정렬기법을 이용한 미등록 대역어의 자동 추출

김 재 훈[†] · 양 성 일[††]

요 약

이 논문은 정렬 기법을 이용한 미등록 대역어 추출 모델을 제안하고 그 추출 시스템을 구현한다. 제안된 미등록 대역어 추출 모델은 일종의 구절정렬 모델로서 경계모델과 언어모델 그리고 번역 모델로 구성된다. 제안된 추출 시스템은 병렬말뭉치 구축, 단어정렬, 미등록어 추출로 구성된다. 이 논문에서는 제안된 시스템을 평가하기 위해서 약 1,500여 개의 미등록어가 포함된 2,220문장의 평가말뭉치를 구축하여 다양한 실험을 수행하였다. 실험을 통해서 제안된 모델이 미등록 대역어 추출에 매우 유용함을 알 수 있었다. 앞으로 좀 더 객관적인 평가를 위해 대량의 평가말뭉치 구축이 선행되어야 하며 좀 더 양질의 병렬말뭉치의 구축이 필요할 것이다. 또한 미등록어 추출 모델을 개선하기 다양한 연구가 추진되어야 할 것이다.

키워드 : 미등록 대역어, 정렬모델, 사전 구축, 병렬말뭉치

Automatically Extracting Unknown Translations Using Phrase Alignment

Kim, Jae-Hoon[†] · Yang, Sung-Il[††]

ABSTRACT

In this paper, we propose an automatic extraction model for unknown translations and implement an unknown translation extraction system using the proposed model. The proposed model as a phrase-alignment model is incorporated with three models: a phrase-boundary model, a language model, and a translation model. Using the proposed model we implement the system for extracting unknown translations, which consists of three parts: construction of parallel corpora, alignment of Korean and English words, extraction of unknown translations. To evaluate the performance of the proposed system, we have established the reference corpus for extracting unknown translation, which comprises of 2,220 parallel sentences including about 1,500 unknown translations. Through several experiments, we have observed that the proposed model is very useful for extracting unknown translations. In the future, researches on objective evaluation and establishment of parallel corpora with good quality should be performed and studies on improving the performance of unknown translation extraction should be kept up.

Key Words : Unknown Translation, Alignment Model, Dictionary Construction, Parallel Corpus

1. 서 론

1960년대 이후 많은 연구자들이 기계번역시스템을 개발하기 위한 연구를 진행하였으나[1], 아직 일반 사용자들이 만족할 만한 기계번역시스템은 거의 없다고 해도 과언이 아니다. NIST(National Institute of Standards and Technology)의 보고[2]에 따르면 기계번역시스템 평가 대회에서 가장 좋은 성능을 보인 기계번역시스템은 Google에서 개발된 아랍어/영어 기계번역시스템이며, 신문기사를 대상으로 0.5060 BLEU[3]를 보였다. 이 측도가 인간의 이해도와 일치한다고는 생각되지 않는다. 다만 이 결과를 볼 때, 약 반세기 동안 기계번역에 대한 연구가 꾸준히 진행되고 있지만, 번역의 질은 그다지 높지 않음을 알 수 있다. 기계번역을 어렵게 하는 요인으로 모호성, 문화와 언어 간 차이, 다중단어 등이 있다[4]. 그러나 언어 변이도 무시할 수 없는 요인 중에 하나이다. 즉 시간이 지남에 따라 새로운 단어가 생성되고 새로운 형태의 어법이 자연스럽게 사용된다. 이와 같이 새로 생성된 단어를 기계번역시스템에서는 미등록어라고 하는데, 이 논문에서는 이와 같은 미등록어의 대역어를 자동으로 추출하는 방법을 제안한다. 새로 생성된 단어(신조어)는 아래와 같은 몇 가지의 특성을 가지고 있다[5, 제4장]. 첫째, 언어학적으로 신조어는 단어 형성 과정을 통하여 생성된다. 둘째, 시간적으로 신조어는 일정 기간 내에서 자주 등장한다. 예를 들면 "북핵"이라는 신조어는 북한이 핵실험을 한 지난해 10월 3일 이후, 일정 기간 동안 집중적으로 등장했

† 종신회원 : 한국해양대학교 컴퓨터공학과 부교수
†† 정 회 원 : 한국전자통신연구원 언어처리연구팀 선임연구원
논문접수 : 2007년 2월 21일, 심사완료 : 2007년 4월 17일

었다. 셋째, 공간적으로 신조어는 일정한 지역이나 같은 영역에 흔히 등장된다. 예를 들면 "6자회담"이라는 신조어는 한국지역의 정치면에서 주로 나타나는 단어이다. 넷째, 사회적으로 신조어는 특정 집단에서 주로 사용된다. 예를 들면 "된장녀"라는 용어는 인터넷을 사용하는 누리꾼들에 의해서 집중적으로 사용된 용어이다. 이 논문은 신조어의 이와 같은 성질을 이용해서 특정 시간 동안 일정 분야의 신문기사의 병렬말뭉치를 구축하고, 구축된 병렬말뭉치로부터 정렬기법을 이용해서 신조어(혹은 미등록어)의 대역어를 자동으로 추출하는 방법을 제안한다. 자동번역에서 미등록 대역어 문제는 많은 경우 경험적인 규칙으로 해결하였다[6-7]. 경험규칙의 경우, 개발자의 의해서 정의된 규칙을 벗어날 경우는 여전히 제대로 번역할 수 없다. 이 문제를 다소 완화하기 위해 이 논문에서는 병렬말뭉치와 정렬방법을 이용한다. 뉴스 기사는 종종 하나 이상의 언어로 인터넷에 공개되므로 손쉽게 병렬말뭉치를 구축할 수 있다.

본 논문의 구성은 다음과 같다. 2절에서 관련 연구로 병렬말뭉치, 정렬방법, 사전구축, 미등록어 처리에 대해 간단히 살펴본다. 3절에서 미등록 대역어 추출 시스템을 기술하고 4절에서 실험 및 평가를 기술한다. 끝으로 5절에서 결론을 맺고, 향후 연구에 대해서 간단히 기술한다.

2. 관련 연구

이 장에서는 병렬말뭉치, 정렬 방법, 대역사전 자동 구축, 미등록어 처리에 대해서 살펴본다.

2.1 병렬말뭉치

병렬말뭉치란 같은 내용을 두 개 언어로 구축된 말뭉치를 말하며, 정렬 단위에 따라 문서병렬말뭉치(document-level parallel corpus), 문장병렬말뭉치(sentence-level parallel corpus), 단어병렬말뭉치(word-level parallel corpus)로 나눌 수 있다. 일반적으로는 문장병렬말뭉치를 병렬말뭉치라고 하며, (그림 1)은 병렬말뭉치의 예로서 중영한 병렬말뭉치[2]를 보이고 있다. 지난 몇 십 년 동안 다양한 방법으로 다양한 병렬말뭉치들이 구축되었으며, 이용 가능한 병렬말뭉치로는 UN Parallel Text[3], Canadian Hansards[4], Europarl[5], 세종 병렬말뭉치[6], 언어자원은행 병렬말뭉치[7] 등이 있다. 병렬말뭉치는 기계번역뿐 아니라 다국어 연구에도 중요한 역할을 담당하고 있다. 더구나 컴퓨터의 처리 능력 향상에 따라 병렬말뭉치에 대한 관심과 사용이 크게 증가하게 되었다.

```
5636:命令\颁布\如下\
#The order went forth that...
#명령은 아래와 같이 반포되었다.

5637:命令\必须\严格\遵守\.\
#The orders must be strictly obeyed.
#명령은 반드시 엄격히 준수해야 한다.

5638:命运\的\女神\垂青\勇士\.\
#fortune favors the brave .
#운명의 여신은 용사를 특별히 애호하신다.
```

(그림 1) 중영한 병렬말뭉치의 예

2.2 정렬 기법

정렬은 정렬 대상에 따라 문서, 문장, 단어/구절 정렬이 있다. 문서(문장, 단어/구절) 정렬이란 서로 다른 언어로 작성된 같은 내용의 문서(문장, 단어/구절)를 정렬하는 것을 말한다[8]. 문서정렬 기법은 웹으로부터 문서를 추출하여 다른 언어로 작성된 같은 내용의 문서인지를 결정하는 방법이며, 주로 기계학습 방법과 경험 규칙을 이용한다[9-10]. 문장정렬 방법에는 길이기반 정렬 기법[11], 편차 정렬 기법, 어휘 정렬 기법이 있다[8]. 길이기반 정렬 기법은 "번역문의 길이는 원시문장의 길이에 비례한다"라는 가정에서 시작되고, 편차 정렬 기법은 병렬말뭉치의 단어(혹은 문자 n-그램)가 나타난 위치의 차이를 이용해서 정렬하는 방식으로 문장정렬에는 그다지 많이 사용되지는 않았다. 어휘 정렬 기법은 대역사전 등과 같은 어휘정보를 이용해서 문장을 정렬한다. 단어정렬은 병렬말뭉치로부터 단어단위의 대역을 찾는 것을 말한다. IBM Model 1-5[12]를 시작으로 많은 연구들이 진행되었으며 대체로 자율학습을 통한 단어정렬 모델을 학습한다. 단어의 정렬은 1:1 대응이 아닐 뿐 아니라 선형정렬이 아니므로 모델 자체가 다소 복잡하다. 단어정렬은 그 자체만으로도 많은 응용분야를 가지고 있다. 예를 들면 대역사전의 자동 생성[13], 의미모호성 해소[14] 등이 있다. 구절정렬의 기본 개념은 정렬 대상의 단위가 단어에서 구절로 확장된 것 이외에는 큰 차이가 없다. 그러나 성능 면에서는 단어정렬보다 우수하다[14]. 가장 대표적으로 사용되는 방법은 단어정렬 확률이나 정렬 결과를 이용해서 두 언어의 구절을 정렬한다[15-16]. 또 다른 방법으로는 병렬 문장을 구문분석 결과(구문트리)를 대응시킴으로써 정렬하는 방법들이 있으며[17-19], 병렬말뭉치에서 직접 학습하는 방법[17]과 대응관계를 직접 작성하는 방법[18] 등이 있다.

2.3 대역 사전 자동 구축

일반적으로 다국어 정보검색 및 기계번역 시스템에 사용되는 번역 사전을 수동으로 구축되나 이 작업은 시간과 경비가 매우 많이 소요된다. 이와 같은 문제를 해결하기 위해서 병렬말뭉치를 이용한 자동으로 번역 사전을 구축하는 연구들이 시작되었다[20-24]. 이 연구들의 대부분은 단어의 공기양상, 상호정보, t-점수 등을 이용하며, 대량의 병렬말뭉치

2) 출처: 언어자원은행 중영한 다국어 코퍼스(http://bola.or.kr/)
3) 영어, 불어, 스페인어 병렬말뭉치, http://www.ldc.upenn.edu/Catalog/readme_files/un.readme.html
4) 영어, 불어 병렬말뭉치, http://www.isi.edu/natural-language/download/hansard/
5) 10개의 유럽어와 영어의 병렬말뭉치, http://www.statmt.org/europarl/
6) 한영, 한일 병렬말뭉치, http://sejong.or.kr/sejong_kr/index.html
7) 중국어, 영어, 한국어, 언어자원은행 다국어 코퍼스, http://bola.or.kr/

로부터 단어 대역사전을 구축하였다. 많은 경우 단어 대역사전 구축에서는 단어정렬의 결과로부터 여러 가지의 여과 기법을 이용해서 신뢰성이 높은 대응 쌍을 구한다. 한편 구절 단위의 사전 구축에 대한 연구는 아직 매우 초기 단계라고 할 수 있다.

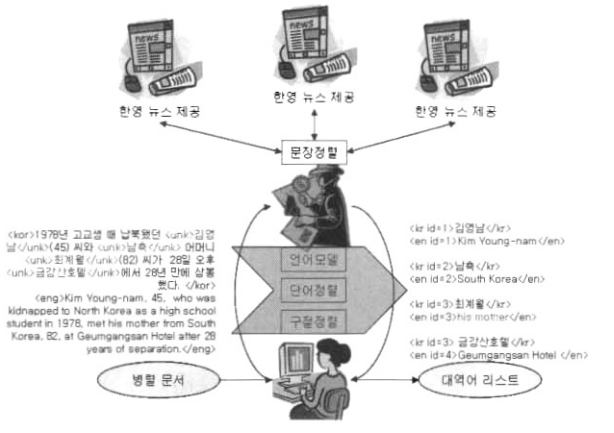
2.4 미등록어 처리

기계번역에서 미등록어 문제는 아주 고질적인 문제이며 피할 수 없는 문제이다. 일반적으로 기계번역에서 미등록어로 인식되면 미등록어를 그대로 번역어로 출력한다. 목적언어에 대해서 어느 정도 알고 있을 경우에는 이렇게 하더라도 큰 문제는 없을 것이다. 예를 들어 영한번역에서 "California"이라는 단어가 미등록어로 인식되었고 번역문장(한국어문장)의 독자는 영어를 어느 정도 이해한다면 번역문장에 "California"가 들어 있어도 이해하는 데는 큰 문제가 없을 것이다. 그러나 대부분의 독자는 목적언어를 전혀 모르는 경우이다. 더구나 문명의 발전이 빠르게 성장하면서 새로운 단어들이 급속히 등장하여 빠른 속도로 전파된다. 그러나 기계번역에서 미등록어에 대한 처리 방법은 많은 경우가 경험적인 방법에 의존하고 있다. 즉 미등록어 문제를 체계적으로 이론적으로 완전하게 정립되지 않았다. 최근에 웹 문서를 이용하거나 복합어 분리 등의 방법[25]과 바꿔쓰기(paraphrasing)[26]를 통해서 미등록어 문제를 해결하려고 하였으나, 여전히 고질적인 문제로 간주되고 있다. 그러나 웹 문서에는 이를 해결할 수 있는 많은 가능성이 있다.

웹은 정보의 보고일 뿐 아니라 자동 번역에 있어서도 필수적인 도구이다[9]. 웹을 통해서 자동번역에 필요한 병렬말뭉치를 자동으로 구축하고 구축된 병렬말뭉치를 이용해서 문장정렬과 단어정렬 과정을 통해서 미등록어 문제를 어느 정도 극복할 수 있을 것으로 판단된다. 그러나 이것이 미등록어 문제를 완전히 해결할 수 있을 것으로 기대하지는 않는다.

3. 정렬기법을 이용한 미등록 한영 대역어의 자동 추출

이 장에서는 정렬 기법을 이용해서 미등록 한국어에 대한 대역어인 영어 단어를 자동으로 추출하는 방법에 대해서 기술한다. 정렬 기법은 2장에서 소개한 문장정렬과 단어정렬 등의 기술을 의미한다. 미등록어는 시스템 사전에 포함되지 않은 단어 혹은 구절을 의미하며 주로 고유명사나 신조어가 여기에 속한다. 이 논문에서 시스템 사전은 한국어 분석 시스템[27]의 사전을 의미하며 미등록어는 결국 한국어 미등록어를 의미한다. 자동 추출이란 프로그램을 통해서 미등록어로 지정된 한국어 구절에 대해서 이에 대응하는 가장 적합한 영어 구절을 찾아주는 것을 의미한다. (그림 2)는 정렬기법을 이용한 미등록 한영 대역어의 자동 추출 시스템의 개념도이다. 첫 번째 단계는 웹으로부터 한영 뉴스 기사들을 수집하고, 수집된 병렬문서와 문장정렬 도구[11]를 이용해서 병렬 말뭉치를 구축한다[28]. 두 번째 단계는 구축된



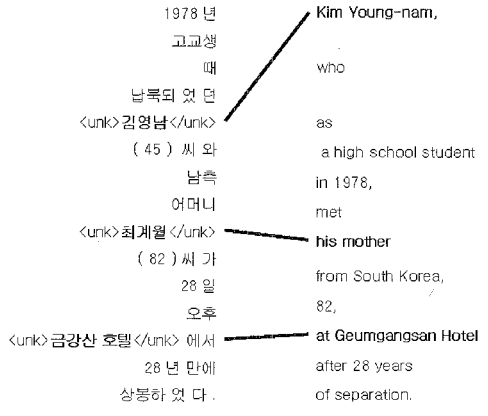
(그림 2) 정렬기법을 이용한 한영 대역어 추출 시스템의 개념도

병렬 말뭉치와 단어정렬 도구[16]를 이용해서 단어정렬 모델을 구하고[29], 구축된 병렬 말뭉치의 영어 말뭉치와 언어모델 구축 도구[30]를 이용해서 목적언어인 영어의 언어모델을 구한다. 이 두 단계를 비실시간 학습(offline learning)이라고 한다. 세 번째 단계는 두 번째 단계와 같은 방법으로 입력 문서에 대해서 단어정렬 모델과 영어 언어 모델을 구한다. 이 단계는 실시간 학습(online learning)이라고 한다. 네 번째 단계는 학습된 단어정렬 모델과 언어 모델을 이용해서 미등록 대역어를 추정하는데, 기본 개념은 구절정렬 알고리즘을 이용하며 입력은 한국어 미등록어가 표시된 한영 병렬문서이고, 출력은 표시된 미등록어에 대한 영어 대역어이다.

3.1 미등록 대역어 추정의 기본 가정

이 논문에서 미등록 대역어의 자동 추출 문제를 정의해보자. 입력으로는 소규모의 병렬문서이다. 이 병렬문서에는 원시 언어의 문장 내에 미등록어가 표시되어 있으며, 이 미등록어의 대역어를 대역 언어의 문장에서 찾는다. 하나의 미등록 대역어는 여러 개가 병렬 문장에서 존재할 수 있으며, 여러 개의 미등록 대역어들 중에서 가장 적합한 미등록 대역어를 추출해야 한다. 이와 같은 문제가 해결되기 위해서는 몇 가지의 가정이 필요하다. 첫째, 입력은 하나 이상의 병렬 문장이며 원시문장과 목적문장은 정렬되어야 한다(다문장 입력). 둘째, 원시언어의 미등록어의 대역어가 대역문장에 출현되어야 한다(출현성). 이 가정이 하나 이상의 미등록 대역어가 반드시 추출되어야 하는 것을 의미하는 것은 아니다. 셋째, 미등록어와 대역어는 연속된 단어나 형태소들이다(인접성). 넷째, 미등록어와 대역어는 하나 이상의 단어로 구성되므로 그 대응관계는 n:m이다(다대다 대응). 다섯째, 미등록 대역어가 여러 개일 경우에는 가장 적합한 미등록 대역어만 출력한다(최적성).

(그림 3)은 미등록 대역어의 자동 추출 예를 보이고 있다. (그림 3)에서 태그 <unk>와 </unk>는 원시언어에서 표시된 미등록어의 경계를 나타내고 줄은 미등록어와 미등록 대역어의 대응관계를 표현한다. 예를 들면 한국어 미등록어



(그림 3) 미등록 대역어의 자동 추출 예

"김영남"의 미등록 대역어는 "Kim, Young-nam"로 매우 적절하게 잘 선택되었다. 반면에 한국어 미등록어 "최계월"에 대해서는 주어진 문장에서는 의미적으로 "his mother"가 적합할 것이다. 그러나 미등록어로서 "최계월"에 대해서 "his mother"가 적합한 대역어라고 말할 수 없다. 이런 경우에는 임계값 등을 이용해서 여과될 것이다. 이하의 절에서는 정렬기법을 가능한 미등록 대역어 자동 추출 모델에 대해서 살펴볼 것이다.

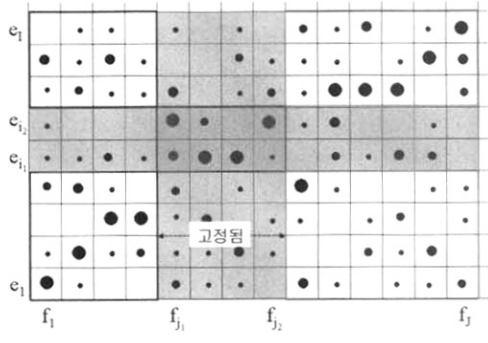
3.2 미등록 대역어 추출 모델

이 논문에서 미등록 대역어 추출 모델은 일종의 구절정렬 모델이며, 주어진 원시문장 f_1^J과 원시문장 내에 포함된 미등록 구절 $f_{j_1}^{j_2}$에 대해서 목적문장 e_1^I에 포함된 미등록 대역어 $e_{i_1}^{i_2}$를 선택하도록 모델링하면 식 (1)과 같다. 여기서 원시문장 f_1^J는 원시문장에 속한 단어열 $f_1 f_2 \cdots f_J$를 줄여서 표기한 것이며 위에서 소개된 다른 표기도 같은 방법으로 적용된다.

$$\begin{aligned} e_{i_1}^{i_2} &= \text{argmax}_{e_{i_1}^{i_2}} P(e_1^I | f_{j_1}^{j_2}) \\ &= \text{argmax}_{e_{i_1}^{i_2}} P(e_1^I) P(f_{j_1}^{j_2} | e_1^I) / P(f_{j_1}^{j_2}) \\ &= \text{argmax}_{e_{i_1}^{i_2}} P(e_1^I) P(f_{j_1}^{j_2} | e_1^I) \\ &= \text{argmax}_{e_{i_1}^{i_2}} P(e_1^{i_1-1}, e_{i_1}^{i_2}, e_{i_2+1}^I) P(f_{j_1}^{j_2} | e_1^{i_1-1}, e_{i_1}^{i_2}, e_{i_2+1}^I) \end{aligned} \quad (1)$$

식 (1)은 (그림 4)와 같이 개념적으로 설명할 수 있다. (그림 4)에서 가로축은 원시문장 f_1^J과 미등록어 $f_{j_1}^{j_2}$를 표현하고, 세로축은 목적문장 e_1^I와 미등록 대역어 $e_{i_1}^{i_2}$를 표현한다. 상자 안에 그려진 동그라미는 대응 정도를 표현한다. 미등록어 $f_{j_1}^{j_2}$에 대응하는 미등록 대역어 $e_{i_1}^{i_2}$들 중에서 대응 정도가 가장 높은 미등록 대역어를 선정하도록 모델링될 것이다.

일반적으로 식 (1)에서 오른쪽 앞부분을 언어모델이라 하고, 뒷부분을 번역모델이라고 한다. 이 논문에서는 이 두 모델은 식 (3)과 (4)와 같이 다시 정의한다.



(그림 4) 미등록 대역어 추출 모델의 개념도

$$P(e_1^{i_1-1}, e_{i_1}^{i_2}, e_{i_2+1}^I) \simeq (1 - P(e_{i_1} | e_{i_1-1})) P(e_{i_1}^{i_2}) (1 - P(e_{i_2+1} | e_{i_2})) \quad (3)$$

$$P(f_{j_1}^{j_2} | e_1^{i_1-1}, e_{i_1}^{i_2}, e_{i_2+1}^I) \simeq P(f_{j_1}^{j_2} | e_{i_1}^{i_2}) \quad (4)$$

식 (3)에서 $1 - P(e_{i_1} | e_{i_1-1})$와 $1 - P(e_{i_2+1} | e_{i_2})$는 각각 $e_{i_1}^{i_2}$가 구절의 왼쪽 경계와 오른쪽 경계가 될 확률이다(경계확률). 즉 왼쪽 경계에서는 e_{i_1-1}와 e_{i_1}가 같은 구절에 포함되지 않아야 하고, 오른쪽 경계에서는 e_{i_2}와 e_{i_2+1}이 같은 구절에 포함되지 않아야 한다. 이하에서 경계 확률의 로그값을 $b_{(i_1,i_2)}(e)$로 표기한다. 또한 식 (3)에서 $P(e_{i_1}^{i_2})$은 구절 $e_{i_1}^{i_2}$의 언어모델이며, 이하에서는 언어모델의 로그값을 $l_{(i_1,i_2)}(e)$로 표기한다. 또한 이 논문에서 식 (4)의 구절 번역 확률 $P(f_{j_1}^{j_2} | e_{i_1}^{i_2})$은 식 (5)과 같이 정의한다[15, 31]. 구절 번역확률 $P(e_{i_1}^{i_2} | f_{j_1}^{j_2})$도 식 (5)와 같은 방법을 정의할 수 있다.

$$P(f_{j_1}^{j_2} | e_{i_1}^{i_2}) = \prod_{j=j_1}^{j_2} \left(\sum_{i=i_1}^{i_2} \frac{1}{i_2 - i_1 + 1} P(f_j | e_i) \right) \quad (5)$$

이 논문에서 [15]에서와 마찬가지로 전체 번역 확률을 식 (6)과 같이 정의한다.

$$t_{(i_1,i_2)}(e,f) = \log(\alpha P(e_{i_1}^{i_2} | f_{j_1}^{j_2}) + (1-\alpha) P(f_{j_1}^{j_2} | e_{i_1}^{i_2})) \quad (6)$$

이들을 종합하여 미등록어 대역어 추출 모델은 식 (7)과 같으며, 매개변수 λ_i는 $0 \le \lambda_i \le 1$와 같은 조건을 가진다.

$$e_{i_1}^{i_2} = \text{argmax}_{e_{i_1}^{i_2}} \lambda_1 t_{(i_1,i_2)}(e,f) + \lambda_2 l_{(i_1,i_2)}(e) + \lambda_3 b_{(i_1,i_2)}(e) \quad (7)$$

또한 $f_{j_1}^{j_2}$에 대한 $e_{i_1}^{i_2}$가 대응되는 정도 $v(e_{i_1}^{i_2} | f_{j_1}^{j_2})$은 식 (7)로부터 식 (8)과 같이 정의된다.

$$v(e_{i_1}^{i_2} | f_{j_1}^{j_2}) = \max_{e_{i_1}^{i_2}} \lambda_1 t_{(i_1,i_2)}(e,f) + \lambda_2 l_{(i_1,i_2)}(e) + \lambda_3 b_{(i_1,i_2)}(e) \quad (8)$$

3.3 미등록대역어 추출 모델의 학습

미등록 대역어 추출 모델을 학습하는 방법은 여러 가지가 가능할 것이다. 가장 간단한 방법으로는 매일 뉴스 기사로부터 병렬문서를 수집하고, 수집된 병렬문서로부터 추가로 병렬말뭉치를 생성하고, 생성된 병렬말뭉치를 점진적으로 확장시킨다. 이렇게 함으로써 미등록의 발생 가능성이 점점 줄어들게 될 것이다. 또 다른 방법으로는 비실시간적으로 대량의 병렬말뭉치로부터 단어정렬 모델을 학습하고, 실시간적으로 작은 양의 입력 병렬문서로부터 단어정렬모델을 학습한다. 이렇게 학습된 두 모델을 결합하여 간단한 방법으로 미등록 대역어 추출 모델을 학습한다. 이 논문에서는 후자의 방법을 이용한다.

(그림 5)는 3.2절에서 설명한 미등록 대역어 추출 모델의 학습 시스템 구조이다. (그림 5)의 왼쪽은 대량의 병렬말뭉치를 이용해서 모델1(단어정렬모델1과 언어모델1)을 비실시간적으로 학습한다[29]. 이 논문에서는 이를 비실시간 학습이라고 한다. 한편 (그림 5)의 오른쪽은 입력 병렬문서를 이용해서 실시간적으로 모델2(단어정렬모델2와 언어모델2)를 학습한다. 이 논문에서는 이를 실시간 학습이라고 한다. 실시간 학습에서 문장정렬과 언어모델은 앞에서 언급한 비실시간 학습과 같은 방법으로 학습되나, 단어정렬은 단순한 단어정렬 모델인 IBM 모델1[12]만 사용한다. 이렇게 학습된 모델1과 모델2는 (그림 6)의 알고리듬 mergeProbability를 이용해서 하나의 모델로 결합한다[8]. 이 알고리즘의 기본 개념은 모델1이 대량의 말뭉치에서 학습되었기 때문에 최대한 신뢰하고 모델2는 평활화(smoothing)의 수단으로 사용된다. 이 논문에서 모델2는 평활화뿐 아니라 미등록어에 대한 확률 정보를 획득하는데 매우 중요한 역할을 한다. 왜냐하면 대부분의 미등록어는 모델1에서 필요한 정보를 얻을 수 없기 때문이다.

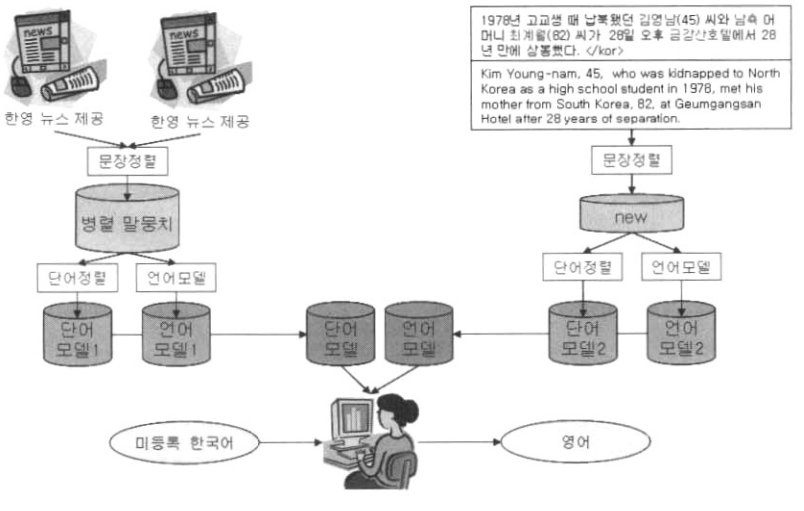
```
알고리즘: mergeProbability(p1, p2)

// 입력: p1와 p1는 확률이며 정의되지 않았을 경우는 undef라는 값을 갖는다.

// 출력: 결합된 확률

1 return p1 if (p1 ≠undef);

2 return p2 if (p2 ≠undef);

3 return MIN_PROBABILITY;
```

(그림 6) 알고리즘 mergeProbability

3.4 미등록 대역어 추출 알고리즘

이 절에서는 앞에서 설명한 미등록어 추출 모델과 학습을 이용해서 미등록 대역어 추출 알고리즘 getUnknownTranslation을 기술한다(그림 7). 이 알고리즘의 입력은 병렬말뭉치 D_{KE}와 한국어 미등록어 리스트 U_K이며, D_{KE}는 한국어 말뭉치 D_K와 영어 말뭉치 D_E로 구성되어 있다. 여기서 $|D_K| = |D_E|$이며, 단어 혹은 형태소 단위로 분석되어 있다고 가정한다. 출력은 한국어 미등록어 U_K에 대응하는 영어 대역어 리스트 $U_{K\rightarrow E}$이며, 각 한국어 미등록어 u_k에 대해 미등록 대역어 $U_{K\rightarrow E}(u_k)$가 존재하지 않으면 nil을 출력한다. 알고리즘에서 자료구조 $U(u_k)$의 초기값은 nil이며 원소는 순서쌍 $(u_e, v(u_e|u_k))$이다.

알고리즘 getUnknownTranslation의 chooseBestTranslation은 (그림 8)과 (그림 9)와 같은 두 종류의 알고리즘을 사용할 수 있다. 두 알고리즘의 차이는 입력 병렬말뭉치 D_K에서 한국어 미등록어 u_k가 두 번 이상 출현해야 한다는 성질을 반영하는 경우(chooseBestTranslation1)와 그렇지 않은 경



(그림 5) 미등록 대역어 추출 모델의 학습 시스템

8) 이 논문에서 알고리즘의 형식의 일부는 perl 언어의 구조를 따른다.

알고리즘: getUnknownTranslation(D_{KE}, U_K)

입력: 병렬말뭉치 D_{KE}, 한국어 미등록어 리스트 U_K

D_{KE}는 D_K(병렬말뭉치의 한국어 문장들)와 D_E(병렬말뭉치의 영어 문장들)로 구성

출력: 한국어 미등록어에 대응하는 영어 대역어 리스트 $U_{K\to E}$

1 단어정렬모델1 $\Pr^1(e_i|f_j)$과 $\Pr^1(f_j|e_i)$을 적재한다.
언어모델1 $\Pr^1(e_i|e_j)$을 적재한다.

2 D_{KE}와 EM 알고리즘을 이용해서 단어정렬모델2 $\Pr^2(e_i|f_j)$와 $\Pr^2(f_j|e_i)$를 학습한다.

3 D_E와 SRILM를 이용해서 언어모델2 $\Pr^2(e_i|e_j)$를 학습한다.

4 $\Pr(e_i|f_j)$=mergeProbability($\Pr^1(e_i|f_j)$, $\Pr^2(e_i|f_j)$) $\forall i, j$;
$\Pr(f_j|e_i)$=mergeProbability($\Pr^1(f_j|e_i)$, $\Pr^2(f_j|e_i)$) $\forall i, j$;
$\Pr(e_i|e_j)$=mergeProbability($\Pr^1(e_i|e_j)$, $\Pr^2(e_i|e_j)$) $\forall i, j$;

5 foreach $(s_K^i, s_E^i) \in D_{KE}$ do

5.1 한국어 문장 s_K^i에 속한 어떤 구절 $f_{j_1}^{j_2}$이 U_K에 포함되면 미등록어로 표시한다.
$f_{j_1}^{j_2}$는 $u_k(\in U_K)$라고 하자.

5.2 식 (7)과 (8)을 이용해서 주어진 병렬문장 s_K^i와 s_E^i에 대해서 $f_{j_1}^{j_2}$에 가장 적합한 대역어 $e_{i_1}^{i_2}$와 $v(e_{i_1}^{i_2}|f_{j_1}^{j_2})$를 구한다.

5.3 $U(u_k) = U(u_k) \cup \{ (e_{i_1}^{i_2}, v(e_{i_1}^{i_2}|f_{j_1}^{j_2})) \}$;

6 enddo

7 foreach $u_k \in U_{K\to E}$ do

7.1 $U_{K\to E}(u_k)$ = bestTranslation($U(u_k)$, u_k);

8 enddo

9 return $U_{K\to E}$;

(그림 7) 알고리즘 getUnknownTranslation

우(chooseBestTranslation2)가 있다. 즉 이 성질을 고려하면 u_k는 적어도 두 번 이상 출현되어야 한다. 두 번 이상 출현된 u_k에 대해서 가장 여러 번 추출된 미등록 대역어 u_e^1을 가장 우선적으로 추정한다. 그렇지 않으면 대응 정도가 가장 높은 미등록 대역어 u_e^2를 후보로 선정하고, 이 대응 정도 $v(u_e^2|u_k)$가 특정 임계값 θ보다 클 경우에 미등록 대역어로 결정한다.

알고리즘: chooseBestTranslation1(S, u)

// 입력: $S = \{(u_e, v(u_e|u_k)) \mid u_k = u\}$

// 출력: u에 가장 적절한 대역어 u_e,

1 return nil if ($|S| < 2$)

2 집합 S에서 u_e의 빈도수를 세어서 가장 높은 빈도를 가지는 대역어를 u_e^1이라고 하자.

3 return u_e^1 if (count(u_e^1) > 2);

4 $v(u_e|u_k)$의 값이 가장 높은 u_e를 u_e^2라고 하자.

5 return u_e^2 if ($v(u_e^2|u_k) > \theta$);

6 return nil

(그림 8) 알고리즘 chooseBestTranslation1

알고리즘: chooseBestTranslation2(S, u)

// 입력: $S = \{(u_e, v(u_e|u_k)) \mid u_k = u\}$

// 출력: u에 가장 적절한 대역어 u_e,

1 $v(u_e|u_k)$의 값이 가장 높은 u_e를 u_e^2라고 하자.

2 return u_e^2 if ($v(u_e^2|u_k) > \theta$);

3 return nil

(그림 9) 알고리즘 chooseBestTranslation2

4. 실험 및 평가

미등록 대역어 추출 모델을 평가하기 위해 공개 말뭉치(public corpus)는 없을 뿐 아니라 미등록어는 응용 시스템에 따라 매우 다르게 때문에 평가하는 데 여러 가지의 어려움이 있었다. 이 장에서는 먼저 실험에 사용된 대량의 병렬말뭉치 구축에 대해서 간단히 기술하고, 미등록 대역어의 자동 추출 시스템을 평가하기 위한 평가말뭉치 구축에 대해서 기술한다. 그리고 나서 미등록 대역어 추출 시스템의 평가 측도와 성능 평가에 대해서 자세히 기술한다.

4.1 병렬말뭉치의 구축

병렬말뭉치 구축 시스템[28]을 이용해서 실험에 필요한 한영 병렬말뭉치를 구축하였다. 구축된 병렬말뭉치의 크기는 약 42만 문장이며 자세한 내용은 <표 1>과 같다. 구축된 말뭉치에 포함된 단어 혹은 형태소 수는 한국어와 영어에 대해서 각각 17,888,617개와 15,581,136개이다. 이 말뭉치는 3장에서 설명한 모델1을 학습하기 위해 사용되며, 학습된 단어쌍은 모두 4,710,654개이며, 여기에 포함된 한국어 및 영어

〈표 1〉 구축된 병렬말뭉치의 크기

출 처	크 기
동아일보 http://english.donga.com/	210,455
전자신문 http://www.etnews.co.kr/	168,460
중앙일보 http://joins.com/cnn/	6,347
VOA http://www.voanews.com/korean/	21,979
기타	17,745
합계	424,986

단어는 각각 158,122개와 113,508이다. 따라서 한국어 단어 당 대응되는 영어 단어 수는 29.7개이고 영어 단어 당 대응되는 한국어 단어 수는 41.5개이다. 한영에 대한 혼잡도(perplexity)보다 영한에 대한 혼잡도가 좀 더 큼을 알 수 있다.

4.2 평가말뭉치의 구축

미등록 대역어 추출 모델을 평가하기 위해 공개 말뭉치는 존재하지 않는다. 따라서 이 논문에서 제안한 미등록 대역어 추출 시스템을 평가하기 위해 평가말뭉치(reference corpus)를 구축하였다. 평가말뭉치의 예는 <표 2>와 같다. <표 2>에서처럼 대역어는 여러 개가 가능하며 주어진 문장에서 찾을 수 없는 경우에는 "정답없음"으로 표시하였다. 이와 같은 방법으로 구축된 한국어 미등록어 수는 1,523개이며 대역어 리스트를 가지는 것이 681개이고 "정답없음"이 840개이다. 한국어 미등록어의 최대 형태소 수는 9이며 평균은 2.7개이다.

〈표 2〉 평가말뭉치의 예

문자열	형태소 분리	대역어 리스트
6자회담	6 자 회담	six-party talks, 6-party talks, 6-way talks
원내대표	원내 대표	floor leader
남자간호사	남자 간호사	male nurses
부동산정책	부동산 정책	real estate policy
정연욱	정연욱	정답없음
수급계획	수급 계획	정답없음

입력 병렬말뭉치는 웹 문서로부터 자동으로 2,220 문장을 구축하였으며, 한국어 형태소 수는 74,864개이고, 영어 단어 수는 64,821개이다. 이 말뭉치에는 자동으로 구축되어 문장 정렬에 관련된 약간의 오류가 포함되어 있으나, 전체적으로 시스템의 성능을 분석하는 데는 큰 문제가 없었다.

4.3 성능 평가 측도

<표 3>에서 A_O는 시스템이 제시한 미등록 대역어를 정확히 맞춘 경우의 수이고, A_X는 시스템이 제시한 미등록 대역어가 완전히 일치하지 않은 경우(A_S는 제외)의 수이고 A_S은 시스템 제시한 미등록 대역어가 정답의 부분문자

〈표 3〉 시스템 성능 평가를 위한 2차원 분할표

정답 / 시스템	있음	없음	
후보 제시	$A = A_O+A_X+A_S$	B	$A+B$ (미등록어 추출된 수)
후보 미제시	C	D	$C+D$ (미등록어 미추출된 수)
	$A+C$ (정답이 있는 미등록어 수) (681)	$B+D$ (정답이 없는 미등록어 수) (840)	$A+B+C+D$ 대상 미등록어 수 (1,521)

열이 되는 경우의 수이다. A_S는 영어의 복수 등의 차이에 의해서 불일치되는 경우이다. 예를 들면 "six-party talk"와 "six-party talks"가 불일치하는 경우의 수이다. 즉 정답에 "six-party talk"가 존재하고 시스템이 "six-party talks"를 찾았을 경우의 수이다. A는 $A_O+A_X+A_S$을 의미한다. 엄밀하게 말하면 A_X와 A_S는 틀린 결과이다. B는 시스템이 미등록 대역어를 제시했지만 정답에는 "정답없음"일 경우의 수이고, C는 시스템이 미등록 대역어를 제시하지 않았지만 정답에는 후보가 있는 경우의 수이다. D는 시스템이 미등록 대역어를 제시하지 않았고 정답에도 "정답없음"일 경우의 수이다. 이 논문에서 성능 평가를 위해 사용된 측도는 정확도$_1$ A_1(accuracy), 정확도$_2$ A_2(accuracy including substring), 정밀도 P(precision) 그리고 재현율 R(recall) 이며 각각 식 (10)-(13)과 같다. 정확도$_1$은 순수한 의미의 정확도이고, 정확도$_2$는 A_S는 맞다고 가정할 경우의 정확도이다.

$$A_1 = \frac{A_O + D}{A+B+C+D} \tag{10}$$

$$A_2 = \frac{A_O + A_S + D}{A+B+C+D} \tag{11}$$

$$P = \frac{A_O}{A+B} \tag{12}$$

$$R = \frac{A_O}{A+C} \tag{13}$$

4.4. 매개변수 최적화

3장에서 설명했듯이 시스템에는 다섯 종류(α, λ_1, λ_2, λ_3, θ)의 매개변수가 있다. α는 한영번역확률, λ_1은 번역확률, λ_2는 경계확률, λ_3은 언어모델의 반영 정도를 의미한다. θ는 임계값으로 미등록어와 미등록 대역어의 대응 정도를 의미한다. 이들의 매개변수의 최적치를 찾는 것은 대단히 어려운 일이다. 이 논문에서는 언덕오르기 탐색 방법(hill-climbing search)을 이용한다. 즉 여러 개의 매개변수들 중에서 하나의 매개변수를 제외하고 모두 고정시키고 그 매개

〈표 4〉 매개변수 ($\alpha, \lambda_3, \lambda_2$)의 최적화 과정

$\alpha/\lambda_3/\lambda_2$	A_1		
	(*, 1.0, 1.0)	(0.3, *, 1.0)	(03, 0.4, *)
0.0	0.5489	0.5561	0.5561
0.1	0.5522	0.5561	0.5601
0.2	0.5555	0.5561	0.5614
0.3	0.5568	0.5561	0.5627
0.4	0.5561	0.5568	0.5588
0.5	0.5555	0.5568	0.5588
0.6	0.5568	0.5568	0.5568
0.7	0.5561	0.5568	0.5575
0.8	0.5561	0.5568	0.5568
0.9	0.5548	0.5568	0.5561
1.0	0.5515	0.5568	0.5568

변수의 최적치를 정한다. 이와 같은 과정을 반복하여 모든 매개변수의 최적치를 결정한다. 이 논문에서 매개변수 (λ_1, λ_2, λ_3)는 λ_1을 기준으로 한 상대적인 값으로 간주하여 λ_1을 항상 1.0으로 설정하였으며, 나머지 매개변수들의 결정 순서는 (α, λ_3, λ_2, θ)로 정했다. 물론 이 순서를 어떻게 정하느냐에 따라 최적치가 또 달라질 수 있다. 또한 세 개의 매개변수 (α, λ_3, λ_2)는 0.0에서 1.0까지 0.1 간격으로 변화시켰으며, 값이 0.0일 경우는 해당하는 속성이 모델에서 불필요함을 의미한다. 임계값 θ는 5에서 50까지 개략적인 로그 간격(log scale)으로 변화시켰다. <표 4>는 임계값 θ가 -7일 때, (α, λ_3, λ_2)의 최적화 과정을 보인 것으로 최종적으로 (0.3, 0.4, 0.3)로 조절되었다. 이와 같은 방법으로 여러 임계값 θ에 대해서 조절하였다. 자세한 결과는 4.5절에서 다룰 것이다.

4.5 성능 평가: 미등록 대역어 추정

<표 5>는 알고리즘 chooseBestTranslation1을 이용한 성능을 보이고 있다. 첫 번째 열의 매개변수들은 4.4절에 기술한 방법으로 최적화되었으며 하나를 제외한 모든 매개변수가 0.0이 아닌 값을 가지고 있다. 이것이 의미하는 바는 3장에서 제안한 모델에서 모든 확률(번역확률, 경계확률, 언어모델)이 미등록 대역어 추출을 위해서 유용함을 의미한다. $A1$을 기준으로 각 매개변수 (λ_2, λ_3, α, θ)는 (0.3, 0.4, 0.3, -7)일 때, 가장 좋은 성능을 보였다. 경계확률과 언어모델은 비슷한 비율로 반영되어야 함을 알 수 있었다. 일반적으로 임계값 θ가 증가하면 많은 결과를 제시하지만 잘못된 미등록 대역어를 제시하는 비율이 점점 더 높아진다. A_2의 경우는 임계값 θ가 증가할수록 좋은 결과를 보였다.

<표 6>는 알고리즘 chooseBestTranslation2를 이용한 성능을 보이고 있다. 이 경우는 매개변수가 0.0인 경우가 있었으나 대부분의 경우가 역이 0.0이 아닌 값을 가지고 있다. 따라서 이 경우에서 제안된 모델에서 모든 확률(번역확률, 경계확률, 언어모델)이 미등록 대역어 추출을 위해서 유용함을 의미한다. A_1을 기준으로 각 매개변수 (λ_2, λ_3, α, θ)는 (0.3, 0.0, 0.7, -8)일 때, 가장 좋은 성능을 보였다. 임계값 θ가 증가하면 거의 모든 미등록어 후보를 제시하게 된다.

알고리즘 chooseBestTranslation2를 이용하는 경우가 좀 더 좋은 성능을 보이지만 반자동으로 미등록 대역어를 획득하려고 할 경우에는 알고리즘 chooseBestTranslation1이 훨씬 더 안정적일 것으로 판단된다.

〈표 5〉 chooseBestTranslation1을 사용했을 때 성능

(λ_2, λ_3, α, θ)	A_O	A_X	A_S	B	C	D	A_1	A_2	P	R
(0.1, 0.4, 0.3, -5)	29	30	5	15	617	827	0.5620	0.5883	0.3671	0.0426
(0.1, 0.4, 0.3, -6)	29	32	7	16	613	826	0.5614	0.5916	0.3452	0.0426
(0.3, 0.4, 0.3, -7)	28	25	9	13	619	829	0.5627	0.5909	0.3733	0.0411
(0.9, 0.0, 0.7, -8)	24	18	14	12	625	830	0.5607	0.5909	0.3529	0.0352
(1.0 0.4, 0.5, -9)	21	18	21	11	621	831	0.5594	0.5988	0.2958	0.0308
(1.0, 0.4, 0.4, -10)	24	23	27	17	607	825	0.5575	0.6080	0.2637	0.0352
(0.8, 0.6, 0.8, -15)	44	61	27	74	549	768	0.5332	0.6087	0.2136	0.0646
(0.4, 0.3, 1.0, -20)	47	100	20	96	514	746	0.5207	0.6126	0.1787	0.0690
(0.4, 0.8, 0.7, -30)	59	86	45	115	491	727	0.5161	0.6316	0.1934	0.0866
(0.4, 0.8, 0.7, -40)	59	89	45	118	488	724	0.5141	0.6316	0.1897	0.0866
(0.4, 0.8, 0.7, -50)	59	89	45	118	488	724	0.5141	0.6316	0.1897	0.0866

〈표 6〉 chooseBestTranslation2을 사용했을 때 성능

(λ_2, λ_3, α, θ)	A_O	A_X	A_S	B	C	D	A_1	A_2	P	R
(0.0, 0.4, 0.3, -5)	27	12	11	7	631	835	0.5660	0.5883	0.2000	0.0117
(0.0, 0.4, 0.3, -6)	41	22	16	16	602	826	0.5693	0.6047	0.3772	0.1263
(0.2, 0.4, 0.3, -7)	47	23	29	24	582	818	0.5680	0.6211	0.3991	0.1307
(0.3, 0.0, 0.7, -8)	108	64	46	68	463	774	0.5791	0.6815	0.3855	0.1557
(0.7 0.4, 0.5, -9)	69	31	48	40	533	802	0.5719	0.6553	0.3680	0.1454
(1.0, 1.0, 0.6, -10)	60	39	78	51	504	791	0.5588	0.6868	0.3312	0.1557
(0.7, 0.6 0.8, -15)	170	147	111	306	253	536	0.4636	0.7058	0.2243	0.2276
(0.9, 0.3, 1.0, -20)	161	221	106	395	193	447	0.3992	0.6835	0.1819	0.2364
(1.0, 1.0, 1.0, -30)	149	313	163	685	56	157	0.2009	0.6205	0.1456	0.2834
(0.3, 0.8, 0.7, -40)	220	310	147	826	4	16	0.1550	0.5515	0.1309	0.2849
(0.3, 0.8, 0.7, -50)	220	311	150	834	0	8	0.1497	0.5509	0.1460	0.3245

4.6 분석

이 논문에서 제안한 미등록 대역어 추출 시스템의 성능을 다른 시스템과 직접적으로 비교할 수 없었다. 왜냐하면 미등록 대역어 추출에 대한 연구가 거의 진행되지 않았으며, 미등록어라는 개념이 시스템에 매우 의존적이기 때문에 더욱 그러하다. 이 논문에서 제안된 미등록 대역어 추출 시스템의 평가에서 몇 가지 문제점을 관찰할 수 있었다. 첫째, 병렬말뭉치가 자동으로 구축됨으로 약간의 오류를 포함하고 있었다. 더구나 입력 병렬말뭉치에도 약간의 오류를 포함하고 있었다. 둘째, 한국어 형태소 분석에서 미등록어가 발생할 경우 미등록어를 추정하여 형태소를 분리한다. 같은 형태소라도 주변 문맥에 따라서 서로 다르게 해석되는 경우가 종종 있다. 예를 들면 "무호흡증"은 주변 문맥에 따라서 "무호흡 증" 혹은 "무호흡 증"으로 분리되었다. 이 경우 두 해석은 서로 다른 결과를 추출하게 된다. 따라서 한국어 형태소 분석에서 미등록어 추정의 일관성 유지가 미등록 대역어 추출 시스템에 영향을 줄 수 있다. 셋째, 평가말뭉치를 작성할 때, 시간적인 속성은 어느 정도 반영되었으나 공간적(영역적)이거나 사회적인 측면은 그다지 고려하지 않았다. 따라서 입력 병렬말뭉치를 만들 때 좀 더 신중함이 있어야 할 것이다.

5. 결론 및 앞으로의 연구 과제

이 논문은 정렬기법을 이용한 한국어 미등록어에 대한 영어 대역어의 자동추출 방법에 대해서 제안하였다. 제안된 시스템의 입력은 병렬말뭉치와 원시언어의 미등록어 리스트이고, 출력은 주어진 미등록어들에 대한 미등록 대역어들이다. 제안된 미등록 한영 대역어 자동 추출 시스템은 크게 네 단계로 구분된다. 첫 번째 단계는 웹으로부터 한영 뉴스 기사들을 수집하고, 수집된 병렬문서와 문장정렬 도구를 이용해서 병렬 말뭉치를 구축한다. 두 번째 단계는 구축된 병렬 말뭉치와 단어정렬 도구를 이용해서 단어정렬 모델을 구하고, 구축된 병렬말뭉치의 영어 말뭉치와 언어모델 구축 도구를 이용해서 목적언어인 영어의 언어모델을 구한다. 세 번째 단계는 두 번째 단계와 같은 방법으로 입력 문서에 대해서 단어정렬 모델과 영어 언어모델을 구한다. 네 번째 단계는 학습된 단어정렬 모델과 언어모델을 이용해서 미등록 대역어를 추정하는데, 기본 개념은 구절정렬 알고리즘을 이용하며, 입력은 한국어 미등록어가 표시된 한영 병렬문서이고 출력은 표시된 미등록어에 대한 영어 대역어이다.

제안된 시스템의 성능을 평가하기 위해 이 논문에서는 약 1500여개의 미등록어 리스트를 포함하는 병렬말뭉치를 평가말뭉치로 구축하였다. 여러 실험을 통해서 제안된 모델이 미등록 대역어를 추출하는 데 유용함을 알 수 있었다. 또한 제안된 시스템의 성능이 약 60%의 정확률을 보였으나, 여러 가지 면에서 개선할 점이 있는 것으로 관찰되었다.

앞으로 제안된 미등록 대역어 추출 시스템의 성능을 개선하기 위해 좀 더 양질의 병렬말뭉치 구축이 필요하며, 미등록 대역어 추출 모델의 문제점을 파악하기 위해 좀 더 다양한 형태의 실험이 이루어져야 할 것이다. 또한 미등록 대역어를 추출하기 위해서 정렬 방법 이외에 서론에서 언급했든 미등록어의 다양한 성질을 고려한 종합적인 연구가 필요할 것이다. 한 예를 들면 언어학적인 면에서 고유명사, 차용어, 전문용어 등으로 분류하여 각각에 대해서 서로 다른 방법으로 미등록 대역어를 추출할 수 있을 것이다. 또한 평가말뭉치를 작성할 때, 시간적인 속성은 어느 정도 반영되었으나 공간적(영역적)이거나 사회적인 측면은 고려한다면 좀 더 좋은 성능을 보일 수 있을 것이다.

참 고 문 헌

[1] Hutchins, W. J. and Somers, H. L., *An Introduction to Machine Translation*, Academic Press Limited, 1992.

[2] Papineni, K. Roukos, S. Ward, Todd, Zhu, W. J., *BLEU: A Method for Automatic Evaluation of Machine Translation*, IBM Research Report RC22176, 2001.

[3] NIST 2006 Machine Translation Evaluation Official Results, http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html, 2006.

[4] Arnold, D. J., Balkan, L., Meijer, S., Humphreys, R. L. and Sadler, L., *Machine Translation: an Introductory Guide*, Blackwells-NCC, London, 1994.

[5] Rey, A., *Eassys on Terminology*, John Benjamins, 1997.

[6] Sinha, R. M. K., "Interpreting Unknown Words in Machine Translation from Hindi to English", *Proceeding of Computational Intelligence,* pp.278-282, 2005.

[7] 이연호, 김금희, 이홍윤, 유병기, 김규웅, 이영교, 임인칠, "한-일 기계번역 시스템의 관용구 및 미등록어 처리 알고리즘" 대한전자공학회 학술대회 논문집, 제14권 1호, pp.201-204, 1991.

[8] Manning, C. D. and Schütze, H., *Foundation of Statistical Natural Language Processing*, The MIT Press, 1999.

[9] Resnik, P. and Smith N.A., "The web as a parallel corpus", *Computational Linguistics*, vo. 29, no. 3, pp.349-380, 2003.

[10] Kilgarriff, A. and Grefenstette, G., "Introduction to the Special Issue on the Web as Corpus". *Computational Linguistics*, vol. 29, no. 3, pp.333-347, 2003.

[11] Gale, W. A. and Church, K. W., "A program for aligning sentences in bilingual corpora", *Computational Linguistics,* vol. 19, no. 1, pp.75-102, 1993.

[12] Brown, P., Della Pietra, V., Della Pietra, S., and Mercer, R., "The mathematics of statistical machine translation: Parameter estimation", *Computational Linguistics,* vol. 19, no. 2, pp.263-311, 1993.

[13] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V., "Translating collocations for bilingual lexicons: A statistical approach", *Computational Linguistics*, vol. 22, no. 1, pp.1-38, 1996.

[14] Diab, M. "An unsupervised method for word sense tagging using parallel corpora: A preliminary investigation", *Special Interest Group in Lexical Semantics Workshop*, Association for Computational Linguistics, 2000.

[15] Zhang, Y. and Vogel, S., "An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora", *Proceedings of the Tenth Conference of the European Association for Machine Translation*, pp.294-301, 2005.

[16] Och, F. J. and Ney, H., "The alignment template approach to statistical machine translation", *Computational Linguistics*, vol. 30, no. 4. pp.417-449, 2004.

[17] Wu, D. "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora", *Computational Linguistics*, vol. 23, no. 3, pp.377-403, 1997.

[18] Yamada, K. and Knight, K. "A syntax-based statistical translation model", *Proceedings of the 39th Annual Conference of the Association for Computational Linguistics*, pp.523-530, 2001.

[19] Ion, R., Ceausu, A. and Tufs, D. "Dependency-based phrase alignment", *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pp.1290-1293 2006.

[20] Gale, W. and Church, K. "Identifying word correspondence in parallel text", *Proceedings of the workshop on Speech and Natural Language*, pp.152-157, 1991.

[21] Fung, P. and Church, K. "K-vec: A new approach for aligning parallel texts", *Proceedings of COLING 94*, pp.1096-1102, 1994.

[22] Hiemstra, D. "Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus", *Proceedings of the 8th CLIN meeting*, pp.41-58, 1998.

[23] Fung, P. "A statistical view of bilingual lexicon extraction: From parallel corpora to nonparallel corpora", *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, pp.1-16, 1998.

[24] Varma, N. *Identifying Word Translation in Parallel Corpora Using Measures of Association*, Master Thesis, Department of Computer Science, University of Minnesota, USA, 2002.

[25] Koehn, P. *Noun Phrase Translation*, PhD. Thesis, University of Southern California, 2003.

[26] Callison-Burch, C., Koehn, P. and Osborne, M. "Improved statistical machine translation using paraphrases", *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp.17-24, 2006.

[27] Kim, C.-H. and Hong, M. "A Korean syntactic parser customized for Korean-English patent MT system", *Proceedings of the 5th International Conference on Natural Language*, pp.44-55, 2006.

[28] 서형원, 김형철, 조희영, 김재훈, 양성일, "웹 문서로부터 한영 병렬말뭉치의 자동 구축", 제26회 한국정보처리학회 추계학술발표대회 논문집, 제13권, 제2호, pp.161-164, 2006.

[29] 조희영, 서형원, 김재훈, 양성일, "한영 명사구 기계 번역", 제18회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.273-278, 2006.

[30] Stolcke, A. "SRILM-An extensible language modeling toolkit", *Proceedings of Intl. Conf. on Spoken Language Processing*, vol. 2, pp.901-904, 2002.

[31] Crego, J.M., Marino, J. B., Gispert, A. "An ngram-based statistical machine translation decoder", *Proceedings of the 9th EuropeanConference on Speech Communication and Technology*, pp.3193-3196, 2005.

김 재 훈

e-mail : jhoon@hhu.ac.kr

1986년 계명대학교 전자계산학과 학사

1988년 한국과학기술원 전산학과 공학석사

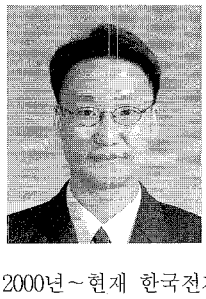
1996년 한국과학기술원 전산학과 공학박사

1988년~1997년 한국전자통신연구원, 선임연구원

1997년~현재 한국해양대학교 컴퓨터공학과 부교수

2001년~2002년 USC, Information Sciences Institute, 방문연구원

관심분야 : 자연언어처리, 한국어정보처리, 정보검색, 정보추출



양 성 일

e-mail : siyang@etri.re.kr

1994년 연세대학교 전산학과 학사

1996년 연세대학교 전산학과 석사

1998년 연세대학교 컴퓨터공학과 박사수료

1996년~2000년 아시아나항공 소프트웨어연구소 주임연구원

2000년~현재 한국전자통신연구원 언어처리연구팀 선임연구원

관심분야 : 자연언어처리, 한국어정보처리, 기계번역, 기계학습