극한 언어 환경에 대응 가능한 영한 자동 주소번역 시스템

김경식○, 황명진, 이승필
시스트란 인터내셔널
{jingzhi.jin, myeongjin.hwang, seungphil.lee}@systrangroup.com

Automatic English-Korean Address Translation System for Extremely Unpredictable Error Generating Language Environments

Jingzhi Jin○, Myeongjin Hwang, Seungphil Lee
SYSTRAN International

요 약

데이터베이스 기반 자동 주소번역은 입력 오류에 취약하며 범용 기계번역을 이용한 주소번역은 입력 및 번역 주소에 대한 품질 평가가 어렵다. 본 논문에서는 예측할 수 없는 입력 오류에도 대응할 수 있는 자동 주소번역 시스템을 제안한다. 제안 시스템은 n-gram 기반 검색, 미검색/오검색 분류, 번역, 신뢰도 자동평가로 구성된다. 신뢰할 수 있는 입력으로 자동 분류한 영문 국내주소를 국문으로 번역한 결과 95% 이상의 정확도를 보였다.

주제어: 자동, 주소번역, 정보검색

1. 서론

자*동 주소번역이란 원시 언어로 된 주소를 목적 언어로 번역하는 자동 기계번역 기법이다. 글로벌 시장경제의 발전에 따라 일상생활에서 해외 직접구매 수요는 나날이 증가하고 있고, 그에 따른 국제 물류유통 수요도 폭발적으로 증가하고 있다. 정확한 주소번역은 물류유통에서 가장 중요한 부분이지만 사람이 일일이 수작업으로 번역하는 것이 국내 현실이다. 인력과 시간, 비용 절감을 위해 자동 주소번역이 필요한 시점이다.

자동 주소번역에 사용되는 원시 언어는 극한 언어 환경(extremely unpredictable error generating environments)에서 생성된다. 극한 언어 환경이란 예측할 수 없는 다양한 오류가 발생하는 환경을 말하며, 이런 환경에서 생성된 원시 언어는 자동 주소번역에서 해결해야 가장 큰 문제다.

본 연구는 자동 주소번역의 초기 연구로서 영문 국내주소를 국문으로 자동 번역하는 방법을 제안한다. 논문의 구성은 다음과 같다. 2장은 자동 주소번역의 연구현황을 분석하고, 3장은 본 논문에서 제안하는 영한 자동 주소번역 방법을 설명한다. 4장은 실험결과를 제시하고 5장은 결론과 향후 연구에 대해 설명한다.

2. 관련 연구

자동 주소번역에 대한 연구는 활발히 진행되고 있지 않다.

한국어 관련 자동 주소번역 서비스는 네이버가 제공하고 있는 영문주소번역기[1]가 유일하다. 이 시스템은 국문 국내 주소를 입력하면 영문 주소와 우편번호를 알려주지만 비정형 입력에 대해서는 결과를 도출하지 못한다.

입력: "서울특별시 서초구 양재동 275-5"
번역: "275-5, Yangjae-dong, Seocho-gu, Seoul, Korea"

web-burger.com[2]는 일본지역에 특화된 자동 주소번역 결과를 알려준다. 입력 양식은 제한적이며 우편번호와 번지수를 정확히 입력하면 그에 상응하는 일문 주소와 영문 주소를 알려준다.

구글번역기[3] 등에서 제공하는 자동번역기도 있지만, 주소번역에 특화되어 있지는 않아서 번역 품질도 낮으며 입력 오류에도 대응하지 못한다.

입력: 5, 10-gil, Mabang-ro, Seocho-gu, Seoul, Korea
번역: 5, 10 길 , Mabang 특별시 , 서초구 , 서울, 한국

현업에서 사용할만한 주소번역기는 찾아보기 힘들며, 활용할 수 있는 시스템은 앞서 살펴본 것처럼 데이터베이스기반이거나 범용 기계번역기다. 두 가지 모두 입력에서 발생할 수 있는 오류에 대응하지 못하며, 범용 번역기는 번역 결과에 대한 신뢰도도 자동 판단하지 못한다.

이러한 상황으로 인해 극한 언어 환경에 대응 가능한 자동 주소번역 시스템이 필요하다. 본 논문에서는 국내 주소를 영문에서 국문으로 자동 번역하는 시스템을 제안한다.

3. 영한 자동 주소번역 시스템

본 논문에서 제안하는 자동 주소번역 시스템은 예측할

* 본 논문은 한국산업기술진흥원의 국제공동기술개발사업 '아시아어와 유럽어 임베디드 음성 번역기'(과제번호: N0001245) 과제의 성과물임

수 없는 입력오류에 대응한다. [표 1]은 주소번역의 입력 주소에는 철자오류, 띄어쓰기 오류, 비표준표기법, 약어, 생략 및 누락, 입력 순서 오류 등의 문제가 존재한다.

[표 1] 주소번역 입력 예문

입력 예문	문제 유형
220 Gung-dong ***Youseong-gu*** Chungnam NationalUniv. EngineeringBuilding #1 Room302-2.Daejeon	비표준표기법
17 WORLD CUP BUK RO , 60 ***GL MAIRO GU*** SEOUL 121921	철자오류, 띄어쓰기 오류
SIX NAMSAN TOWER 98 HUAMRO	OCR 인식 오류
15TH FL., HANJIN BLDG, 118, 2-GA NAMDAEMUM-RO, JUNG-GU	약어
KDI School 15 Giljae-gil SejongSejong***SE***	생략 및 누락

본 논문에서는 이런 예측할 수 없는 입력오류를 [그림 1]과 같이 처리하고 번역한다.

[그림 1] 자동 주소번역 흐름

3.1 전처리

전처리 단계에서는 검색 복잡도를 낮추고 검색 정확도를 높이는 것을 목표로 입력 문자열을 정제하는 작업을 한다. 이 단계를 거친 결과물은 최종 번역결과가 나올 때까지 번복 없이 사용될 것이므로 정보손실 없이 정제되도록 하는 것이 중요하다. 따라서 주소 표기상의 특성 및 약어 정규화, 확실한 띄어쓰기 복원과 극히 일부의 오타교정만 처리한다.

입력: "Gangnam-gu Apgujung-ro 29-gil 71Hyundai Apt.32-302Seoul06002"
결과: "GANGNAMGU APGUJUNGRO 29GIL 71 HYUNDAI APT. 32-302 SEOUL 06002"

3.2 주제어(keyword) 확장

주제어 확장 단계에서는 색인 검색 시 원하는 결과를 더 잘 찾을 수 있도록 원래 입력에 없던 검색어를 추가하는 작업을 한다. 추가되는 검색어는 입력과 데이터베이스의 실제 데이터 간 오차를 줄이는 역할을 한다.

입력: namdaemoonro
결과: namdaemoonro namdaemunro namdaemoonno
입력: 12 28
결과: 12 28 12-28

주소 표기에서 '문로'를 영문으로 표기할 경우 발음대로 '문노'로 표기하는 것이 표준이며 데이터베이스도 그렇게 구성되어 있다. 하지만 건물명 등에서는 이를 지키지 않는 경우도 있다.

3.3 색인 검색

주소표기 형태의 비일관성이나 비표준을 벗어나는 오류도 있다. 행정구역의 표기 순서가 뒤죽박죽이거나 누락 및 생략, 오탈자 및 띄어쓰기 오류가 그러한 예이다. 특히 OCR로 인식된 주소의 경우 예측할 수 없는 오타나 띄어쓰기 오류를 포함하고 있다. 이와 같은 형태의 입력은 일반적인 데이터베이스 기반으로 방법으로는 처리가 불가능하다. 본 논문에서는 이 문제를 극복하기 위해 정보검색[4]에서 흔히 쓰이는 n-gram 색인 검색 방법을 사용한다.

색인에 사용되는 학습데이터(training data)는 정부에서 공개한 전국 주소데이터[5]와 영한 주소사전을 사용하여 구축하였다.

3.4 최적후보 선택

우리가 색인 검색을 통해 찾고자 하는 주소가 항상 첫 번째로 나타나지는 않는다. 원인은 입력에 포함된 불필요한 잡음때문일 수도 있고 n-gram을 사용하는 방법론 자체의 문제일 수도 있다. 이를 극복기 위해 상위 후보 주소(10~20%)들을 대상으로 입력 주소와의 유사도를 정밀하게 재측정한다. 방법은 [수식 1]과 같다.

$$best = \mathrm{argmax}_{address_i} \sum_{n=2,3,4} score(n, address_i) \times n \quad \text{[수식 1]}$$

[수식 1]에서의 $score(n, address_i)$는 입력 주소와 후보 주소 간의 n-gram 유사도이다.

3.5 오검색/미검색 판별

[표 3] 색인 저장 양식

저장 양식
03921|서울특별시|SEOUL|마포구|MAPOGU||||월드컵북로60길|WORLD CUP BUNGNO 60GIL||||||우리금융상암센터|WOORI FINANCIAL SANGAM CENTER||상암동|SANGAMDONG|||상암동|SANGAMDONG||||||||17|1585|121-835

[표 3]은 최종 선택된 후보주소(검색결과)의 예이다. 후보주소는 번역 및 신뢰도 측정의 길잡이 역할로 사용된다. 후보주소를 통해 번역 및 신뢰도 측정을 하기 위해서는 오검색/미검색 판별이 선결되어야 한다. 오검색/미검색 판별은 후보주소와 입력주소의 행정단위별 비교를 통하여 이뤄진다.

오검색: 후보주소에 있지만 입력주소와 불일치하여 버려야 할 부분
미검색: 입력주소에 있지만 후보주소에는 관련 정보가 없는 부분

비교가 끝난 후 오검색과 미검색을 제외하면 입력주소와 후보 주소 간 '정렬정보'를 얻을 수 있다. 즉, 입력 주소의 정확한 표기 및 국문 대역 정보를 후보주소로부터 알 수 있게 되는 것이다. [그림 2]는 그 과정이다.

3.6 번역 및 생성

번역은 데이터베이스 기반 번역과 일반적인 기계번역 방법을 사용한다.

오검색/미검색 판별로 얻은 정렬정보를 통해 국문 대역어를 [그림 2]와 같이 구할을 수 있다. 즉, 정렬된 주소는 데이터베이스 기반 번역을 수행할 수 있 한다.

정렬된 결과: 'deoksan villa' ==> 'DEOKSAN VILLA'
대역 결과: 'deoksan villa' ==> '덕산빌라'

그리고 미검색 부분은 일반적인 기계번역을 사용한다. 신뢰할 수 있는 입력주소의 경우 주소지 확인에 중요한 주소정보는 후보주소를 통해 고품질의 데이터베이스기반 번역을 하고, 상대적으로 덜 중요한 나머지 정보, 예를 들어 건물명이나 기관명 등은 저품질의 기계번역을 한다. 정확성과 유연성을 모두 확보할 수 있는 구성이다.

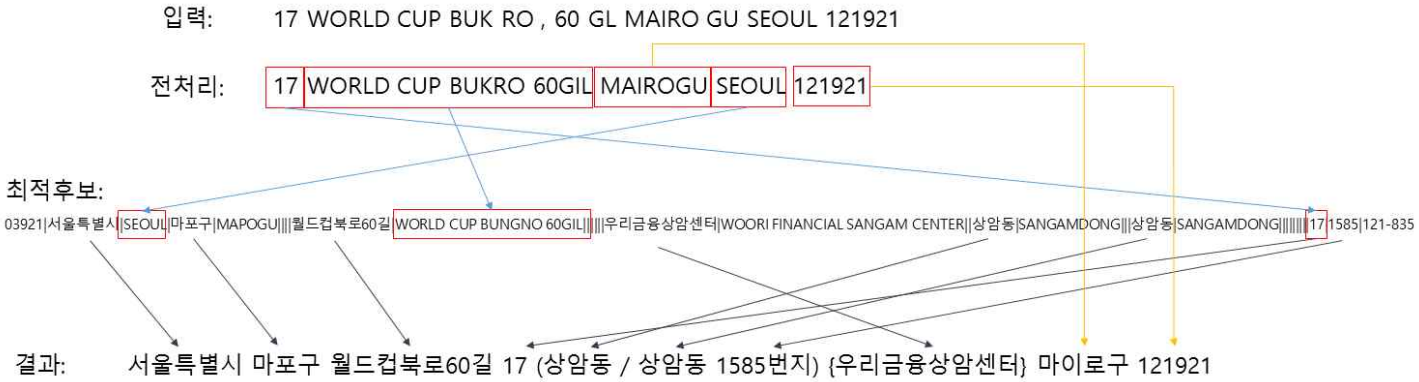
3.7 신뢰도 측정

입력주소와 최종 결과주소를 생성하는 과정에서 본 시스템은 신뢰도 평가를 진행한다. 신뢰도는 평가 등급에 따라 [표 4]와 같이 분류한다.

[표 4] 신뢰도

LEVEL		신뢰도
신뢰	A	완벽한 번역
	B	번지수까지 신뢰할 수 있는 번역
	C	신뢰할 수 있는 번역(지번 누락)
비신뢰	D	오검색으로 의심되는 번역
	E	정보 부족한 번역
	F	번역 실패

A,B,C 등급은 신뢰할 수 있는 수준의 번역결과이고, D,E,F는 신뢰할 수 없는 수준의 번역결과이다.



[그림 2] 최종 결과 출력의 예

신뢰도는 '정렬정보'로부터 얻을 수 있다. 예를 들어 도로명과 건물번호가 확인되면 건물까지의 주소는 신뢰할 수 있으며 검색된 후보주소의 정보는 그대로 사용해도 된다.

신뢰도 측정은 번역 결과에 대한 신뢰 수준을 알려줄 뿐만 아니라 입력에서 누락된 정보도 확인해 번역 결과에 적용할 수 있다. 예를 들어 [그림 2]의 경우 서울특별시나 마포구, 우편번호 등이 입력에서 누락되어 있었어도 복원해 번역할 수 있다.

4. 실험 및 평가

객관적인 성능 평가를 위해 무작위로 추출한 960개 입력주소를 평가 데이터로 사용했다. 그 결과는 [표 5]와 같다. [표 5]에서 "정상"은 번역결과가 정확한 것이

고, “불량”은 번역결과가 부분적으로 정확한 것이고, “오류”는 틀린 번역결과이다.

[표 5] 성능 평가

신뢰도	정상	불량	오류
A	264	0	3
B	397	0	31
C	42	0	0
E	35	29	4
G	105	19	22
Z	6	1	2
합계	849	49	62

[표 5]에서 알 수 있는 바와 같이 신뢰할 수 있는 번역결과의 정확도는 95.39%이다.

5. 결론

본 연구는 n-gram 기반 검색, 미검색/오검색 분류, 번역, 신뢰도 자동평가로 구성된 영한 자동 주소번역 시스템을 제안하였다. 성능 평가 결과, 고신뢰로 분류된 번역 결과의 정확도는 95.39%였다. 현업에서의 활용성을 높이기 위해서는 고신뢰로 분류된 번역 결과의 정확도를 더 높여야 한다. 또 물류유통을 대상으로 한 시스템이기에 휴리스틱(heuristic)한 기법을 추가하여 번역 성능을 올릴 예정이다.

참고문헌

[1] 네이버 영문주소번역기 [Online]. Available: https://search.naver.com/search.naver?where=nexearch&query=영문주소번역기 (accessed 2016, September 07)

[2] Address Translation System [Online]. Available: http://sys.web-burger.com/zip/ (accessed 2016, September 07)

[3] 구글 번역기 [Online]. Available: https://translate.google.co.kr/ (accessed 2016, September 07)

[4] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[5] 도로명주소 안내시스템 [Online]. Available: http://www.juso.go.kr/ (accessed 2016, September 07)