된 외래어표기 사전의 예

	수동	PA-0.01	PA-MAN
1	그래프 graph	그래프 graph	그래프 graph
2	그래픽 graphic	그래픽 graphic	그래픽 graphic
3		너비 have	
4	네트 net	네트 net	네트 net
5	네트워크 network	네트워크 network	네트워크 network
6	네트워크 networks	네트워크 networks	네트워크 networks
7		노드 mode	노드 mode
8	노드 node	노드 node	노드 node
9	노드 nodes	노드 nodes	노드 nodes
10		노드 not	노드 not
11		다이어 data	
12	다중컴퓨터 multicomputer		
13	다중프로세서 multiprocessing		
14	다중프로세서 multiprocessor		
15	데드라인 deadline	데드라인 deadline	데드라인 deadline
16	데이타 data	데이타 data	데이타 data
17		데이타 that	데이타 that
18	데이타베이스 database	데이타베이스 database	데이타베이스 database
19	데이타베이스 databases	데이타베이스 databases	데이타베이스 databases
20		데이타베이스내 database	
21	데이터 data	데이터 data	데이터 data
22		데이터 that	
23	데이터_플로우 dataflow		
24	도큐멘트 document		도큐멘트 document
25	드라이버 driver	드라이버 driver	드라이버 driver
26	드라이버 drivers	드라이버 drivers	드라이버 drivers

리해야 한다.

5) 영어의 외래어 표기는 반드시 그 단어에 충실하지는 않다.

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. "relational database (릴레이션 데이터베이스)"

6) 단어단위의 외래어 판별방법에 한계가 있다.

외래어 판별은 현재 한 단어에 한해서 판단하고 있다. 이 방법으로는 그 문맥에 따라 여러가지로 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 문맥내에서의 의미에 따라 외래어와 한국어로 판단해야 한다.

7) 외래어가 있더라도 반드시 대응되는 원어가 존재하지는 않는다.

번역문의 경우, 독자에게 그 의미를 정확하게 전달하기위해 의역을 하는 경우가 많다. 외래어 표기에 관련되어서도 이러한 의역을 고려해야 한다. 예를 들어 영어문장에서 "Bill Gates"는 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 음운정렬로 찾아낼 수는 없으면, 이에 따라 잘못된 정렬결과를 내 놓을 수도 있다.

6. 결론 및 향후 연구

본 논문에서는 통계적 방법을 이용하여 확률적으로 음운정렬을 하는 방법을 제안하고 실험하였다. 실험결과 수동으로 한국어 중 외래어 부분을 추출한 후, 영어 문장에서 대응되는 단어를 뽑아냈을 경우, 평균 재현률 86%, 평균 정확률 91%로 비교적 높았다. 문제는 한국어에서 순수한 외래어를 추출하는 단계에 오류가 많이 포함되었는데, 이 부분은 미등록어 추정기술과 외래어 구분기술을 적절히 조합하여 보완할 수 있을 것으로 보인다. 또한 실험결과, 다중어 처리시 오히려 성능이 저하되었는데, 이를 효과적으로 처리할 수 있는 방법에 대한 연구도 필요하다.

참고문헌

강승식, 1995, "한국어 자동 색인을 위한 형태소 분석기능," 제 22회 한국정보과학회 봄 학술발표논문집, 22권 1호, pp. 929-932.

김병혜, 1991, "영어단어의 알파벳표기로부터 한글표기로의 자동변환," 석사학위 논문, 서강대학교 공공정책대학원.

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

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

박영찬, 최기선, 김재균, 김영환, 1996, "한국어 정보검색 연구를 위한 시험용 데이터 모음 2.0 (KTSET 2.0) 개발," 한국정보과학회 인공지능연구회 춘계학술발표대회 논문집, 서울, pp. 59-65.

신중호, 1996, "한국어/영어 병렬 코퍼스에 대한 단어단위 및 구단위 정렬 모델," 석사학위 논문, 한국과학기술원.

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

이현복, 1979, "외래어 표기법 개정 시안의 문제점," 어학연구 15.1, pp. 39-59.

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.

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

E. Charniak, 1993, "Statistical Language Learning," The MIT Press, pp. 21-38.

S. F. Chen, 1993, "Aligning sentences in bilingual corpora using lexical information," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*," Ohio, pp. 9-16.

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.

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.

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.

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.

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.

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

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.

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

K. Jeong, Y. Kwon and S. H. Myaeng, 1997, "Construction of equivalence classes of foreign words through automatic identification and extraction," *Natural Language Processing Pacific Rim Symposium '97*, pp. 335-340.