http://dx.doi.org/10.15749/jts.2019.20.2.005

번역학연구
2019 여름호 제20권2호

문학번역에서의 기계번역과 인간번역 문체에 대한 전산문체학적 비교 연구*

이 창 수
(한국외대)

1. 들어가는 말

70년대 규칙기반기계번역(rule-based machine transaltion: RBMT)이 기계번역 연구의 물꼬를 튼 이후 IBM 모델이라고 불리는 단어 차원의 통계기계번역(statistical machine translation: SMT)과 구기반기계번역(phrase-based machine translation: PBMT)을 거치면서 기계번역이 발전하였다. 최근에는 인공지능을 활용한 신경망기계번역(neural machine translation: NMT)의 등장으로 기계번역의 품질이 크게 높아졌다(cf. 가르그와 아가왈(Garg & Agarwal 2018)). 그렇지만 신경망기계번역의 품질에 관해서는 긍정적인 평가와 부정적인 평가가 엇갈

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2018S1A5A2A01030227)
본 연구는 2019년도 한국외국어대학교 교내연구비 지원을 받아 작성되었음.

리고 있다. 가령, 신경망기계번역에 대한 자동평가와 인간평가 연구를 수행한 카스틸호 외(Castilho et al. 2017)는 신경망기계번역이 과거 모델에 비하여 진일보한 것은 맞지만 그 능력을 과대평가하지 말아야 한다고 하였다.

기계번역의 품질이 눈에 띄게 개선되면서 최근에는 기계번역을 문학번역에 적용할 수 있을 것인가 하는 문제가 관심을 끌고 있다. 이 문제에 관해서도 기계번역의 품질이 크게 개선된 연구 결과를 바탕으로 긍정적 평가를 내놓는 측과(cf. 토랄과 웨이(Toral & Way 2018); 토랄 외(Toral et al. 2018)) 문학번역 적용의 한계를 지적하는 목소리(cf. 타이발코스키-실로프(Taivalkoski-Shilov) 2018; 무어켄스 외(Moorkens et al. 2018)가 공존하고 있다.

국내 연구의 경우 최근 들어 기계번역에 관한 연구가 크게 활성화되고 있는데(cf. 김순미 외 2019; 이성화와 김세현 2018; 서보현과 김순영 2018), 문학번역을 소재로한 연구로 국한할 경우 대부분의 연구는 기계번역과 인간번역 결과물을 일대일로 비교하여 기계번역의 한계와 오류를 지적하는 내용이다(cf. 마승혜 2018; 이준호 2019). 그러나 과거에 비하여 번역품질이 아무리 발전하였다고 하더라도 현 시점에서 기계번역을 문학번역에 바로 적용하는 것은 여러 면에서 무리가 있으며, 그나마 어느 정도 가독성을 갖춘 번역물을 생산하려면 번역 후 에디팅(post-editing)과 교정(revision)을 거쳐야 하는 상황이다(cf. 베사시어 & 슈와츠(Besacier & Schwartz 2015)).

이 같은 점을 고려하여 본 연구에서는 기계번역을 문학번역에 적용했을 때 발생하는 문제를 논하는 평가적 관점에서 벗어나서 전산문체학 관점에서 다음과 같은 세 가지 연구문제를 다루고자 한다. 첫째, 기계번역사는 인간번역사처럼 나름대로의 문체가 있는가? 둘째, 기계번역사와 인간번역사의 문체는 확실히 구분되는가? 셋째, 기계번역사의 문체는 시간을 두고 인간번역사와 유사해지고 있는가? 본 연구는 기계번역과 인간번역의 문체 차이를 규명하는 다단계 프로젝트의 초기 연구로 기획되었다.

2. 기계번역을 이용한 문학번역

최근에 신경망기계번역의 등장으로 기계번역의 품질이 눈에 띄게 향상되면

서 '기계번역이 문학번역을 해낼 수 있을까?'하는 질문이 문학번역가나 번역학자들 사이에 관심을 끌고 있다. 2017년 1월에 한국문학번역원이 문학번역가 30명을 대상으로 한 이메일 설문조사에 따르면 응답자 30명 전원이 아직은 기계번역이 출판 가능한 품질의 번역물을 내놓기에는 시기상조라는 점에 의견이 일치하였다. 그렇지만 상당수가 번역에서 기계번역을 사용해 본 경험이 있다고 하였고, 기계번역이 생각지도 못한 신선한 번역문을 내놓는다든지, 기계번역이 인간번역사를 뛰어 넘게 될 날을 걱정하는 목소리도 있었다(정상혁 2017).

신경망기계번역이 등장하기 전에도 기계번역을 문학번역에 적용하는 가능성을 타진한 연구들이 있었다. 토랄과 웨이(Toral & Way 2014)는 실험 연구에서 당시의 통계기계번역기(SMT)을 활용하여 스페인어 원작 소설 1권을 카탈리아어로 번역한 후에 결과물에 대한 자동평가를 실시한 결과 인간번역과 유사성을 측정하는 BLEU 지수가 66.2가 나왔다고 보고하였다. 이는 단어나 구 선택에서 인간번역과 66.2퍼센트가 일치했다는 의미로, 동 연구자들은 기술번역에서와 마찬가지로 문학번역에서도 기계번역이 인간번역을 보조하는데 유용한 도구가 될 수 있다고 보았다. 뒤 이어 동 저자들은(토랄과 웨이 2015) 원작과 번역문으로 학습시킨 통계기계번역기를 활용하여 불어 소설 1권을 영어와 이태리어로 번역하는 실험을 실시하였는데, 불어와 언어구조가 유사한 이태리어 기계번역이 BLEU 지수에서 더 높은 점수를 받았다. 이어 스페인어 소설 1권을 카탈리이아로 번역하여 정성적 번역 품질 평가를 실시하였다. 그 결과 언어 유창성과 문법 및 어휘 사용에서 많은 오류가 발견되었다. 이를 토대로 향후 문학번역에 기계번역을 적용하려면 (1) 기계번역 자체의 성능이 향상되어야 하며 (2) 보통 기술번역에 적용되는 기계번역-포스트 에디팅의 작업흐름은 문학번역에 적합하지 않기 때문에 인간번역사가 번역을 하면서 기계번역의 제안을 검토하는 상호작용적 기계번역(interactive MT)과 같은 다른 작업흐름을 개발해야 한다고 제안하였다.

그 후 신경망기계번역이 등장하면서 기계번역의 품질이 크게 개선되었다. 토랄과 웨이(2018)는 1억 단어 규모의 영어-카탈리이어 문학번역코퍼스를 사용하여 통계기계번역인 문구기반기계번역기와 신경망기계번역기를 학습시킨 후 12개 소설의 번역물에 대한 BLEU 자동평가를 실시하였다. 그 결과 신경망기계번역에서 11% 향상된 결과가 도출되었다. 실험 테스트 중 3개에 대하여 인간평가

를 실시한 결과 전문번역사에 버금가는 품질을 보인 문장 수가 문구기반번역에서는 8%~20%인데 반하여 신경망기계번역에서는 17~34%로 더 높게 나타났다. 신경망기계번역은 번역물의 포스트-에디팅에서도 문구기반기계번역에 비하여 시간, 기술, 정신적 노력 측면에서 생산성이 더 높게 나타났다(토랄 외 2018).

그런데 위와 같은 결과를 뒤집어 말하면 기계번역의 품질이 아무리 좋아져도 문학번역의 경우 아직은 포스트-에디팅이 없이는 어느 정도 가독성이 있는 번역물이 나오기 어렵다는 것을 보여준다. 그런데 최근에는 기계번역→ 포스트-에디팅으로 이어지는 작업 자체가 문학번역에 맞지 않는다는 지적이 나오고 있다. 무어켄스 외(2018)는 6명의 영어-카탈리아어 문학번역가에게 수작업 번역, 통계기계번역 후 포스트 에디팅, 신경망기계번역 후 포스트-에디팅 등 3개의 작업을 수행케 한 후에 설문조사와 인터뷰를 통해 번역사들의 의견을 취합하였다. 그 결과 6명 모두 급히 번역을 해야할 경우에는 유용할지 모르지만 기계번역 후 포스트-에디팅하는 과정이 복잡하고 번역결과도 무미건조하여(bland) 처음부터 직접 번역하는 것을 선호했다. 또한 경험이 많은 번역사 일수록 기계번역의 한계를 더 많이 지적하였다. 결론적으로 인간번역물을 학습시킨 비의식적 알고리즘을 그런 번역물을 생산한 인간번역의 품질과 비교하는 것 자체에 문제가 있으며, 현재 상태로는 기계번역이 인간번역사에게 위협이 될 날은 멀었다고 주장하였다. 타이발코스키-실로프(2018)는 문학번역은 설화구조를 통합적으로, 다른 텍스트들과의 관계(가령, 간텍스트성) 속에서 이해해야 하는데 구나 문장 단위로 번역하는 기계번역의 분절번역방식은 포스트-에디팅을 하더라도 그런 번역 단위를 벗어날 수 없기 때문에 원문의 의미와 구조를 왜곡시킬 수밖에 없다고 보았다. 또한 문학에는 다양한 저자와 번역가의 목소리가 담기게 마련인데 기계번역은 이와 같은 멀티-보이스적인 문학작품의 성격을 동질화시켜버리는 위험을 앉고 있다고 지적하였다. 베사시어와 슈와츠(2015)도 소설 번역에서 기계번역 후 포스트-에디팅을 하는 실험을 실시하였는데 포스트-에디팅이 세그먼트 별로 이뤄지기 때문에 문학텍스트로서의 설화적 일관성이 심각할 정도로 훼손되는 문제를 지적하였다. 이에 대한 보완책으로 포스트-에디팅 후 추가 감수(revision)작업을 실시하였는데 그 결과물도 BLEU 지수에서 포스트-에디팅만한 결과물과 큰 차이가 나지 않았다.

여기서 본 연구와 관련하여 한 가지 관심이 가는 대목은 타이발코스키-실로

프(2018)가 제기한 기계번역에서의 번역사의 목소리(translator's voice)의 문제이다. 동 저자는 기계번역사를 원저자나 장르 별로 구축된 번역메모리로 학습시키지 않는 한 기계번역은 저자나 장르를 구분하지 않고 모든 텍스트를 동일하게 번역하여 원저자 텍스트 소유권(textual ownership)을 훼손하는 결과를 낳을 것이라고 지적하였다. 그런데 전산문체학의 관점에서 번역사의 목소리는 번역사 고유의 문체적 특징, 또는 베이커(Baker 2000: 245)가 언급한 번역사의 문체적인 지문(thumb-print)과 관련 있다. 베이커는 이와 같은 '번역사의 문체'를 번역사의 특징적 언어사용, 언어적 습관의 개인적 프로파일로 정의하였으며 이는 언어적 행위의 반복적 패턴을 통해 나타난다고 주장하였다. 이런 관점에서 본다면 기계번역사도 인간번역사처럼 번역 결과물에 자기 고유의 문체적 지문을 남기는가 하는 의문을 갖게 된다. 이와 더불어 기계번역사들과 인간번역사들의 문체가 상호 뚜렷이 구분되는지도 관심사항이다.

3. 분석 데이터 및 방법

번역사 문체를 연구하는데 가장 큰 걸림돌은 번역문이 원문의 영향을 받는다는 점이다. 번역문에는 원저자의 문체와 번역사의 문체가 섞여 있기 때문에 번역문 간에 문체 차이가 누구의 문체를 반영하는지가 애매하다. 이런 문제를 극복하기 위하여 동일 원문에서 출발한 번역문을 비교 분석하는 방법이 흔히 사용된다. 이 경우 원문이란 변수가 통제되기 때문에 번역문 간의 문체 차이는 번역가의 선택에서 비롯되었다고 확신할 수 있기 때문이다(먼데이(Munday 2007: 20)).

본 연구에서도 이와 같은 연구 방법을 채택하여, 표 (1)과 같이 황석영의 한국어 소설『삼포 가는 길』에 대한 2명의 인간번역사와 3개의 온라인 기계번역사의 영어 번역 결과물을 수집 비교하였다. 기계번역사의 결과물은 1년의 시차를 두고 두 번 수집하였기 때문에 실제 번역물 수는 8개이다. 분석 데이터로『삼포 가는 길』의 영문번역본을 선택한 이유는 무엇 보다 인간번역사의 번역물로 2종이 존재하기 때문에 기계번역 대 인간번역을 집단으로 비교할 수 있는 최소의 조건을 충족하기 때문이다. 또한『삼포 가는 길』이 단편이기 때문에 장편에 비하여 기계번역 웹사이트에서 번역을 수행하여 결과물을 얻어내는 것이

좀 더 용이하는 기술적 측면도 고려되었다.

〈표 1〉 분석 코퍼스 정보

번역사 분류	세부 내역	파일명
인간번역사	Kim Dahee (2008)	HT1
	Kim U-Chang (2012)	HT2
기계번역사	Google (2018, 4)	MT_Google1
	Google (2019, 3)	MT_Google2
	Papago (2018, 4)	MT_Papago1
	Google (2019, 3)	MT_Papago2
	Bing (2018, 4)	MT_Bing1
	Bing (2019, 4)	MT_Bing2

전산문체학 차원에서 상기 8개의 번역문 간의 문체적 유사성 또는 차이를 밝히기 위하여 본 연구에서는 컴퓨터언어 R을 활용한 통계적 분석을 실시하였다. 먼저, stylo 패키지를 사용하여 위 코퍼스에서 1-gram, 2-gram, 3-gram 등 총 3개의 어휘다발의 최빈도 어휘를 각각 500개씩 추출하였다. 어휘다발이란 단어가 연쇄적으로 연결된 묶음으로 가령, 3-gram이라고 하면 in.front.of, by.the.way와 같이 3개의 단어가 연결된 다발을 일컫는다. 이와 같은 어휘 다발 데이터를 사용하여 저자판별(authorship attribution)에서 많이 활용되는 버로우즈의 델타값(버로우즈(Burrows) 2002a, 2003)을 구한 후, 이를 바탕으로 다차원 척도분석(multidimensional scaling: MDS)과 위계적 군집분석(hierarchical cluster analysis: HCA)을 실시하였다.

버로우즈의 델타는 최빈도 어휘를 이용하여 문서 간의 거리를 측정하는 방법으로 간단하게 설명하면 (1) 개별 어휘에 대하여 전체 평균에 대한 표준편차를 z값으로 변환한 후 (2) 거리를 측정하고자 하는 두 문서의 z값의 차이를 구한다. (3) 모든 어휘에 대하여 이와 같이 z값의 차이를 구한 후 (4) 이를 합해서 평균을 낸 값이 델타 값이 된다. 이를 수식으로 나타내면 아래와 같다.

$$\Delta_c = \sum_i \frac{\left| Z_{c(i)} - Z_{t(i)} \right|}{n}$$

이와 같이 계산된 문서 간 거리는 거리행렬표(distance matrix)로 나타난다. 이를 이용하면 문서 간의 군집(cluster)을 그래프로 표시할 수 있는데 이와 같은 목적으로 사용되는 대표적인 통계분석이 MDS이다. 거리행렬표 상의 문서 간 거리를 그래프로 표현하려면 문서 수만큼의 다차원적 공간이 필요한데 이를 2차원 공간으로 축소하여 보여주는 분석법이다. 이와 더불어 HCA를 실시하면 문서들이 서로 어떤 순위로 군집을 형성하는지를 알 수 있다.

분석에는 1-gram의 경우는 어휘목록에서 상위 빈도 100개 어휘를 사용하였으며, 2-gram의 경우는 70개, 3-gram의 경우에는 40개를 사용하였다. 2-gram과 3-gram에서 분석 어휘수를 순차적으로 줄인 이유는 다음과 같다. 전산문체학에서 저자분별에 사용하는 언어적 기재는 수백 가지가 넘지만 그 중에 가장 많이, 성공적으로 사용된 기재는 기능어휘(function words)이다. 전치사, 조동사, 대명사 같은 기능어는 내용어휘(content words)를 연결하여 구문을 형성하는 문법적 기능을 수행하기 때문에 내용에 독립적인 저자 문체를 형성하는데 중요한 역할을 한다(주올라(Juola 2006: 265)). 이런 기능어들은 항상 어휘빈도수에서 상위를 차지하기 때문에 어휘목록에서 일정 수의 최빈도어휘를 사용하면 기능어휘만 따로 뽑아 분석하지 않더라도 기능어휘 중심적인 분석이 가능하다. 버로우즈(Burrows 2002a)는 저자분별에 사용할 최빈도 어휘를 1-gram의 경우 50~100개로 제안하였다. 그런데 2-gram이나 3-gram으로 올라갈수록 어휘조합 특성상 빈도수가 줄어들어 어휘목록을 추출했을 때 데이터 희소성의 문제가 발생할 수 있다. 이런 문제는 텍스트 길이가 짧을수록 심각해진다(스타마토스(Stamatatos) 2009: 541). 따라서 기능어휘 위주의 분석을 유지하기 위해서는 분석에 사용할 최빈도 어휘 수도 줄여나가는 것이 합당하다.

4. 분석 결과

본 장에서는 3절에서 설명한대로 다차원척도분석의 결과를 살펴 본 후에 위계적 군집분석의 결과를 논의하도록 한다.

4.1. 기초통계분석

본격적인 통계분석에 앞서서 분석텍스트에 대한 기초 통계적 특징부터 살펴보도록 한다. <표 2>에는 분석텍스트 8개에 대하여 총어휘수(tokens), 표준어휘다양도(STTR), 문장수(sentences)와 단어 수로 나타낸 평균문장길이(sentence length) 등의 정보가 나와 있다. 이들 정보는 워드스미스 툴즈(Wordsmith Tools) 버전 7을 사용하여 추출하였다.

총어휘수를 보면 기계번역사들은 전부 6,000대인데 반하여 인간번역사인 HT1와 HT2는 각각 7,000대와 8,000대로 크게 차이가 난다. 그 원인은 구체적으로 조사해봐야겠지만 원문에 명시적으로 주어진 단어와 구만 번역하는 기계번역사와 달리 인간번역사들은 필요에 따라 원문에 명시적으로 없는 내용도 추가하기 때문에 나타난 결과일 수 있다. 또는 기계번역이 어려움을 겪는 문학번역의 특성 상 기계번역에서 누락이 발생한 결과일 수도 있다.

〈표 2〉 분석텍스트의 기초 통계

text file	tokens	STTR	sentences	sentence length
Sampo_Bing1.txt	6,493	41.77	657	9.88
Sampo_Bing2.txt	6,663	41.08	681	9.78
Sampo_Gogole1.txt	6,144	40	558	11.01
Sampo_Google2.txt	6,480	40.82	648	10
Sampo_HT1.txt	7,393	45.59	637	11.61
Sampo_HT2.txt	8,085	41.83	710	11.39
Sampo_Papago1.txt	6,075	43.77	632	9.61
Sampo_Papago2.txt	6,559	43.53	632	10.38

STTR은 어휘 유형(타입) 대 전체 어휘(토큰) 수의 비율을 표준화한 수치이다. 이 수치가 높을수록 해당 텍스트에서 다양한 유형의 어휘가 사용되었다는 것을 의미한다. <표 2>에서는 인간번역사인 HT1인 가장 높은 STTR 값을 보인데 반하여 HT2는 4위에 올라있고, 2, 3위를 기계번역사인 Papago가 차지하고 있다. 문장 수에서도 인간번역사인 HT2가 가장 높고 HT1은 4위에 올라 있어 STTR과 비슷한 양상을 보이고 있다. 이 결과만 놓고 보면 현 시점에서 문학번역에 관한한 STTR과 문장수는 기계번역과 인간번역을 구분하는 기준이 되지 못한다고 할 수 있다. 마지막으로 평균 문장길이에서는 인간번역사들이 상위 1,

2를 차지하고 있어, 인간번역사에 비하여 문장 길이가 짧은 것이 기계번역사의 특징이 될 수 있음을 보여준다.

인간번역사와 기계번역사가 어느 정도 구분되는 총어휘수와 평균문장길이를 놓고 보면 흥미로운 점이 발견된다. 기계번역사의 경우 2018년도 버전과 2019년도 버전을 비교해보면 총 어휘수의 경우는 모든 기계번역사에서 어휘수가 증가하였다. 이는 신경망번역이 시간이 지나면서 번역하는 어휘수가 늘어나고 있다는 것을 보여준다. 평균문장길이의 경우에는 Papago의 경우에만 증가하였다. 이와 같은 어휘수나 평균문장길이의 증가는 기계번역사가 인간번역사와 유사해지는 방향으로 움직이고 있다는 것을 의미할 수 있기 때문에 좀 더 심도 있는 연구와 분석이 필요한 부분이다.

4.2. 다차원척도분석(MDS)

이번에는 본격적인 문체 분석으로 들어가서 먼저 다차원척도분석의 결과를 살펴보도록 한다. <표 3>은 1-gram 어휘를 사용하여 계산해 낸 버로우즈의 델타값 행렬표로 8개의 TT간의 거리를 나타낸다. 이 표를 보고 각 번역문 간의 거리를 분석하는 것은 숫자를 일일이 비교해야하기 때문에 쉽지 않다. 이런 거리를 그래프로 나타낼 수 있다면 직감적인 분석이 가능할 것이다. 8개의 TT간의 상호 거리를 그래프로 표시하려면 8개의 축이 있는 8차원 공간이 필요한데 이는 물리적으로 나타낼 수 없다. MDS는 이를 <그림 1>의 그래프에서 보듯이 2개의 축(Coordinate)으로 축소해서 보여준다.

<그림 1>의 그래프를 보면 오른편 상단에 인간번역사 결과물인 HT1과 H2가 위치해 있고 기계번역사의 결과물들이 왼쪽 하단에 위치해있다. 기계번역의 경우 2018년도의 결과물과 2019년도의 결과물이 서로 인접해서 하나의 군집을 형성하고 있다. 이 결과만 본다면 그래프 상에서 인간번역사에 가장 가깝게 위치한 기계번역사는 Papago이다. 이에 비하여 Google과 Bing은 큰 거리를 두고 떨어져 있다.

앞서 언급한 연구 질문과 관련하여 그래프 상에서 눈여겨 볼 점은 인간번역사와 기계번역사 모두 서로 큰 거리를 두고 떨어져 있다는 점이다. 이는 번역사들 간에 문체 차이가 뚜렷하다는 것을 의미한다. 이를 보면 기계번역사도 인

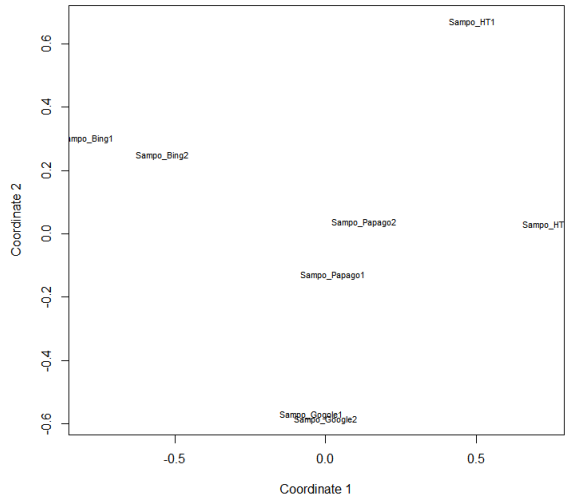
간번역사만큼 나름대로의 독립된 문체를 가지고 있다고 판단할 수 있다. 동시에 <그림 1>에서 기계번역사는 x축을 따라서 왼쪽에 몰려 있고 인간번역사는 오른쪽 끝에 위치해있다. 이는 x축이 기계번역사와 인간번역사를 구분하는 차원(dimension) 역할을 하고 있다는 것을 보여준다.

각 기계번역사를 보면 2018년도 버전과 2019년도 버전이 하나의 군집을 형성할 정도로 가깝게 붙어 있어 1년이란 시간이 지났지만 각 기계번역사의 문체가 크게 달라지지 않았음을 보여준다. 이는 1년이란 단위는 기계번역이 괄목할 정도로 인간번역에 다가설 수 있는 충분한 시간이 아니란 점을 시사한다. 다만 Bing과 Papago의 경우는 2019년도 결과물이 인간번역사 결과물이 있는 오른쪽으로 조금 더 이동해있다. 이에 반하여 Google은 두 번역문이 거의 겹쳐있다. 그렇다면 이와 같은 해석이 2-gram 분석에도 유용한가?

〈표 3〉 1-gram 버로우즈의 델타 값 행렬도

	Sampo_Bing1	Sampo_Bing2	Sampo_Gogole1	Sampo_Google2	Sampo_HT1	Sampo_HT2	Sampo_Papago1
Sampo_Bing2	0.7704752						
Sampo_Gogole1	1.2751857	1.1176282					
Sampo_Google2	1.2783161	1.1001203	0.4551485				
Sampo_HT1	1.4734815	1.2733838	1.3795992	1.3957691			
Sampo_HT2	1.6000339	1.4366908	1.2345324	1.1285255	1.1055639		
Sampo_Papago1	1.2414146	1.0815394	1.0230740	0.9575655	1.2671343	1.1776366	
Sampo_Papago2	1.2847180	1.0343907	1.0284880	1.0167589	1.1229206	1.1356775	0.8894666

〈그림 1〉 1-gram MDS

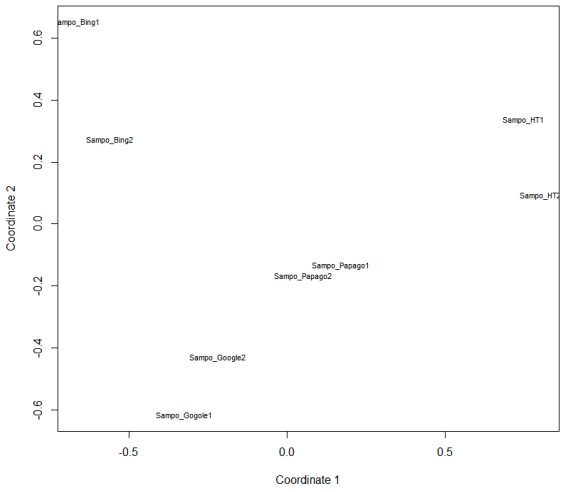


<그림 2>는 2-gram에 대한 MDS 결과이다. 이를 보면 기계번역사들 간의 거리는 여전히 크게 벌어져 있어 <그림 1>에서 분석한대로 기계번역사들이 나름대로의 문체를 유지하고 있다. 이에 반하여 인간번역사의 경우는 거리가 크게 좁아져 하나의 군집을 형성하는 양상을 보이고 있다. 동시에 x축을 따라서 좌-기계번역사, 우-인간번역사로 더 명확히 나타나고 있다. 각 기계번역사의 2018과 2019년도 버전을 검토해 보면 Bing의 경우는 두 버전 사이의 거리가 더 벌어져 있고 반대로 Google의 경우는 거의 겹쳐 있던 두 버전이 조금 거리를 두고 떨어져 있다. 또 Bing과 Google의 경우는 2019년도 버전이 좀 더 오른쪽으로 이동하면서 인간번역사와 가까워진데 반하여 Papago는 거의 하나로 뭉치면서 그와 같은 차이가 사라졌다. 이런 상태에서 x축을 기준으로 볼 때 2019년 버전이 눈에 띄게 인간번역사가 있는 오른쪽으로 이동한 경우가 없기 때문에 1년 사이에 기계번역사의 문체가 크게 달라지지 않았다는 1-gram의 분석도 유효하다.

〈표 4〉 2-gram 버로우즈의 델타 값 행렬도

	Sampo_Bing1	Sampo_Bing2	Sampo_Gogole1	Sampo_Google2	Sampo_HT1	Sampo_HT2	Sampo_Papago1
Sampo_Bing2	0.9653040						
Sampo_Gogole1	1.3744226	1.1206024					
Sampo_Google2	1.2539330	1.0920996	0.5567243				
Sampo_HT1	1.5654785	1.4943492	1.4979863	1.3180676			
Sampo_HT2	1.6583049	1.5145375	1.4243304	1.2973895	0.8496871		
Sampo_Papago1	1.2987809	1.0759169	0.9044335	0.8680441	1.0268516	0.9407942	
Sampo_Papago2	1.3088505	1.1284906	0.9824258	0.9472898	1.2079820	1.1167473	0.8105275

<그림 2> 2-gram MDS

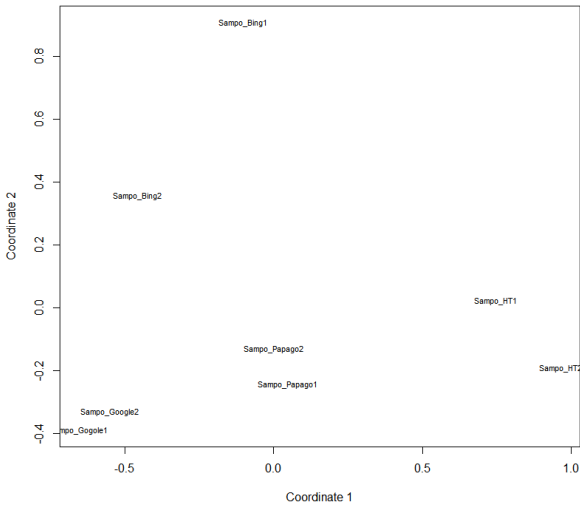


<그림 3>의 3-gram 결과를 보면 1-gram 및 2-gram과 거의 비슷한 양상을 보이고 있다. 3개의 기계번역사간의 거리가 여전히 크게 벌여져 있어 문체의 차이가 뚜렷이 나타나는 동시에 x축을 따라 기계번역사와 인간번역사가 좌-우로 나뉘어진 배치가 간격이 더 벌어져 있다. 기계번역사들의 2018년과 2019년 버전을 비교해봤을 때는 두 버전 사이가 크게 벌어진 Bing을 제외하면 1-gram 및 2-gram과 큰 차이가 없다. Google의 경우는 2-gram에 비하여 양 버전의 거리가 좀 더 좁혀져 있고, Papago의 경우는 조금 더 떨어져있지만 여전히 인간번역사와는 큰 거리를 두고 떨어져 있다.

〈표 5〉 3-gram 버로우즈의 델타 값 행렬도

	Sampo_Bing1	Sampo_Bing2	Sampo_Gogole1	Sampo_Google2	Sampo_HT1	Sampo_HT2	Sampo_Papago1
Sampo_Bing2	0.9977741						
Sampo_Gogole1	1.4388222	1.0430139					
Sampo_Google2	1.3699528	0.9337627	0.5148307				
Sampo_HT1	1.3054166	1.3757199	1.4841983	1.4382031			
Sampo_HT2	1.5797238	1.6059145	1.6978315	1.5379266	0.6047415		
Sampo_Papago1	1.2617058	0.9841161	0.8871022	0.9003478	0.9151785	1.0780671	
Sampo_Papago2	1.2252722	1.0318321	1.0460441	0.9593988	1.0894743	1.1907861	0.6669855

〈그림 3〉 3-gram MDS



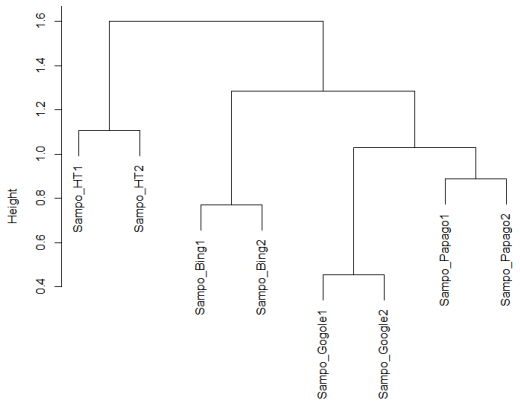
이상의 결과를 종합하여 연구 질문에 답을 한다면, 첫째, 기계번역사들도 인간번역사들과 마찬가지로 자신의 문체를 가지고 있다. 둘째, 기계번역사와 인

간번역사는 문체적으로 확실히 구분된다. 다만 기계번역사 중에는 Papago가 인간번역사 쪽으로 가장 근접해 있다. 셋째, 2018~2019년 1년 사이에 기계번역의 문체가 인간번역에 근접하는 방향으로 크게 달라지지 않았다.

4.3. 위계적 군집분석(HCA)

이번에는 4.2절의 MDS 분석의 결과를 검증하는 차원에서 HCA분석 결과를 살펴보자. 위계적 군집분석은 <표 3>, <표 4>, <표 5>의 거리 데이터를 사용하여 맨 아래 개별 텍스트 차원에서부터 시작해서 위로 올라가면서 거리가 가장 가까운 문서나 군집을 차례로 묶어가는 분석법이다. 이렇게 해서 만들어진 나뭇가지 형태의 그래프를 덴드로그램(dendrogram)이라 한다. <그림 4>, <그림 5>, <그림 6>은 각각 1-gram, 2-gram, 3-gram에 대한 HCA 덴드로그램이다.

〈그림 4〉 1-gram HCA

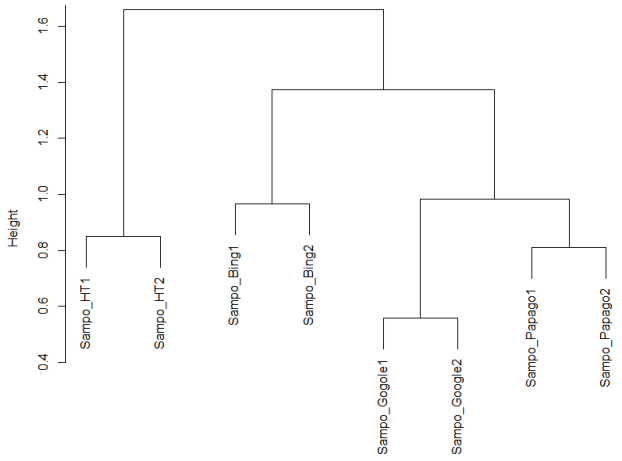


<그림 4>의 그래프를 맨 아래쪽에서부터 올라가면서 검토하면 <그림 1>의 MDS 그래프에서 가장 가까운 거리에 위치한 Google의 두 번역문이 먼저 하나의 군집으로 묶여 있는 것을 볼 수 있다. 다음으로는 Bing과 Papago의 두 번역문이 차례대로 군집을 이루고 있다. 마지막으로 인간 번역물 HT1과 HT2가 하

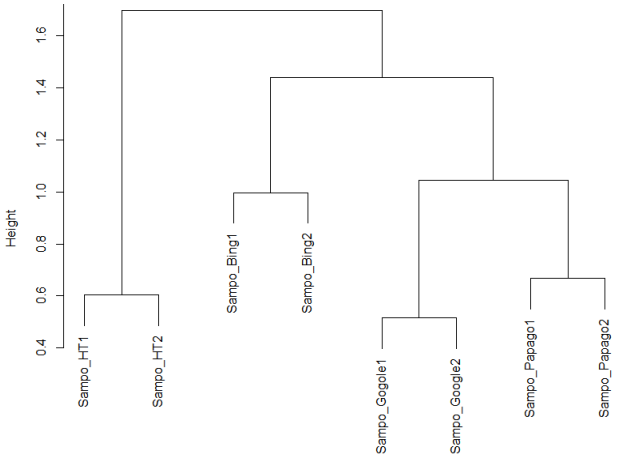
나의 가지를 형성하고 있다. 먼저, 기계번역사 별로 2018과 2019년 번역물이 제일 먼저 군집을 형성하는 것은 두 번역물의 문체가 그 만큼 가깝다는 것을 뜻한다. 즉, 1년 동안 문체적으로는 큰 변화가 없었음을 보여준다. 반대로 각 기계번역사별로 군집을 형성한 것은 기계번역사 간에도 뚜렷한 문체의 차이가 존재한단 것을 확인시켜 준다. 4개의 하위 군집이 위로 올라가면 Google과 Papago 군집이 먼저 묶이고 있다. 이렇게 묶인 Google-Papago 군집이 다음 단계에서 Bing과 합쳐지고 있다. 이는 기계번역사들 중에는 Google과 Papago가 문체적으로 좀 더 가깝고 Bing은 이들과 떨어져 있음을 보여준다. 덴드로그램 맨 위쪽에서는 최종적으로 기계번역사 전체를 하나로 묶은 군집과 인간번역사를 묶은 군집이 합쳐지고 있다. 이와 같은 군집 양상은 <그림 5>과 <그림 6>에서도 거의 동일하게 나타나고 있다.

이와 같은 HCA 분석 결과는 개별 기계번역사들이 문체적으로 구별되며 동시에 기계번역사와 인간번역사 간에도 각자 별도의 군집을 형성할 만큼 뚜렷한 문체의 차이가 존재한다는 의미로 4.2절에서 분석한 내용과 일치한다.

〈그림 5〉 2-gram HCA



〈그림 6〉 3-gram HCA



5. 결론

본 연구에서는 동일 한국어 단편소설에 기반을 둔 인간번역사와 기계번역사의 영어 번역물의 문체를 전산문체학 관점에서 분석해보았다. 그 결과 앞서 제기했던 연구문제에 대하여 (1) 기계번역사들도 인간번역사처럼 나름대로의 문체를 가지고 있으며, (2) 동시에 기계번역사와 인간번역사의 문체 간에는 분명한 차이가 있고, (3) 1년이란 시간 동안 기계번역사의 문체가 크게 달라지지 않았다는 결론을 도출하였다. 기계번역사들이 서로 구별되는 문체를 가지고 있는 것은 번역과정을 제어하는 알고리즘의 차이 때문일 수도 있고 학습에 사용된 코퍼스의 차이 때문일 수도 있다. 그러나 동시에 기계번역사들이 인간번역사들과 문체적으로 명확하게 분리되는 점은 기계번역사들이 공통적으로 사용하는 규칙이 존재한다는 것을 암시한다. 마지막으로 연속 학습을 통해 지속적으로 발전한다는 NMT이 1년 동안 문체 면에서 눈에 띄게 인간번역사 쪽으로 이동하지 않았다는 점은 문학번역에서 기계번역사가 인간번역사의 문체에 근접하기 까지는 (근접할 수 있다는 전제하에) 현재 기술 수준에서는 상당한 시간이 소요될 수 있음을 암시한다.

기계번역과 인간번역 간의 문체 차이는 이제 초기 연구가 진행될 만큼 아직 많은 것이 규명되지 않은 주제이다. 단순하게 생각하면 기계번역이 제대로 된 문학번역을 할 수 있을 때는 번역의 완성도(품질) 못지않게 기계번역사가 인간번역사의 문체를 흉내 낼 수 있을 때라고 할 수 있다. 그렇지만 기계번역이 인간번역과 문체가 유사해진다는 것이 무엇을 의미하는지는 숙고가 필요하다. 왜냐하면 인간번역 문체라고 하지만 그 안에서도 번역사들 간에 큰 차이를 보이기 때문이다. 또 번역사에 따라 자신의 문체를 적극적으로 드러내기도 하고 원문 저자의 문체 뒤에 숨기도 한다(버로우즈 2002b). 따라서 기계번역이 인간번역의 문체를 가진다는 것은 인간번역 문체라는 어떤 점을 향해 하나로 수렴되는 과정이 아니라 나름대로의 독자적인 문체 거리를 유지한 상태에서 인간번역사들 군집 안으로 들어 온 상황이라고 할 수 있다. 그것은 전산문체학 차원에서는 MDS나 HCA 같은 분석에서 두 집단이 더 이상 별도의 군집으로 나눠지지 않고 서로 뒤섞이는 시점이 될 것이다.

본 연구는 문학번역에서 기계번역과 인간번역의 문체 차이를 비교 분석하는 장기 연구프로젝트의 기초연구로 기획되었다. 따라서 몇 가지 연구의 한계를 앉고 있다. 가령, 특정 작품에 대한 번역물을 비교하였기 때문에 본 연구결과를 일반화하기에는 무리가 있다. 또한 기계번역 기술이 계속 발전하고 있다는 점을 고려할 때 본 연구의 결과는 연구 데이터가 수집된 시점에서만 유효하다. 이와 같은 문제는 향후 연구프로젝트를 수행하면서 분석 데이터의 양과 종류를 늘리고 여러 시점에서 데이터를 수집 분석하는 통시적 연구를 통해 보완할 계획이다.

참고문헌

김순미, 신호섭, 이준호 (2019).「번역학계와 언어서비스업체(LSP)간 산학협력연구: ‘포스트에디팅 생산성’과 ‘기계번역 엔진 성능 비교’」,『번역학연구』20(1): 41-76.

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

서보현, 김순영 (2018)「기계번역 결과물의 오류유형 고찰」,『번역학연구』19(1): 99-117.

이성화, 김세현 (2018)「영-한 및 한-영 기계번역 품질향상을 위한 프리에디팅 기법 제안」,『번역학연구』19(5): 121-154.

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

정상혁 (2017)「진화하는 번역기 ... 사라지는 번역가?」,『조선일보』, 2018년 10월 12일 검색.
(news.chosun.com/site/data/html_dir/2017/01/18 /2017011800020.html)

Baker, Mona (2000) 'Towards a Methodology for Investigating the Style of a Literary Translator', *Target* 12(2): 241-266.

Besacier, Laurent and Lane Schwartz (2015) 'Automated Translation of a Literary Work: A Pilot Study', *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature* (Denver, Colorado, June 4, 2015), 114-122.

Burrows, John F (2002a) ''Delta': A measure of Stylistic Difference and a Guide to Likely Authorship', *Literary and Linguistic Computing* 17(3): 267-87.

Burrows, John. F (2002b) 'The Englishing of Juvenal: Computational Stylistics and Translated Texts', *Style* 36(4): 677-699.

Burrows, John F (2003) 'Questions of Authorship: Attribution and Beyond', *Computers and the Humanities* 37(1): 5-32.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley and Andy Way (2017) 'Is Neural Machine Translation the New State of the Art?', *Prague Bulletin of Mathematical Linguistics* 108(1): 109-120.

Garg, Ankush and Mayank Agarwal (2018) *Machine Translation: A Literature Review*, arXiv:1901.01122v1.

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

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.

Munday, Jeremy (2007) *Style and Ideology in Translation: Latin American Writing in English*, New York: Routledge.

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

Stamatatos, Efstathios (2009) 'A Survey of Modern Authorship Attribution Methods', *Journal of the American Society for Information Science and Technology* 60(3): 538-556.

Taivalkoski-Shilov, Kristiina (2018) Ethical Issues Regarding Machine(-assisted) Translation of Literary Texts, *Perspectives*, DOI: 10.1080/0907676X.2018.1520907.

Toral, Antonio and Andy Way (2014) 'Is Machine Translation Ready for Literature?', *Proceedings of Translating and the Computer* 36 (London, 27-28 November 2014), 174-176.

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

Toral, Antonio, Martijn Wieling and Andy Way (2018) 'Post-editing Effort of a Novel With Statistical and Neural Machine Translation', *Frontiers in Digial Humanities* 1: 1-11.

[Abstract]

Stylometric Comparative Analysis of Style in Human vs. Machine Literary Translations

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

The current research is designed as a pilot study under a project aimed at investigating differences in style between human and machine translators in Korean-English literary translation. The research seeks to address three questions from a stylometric or computational linguistic perspective. (1) Do machine translators have their own unique styles? (2) Are they clearly distinguishable from human translators in style? (3) Are they progressing over time in such a direction that they are becoming more like human translators? These questions are tackled by analyzing the English translations by two human translators and three machine translators of a single Korean short novel. The translations by the machine translators were collected at two points separated by a span of one year, providing us with a total of eight translated texts. Burrows' delta scores, a popular measure of textual distance, were extracted from the texts and analyzed by two unsupervised statistical methods – multidimensional scaling and hierarchical cluster analysis. The machine translators displayed interdependent styles clearly distanced from one another, while they as a whole were distinctly separated from the human translators. The machine translators showed no evidence of having narrowed the distances between them and the human translators over one year.

▸ Keywords: machine translation, translating style, literary translation, stylometry (computational stylistics)
▸ 주제어: 기계번역, 번역 문체, 문학 번역, 전산문체학

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

논문투고일: 2019년 4월 30일
심사완료일: 2019년 5월 25일
게재확정일: 2019년 5월 28일