http://dx.doi.org/10.15749/jts.2021.22.1.008

번역학연구
2021 봄호 제22권1호

기계학습 알고리즘을 활용한 문학번역에서의 기계 번역과 인간 번역 결과물 분류 연구*

이 창 수
(한국외국어대학교)

1. 서론

2016년은 구글 번역이 신경망기계번역(NMT)을 채택하면서 번역 역사에 획기적 변혁이 일어난 한 해였다. 구글신경망번역기를 개발한 연구팀은 이전의 통계기반번역에서 NMT로 옮겨 오면서 기계번역의 품질이 이중 언어 사용자의 번역 품질에 근접했다고 주장하였다(Wu et al. 2016). 마이크로소프사의 신경망 기계번역기를 영-중 번역에 적용하여 품질을 평가한 하산 외(Hassan et al. 2018)도 기계번역이 전문번역가에 버금가는 번역 결과물을 내고 있다고 주장하였다. 일부 반론은 있지만 NMT가 품질과 생산성 측면에서 기계번역의 새로운

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2018S1A5A2A01030227)
본 연구는 2021년도 한국외국어대학교 교내연구비 지원을 받아 작성되었음.

지평을 열었다는 평가가 지배적이다(Castilho et al. 2019)).

기계번역의 품질이 높아지면서 한편으로 기계번역을 문학번역에 적용할 수 있는 가능성을 타진하는 연구도 등장하였다. NMT 이전에도 기계번역을 문학번역에 적용한 연구가 드물게 있기는 했지만 스페인어-카탈란어 같이 매우 상이한 언어 사이에 적용해본 결과 텍스트의 응집성, 구문 구조, 비유적 언어 번역 등에서 많은 오류가 발생하여 기계번역 자체의 품질이 크게 높아져야 한다는 지적이 있었다(Toral and Way 2015). NMT 등장 이후 기계번역을 문학작품 번역에 적용해본 연구에서는 기존 통계기반 기계번역에 비하여 품질이 크게 향상된 것으로 나타났다. 상기 저자들이 2018년도에 영어-카탈란어를 대상으로 수행한 연구에서는 NMT 번역이 기존 구단위 기계번역(PBMT)에 비하여 BLEU 자동 평가에서는 11퍼센트 향상된 결과를 기록하였고, 전문가 평가에서는 17~34퍼센트의 텍스트에서 번역 품질이 인간 번역 수준에 버금가는 것으로 평가되었다(Toral and Way 2018). 또한 토랄, 위에링 외(Toral, Wieling et al. 2018)의 연구에서도 PBMT에 비하여 NMT의 문학번역 품질 향상이 포스트-에디팅 생산성 증가에 기여한다는 결과를 내놓았다.

문학번역에서 기계번역과 인간번역을 비교한 이런 연구들은 BLEU 같은 자동평가나 인간평가 방식을 통해 기계번역 결과물이 인간번역가의 번역물에 근접한 정도를 평가한다. 그런데 BLEU는 기본적으로 기계번역과 인간번역가의 결과물을 일대일로 비교하여 어휘 패턴이 일치한 비율을 계산하는 방식(Papineni et al. 2002)으로 평가 결과를 일반화시키기 어렵다. 또한 인간평가는 토랄, 카스틸호 외(Antonio, Castilho et al. 2018)가 지적하였듯이 평가자에 따라 평가 결과가 큰 오차를 보이기 때문에 객관성 면에서 역시 일반화시키기 어려운 점이 있다. 본 연구에서는 전산언어학 관점에서 문학번역에서 기계번역과 인간번역 결과물의 차이를 문서분류(text classification)의 문제로 설정하고 기계학습 알고리즘이 두 번역 모드의 번역 샘플들을 얼마나 정확히 분류해내는지를 중심으로 연구를 진행하였다. 어휘 사용 패턴을 학습한 기계학습 알고리즘이 두 번역 모드의 결과물을 집단 별로 명확히 분류한다면 두 번역 모드 사이에는 최소한 어휘 사용 면에서 분명한 차이가 존재한다고 할 수 있다. 이런 관점에서 본 연구는 28편의 한국 소설에 대한 인간번역과 3종의 기계번역기의 영어 번역 결과물을 분석 데이터로 사용하여 다음과 같은 연구 질문에 대한 답을 구하는

것을 목적으로 한다. (1) 컴퓨터 알고리즘이 문학번역에서 인간번역과 기계번역 샘플을 명확히 구분하는가? (2) 컴퓨터 알고리즘이 3종의 기계번역기의 번역 샘플을 구분하는가? (3) 딥러닝을 통해 기계번역기들이 계속 진화한다는 주장을 고려할 때 위와 같은 양상은 1년이란 시차를 두고 어떻게 변화하였는가?

2. 학문적 배경

2.1. 기계 번역의 문학번역 적용 선행 연구

앞서 NMT를 문학번역에 적용하였을 때 긍정적 결과를 도출한 국외 연구 사례를 소개하였는데, 반대로 부정적인 견해를 내놓은 연구도 있다. 가령, 타이발코스키-쉴로브(Taivalkoski-Shilov 2018)는 문장이나 구 단위의 기계번역 방식은 문학 작품의 서사 구조를 왜곡시킬 위험이 있으며, 포스트 에디팅을 거치더라도 기술 번역에 비하여 에디터에 대한 인지적 부담이 훨씬 크기 때문에 이는 품질에 부정적 영향을 끼친다는 연구 결과를 내놓았다. 동 저자는 기계번역은 문학작품의 특징인 다양한 목소리를 구별하여 번역하는데 어려움이 있고, 다양한 작가들의 문체를 동질화시킬 수 있는 위험이 있다고 지적하였다. 무어켄스 외(Moorkens et al. 2018)의 연구에서는 경험이 많은 문학번역가 일수록 기계번역 결과물을 활용한 번역 품질에 만족하지 못하여 이를 포스트 에디팅하는 것보다는 처음부터 독자적으로 번역하는 것을 선호하는 것으로 나타났다.

국내 관련 연구로 눈을 돌리면 기계번역을 문학번역에 적용하는 것에 대하여 아직은 부정적인 견해가 압도적이다. 2017년 한국문학번역원이 문학번역가들을 대상으로 한 설문조사에서 일부 위기감을 표명한 번역가도 있었지만 조사 대상 전원이 아직 독자들이 읽을 만한 번역을 기대하기는 시기상조라는데 의견일치를 보였다(정상혁 2017). 번역연구자들의 의견도 대동소이하다. 이준호(2019)는 4종의 한국 소설을 2개의 기계번역기로 번역한 후 인간번역가의 결과물과 비교하였는데 기계번역기는 번역이 누락된 경우가 많고, 문맥을 반영하거나, 긴 문장을 처리하거나, 처음 접한 단어를 번역하는데 어려움을 갖고 있는 것으로 분석하였다. 마승혜(2018)는 한강의 『채식주의자』를 영어로 번역한 데보라 스미스의 영역본과 구글 기계번역본을 비교하였는데 기계번역은 문맥 해석 능력, 그에 따른 어휘 및 표현 선택 능력, 원문 재현 및 창의력 등에서 인간번역에 한참 못 미친다고 평가하였다. 전혜진(2019)은 똘스또이의 『유년시절』을 대상으로 기계번역과 인간번역 결과물을 언어학과 문화적 측면에서 비교하였다. 그 결과 기계번역은 어휘, 문법, 화용, 문체 등 모든 면에서 여러 문제를 앉고 있는 것으로 분석되었다. 이를 바탕으로 저자는 고도의 언어수행 능력, 텍스트 분석 능력, 맥락, 상황 이해 능력, 전략적 선택 능력, 소통적 번역 능력, 창의적 번역 능력과 감성을 요구하는 문학번역에서 기계번역은 인간번역에 위협이 될 수 없다고 주장하였다.

2.2. 기계학습 알고리즘을 활용한 문서분류 및 저자분류

문서분류(text classification)란 컴퓨터 알고리즘을 통해 장르, 저자, 내용 등 다양한 기준으로 문서를 분류하는 작업으로 텍스트 마이닝의 핵심 분야이다. 이런 문서분류 기술은 수많은 디지털 문서가 쏟아져 나오는 오늘날 수작업으로 하기 힘든 분류작업을 자동화하는데 사용되고 있다. 가령, 뉴스 포털에서 여러 언론사에서 들어오는 기사를 주제별로 분류하거나, 디지털 라이브러리에서 다량의 문서를 장르 별로 분류하거나, 이메일에서 스팸 메일을 걸러내는 등 다양한 문서분류 상황에 적용된다(Aggarwal and Zhai 2012: 164-165)).

문서분류의 한 분야로 저자분류(author attribution)가 있는데 이 경우 목적은 문서를 저자별로 분류하는 것으로, 자연언어처리(NLP)와 밀접한 연관이 있으며 이 주제에 대한 연구는 전산문체학(stylometry)이란 분야의 탄생으로 이어졌다(Bokka et al. 2019: 152). 저자분류의 일반적 경우는 여러 명의 저자들의 문서를 사용하여 컴퓨터 알고리즘에 저자들의 문체를 학습시킨 후, 이렇게 학습된 모델을 사용하여 저자 미상의 문서를 식별하는 것이다. 여기서 저자는 저자 집단일 수도 있으며 이 경우는 저자 집단 별로 문서를 분류 식별하게 된다. 현 연구에서는 인간번역가들이 번역한 문서를 하나로 묶어 단일 번역가가 번역한 문서로 취급하고, 구글, 네이버, 마이크로소프트사의 NMT를 각각 단일 번역가로 취급하여 4명의 번역가 간에 번역결과물이 어떻게 분류되는지를 실험할 것이다. 이 경우 3종의 기계번역기에 대비하여 인간번역가는 다수의 번역가들로 구

성되었는데 이들이 마치 단일 번역가처럼 하나의 집단으로 분류된다면 기계번역과 인간번역 간에는 어휘 사용 문체에서 분명한 차이가 있다는 결론을 내릴 수 있다.

저자판별에 사용되는 언어 자질은 다양하지만 본 연구에서는 가장 많이 사용되는 어휘별 빈도를 사용할 것이다. 이를 최빈도어휘(MFW)라고 부르는데 모든 종류의 영어 텍스트에서 가장 높은 빈도수를 차지하는 어휘들은 대명사, 전치사, 조동사 같은 기능어로 이들은 문서 내용에 독립적이면서도 구문의 뼈대를 형성하기 때문에 개인 문체 형성에 중요한 역할을 한다(Juola 2006: 265).

이 같은 분석 요소를 사용하여 실제 분류 작업을 시행하는 컴퓨터 알고리즘은 분류기(classifier)라고 한다. 문서분류에는 매우 다양한 분류기가 사용되는데 크게 자율학습(unsupervised learning)과 지도학습(supervised learning)으로 나뉜다. 자율학습이란 주어진 텍스트 데이터 내에서 자율적 패턴 인식을 통해 주어진 문서를 집단으로 묶어주는 방식으로 군집분석(clustering)이라고도 한다. 이에 반하여 지도학습은 주어진 텍스트 데이터를 훈련용과 실험용으로 나눈 후 훈련용을 사용하여 컴퓨터 알고리즘을 학습시켜 모델을 구축한 후, 해당 모델을 실험용에 적용하여 텍스트 소속 집단을 얼마나 정확히 예측하는지를 측정하는 방식으로 기계학습이라고도 불린다(Alloghani et al. 2019: 4). 본 연구에서는 상대적으로 최신 기술인 지도학습 방식의 기계학습 분류기를 사용할 것이다. 기계학습 알고리즘도 의사결정 트리(Decision Tree), 선형 판별 분석(Linear Discriminant Analysis: LDA), 로지스틱 회귀분석(Logistic Regression), 서포트 벡타 머신(SVM), 신경망(Neural Networks), 랜덤 포레스트(Random Forest: RF) 등 종류가 매우 많다. 각각 다른 알고리즘을 사용하기 때문에 정확도에 차이가 있고 이런 차이는 분류하고자 하는 문서 종류에 영향을 받기도 한다. 본 연구에서는 교차 분석을 위해서 LDA와 RF를 사용하였다. 이들 알고리즘에 대한 추가 언급은 다음 절에서 논하기로 한다.

3. 분석 데이터 및 방법

본 연구에서 분석한 데이터는 28편의 한국어 장·단편 소설을 인간번역가가 영어로 번역한 29 번역본(원 소설 한 편은 2명이 번역하여 29편임)과 구글(translate.google.com), 마이크로소프트(www.bing.com/translator) 및 네이버(papago.naver.com) 온라인 번역기를 사용하여 1년 간격을 두고 2차례에 걸쳐 수집한 번역기 당 28개의 번역본을 합쳐 총 197개의 번역 샘플로 구성되었다. 데이터에서 인간번역본은 Human, 구글은 Google, 마이크로소프트는 Bing, 네이버는 Papago로 표기하였다. 상기 4 종의 번역문은 원문이 똑같기 때문에 원문 내용이 연구에 미칠 가능성이 배제되어 4 종간의 언어적 차이는 번역가들의 선택의 차이로 귀결된다.

인간번역가는 실제로는 여러 명의 다른 번역가들로 구성되어 서로 문체가 다르겠지만 단일 번역가의 작품처럼 취급하였다. 이와 같은 데이터 구성에 기초하여 인간번역가 결과물이 기계번역 결과물에 맞서 별도의 군집으로 분류되는지에 관한 첫 번째 연구 질문과 상기 3종의 기계번역기의 결과물 또한 별도로 분류되는지에 관한 두 번째 연구 질문에 대한 답을 도출하였다.

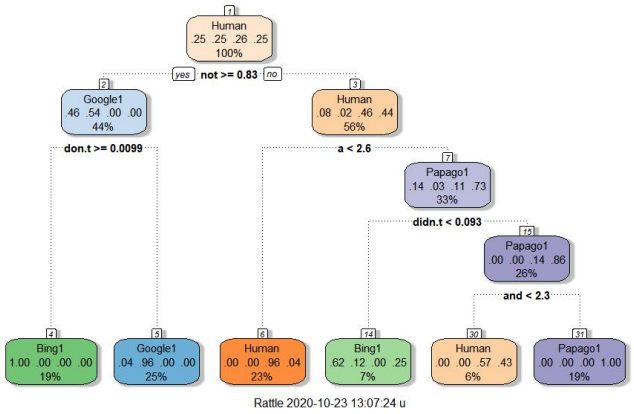
인간번역본은 1980년 중반대에서 2000년대까지 출판연도가 다양하다. 기계번역본은 2019년 2월과 1년 후인 2020년 2월, 두 차례에 걸쳐 수집되었다. 이는 딥러닝을 통해 계속 진화한다는 NMT가 1년 동안에 본 연구에서 분석한 어휘사용 패턴에서 어떤 변화를 보였는지를 확인하기 위한 용도로 본 연구의 세 번째 연구 질문과 관련 있다. 분석데이터에서 기계번역기의 1차 번역본은 Google1, Bing1, Papago1로 표기하였고 2차 번역본은 Google3, Bing3, Papago3로 표기하였다.

분석데이터 총 규모는 2,550,408단어로 가장 작은 번역본은 3,691 단어, 가장 큰 번역본은 62,056단어였다. 번역본 간에 크기에서 차이가 나는 것은 기본적으로 원문의 크기가 다른데서 비롯된다. 원문 크기의 차이는 4종의 번역문에 동일하게 반영되기 때문에 연구에 미치는 영향은 미비한 것으로 판단하였다. 여기서 중요한 것은 샘플 사이즈가 작은 번역본이 문서분류 실험을 하는데 적합하냐는 문제인데, 최근에는 트위터 글처럼 크기가 매우 작은 텍스트를 가지고도 문서분류 연구를 하고 있으며(cf. Harjule et al. 2020), 에더(Eder 2013)는

영문 저자판별에서 3,000단어를 최저 크기로 설정하였는데 본 연구 분석데이터는 이 기준을 모두 충족하여 샘플 크기 면에서 큰 문제는 없는 것으로 판단하였다.

앞서 언급하였듯이 본 연구의 문서분류 실험에는 교차분석을 통한 분석 신뢰도 확보를 위하여 선형판별분석(LDA)과 래덤포리스트(RF) 등 두 종의 기계학습 알고리즘을 사용하였다. RF는 기본적으로 트리기반 분류 알고리즘으로 문서분류에서 처리 속도가 빠르고 정확하다는 평가를 받는다(Kowsari et al. 2018: 21). 의사결정트리 알고리즘은 톱다운 방식으로 위에서부터 차례대로 문서를 두 집단으로 분류하여 내려간다. 본 연구에서는 100개의 최빈도 어휘를 분석 어휘로 사용하였는데 의사결정트리는 그림 1에서 보듯이 분류기여도가 가장 높은 어휘부터 사용하여 한 단계씩 내려가며 각 노드에서 두 개의 가지로 나뉘며 내려간다. 일반적 의사결정트리는 이런 과정을 한번만 수행하는데 반하여 RF는 각 노드마다 여러 개의 분류 어휘를 랜덤하게 사용하거나 조금씩 다른 데이터를 사용하여 여러 개의 나무를 형성한 후 이렇게 랜덤하게 형성된 나무들을 통합하여 평균을 내는 앙상블 방식을 취하기 때문에 신뢰도 및 정확도가 높다는 것이 장점이다(Aggarwal 142-147).

그림 1 R의 rpart 패키지를 사용한 의사결정트리 분석 예



LDA는 로지스틱 회귀분석법과 비슷하지만 후자는 분석 집단(또는 클래스)이 두 개인 데이터에 한정된데 반하여 LDA는 본 연구 데이터처럼 다 집단 데이터에 적용되는 특징이 있다. 이 같은 특징 외에도 분석 결과에 대한 신뢰도가 높기 때문에 다집단 분류 문제에서 벤치마킹 분석법으로 사용되는 기계학습 알고리즘이다(Brownless 2016: 61-64). LDA는 일반 회귀분석처럼 각 집단에 대한 평균값과 분산값을 계산한 후 전체 집단에 대한 평균값을 계산해서 텍스트 샘플이 각 집단에 속할 확률을 예측한다. 이 과정에서 LDA는 집단 내 분산값은 최소화하고 집단 간 분산값은 최대화하여 집단 내 샘플간 거리는 최대한 좁히고 집단 간 거리를 최대한 늘어나도록 한다. 따라서 LDA와 비슷한 자율학습 차원축소법인 주성분분석(PCA)에 비하여 집단 간 분류를 더 명확히 하는 특징이 있다.

본 연구에서는 R통계언어의 randomForest와 lda 패키지를 사용하여 RF와 LDA 분석을 시행하였다. 197개 번역 샘플을 R에 탑재한 후 일련의 전처리과정을 거쳐 그림 2와 같은 분석데이터를 추출하였다. 그림 2에서 행은 번역 샘플을 나타내는데 그림에는 197개 중 79~89번 사이의 샘플만 예시되어 있다. 열 방향으로는 번역 샘플을 인간과 기계로 분류한 TR 변수, 인간과 각 기계번역기로 구분한 Corpus 변수가 있고 그 이후로는 최빈도어휘 100개가 나열되어 있는데 그림에는 총 빈도 상위 10개만 표시하였다. 분석에서는 197개 번역 샘플을 랜덤하게 반으로 나눈 후에 하나는 알고리즘 훈련용으로 다른 하나는 테스트용으로 사용하였다. 다음 절의 분석 결과에서는 테스트용 데이터를 사용한 분류 예측 결과를 제시하고 그 의미를 논하도록 한다.

그림 2 분석데이터

	TR	Corpus	the	a	and	to	of	i	was	in.
79	HT	Human	4.695687	2.693806	2.196695	2.633347	1.961575	3.7417708	1.4711810	1.330109
80	HT	Human	4.296624	2.804939	2.255371	3.061880	1.862822	4.2538006	1.9341946	1.327528
81	HT	Human	4.557004	2.126602	2.645824	3.142952	1.745471	0.1546620	2.4359258	1.325674
82	HT	Human	5.186636	1.988013	2.522106	3.079936	1.222479	0.5281586	1.5963444	1.572607
83	HT	Human	5.598160	1.904397	2.466769	2.914110	1.942740	4.0452454	1.9299591	1.514571
84	HT	Human	5.269530	2.600173	1.620063	2.715480	2.208129	2.5944076	2.0639954	2.450274
85	HT	Human	3.603869	2.283866	2.100895	3.032085	1.708815	3.2771352	1.6058943	1.288963
86	MT	Papago1	9.562269	3.220660	2.452887	2.004188	2.233523	1.0170506	1.5355469	1.774853
87	MT	Papago1	7.710896	4.130053	3.075571	2.196837	2.108963	0.7688928	1.8673111	1.691564
88	MT	Papago1	7.623412	3.179893	2.867458	2.221759	3.096577	0.8887037	2.2009304	1.763521
89	MT	Papago1	6.449266	3.783569	2.328350	2.619394	2.599550	1.4882921	0.6879217	1.865326

4. 분석 결과

4.1. RF 분석 결과

먼저 RF를 사용한 분석 결과부터 논하도록 한다. 분석은 2019년도와 2020년도 데이터를 나눠서 2차례에 걸쳐 시행하였다. 먼저 그림 3과 그림 4에 나와 있는 2019년도 데이터에 대하여 시행한 예측 분류 결과를 살펴보자. 그림 3은 샘플 숫자 별로 예측된 결과를 보여주며 그림 4는 같은 수치를 백분율로 나타낸 것이다. 행은 실제 샘플 소속 집단을 열은 RF가 예측한 결과를 보여준다. 먼저 Bing1부터 보면 총 13개의 테스트 샘플 중 한 개를 제외한 12가 정확히 예측되었다. 백분율로 보면 예측 정확도가 92퍼센트이다. Google1의 경우는 13 샘플 모두 100퍼센트 정확하게 예측되었다. 인간번역 샘플인 12개도 100퍼센트의 예측 정확도를 기록했다. 마지막으로 Papago1은 예측 정확도가 가장 떨어졌는데 19개 중 15를 맞게 분류하여 79프로의 정확도를 보였지만 여전히 높은 수준이다. 이와 같은 결과를 바탕으로 다음과 같은 결론을 도출할 수 있다.

그림 3 2019년 RF 예측 결과 (샘플 수)

	predicted			
observed	Bing1	Google1	Human	Papago1
Bing1	12	1	0	0
Google1	0	13	0	0
Human	0	0	12	0
Papago1	0	0	4	15

그림 4 2019년 RF 예측 결과 (백분율)

	predicted			
observed	Bing1	Google1	Human	Papago1
Bing1	0.92307692	0.07692308	0.00000000	0.00000000
Google1	0.00000000	1.00000000	0.00000000	0.00000000
Human	0.00000000	0.00000000	1.00000000	0.00000000
Papago1	0.00000000	0.00000000	0.21052632	0.78947368

첫째, 인간번역 샘플은 기계번역 샘플과 명확히 구별되고 있다. 이는 인간번역 샘플이 다른 집단의 샘플로 잘못 예측된 케이스가 한 건도 없기 때문에 인간번역 샘플은 독자적으로 완전히 독립된 집단을 형성한다고 볼 수 있다.

둘째, 기계번역기들도 나름대로 독자적인 집단을 형성하고 있다고 볼 수 있다. Bing1에서 1건이 Google1의 샘플로 예측되어 92퍼센트 예측 정확도를 보였지만

문서분류에서 90퍼센트면 예측 정확도가 매우 높은 수치이다. Papago1의 경우는 예측율이 79퍼센트로 상대적으로 떨어지지만 5건 중에 4건을 정확히 예측했다는 의미이기 때문에 결코 낮은 정확도는 아니다. 이는 컴퓨터 알고리즘이 기계번역 샘플이 어떤 기계번역기의 결과물인지를 상당히 정확하게 분류할 수 있다는 뜻으로 기계번역기 사이에도 어휘사용 패턴에 뚜렷한 차이가 있음을 보여 준다.

셋째, Papago1의 경우 20퍼센트 샘플이 Human으로 분류되어 기계번역기 중 인간번역에 가장 근접한 것으로 평가된다. 특히 다른 기계번역기에서는 번역 샘플이 Human으로 분류된 케이스가 없다는 점에서 상대적 거리로 평가하자면 Papago1이 Human과 거리가 가장 가깝다고 볼 수 있다.

그렇다면 1년 후인 2020년 데이터에선 어떤 결과가 나왔을까? NMT는 계속적인 딥러닝 학습을 통해 진화한다고 설명되는데 과연 1년 후에는 기계번역기들이 Human에 더 가까워졌을까? 이 의문점에 대한 대답을 얻기 위하여 그림 5와 그림 6의 2020년 분석 결과를 살펴보자. 1차와 2차 분석 데이터는 별도로 랜덤하게 나눠 실험하였기 때문에 2차에서 테스트용으로 사용된 샘플 수가 1차와 차이가 난다. 먼저 Bing3을 보면 총 14개 샘플 중 10개가 정확히 예측되어 71퍼센트의 예측 정확도를 보였다. Google3은 1개를 제외한 13개 샘플이 정확히 분류되어 93퍼센트의 예측율을 기록하였다. Human의 경우도 1개를 제외하고 16개가 정확히 예측되어 예측 정확성은 94퍼센트에 달했다. Papago3은 12개 중 5개만 정확히 분류되고 나머지는 Bing3와 Google3 샘플로 잘못 분류되어 예측 정확도는 42퍼센트를 나타냈다. 이를 2019년도 분석 결과와 비교하면 다음과 같은 결론을 도출할 수 있다.

그림 5 2020년 RF 예측 결과 (샘플 수)

	Reference			
Prediction	Bing3	Google3	Human	Papago3
Bing3	10	2	0	2
Google3	1	13	0	0
Human	0	1	16	0
Papago3	4	3	0	5

그림 6 2020년 RF 예측 결과 (백분율)

	predicted			
observed	Bing3	Google3	Human	Papago3
Bing3	0.71428571	0.14285714	0.00000000	0.14285714
Google3	0.07142857	0.92857143	0.00000000	0.00000000
Human	0.00000000	0.05882353	0.94117647	0.00000000
Papago3	0.33333333	0.25000000	0.00000000	0.41666667

첫째, 인간번역 샘플은 여전히 기계번역 샘플과 확실히 구분되는 독자적인 집단을 형성하고 있다. 예측 정확율이 100퍼센트에서 94퍼센트도 낮아졌지만 반대로 예측 오차율이 0.6퍼센트일 정도로 여전히 컴퓨터 알고리즘이 인간번역 샘플을 정확히 구별해내고 있다.

둘째, 기계번역기들 간에는 샘플을 잘못 분류하는 비율이 전반적으로 높아졌다. Bing은 92퍼센트에서 71퍼센트로, Google은 100퍼센트에서 92퍼센트로 낮아졌고, 특히 Papago의 경우는 79퍼센트에서 42퍼센트로 크게 떨어졌다. 여기서 중요한 점은 잘못 분류된 샘플 중 Human으로 분류된 것은 한건도 없고 전부 다른 기계번역기의 샘플로 분류되었다는 점이다. 이는 1년 동안 기계번역기들이 서로 유사한 방향으로 변화했음을 시사한다.

셋째, Papago를 Human에 가장 근접한 기계번역기로 평가했던 2019년 결과가 유지되지 않았다. 대신에 Papago 번역 샘플은 다른 기계번역기 결과물로 잘못 분류된 경우가 늘어나 다른 기계번역기와 간격이 좁아진 상황이 되었다.

위 결과를 요약하면 다음과 같다. 첫째, RF 알고리즘은 인간번역과 기계번역의 결과물을 정확히 구분해내며 2019년에서 2020년 사이에 이런 결과에 유의미한 변화가 없었다. 둘째, 이와 대조적으로 기계번역기들 간에는 2019년도에 명확했던 상호간 변별력이 약화되어 2020년에는 번연결과물이 서로 유사해지는 방향으로 변화하였다.

4.2. LDA 분석 결과

이번에는 LDA 분석 결과를 통해 앞서 RF 분석에서 도출한 결론의 유효성을 교차 검증해보도록 한다. 먼저 그림 7과 그림 8의 2019년도 예측 결과를 살펴보면 Bing1과 Human에서 각각 샘플 한 개가 Google1과 Papago1로 잘못 예측된 것을 제외하면 전체적으로 90퍼센트가 넘는 매우 정확한 예측율을 기록하였다. 이는 그림 3과 그림 4의 RF 분석과 거의 일치하는 결과로 2019년도에는 인간번역과 기계번역 뿐만 아니라 기계번역기들 간에도 어휘 사용 패턴에서 명확한 차이가 있었음을 보여준다. 특히 그림 3과 그림 4의 RF 분석에서 Papago의 일부 샘플이 Human으로 잘못 분류되었던 것과 달리 LDA는 Papago 샘플을 모두 정확하게 식별하였다. 이는 LAD가 예측 정확도가 좀 더 높다는 것을 보

여주며, 동시에 RF의 예측 결과에 따라 다른 기계번역기에 비하여 Papago가 Human에 좀 더 가깝다고 봤던 해석이 유효하지 않음을 뜻한다.

그림 7 2019년 LDA 예측 결과 (샘플 수)

	Reference			
Prediction	Bing1	Google1	Human	Papago1
Bing1	12	1	0	0
Google1	2	11	0	0
Human	0	0	11	1
Papago1	0	0	0	19

그림 8 2019년 LDA 예측 결과 (백분율)

	predicted			
observed	Bing1	Google1	Human	Papago1
Bing1	0.92307692	0.07692308	0.00000000	0.00000000
Google1	0.15384615	0.84615385	0.00000000	0.00000000
Human	0.00000000	0.00000000	0.91666667	0.08333333
Papago1	0.00000000	0.00000000	0.00000000	1.00000000

다음에는 그림 8과 그림 9에 나와 있는 2020년도 데이터에 대한 LDA 예측율을 살펴보자. 2020년도에는 Human은 예측율이 2019년도의 92퍼센트에서 94퍼센트로 높아져 17건의 인간번역 샘플 중 16건이 정확히 분류되었다. 이에 반하여 기계번역기 샘플에 대한 예측 정확도에서는 Google3을 제외하고 Bing3와 Papago3는 각각 92퍼센트에서 71퍼센트, 100퍼센트에서 42퍼센트로 하락했다. 이는 그림 5와 그림 6의 RF 결과와 정확히 일치하는 결과다. 즉, 2019년에서 1년의 시간이 경과한 2020년에도 인간번역과 기계번역은 매우 명확하게 구분되었다. 반대로 기계번역기 간에는 상호 잘못 예측된 비율이 높아져서 점차 동질화되는 방향으로 진화하였다.

그림 9 2020년 LDA 예측 결과 (샘플 수)

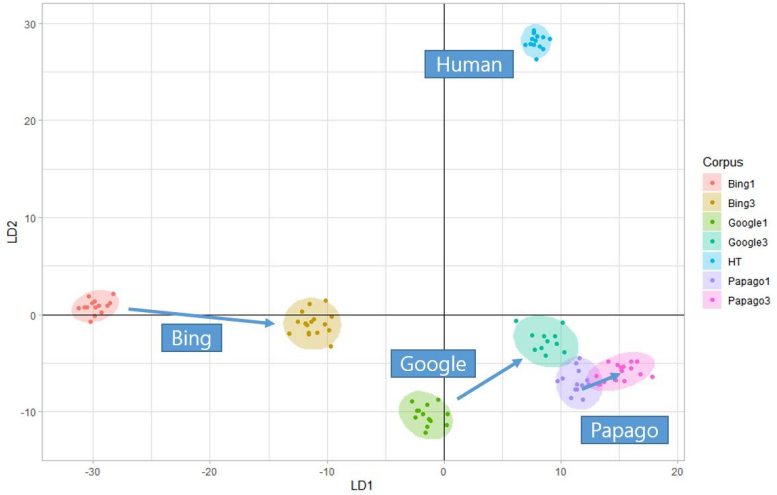
	Reference			
Prediction	Bing3	Google3	Human	Papago3
Bing3	10	2	0	2
Google3	1	13	0	0
Human	0	1	16	0
Papago3	4	3	0	5

그림 10 2020년 LDA 예측 결과 (백분율)

	predicted			
observed	Bing3	Google3	Human	Papago3
Bing3	0.71428571	0.14285714	0.00000000	0.14285714
Google3	0.07142857	0.92857143	0.00000000	0.00000000
Human	0.00000000	0.05882353	0.94117647	0.00000000
Papago3	0.33333333	0.25000000	0.00000000	0.41666667

LDA는 PCA와 유사한 차원축소법이란 특징을 갖고 있어 분류 예측 결과를 그림 11과 같이 2차원 평면 그래프로 나타낼 수 있다. 이 그래프는 2019년과 2020년 데이터를 합쳐 분석한 결과를 나타낸 것으로 양 연도 사이에 어떤 변화가 있었는지를 시각적으로 관찰할 수 있다. 그림 11에서 인간번역 샘플은 HT, 각 기계번역기의 1차와 2차 번역 샘플은 Bing1, Bing3, Google 1, Google3, Papago1, Papago3로 각각 표시되어 있으며 화살표는 각 기계번역기의 1차와 2차 샘플 군집의 이동 방향을 보여준다. 그림 11의 그래프는 앞서 분석한 결과를 좀 더 명확하게 보여준다.

그림 11 LDA 2019, 2020년 통합 분석 결과



첫째, 인간번역과 기계번역은 어휘 사용면에서 명확히 구분된다. 이는 그래프 상에서 LD2축을 따라 수직으로 상호간에 큰 거리를 두고 인간번역은 +영역에 기계번역은 -영역에 포진된 것을 보면 알 수 있다. 이와 같은 거리는 2019년도와 2020년도 사이에 크게 달라지지 않았다. 둘째, 기계번역기 간에도 어느 정도 분류가 명확하지만 시간이 가면서 상호간 차별성이 약화되는 경향을 보인다. 그래프 상에서 기계번역기 샘플들은 LD1 축을 중심으로 좌우로 배열되어

있다. 2019년에서 2020년으로 넘어오면서 기계번역기들은 LD1축을 중심으로 좌에서 우로 이동하였는데 그 결과 1차에 비하여 2차에서 기계번역기간의 거리가 크게 좁혀졌다. Papago는 거리 이동이 적은 반면 Bing과 Google이 Papago 쪽으로 거리를 크게 좁혀 이동한 양상이다. 특히 Bing과 Google은 1차와 2차 번역물이 별도의 군집으로 분류될 만큼 변화의 폭이 크다. 이에 반하여 Papago는 1차와 2차가 부분적으로 겹쳐있어 변화의 폭이 상대적으로 적어 보인다. 이를 통해 2019년에서 2020년 사이에 Bing과 Google이 Papago와 거리를 크게 좁히면서 기계번역기의 번역 결과물들이 서로 유사해지는 방향으로 변화했다고 판단할 수 있다. 이는 앞서 살펴봤던 RF 분석 결과를 뒷받침한다.

5. 결론

본 연구에서는 인간번역사와 3종의 기계번역기의 문학작품 번역 결과물을 분석데이터로, 최빈도어휘 100개를 분석 변수로 사용하여 컴퓨터 알고리즘이 번역 샘플의 실제 생산자를 얼마나 정확히 예측할 수 있는지를 실험하였다. RF와 LDA 알고리즘을 사용한 교차 검증 분석에서 컴퓨터는 2019년과 2020년 데이터에서 모두 인간번역과 기계번역을 명확히 구분해냈다. 이는 어휘사용 패턴에서 기계번역과 인간번역은 여전히 매우 다른 영역으로 존재한다는 것을 보여준다. 이에 반하여 기계번역기의 경우는 2019년도에 상호 명확하게 분류되던 것이 2020년에 들어 상호간에 샘플이 잘못 분류되는 경우가 늘어났다. 즉, 1년의 시간 동안 기계번역이 진화한 방향은 인간번역과의 거리를 좁히는 방향이 아니라 상호 간에 거리가 좁아지는 방향이었다. 이는 최소한 어휘 사용 면에서는 인간번역과 기계번역은 여전히 매우 상이한 선택을 하고 있는 반면, 기계번역기들은 서로 동질화되는 방향으로 발전하고 있음을 보여준다.

여기서 한 가지 주목할 점은 인간번역 샘플은 여러 명의 다른 번역가들의 결과물로 그 자체로 다양한 번역 문체를 내포하고 있다는 것이다. 그럼에서 불구하고 기계학습 알고리즘이 기계번역 샘플과 대비하여 인간번역 샘플을 매우 정확하게 구분해 냈고 그런 간극이 1년의 시간을 두고 크게 달라지지 않았다는 본 연구 결과는 두 번역 모드 간에 차이가 구조적일 수 있는 점을 시사한다.

기계번역기는 학습과정에서 구축한 통계적 모델에 기초하여 어휘 선택을 하는 반면 인간번역사는 언어적 지식뿐만 아니라 텍스트에 대한 보다 광범위한 메타데이터와 경험적 지식에 기초하여 직관적 선택을 하는 차이가 있다. 이런 차이를 감안하여 NMT가 인간처럼 참조데이터를 사용할 수 있도록 하는 보완책이 제시되고 있지만(cf. Fu et al. 2019) 그것은 인간번역사의 번역 의사 결정 과정의 일부분을 모방하는 것일 뿐 근본적인 차이를 없애는 것은 아니다. 두 번역 모드 간에 이 같은 의사결정 과정의 구조적 차이가 본 연구에서 인간번역과 기계번역 결과물이 완전하게 분리되는 결과를 가져왔을 수 있다. 물론 이것은 하나의 가설로 앞으로 다양한 추가 연구를 통해 검증되어야 할 것이지만 그 같은 연구는 오류 분석에 머물고 있는 문학번역에서 기계번역과 인간번역 간의 비교연구를 보다 다양한 차원으로 확대시키는데 기여할 것이다.

참고문헌

마승혜 (2018)「문학작품 기계번역의 한계에 대한 상세 고찰」.『통번역학연구』, 22(3): 65-88.

이준호 (2019)「문학번역 적용을 위한 기계번역의 현주소」,『통번역학연구』23(1): 143-167.

전혜진 (2019) AI 시대, 문학번역에서 기계번역과 인간번역 비교분석 연구 －똘스또이의『유년시절』번역 분석을 중심으로」,『노어노문학』31(1): 111-154.

정상혁 (2017)「진화하는 번역기 ... 사라지는 번역가? ,『조선일보』. Available at https://www.chosun.com/site/data/html_dir/2017/01/18/20170118 00020.html.

Aggarwal, Charu C. (2018) *Machine Learning for Text*, Cham, Switzerland: Springer.

Aggarwal, Charu C and Chengxiang Zhai (2012) 'A Survey of Text Classification Algorithms', in Charu C. Aggarwal and ChengXiang Zhai (eds) *Mining Text Data*, Berlin/Heidelberg: Springer, 163-222.

Alloghani, Mohamed, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain and Ahmed J. Aljaaf (2019) 'A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science' in Michael W. Berry, Azlinah Mohamed and Bee Wah Yap (eds) *Supervised and Unsupervised Learning for Data Science*, Cham, Switzerland: Springer, 3-21.

Bokka, Karthiek Reddy, Shubhangi Hora, Tanuj Jain and Monicah Wambugu (2019) *Solve Your Natural Language Processing Problems with Smart Deep Neural Networks*, UK: Packt Publishing.

Brownlee, Jason (2016) *Master Machine Learning Algorithms*, Melbourne, Australia: Brownlee.

Castilho, Sheila, Federico Gaspari, Joss Moorkens, Maja Popovi and Antonio Toral (2019) 'Editors' Foreword to The Special Issue on Human Factors in Neural Machine Translation', *Machine Translation* 33: 1-7.

Eder, Maciej (2013) 'Does Size Matter? Authorship Attribution, Small Samples, Big Problem', *Literary and Linguistic Computing* 30(2): 167-182.

Fu, Han, Chenghao Liu and Jianling Sun (2019) *Reference Network for Neural Machine Translation*, arXiv:1908.09920 [cs.CL]

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann … and Ming Zhou (2018) *Achieving Human Parity on Automatic Chinese to English News Translation*, arXiv:1803.05567 [cs.CL].

Harjule, Priyanka, Astha Gurjar, Harshita Seth and Priya Thakur (2020) 'Text Classification on Twitter Data', in *Proceedings of 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, Jaipur, India, 160-164.

Juola, Patrick (2006) 'Authorship attribution,' *Foundations and Trends in Information Retrieval* 1(3): 233-334.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes and Donald E. Brown (2019) 'Text Classification Algorithms: A Survey', *Information* 10(4): 1-68.

Moorkens, Joss, Antonio Toral, Sheila Castilho and Andy Way (2018) 'Translators' Perceptions of Literary Post-Editing Using Statistical and Neural Machine Translation', *Translation Spaces* 7(2): 240-262.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu (2002) 'BLEU: a Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 311-318.

Taivalkoski-Shilov, Kristiina (2018) 'Ethical Issues Regarding Machine(-assisted) Translation of Literary Texts', *Perspectives* 27(5): 689-703.

Toral, Antonio and Andy Way (2015) 'Machine-assisted Translation of Literary Text: A Case Study', *Translation Spaces* 4: 241-268.

Toral, Antonio and Andy Way (2018) *What Level of Quality can Neural Machine Translation Attain on Literary Text?* arXiv:1801.04962v1 [cs.CL].

Toral, Antonio, Martijn Wieling and Andy Way (2018) 'Post-editing Effort of a Novel With Statistical and Neural Machine Translation', *Front. Digit. Humanit.* Available at https://doi.org/10.3389/fdigh.2018.00009.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. (2018) 'Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation', in *Proceedings of the Third Conference on Machine Translation (WMT) (Volume 1: Research Papers)*, Belgium, Brussels, 113-123.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey … and Jeffrey Dean (2016) *Google's Neural Machine Translation System: Bridgi*ng the Gap between Human and Machine Translation. arXiv:1609.08144 [cs.CL].

[Abstract]

Machine Learning Classification of Literary Translation Samples by Human and Machine Translators

Lee, Chang-soo
(Hankuk University of Foreign Studies)

The current paper reports the results of a text classification experiment on literary translation samples by human and machine translators. The original data consists of the English translations of 28 short and long Korean novels by a set of human translators and 3 Web-based neural machine translators – Google Translate (Google), Bing (Microsoft), and Papago (Naver). Machine translation samples were collected twice in February 2019 and February 2020. One hundred most frequent words were extracted from the data and subjected to supervised classification by two machine learning algorithms – random forest (RF) and linear discriminant analysis (LDA) - for cross-reference tests. The most important findings are as follows. First, Both RF and LDA classified human and machine translation samples from both 2019 and 2020 with high accuracy, with prediction accuracy rates topping 90 percent. This indicated a clear distinction in word use patterns between human and machine translators, which did not change much over the 1-year period. Second, in both RF and LDA tests, most of the 2019 machine translation samples were accurately classified according to their translators with prediction accuracy rates ranging between 78 and 100 percent. Classification accuracy, however, fell visibly for Bing and Papago in 2020, with Papago plunging from 100 and 80 percent to 41 percent. This meant that over the 1-year period the three machine translators moved in closer toward each other, suggesting a trend toward homogeneity in word use patterns over time.

▸ Key Words: machine translation, literary translation, machine learning text classification, random forest, linear discriminant analysis
▸ 주제어: 기계 번역, 문학 번역, 기계 학습 문서 분류, 랜덤 포레스트, 선형 판별 분석

이창수
한국외국어대학교 통역번역대학원 교수
soolee@hanmail.net
관심분야: 코퍼스번역연구, 전산문체학, 디지털 인문학, 비평담화분석, 체계기능언어학

논문투고일: 2021년 2월 7일
심사완료일: 2021년 2월 28일
게재확정일: 2021년 3월 4일