기계 번역기의 언어별 외래어 인식 정확도 비교 연구

김규석*
*한국폴리텍대학 분당융합기술교육원 데이터융합SW과
kyuseokkim@kopo.ac.kr

A Comparative Study on the Machine Translation Accuracy of Loanword by Language

Kyuseok Kim*
*Dept. of Data Convergence Software, Bundang Convergence Technology Campus of Korea Polytechnics

요 약

4차 산업혁명 시대에는 빠른 무선 네트워크와 빅데이터를 기반으로 다양한 기술과 서비스들이 생겨나고 있다. 이런 환경 속에서 우리는 언제 어디서나 스마트폰을 통해 음악을 듣고, 게임을 하며, 웹서핑을 하는 등 PC에 버금가는 다양한 활동을 할 수 있다. 누구든 쉽게 전세계의 웹페이지에 접속하고 SNS를 통해 외국인 친구들과도 쉽게 연락을 할 수 있다. 기계 번역 기술 또한 이렇게 사용자가 늘어나는 만큼 빅데이터를 기반으로 그 정확도가 향상되고 있다. 그러나 일반 명사나 구문과는 다르게 은어, 외래어 등의 사용빈도가 상대적으로 낮은 단어들에 대한 기계 번역 정확도는 여전히 개선이 필요하다. 본 연구에서는 국내에서 가장 많이 사용되는 기계 번역기인 papago 번역기와 Google 번역기의 외래어 인식 정확도에 대한 비교 연구를 진행하였다. 추후, 본 연구 결과를 통해 앞으로의 새로운 연구 방향을 제시한다.

1. 서론

외래어란 고유어가 아닌 외국에서 들어와 자국어처럼 사용되는 말을 의미한다. 외래어와 같은 의미로 차용어라고도 일컬어진다[1].

최근에는 딥러닝, 머신러닝과 같은 인공지능 관련 기술이 크게 각광받으며 발전하고 있다. 그래서 기계 번역 분야에서도 인공지능 기술을 활용하여 그 정확도가 향상되고 있다. 기존에는 통계 기반의 기계 번역(SMT: Statistical Machine Translation)에서 현재는 신경망 기반의 기계 번역(NMT: Neural Machine Translation)으로 동작하여 기계 번역의 정확도가 향상되고 있는 것이다[2].

기존의 기계 번역 기술은 SMT 방식을 활용하여 말뭉치 단위로 번역한 결과를 조합하기 때문에 오역이 많았다. 하지만, 현재의 기계 번역 기술은 NMT 방식을 사용하여 전체 문맥을 파악한 후, 문장 내에서의 의미 차이 등을 반영하기 때문에 더 자연스러운 결과물을 보인다[3].

그러나 딥러닝, 머신러닝과 같은 기술들은 수행하기 위한 학습 데이터가 존재해야 한다. 그래서 기계 번역의 정확한 결과물을 내기 위해서는 관련 학습 데이터의 존재 여부가 그 정확도에 영향을 준다[2][4]. 그리고 일반적인 기계 번역은 학습할 문장이 300만 문장 이상은 되어야 제대로 된 학습이 가능하다고 이야기 한다[3]. 그러나 외래어는 전세계로부터 유래한 것으로 지금도 계속 생겨나고 있다. 따라서, 외래어의 기계 번역 정확도를 높이기 위해서는 그에 따른 데이터가 필요한 상황이다.

기계 번역기의 국내 시장 점유율은 Naver papago와 Google 번역기가 약 6대 4정도로 양분해 있다[5].

따라서, 본 연구에서는 Naver papago와 Google 기계 번역기의 외래어 번역 정확도를 언어별로 비교한다.

본 논문의 2장에서는 네이버 파파고와 구글 기계 번역기의 번역 구조 및 원리에 대하여 설명하고, 3장에서는 본 연구의 방법에 대하여 설명한다. 또한, 4장에서는 이를 바탕으로 한 연구 결과에 대하여 기술하며 마지막 5장에서는 결론과 향후 연구 방향에 대하여 다룬다.

2. 기계 번역기별 구조 및 원리

2.1. Naver papago 번역기

Naver papago 번역기는 문장 전체의 정보를 바탕으로 번역을 수행하는 NMT 방식을 사용한다. 기존의 SMT 방식의 번역보다는 맥락에 더 적합한 결과물을 내놓는다[6].

웹기반의 papago 번역기는 (그림 1)과 같이 한국어, 영어, 일본어 등 14개의 언어를 지원하며, 한 번에 5,000글자까지 번역이 가능하다[7].

(그림 1) Naver Papago 번역기

2.2. Google 번역기

웹기반의 Google 번역기는 100여개의 외국어를 지원하며 (그림 2)와 같다[8]. Google 번역기는 국제기구나 기업 등에서 동일한 문서를 A 언어, B 언어, C 언어 등 여러 언어로 만들어 놓은 것을 검색엔진이 검색하여, 해당 문서를 기반으로 번역한다. 이 기술을 말뭉치 기반 기계 번역(CBMT: corpus-based machine translation)이다[9][10].

(그림 2) Google 번역기

3. 연구 방법

3.1. 연구 환경

가. 기계 번역기 및 자동화 프로그램

본 연구에서는 (그림 1)과 (그림 2)와 같이 웹기반의 papago 번역기와 Google 번역기를 활용하였다. 이 번역기들을 활용하여 대량의 외래어 번역을 위해 Python과 Selenium 모듈을 활용하여 프로그램을 작성하였다.

나. 연구 데이터

본 연구에서 사용된 외래어는 국립국어원의 '외래어 표기 용례 정보 통합 파일'을 활용하였다[11]. 이 파일에 정리된 외래어 총 66,530개는 인명, 지명, 일반 용어로 구성되어 있다. 또한, 기원된 언어로는 영어, 일본어, 중국어 등 100여개로 분류되어 있다.

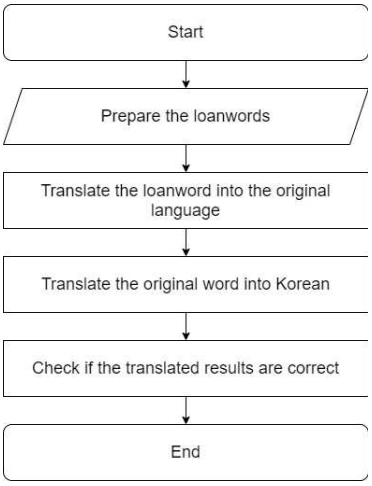
본 연구에서는 수능 제 2외국어 영역에 포함된 언어 중 서양 언어인 독일어, 러시아어, 에스파냐어, 프랑스로부터 기원된 일반 용어를 활용하였다. <표 1>과 같이 4개의 기원된 언어의 외래어 개수는 5,523개이며, 일반 용어는 1,437개이다. 각 언어별 일반용어의 개수는 독일어 576개, 러시아어 121개, 에스파냐어 100개, 프랑스어 640개이다.

<표 1> 언어별 외래어 개수

	일반 용어 개수	전체 외래어 개수
독일어	576	1,230
러시아어	121	1,436
에스파냐어	100	1,159
프랑스어	640	1,698
계	1,437	5,523

3.2. 연구 시나리오

본 연구에서의 실험 과정은 (그림 3)과 같다. 국립국어원에 정리된 외래어 중 독일어, 러시아어, 에스파냐어, 프랑스어로부터 유래한 일반 용어를 준비한다. 이 단어들의 원어와 한국어로 된 외래어를 papago와 Google 번역기에 입력하여 정확하게 번역되는지를 확인하는 것이다. 또한, 마침표, 띄어쓰기 등은 고려하지 않고 단어의 정확도만 확인하였다.



(그림 3) 연구 순서도

4. 연구 결과

(그림 3)의 절차대로 papago와 Google 번역기를 활용하여 한국어인 외래어를 원어로 번역했을 때와 원어를 한국어로 번역했을 때의 정확도를 확인하였다.

papago 번역기를 활용한 외래어 번역 결과의 정확도는 <표 2>와 같다. 한국어인 외래어를 원어로 번역했을 때의 정확도는 평균 10.7%로 원어를 한국어인 외래어로 번역했을 때의 정확도인 평균 12.2%보다 낮았다. 또한, 4개의 언어 중에서 러시아어에서 기원한 외래어의 번역 정확도는 1.0 ~ 6.6%로 나머지 3개 언어의 번역 정확도인 8.9 ~ 15.0%에 비해 낮았다.

<표 2> papago 번역기의 외래어 번역 정확도

	외래어→원어	원어→외래어
독일어	80(13.9%)	74(12.8%)
러시아어	1(1.0%)	8(6.6%)
에스파냐어	16(16.0%)	15(15.0%)
프랑스어	57(8.9%)	78(12.2%)
평균	154(10.7%)	175(12.2%)

Google 번역기를 활용한 외래어 번역 결과의 정확도는 <표 3>과 같다. 한국어인 외래어를 원어로 번역했을 때의 정확도는 평균 11.6%, 원어를 한국어인 외래어로 번역했을 때의 정확도인 평균 9.0%보다 높았다. 또한, 4개의 언어 중에서 러시아어에서 기원한 외래어의 번역 정확도는 0%로 나머지 3개 언어의 번역 정확도인 2.3 ~ 15.8%에 비해 낮았다.

<표 3> Google 번역기의 외래어 번역 정확도

	외래어→원어	원어→외래어
독일어	91(15.8%)	88(15.3%)
러시아어	0(0%)	0(0%)
에스파냐어	15(15.0%)	23(23.0%)
프랑스어	60(9.4%)	19(2.3%)
평균	166(11.6%)	130(9.0%)

5. 결론

우리는 인터넷을 통해 전 세계의 사람들과 자유롭게 의사소통을 할 수 있다. 현재는 스마트폰, 태블릿 등의 모바일 기기 발달과 함께 언제 어디서나 인터넷에 접속할 수 있어 시간적, 공간적 제약이 점점 줄어들고 있다.

전 세계 사람들과 의사소통을 함에 있어 영어 뿐만 아니라 다양한 언어로 의사소통을 하기 위해서는 기계 번역기가 필요하다. 4차 산업 혁명 시대에는 인공지능 기술의 발달로 기계 번역의 정확도도 높아지고 있다. 하지만, 번역의 속도는 사람보다 빠르지만 정확도는 사람만큼 높지 않다. 또한, 은어, 외래어 등의 신조어들이 빠르고 다양하게 생겨나고 있기 때문에 그 정확도를 높이는 기술이 필요한 것이 현실이다.

본 연구에서는 국립국어원에 정리된 외래어 중 독일어, 러시아어, 에스파냐어, 프랑스어에 기원을 두고 있는 외래어들의 인식 정확도에 대한 비교 연구를 진행하였다.

분석 결과, 인식 정확도의 몇 가지 특징을 발견할 수 있었다.

첫 째, papago와 Google 번역기 모두 러시아어에서 기원한 외래어의 번역 정확도가 다른 3개의 언어에 비해 낮았으며, 평균에 못 미쳤다. 둘 째, 두 번역기 모두 번역의 정확도는 평균 9.0 ~ 12.2%로 10% 내외의 수준이었다. 셋 째, 두 번역기 모두 독일어와 에스파냐어에서 기원한 외래어의 번역 정확도가 러시아어와 프랑스어에서 기원한 외래어의 번역 정확도 보다 상대적으로 높았다.

추후 연구에서는 기원 언어별 정확도의 차이가 발생하는 원인을 확인하기 위하여 특정 데이터와의 상관관계를 분석할 수 있을 것이다. 또한, 정확도가 낮은 단어를 분류하여 특정 언어나 특정 단어들의 외래어 번역 정확도를 높일 수 있는 방법을 연구할 것이다.

참고문헌

[1] https://ko.wikipedia.org/wiki/외래어

[2] Koo, M. W., Park, S. G., “Implementation of Word Rejection for a Speech Recognition System Based on HMM”, Proceedings of the Korean Institute of Communication Sciences Conference, Vol. 15, No. 1, pp. 94~97, 199

[3] Byun, Y. H., “Is it possible to use machine translation Indonesian to Korean?”, The Journal of Asian Studies, Vol. 22, No. 4, pp. 51~76, 2019

[4] Park, O. S., “Error Analysis According to the Typological Characteristics of Source Text in Korean-English Machine Translation”, The Journal of Society for Mumanities Studies in

East Asia, pp. 155~183, Vol. 41, 2017
[5] https://www.sedaily.com/NewsVIew/1Z5HAK5IPM
[6] https://www.ncloud.com/product/aiService/papagoNmt
[7] https://papago.naver.com/
[8] https://translate.google.com/
[9] D. Zhou and Y. Wang, “Corpus-based Machine Translation: Its Current Development and Perspectives”, International Forum of Teaching and Studies, Vol. 11, No. 1-2, pp. 90 ~ 95, 2015
[10] Y. Wu, M. Schuster, Z. Chen, Q. Le and M. Norouzi, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”, arXiv:1609.08144, 2016
[11] https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=208&etc_seq=674