http://dx.doi.org/10.15749/jts.2023.24.2.004

번역학연구
24권 2호
(2023.6.30.)

자동화된 기계학습(AutoML)을 활용한 특허 특화 번역엔진의 영한번역 성능 평가

최 효 은 · 이 청 호 · 이 준 호
(이화여대 · 에버트란 · 중앙대)

1. 서론

발명을 보호하는 권리인 특허권과 관련해서 특허를 출원하거나 권리를 보호받기 위한 과정에서 필요한 특허 명세서 번역, 국내외 특허 심사와 관련된 번역 등을 특허번역이라고 한다. 자국어주의가 강조되는 특허의 특성상 국제출원으로 인한 수요가 특허번역의 가장 큰 비중을 차지한다. 최근 10년간 국제출원의 대표격인 PCT(Patent Cooperation Treaty) 출원의 흐름을 보면, 2012년 출원 건수가 195,345건으로 20만 건에 미치지 못했으나 2021년에 이르러 출원 건수는 무려 277,500건으로 30만 건에 육박할 정도로 증가했다(WIPO 2022: 2). 이와 같은 추세를 반영하여 국내 특허번역 시장규모는 2011년 약 410억 원에서 2020년 약 900억 원 규모로 2배 이상 큰 폭으로 성장할 만큼(이성용 2022)[1] 매년 번역 수요 또한 비약적으로 증가하고 있다.

이와 같이 엄청난 수요를 사람이 전적으로 감당하기에는 어려움이 있으며,

1) 특허뉴스. https://www.e-patentnews.com/8398

따라서 특허번역 업계에서는 인공신경망 기계번역(NMT)이 주목받기 훨씬 이전부터 기계번역을 활발하게 활용해 왔다(최효은, 이지은 2017: 140). 특히 NMT 출현 이후 기계번역 결과물의 품질이 비약적으로 좋아지면서 출원된 특허가 권리를 인정받을 만한 발명인지를 심사하는 심사관은 물론이고 특허번역에 종사하는 전문가들도 기계번역을 활용하는 추세다. 번역 전문가들은 물론 아직도 기계번역이 전적으로 의지할 수 있는 수준은 아니나 인간번역에서 있을 수 있는 누락을 방지할 수 있고 전문용어의 번역을 참고할 수 있는 장점이 있다고 말한다.

여기서 주목할 점은 NMT 역시 데이터 학습을 통해 만들어진 하나의 결과물이며, 어떠한 종류의 학습데이터를 어떻게 처리하느냐에 따라 NMT의 품질은 크게 달라질 수 있다는 것이다. 하지만 아직까지 사용자들은 대부분 구글과 같은 범용 기계번역이 일방적으로 제공하는 결과를 수동적으로 받아들이고 후처리 작업을 하는 것이 일반적이다. 따라서 특정 영역의 전문번역사들을 위해 특화된 번역엔진을 제공하는 업체가 있거나, 기계번역을 활용하는 주체들이 각자의 필요에 맞게 맞춤형으로 기계번역엔진의 품질을 개선할 가능성에 주목할 필요가 있다. 이러한 특수 영역에 특화된 기계번역엔진의 도입은 기계번역 사용의 효율성 개선에 크게 기여할 수 있기 때문이다.

이러한 배경하에 본 연구에서는 구글 클라우드 AI(Google Cloud AI)에서 제공하는 자동화된 기계학습(AutoML)을 활용하여 구글 번역엔진을 특허 전문 병렬코퍼스로 학습시킨 후 학습된 번역엔진이 생산하는 번역 결과물이 기존의 구글번역 결과물과 어떻게 달라지는지를 비교 분석하고자 한다. 이를 통해 전문 분야에서 기계번역을 사용하는 전문가들이 AutoML을 활용하여 좀 더 효율적으로 기계번역을 활용할 수 있을지의 여부를 타진해 보고자 한다.

2. 선행연구 분석

2.1 특허 기계번역에 관한 연구

본고의 연구 대상과 동일한 특허 분야에 있어서 기계번역에 관한 연구는

NMT 이전부터 활발하게 이루어져 왔다. 최승권(2007)은 패턴 기반의 영한 특허문서 자동번역 시스템에서 영어 특허문장을 한국어로 자동으로 번역하기 위해 사용된 영한 특허번역패턴에 대해 기술하였다. NMT 이전 단계의 패턴 기반 엔진에 대해서 최승권(2007: 302-306)은 전형적인 영어 특허 문장의 특징을 분석한 후, 이에 대한 고유한 한국어 대응 번역패턴을 대입하여 특허번역의 패턴을 형식화하였다. 이와 같이 형식화한 패턴을 적용한 후 번역률을 평가하였으며, 영한 특허문서의 자동번역에서 특허번역패턴이 포함된 경우 그렇지 않은 경우보다 번역률이 1.20% 높았으며, 체감번역률은 무려 11.60%의 차이가 있었다. 이를 통해 특허번역패턴의 유무가 번역 결과의 품질에 중요한 영향을 미친다는 점을 시사하였다(최승권 2007: 319).

한편 최효은과 이지은(2017)은 최승권(2007)과는 반대의 언어 방향, 즉 한국어 → 영어 방향을 기준으로 한국 특허청의 특허 검색사이트인 '키프리스'[2]에서 제공하는 무료 한국어-영어 기계번역기인 패턴 기반의 'K2E-PAT'의 영어 번역 결과물 100문장을 대상으로 자동평가와 수동평가를 진행하였다. 분석 결과 자동평가와 수동평가 결과 모두 낮은 점수를 기록하여 당시 K2E-PAT의 번역 품질이 충분하지 않다는 결론을 내렸다(최효은, 이지은 2017: 165).

최승권(2007)과 최효은과 이지은(2017)은 NMT 도입 이전에 장기간 사용되었던 패턴 기반의 특허 기계번역엔진을 연구의 대상으로 삼았다는 점에서 공통된다. 패턴 기반 번역엔진은 NMT와는 완전히 다른 방식으로 패턴의 입력값을 바탕으로 하여 출력값을 도출하므로 입력값의 양과 질이 번역 결과물의 품질을 좌우하며, 당시 기준으로 해도 실제 대중에 공개된 패턴 기반의 기계번역엔진의 품질이 썩 만족스러운 수준은 아니었음을 알 수 있다.

해외에서는 차이(Tsai 2017)가 한국의 K2E-PAT과 비슷한 시기에 개발되었으므로 역시 NMT 이전의 통계 기반인 대만특허청(TIPO)의 기계번역엔진에 의한 발명의 명칭의 영어 → 중국어 번역 결과물을 분석하였다. 철자, 형태, 어휘, 의미, 통사를 기준으로 번역 결과물의 오류를 지적한 뒤 이러한 오류가 발명의 명칭당 최소한 하나 이상씩 발견되었으며 이로 인해 번역 결과물의 품질이 저하되고 평가 결과에도 영향을 미쳤음을 밝혔다(Tsai 2017: 154).

---

2) www.kipris.or.kr

로시와 위긴스(Rossi and Wiggins 2013) 역시 NMT 이전 단계의 자체 개발한 기계번역기를 전문 특허 분야에 적용하면서 그 가능성을 타진하기 위해 기계번역 결과물을 대상으로 자동평가 및 수동평가를 수행하였다. 로시와 위긴스(2013: 121)는 특히 사람에 의한 수동평가를 위해서 특허번역에서 중요한 평가기준을 용어, 정보 누락, 정보 추가, 단어 순서의 네 가지로 세분화했다는 데 그 의미가 있다고 할 수 있겠다.

NMT 도입 이후의 특허 기계번역 연구는 NMT 직전의 통계 기반 번역엔진(Statistical Machine Translation, SMT)과 NMT의 비교가 주를 이룬다. 카스틸호 등(Castilho et al. 2017)은 자체적으로 새롭게 개발한 NMT와 기존의 SMT를 대상으로 화학 분야의 발명 명칭과 요약의 번역 결과에 대해 자동평가와 수동평가를 진행했다. 카스틸호 등(2017: 114)에 의하면, BLEU로 대별되는 자동평가의 결과는 발명의 명칭에서는 NMT가, 요약에서는 SMT가 우수한 것으로 드러났으며, 사람에 의한 수동평가에서는 NMT보다 SMT의 결과물이 전반적으로 좀 더 나은 것으로 평가되었다.

이와 유사하게 키노시타, 오시오와 미츠하시(Kinoshita, Oshio and Mitsuhashi 2017)는 일본특허청(JAPIO)이 참여한 프로젝트를 소개하면서 영어 ↔ 일본어, 중국어 → 일본어, 한국어 → 일본어 방향의 특허번역에 대해 자체적으로 개발한 NMT와 SMT에서 학습시킨 코퍼스의 크기에 따라 자동평가의 결과 및 수동평가의 결과가 어떻게 달라지는지를 살펴보았다.

이와 같이 NMT 도입 이후, NMT를 연구 대상으로 한 특허 분야의 기계번역 연구들은 대체로 여전히 공존하고 있는 SMT와 품질 비교가 주를 이루며, 대부분 자체적으로 개발하고 학습시킨 엔진을 연구 대상으로 하는 것을 알 수 있었다. 즉, 아직까지는 실제로 현업에서 주로 사용하고 있는 범용 기계번역엔진, 예를 들어 구글번역 등을 대상으로 한 연구를 찾아보기 어려운 것이 현실이다.

이러한 배경을 고려할 때, 실무에서 광범위하게 사용하고 있는 구글번역을 대상으로 AutoML을 활용한 학습 과정을 통해 학습 전과 학습 후의 특허번역 결과물을 비교해보고자 하는 본 연구는 특허 기계번역 분야에서도 연구의 공백을 메울 수 있으리라 여겨진다.

2.2 기계번역 성능 개선에 관한 연구

다양한 연구를 거듭하면서 기계번역의 성능은 꾸준한 개선을 보여왔다. 특히 NMT 도입 이후에는 인간과 경쟁 혹은 인간의 대체 담론까지 등장할 정도로 일부 영역에서 기계번역 결과물의 품질이 우수성을 보였던 것도 사실이다. 하지만 인간번역사와 유사한 수준의 양질 번역 결과물을 생성하는 것을 목표로 한다면 기계번역 결과물이 우수하다고만 보기는 어렵다. 더욱이 번역 과정에 있어 전문적 지식이 필요하고, 번역 결과물의 오류가 가지는 사회적 함의가 큰 영역에서는 기계번역 사용에 어려움이 있다. 따라서 기계번역 결과물의 품질을 근본적으로 개선하지 않는다면 기계번역 원활한 실무 사용에는 제약이 있을 수 밖에 없다.

이러한 기계번역 사용의 난제를 극복하기 위해 기계번역 모델을 개선하기 위한 노력은 꾸준히 이뤄져 왔다. 예를 들어 전반적인 모델의 구조를 변경하여 기존의 모델 대비 자동평가에서 더 높은 점수를 득할 수 있음을 입증한 연구들이 존재한다(Guo et al. 2019; 정영준 외 2019 등). 이처럼 전체 모델의 성능 개선을 위한 접근법은 기계번역 성능 개선의 근본적 해결책이 될 수 있기에 의미가 있다. 하지만 NMT 결과물에서 빈번하게 발생하는 다양한 문제를 개별적으로 해결하기에는 한계가 있다. 이에 연구의 범위를 NMT의 잘 알려진 단점으로 한정하고, 이를 극복하려는 방안을 제시한 연구들이 등장하고 있다. 예를 들어 완 외(Wan et al. 2022)는 짧은 문장에서 발생하는 NMT의 단점을 극복하기 위해 학습을 위한 데이터에 대한 분산 균형이나 맥락 정보를 보완하는 방안을 제시하였다. 국내의 경우 김정희 외(2022)는 기계번역의 존댓말과 반말 혼용 문제를 해결하기 위해 번역 모델에 사용하는 서브워드를 자모 단위로 구성하고, 코퍼스 문장들에서 존댓말과 반말을 통일하여 모델을 학습하는 방안을 제시하였다.

이와 같은 공학적 설계를 통해 기계번역의 한계를 극복하는 접근 외에, 또 다른 접근은 양질의 학습데이터를 활용하여 기계번역 결과물의 품질을 높이는 것이다. 물론 양질의 병렬코퍼스를 확보하여 비지도 학습을 시행하는 것은 어려운 일이다. 하지만 양질의 데이터를 확보할 수만 있다면 우수한 기계번역 결과물에 기여할 수 있다는 연구가 존재한다. 박찬준과 임희석(2020)은 국내에서

구축한 공공 AI Hub 데이터를 사용하여 기존의 모델보다 우수한 성능의 모델을 구현했으며, 모델 변경보다 데이터의 품질이 더 중요함을 강조하였다.

하지만 양질의 데이터를 확보한다고 해도 기계번역은 학습된 데이터와 다른 영역에서는 성능이 부족할 수 있다. 이러한 단점을 극복하기 위해 기계번역의 학습 및 사용 영역을 특정 주제로 한정하는 접근법이 등장하였다. 에체고옌 외(Etchegoyhen et al. 2018)는 특정 주제에 한정하여 기계번역을 학습시키는 것이 정확한 기계번역 시스템을 만드는 것만큼이나 중요함을 강조하였다. 이러한 주장을 뒷받침하기 위해 세 개의 주제 영역에 특화하여 학습된 기계번역 결과물과 그렇지 않은 기계번역 결과물을 비교하였다. 평가에 있어서는 BLEU, METEOR, TER 등의 자동평가와 유창성, 충분성, 필요성, 용이성, 노력이라는 항목으로 구성된 인간평가를 모두 시행하여 평가의 객관성을 높이기도 했다. 이처럼 특정 영역에 특화된 기계번역 접근법은 이미 그 실효성을 인정받아, 유럽연합에서는 공공 기관 간의 언어장벽을 허물기 위한 프로젝트인 iADAATPA가 2019년 실행되기도 하였다. 왕 외(Want et al. 2019)는 여기서 한 걸음 더 나아가 게임번역에 특화된 번역 모델 생성을 위해 구글플레이에서 병렬코퍼스를 확보하여 모델을 학습시키고, 등장 빈도가 낮은 단어에 대한 처리 메커니즘을 적용하여 범용 기계번역보다 높은 효율성을 보였다고 주장하였다.

국내에는 이러한 사례가 많지는 않지만, 박찬준 외(2020)는 코로나19와 관련된 말뭉치를 활용하여 코로나19라는 특정 주제에 국한해서는 범용으로 사용되는 구글 번역기와 비교하여 더 높은 BLEU 스코어를 확보할 수 있음을 주장하였다. 또한, 김세린과 권혁철(2022)은 오픈소스 모델인 M2M-100을 베이스라인으로 설정한 이후, AI Hub에 공개된 의료 및 보건 분야 병렬코퍼스를 활용하여 모델을 학습시킨 이후, 데이터 증강(data augmentation), 기술 용어 추출(technical term extraction) 등을 활용하여 상대적으로 적은 데이터로 베이스라인 모델보다 높은 성능을 보여준 사례를 소개한 바 있다.

번역학계의 연구를 살펴보면, 이준호(2022)는 법률번역 전문업체가 시스트란의 플랫폼에 구현한 트랜스포머 모델에 계약서 데이터를 학습시켜 구글번역의 한영번역 결과물과 대조 분석을 시행하였다. 그 결과 일반적인 문장에서는 큰 차이가 없었으나, 난제가 있는 문장에서는 특정 주제 영역에 특화된 모델이 정확도, 유창성, 사용 적합성, 필수적 수정의 필요성 등 모든 항목에서 구글번

역 대비 우위를 보였다고 보고하였다. 본 연구와 유사한 접근 방식을 취한 이지은과 최효은(2023)의 연구는 법률 및 특허에 특화된 NMT 엔진인 오트란을 활용하여 한영번역 결과물을 생성하였다. 이후 자동평가와 수동평가를 시행하여 범용 NMT 대비 맞춤형 NMT의 성능 우위를 확인하였다. 다만 맞춤형 엔진 간에도 법률용 엔진이 특허용 엔진보다 성능이 우수하다고 보고하며, 특허 문건의 "길고 복잡한 문장 특성"이 기계번역 성능 저하의 원인일 수 있음을 지적하였다.

이상의 문헌 검토에서 볼 수 있듯 양질의 번역 데이터 구축은 기계번역 품질을 높이는 유효한 방안 중 하나이다. 특히 특정 주제 영역의 대표성을 지니는 데이터는 특화된 기계번역 모델을 만드는데 기여할 수 있다. 이러한 데이터 구축은 번역실무자와 번역학계가 적극적으로 개입할 수 있는 영역이라 할 수 있으며, 본고에서 사용하는 AutoML Translation은 공학적 지식이나 코딩에 대한 지식이 없어도, 양질의 기계번역엔진 생성에 도움을 주는 구글의 범용 서비스이다. 따라서 AutoML Translation의 실효성을 검증할 수 있다면, 앞으로 더 많은 사용자가 자신의 주제 영역에 특화된 기계번역엔진을 구현하고 이로부터 혜택을 볼 수 있을 것이다.

여기에 더해 번역학계의 특화 번역엔진 연구가 영한보다는 한영번역에 더욱 집중하고 있다. 또한, 특허 문건의 기계번역은 쉽지 않은 작업이며, 맞춤형 기계번역엔진의 실효성이 검증된 사례는 매우 드물다.

따라서 본 연구는 양질의 영한 병렬코퍼스를 활용하여 특허 기계번역엔진의 실효성을 추가적으로 검증하고, 데이터 구축에 대한 미래 방향성을 고찰하는 계기를 제공하고자 한다. 이를 통해 본고는 번역학계가 산업계 및 공학계와 협업할 가능성을 모색하고자 한다.

3. 연구 방법

3.1 본 연구에 사용한 기계번역 AutoML의 준비 과정

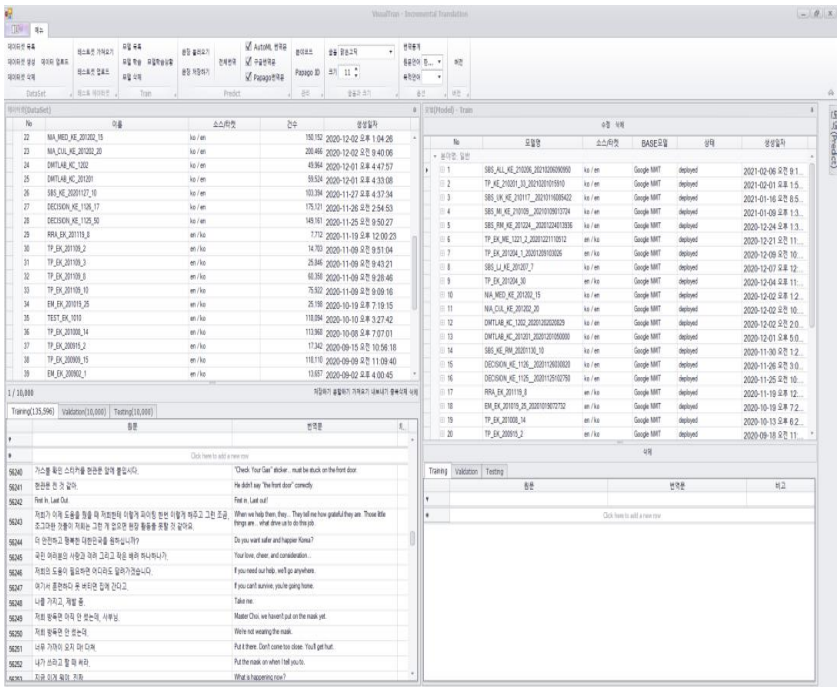
AutoML은 머신러닝의 한 기법으로서 머신러닝 모델의 선택, 구성, 파라미터화를 자동화하여 최소한의 노력으로 최적의 효과를 달성하기 위해 만들어졌다. 구글 클라우드(Google Cloud)는 AutoML을 활용하면 “머신러닝 전문 지식이 부족한 개발자도 비즈니스 니즈에 맞게 고품질 모델을 학습이 가능함”을 홍보하고 있으며 정형데이터, 시각, 언어 등의 영역에서 다양한 서비스를 제공하고 있다. 그중 본 연구에서 사용한 AutoML Translation은 백 개 이상의 방대한 언어 번역을 지원한다. 가장 큰 특징은 특정 분야와 관련해 이미 보유하거나 별도 생성한 원문 및 번역문 쌍을 업로드하여 학습을 진행하면 해당 분야에 특화된 번역 결과를 제공하는 기계번역 모델을 비교적 손쉽게 생성할 수 있다는 것이다.

본 연구는 AutoML에서 학습된 특허 전문 영한 기계번역 모델 확보를 위해 기계번역 전문업체인 에버트란과 협업하였다. 에버트란은 2000년 초부터 기계번역 및 번역작업관리에 특화된 솔루션을 개발하면서 산자부 및 법제처 등 여러 국가기관 및 번역서비스 기업에 솔루션을 납품해왔다. NMT가 출시된 이후에는 다양한 기계번역 서비스의 호출 및 연동을 자사의 솔루션 내에 구현하여 편리한 기계번역 활용 및 포스트에디팅을 지원해왔다. 이후 2019년 한국지능정보사회진흥원(NIA)의 ‘인공지능학습용 데이터 구축’ 사업에 참여하면서 특화형 기계번역 모델 생성을 본격적으로 진행하였다. 또한, 국가 법령 번역 및 지자체 조례 분야에 적합한 기계번역 모델 생성을 위한 법령 분야 병렬말뭉치를 구축하였고, 해당 말뭉치를 학습에 활용하여 법령 전문번역 서비스를 시작하게 되었다. 이 과정에서 일반적인 기계번역 모델에 분야별 병렬말뭉치를 추가하여 학습시키는 AutoML 번역 모델의 품질이 일반 번역 모델보다 우수할 수 있음을 파악하고 특허 분야에도 특화된 모델 학습을 시행하였다.

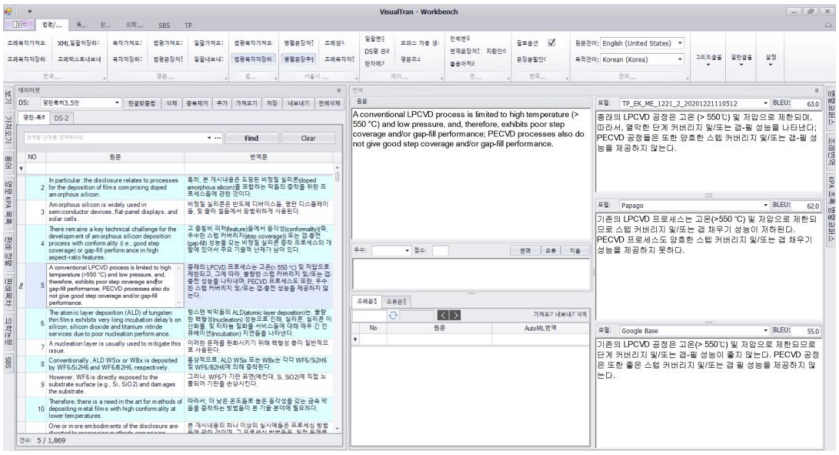
구체적 학습 방법을 살펴보면, AutoML Translation은 사용자 제공 데이터를 자동으로 학습 세트(training set), 검증 세트(validation set), 테스트 세트(test set)로 분류해 주는 기능을 제공한다. 이중 학습 세트(약 80%)만 실제로

AutoML학습에 활용하며, 검증 세트(약 10%)는 AutoML학습 결과 생성된 여러 번역 모델 중 가장 적합한 모델을 선정하기 위해 활용된다. 반면 테스트 세트(약 10%)는 최종 번역 모델을 평가하고 BLEU 스코어 산출하는 과정에 사용된다. 본 연구 역시 동일한 접근법을 사용했으며, 에버트란에서 자체 개발한 워크벤치 프로그램을 사용하여 학습을 진행했으며, <그림 1>과 <그림 2>를 통해 학습데이터 및 평가과정 일부를 확인할 수 있다.

〈그림 1〉 모델 생성 학습에 사용된 데이터 세트



〈그림 2〉 평가 데이터 일부



본 연구에 사용된 모델을 생성하기 위해 에버트란은 2021년 민간 특허번역 전문 기관과의 협력으로 확보한 70만 문장의 병렬말뭉치를 인공지능 학습이 가능한 학습데이터로 전환했으며, 이후 데이터를 정제하여 특허 분야 번역 모델을 생성하였다. 다음으로 전문 특허번역사로서 20년 이상의 경험이 있는 전문가 5인의 평가를 통해 번역 모델의 품질을 지속해서 개선했으며, 에버트란 내부적으로 평가를 진행하였다. 그 결과 BLEU 스코어 평가로는 획득하기 어려운 80점 이상을 획득하는 등 영한 특허의 번역패턴이나 용어를 매우 정확하게 AutoML 모델을 통해 구현할 수 있다는 가정을 지지하는 초기 데이터를 확보할 수 있었다.

3.2 분석 방법

본 연구의 핵심 목표는 주제 영역에 특화된 기계번역엔진이 범용 기계번역엔진 대비 더 높은 품질의 기계번역 결과물을 출력할 수 있으며, 이를 통해 특허번역 업무 효율성 개선에 도움을 줄 수 있는가를 고찰하는 것이다.

텍스트의 선정에 있어 특허번역 업무의 대표성을 지니는 텍스트 선정이 필요하기에, 특허 처리 과정에서 번역 업무가 많이 발생하는 특허 명세서로 범위를 한정하였으며 세부 전문 분야를 전기·전자로 한정하였다. 본 연구에서 사용

한 AutoML Translation 모델은 명세서를 위주로 학습이 진행되었으며, 범용 엔진과의 비교 분석을 위해 동일 주제 영역에서 무작위 샘플 선정을 실시하였다. 에버트란이 초기 분석에 사용한 문장 100문장(이하 데이터 세트 1)과 두 명의 공동연구자가 추가로 100문장(이하 데이터 세트 2)을 선정하여 총 200개의 영어 문장을 분석의 대상으로 삼았다. 데이터 세트 2의 경우 텍스트 선정의 대표성을 위해 전기·전자의 대표 하위 분야인 반도체 관련 특허 신청이 많은 대표적인 다국적 반도체 기업 8개를 선정한 이후, 특허 정보를 찾을 수 있는 구글 특허, WIPO, 유럽 특허청, 미국 특허청의 웹페이지를 통해 8개 기업에서 출원한 특허공보 형태의 영문 명세서를 여러 건 검색한 후 특허공보의 전 부분에서 두 명의 연구자가 각각 특허 문서의 전형성을 지니는 50문장의 샘플링을 2022년 7월 실시하였다. 샘플링 시 명세서를 구성하는 배경기술, 발명의 상세한 설명, 발명의 효과 등 여러 부분에 속하는 문장을 골고루 추출하여 평가를 위한 문장이 명세서의 특정 부분에 치우치지 않도록 함으로써 연구 결과를 일반화할 수 있도록 하였다.

본 연구에서는 자동화된 평가와 인간의 수동평가의 균형을 잡기 위하여 다음과 같은 세 가지 평가 방법을 사용하였다. 첫째, 인간번역과 기계번역의 유사도를 나타내는 BLEU 스코어를 문장 단위로 비교하여 구글번역과 AutoML Translation의 인간번역과 유사도를 비교한다. 이를 위해 특허번역을 다년간 진행해온 특허 전문 번역사 1인이 200문장에 대한 번역을 시행하였고 해당 결과물을 참조 번역으로 사용하였다. 물론 BLEU 스코어가 절대적인 품질의 지표라고 보기는 어렵다. 하지만 BLEU 스코어가 높다면 최소한 포스트에디팅 작업의 기술적 및 인지적 노력이 감소할 가능성이 크다고 볼 수 있다. 둘째, 기계번역의 성능이란 한 모델에서 나오는 오류가 다른 모델에서 동일하게 발생하는가 아니면 해결할 수 있는가로 판단할 수 있다. 이를 위해 데이터 세트 1, 세트 2에서 각각 전반부 50문장에 대해 구글번역의 오류 식별하였고, 후반부 50문장에서는 AutoML Translation의 오류 식별 작업을 진행하였다. 이후 구글번역에서 발생한 오류가 AutoML Translation의 해결되었는지, 반대로 AutoML Translation의 오류가 구글번역에서 해결되었는지 여부를 정량화하여 두 모델의 성능을 가늠해 보고자 시도하였다. 셋째, 특허 명세서의 특성과 기계번역의 한계를 고려한 수동 분석을 진행하였다. 특허 문서의 특성상 주술 구조가 복잡하

고 문장의 길이가 긴 경우가 다수 있으며, 이러한 특성은 기계번역의 결과물 출력에 부정적으로 작용할 가능성이 크다. 따라서 비교적 단순한 단문과 문장 구조가 복잡하고 길이가 긴 장문을 위주로 샘플링을 진행하여 단문인 15단어까지의 42개 문장과 장문인 33단어 이상의 36개 문장의 총 78개 문장에 대해서 품질 평가를 수행하였다. 이와 같이 샘플링을 기반으로 한 수동평가를 통해 AutoML Translation과 구글번역 간 단문과 장문을 처리함에 있어 차이가 있는지 확인해 보고자 한다.

4. 분석 결과

4.1 BLEU 스코어를 활용한 분석

BLEU 스코어는 인간번역과 기계번역의 유사성을 살펴볼 수 있는 지표 중의 하나이다. 구글번역과 AutoML Translation의 BLEU 스코어 평균은 데이터 세트 1에서 AutoML Translation이 유의미하게 높았다(평균 88.05 : 67.29, 표준편차 13.53 : 16.26, $p<0.0001$, $t=9.813$). 문장 단위 스코어의 비교에서도 AutoML Translation은 구글번역 대비 92문장에서 더 높은 BLEU 스코어를 보였으며, 반면 구글이 AutoML Translation 대비 더 높은 BLEU 스코어를 보인 문장은 6문장에 불과했다.

〈표 1〉 데이터 세트 1 BLEU 스코어 비교

구분	AutoML Translation	구글번역
평균	88.05	67.29
AutoML BLEU > 구글번역 BLEU		92문장
AutoML BLEU = 구글번역의 BLEU		6문장
AutoML BLEU < 구글번역 BLEU		2문장

반면 데이터 세트 2에서는 구글번역과 AutoML Translation은 각각 53.87과 60.42의 BLEU 스코어 평균을 기록하였다. 괄목할 부분은 BLEU 스코어의 평균이 데이터 세트 1 대비 확연히 낮아졌다는 점과 데이터 세트 2에서 두 번역

엔진 간의 평균차가 크지 않다는 점이다. 그럼에도 불구하고 두 그룹의 BLEU 스코어 평균은 통계적으로 유의미한 차이가 있었다(표준 편차 12.2 : 14.95, $p<0.0009$, $t=3.38$). 여기에 더해 문장 단위의 BLEU 스코어 비교에서도 AutoML Translation은 구글번역 대비 더 높은 BLEU 스코어를 68문장에서 보였다. 반면 구글이 AutoML Translation 대비 더 높은 BLEU 스코어를 보인 문장은 26문장에 불과했다.

〈표 2〉 데이터 세트 2 BLEU 스코어 비교

구분	AutoML Translation	구글번역
평균	60.42	53.87
AutoML BLEU > 구글번역 BLEU		68문장
AutoML BLEU = 구글번역의 BLEU		6문장
AutoML BLEU < 구글번역 BLEU		26문장

이상의 결과를 요약하자면 AutoML Translation은 두 번의 테스트 모두에서 구글 대비 높은 BLEU 스코어를 기록하였다. 물론 BLUE 스코어의 우위가 번역의 "정확도"를 의미하지는 않는다. 하지만 본 분석에서 사용한 인간번역이 특허번역의 전형성을 반영했다고 전제하에, AutoML Translation이 구글번역 대비 특허번역의 전형성을 더욱 잘 반영한 결과물을 출력했다고 할 수 있다.

분석 과정에서 상기 주장을 지지하는 예시가 다수 발견되었으며, 가장 대표적으로 어미의 통일성에 대한 문제를 들 수 있다. 아래 예시에서 볼 수 있듯 동일한 주체에 대한 번역이지만 구글번역은 존칭형 사용에 있어 통일성을 보여주지 못한 경우가 자주 관찰되었다.

<구글번역 예시>
벅 컨버터는 전류를 승압하면서 주 전원 공급 장치의 입력과 출력 사이의 전압을 강압하거나 낮추는 DC-DC 전력 변환기입니다.
벅 컨버터는 배터리 또는 일부 다른 유형의 DC 전원일 수 있는 입력 전압 소스(110)를 포함한다. (밑줄은 저자가 표시)

하지만 AutoML Translation은 이러한 어미의 불일치가 전혀 관찰되지 않았

다. 여기서 한 단계 더 나아가 특허 문장에서 사용되는 전형적 종결어미인 '-한다'의 형태를 다수의 문장에서 보여주었다. 여기에 더해 특허 문장에서 자주 사용되는 단어와 구가 AutoML Translation에서 더 자주 사용된 것을 볼 수 있었다.

ST: The machine readable instructions described herein may be stored in one or more of a compressed format.
AutoML: 본원에서 설명된 머신 판독가능 명령들은 압축된 포맷 중 하나 이상으로 저장될 수 있다.
Google: 여기에 설명된 기계 판독 가능 명령어는 압축된 형식 중 하나 이상으로 저장될 수 있습니다.

ST: As used herein, the term "source material" may be apprehended as a material that is evaporated and deposited on a surface of a substrate.
AutoML: 본원에서 사용되는 바와 같이, "소스 재료"라는 용어는, 증발되어 기판의 표면 상에 증착되는 재료로서 이해될 수 있다.
Google: 본 명세서에서 "소스 물질"이라는 용어는 기판의 표면에 증발되어 증착되는 물질로 이해될 수 있다.

ST: Figure 1 illustrates an example DC-DC buck converter.
AutoML: 도 1은 예시적인 DC-DC 벅 컨버터를 예시한다.
Google: 그림 1은 DC-DC 벅 컨버터의 예를 보여줍니다.

(밑줄은 저자가 표시)

상기 관찰은 인간번역 수준의 결과물을 지향하는 풀 포스트에디팅 작업에서 AutoML Translation의 결과물이 수용성 및 스타일 개선 작업의 부담이 상대적으로 적을 가능성을 시사한다. 하지만 기계번역 결과물에 대한 객관적 평가를 위해서는 다양한 평가 방식 그리고 자동과 수동 분석의 조화가 중요하다. 이에 4.2와 4.3에서는 표본 수집에 근거한 수동평가 결과를 제시하고자 한다.

4.2 심각도가 높은 전형적 오류 분석

본 분석에서는 누락, 잘못된 주어 처리, 장문 처리의 취약성, 맥락파악 취약

성 등 이미 알려진 전형적이고 심각도가 높은 오류를 먼저 분석하고자 한다. 이미 알려진 취약점으로 인한 문제가 발생한다는 것은 다른 특허 텍스트에서도 유사한 문제가 발생할 확률이 있다는 뜻이다. 또한, NMT의 성능이 높아지면서 사소한 문제점만 수정하면 “사용 가능한 품질”의 문장을 만들 수 있는 경우가 많다. 하지만 심각도가 높은 문제점이 있다면 더 높은 수준의 기술적 노력 및 인지적 노력이 필요할 것이다. 이후 한 엔진에서 발생한 문제가 얼마나 많으며, 해당 문제가 다른 엔진에서 비교 분석을 진행하여 두 엔진 간의 상대적 성능 우위를 논하고자 한다.

오류의 심각성 분류를 위해 로컬라이제이션 표준화 협회(Localization Industry Standards Association)의 품질 기준을 준용하여 중대 오류는 ‘문장이 완성은 되었으나 독자의 의미 이해를 호도할 가능성이 있어 재작업이 필요한 경우’로 정의하였다. 반면 심각한 오류는 ‘문장이 완성되지 않았거나, 누락이 있거나, 특허의 의도 자체를 변경하는 수준 있어 상당 부분의 다시 쓰기가 필요한 경우’로 정의하였다.

오류 분석을 위해 데이터 세트 1에서는 전반 50문장에서 구글번역의 오류를 식별하고, 후반 50문장에서 AutoML Translation의 오류를 식별하였다. 반대로 데이터 세트 2에서는 전반 50문장에서 AutoML Translation의 오류를 식별하고, 후반 50문장에서 구글번역의 오류를 식별하였다.

〈표 3〉 데이터 세트 1 오류 분석

전반	구글번역 문제점	AutoML 해결 빈도	해결 비율
심각	8	5.5	61%
중대	1	0	
후반	AutoML 문제점	구글번역 해결 빈도	해결 비율
심각	6	1	13%
중대	2	0	

데이터 세트 1의 문제점의 해결 빈도를 보면 구글번역에서 발생한 문제의 다수를 AutoML Translation에서 해결하고 있지만(61%), 구글은 AutoML Translation에서 발생한 문제를 거의 해결하지 못하고 있다(13%). 문제의 발생 빈도 역시 구글번역이 상대적으로 더 높았다. 상기 결과에 추가적으로 심각한

오류 2점, 중대한 오류에 1점의 가중치를 부여할 경우, 17대 14로 구글번역이 더 큰 문제점을 보였다.

이와 같은 AutoML Translation의 상대적 우세는 데이터 세트 2의 분석에서도 계속되었다.

〈표 4〉 데이터 세트 1 오류 분석

전반	구글번역 문제점	AutoML 해결 빈도	해결 비율
심각	7	5	50%
중대	6	1.5	
후반	AutoML 문제점	구글번역 해결 빈도	해결 비율
심각	1	0.5	40%
중대	9	3.5	

데이터 세트 2의 문제점 해결 빈도를 보면 구글번역에서 발생한 문제의 절반가량을 AutoML Translation이 해결하고 있지만(50%), 구글은 상대적으로 낮은 해결률을 보였다(40%). 문제의 발생 빈도 역시 구글번역이 상대적으로 더 높았으며, 심각도에 따른 가중치를 부여한다면 20대 11로 구글번역이 더 큰 문제점을 보였다.

가중치를 적용한 데이터 세트 1과 2의 정량적 결과를 종합하자면, 구글의 문제를 AutoML Translation이 해결한 비율은 60%이며, 반대로 AutoML Translation의 문제점을 구글번역이 해결한 비율은 26%에 불과하다. 여기에 더해 전체 심각 및 중대 오류의 가중치 합산 역시 AutoML Translation은 25이지만 구글번역은 37을 기록하였다. 상기 데이터는 AutoML Translation은 작업자의 많은 개입을 요하는 심각한 오류 발생의 빈도가 상대적으로 낮으며, 구글번역에서 발생하는 오류가 발생하지 않을 가능성이 있다고 해석할 수 있다. 따라서 포스트에디팅을 수행하는 작업자로서는 문제의 절대 빈도가 낮기에 문제 식별에 들어가는 노력, 그리고 문제 일부분이 해결되었기에 해결책 제안에 필요한 노력이 낮아질 수 있을 것이다.

4.3 단문 및 장문 샘플링 수동평가 결과

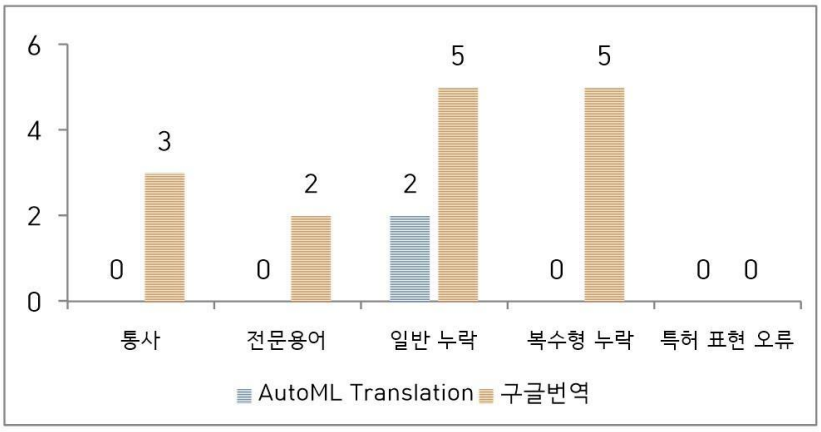
본 분석에서는 수동평가로 샘플링 분석을 진행하였으며 장문이 주요한 특징을 이루는 특허 명세서의 문장 특징을 반영하여 단문 42개 문장과 장문 36개 문장의 번역 품질을 살펴보았다.

단문은 15단어 이내의 문장을 말하며, 구체적인 분석 대상은 데이터 세트 1에서 18개 문장과 데이터 세트 2에서 24개 문장의 총 42개 문장이다. 데이터 세트 1의 18개 문장에 대한 AutoML Translation과 구글번역의 오류 차이를 살펴보면, 우선 구글번역에서 오류 없는 문장, 즉 정확하게 번역한 문장이 전체 18개 문장 중 단 5개 문장으로 27.8%의 정확성을 보인 반면, AutoML Translation에서는 오류 없는 문장이 16개 문장으로 무려 88.9%의 정확성을 보였다. <표 5>는 오류 문장과 오류 없는 문장 간 비중 및 전체 오류의 총합을 제시한다. 또한 구체적인 오류 분류별 오류 개수는 아래의 <그림 3>과 같다.

〈표 5〉 데이터 세트 1 단문의 오류 비교

분류	AutoML Translation	구글번역
오류 있음	2문장(11.1%)	13문장(72.2%)
오류 없음	16문장(88.9%)	5문장(27.8%)
오류의 총합	2(문장당 0.11개 오류)	15(문장당 0.83개 오류)

〈그림 3〉 데이터 세트 1 단문의 오류 개수 비교

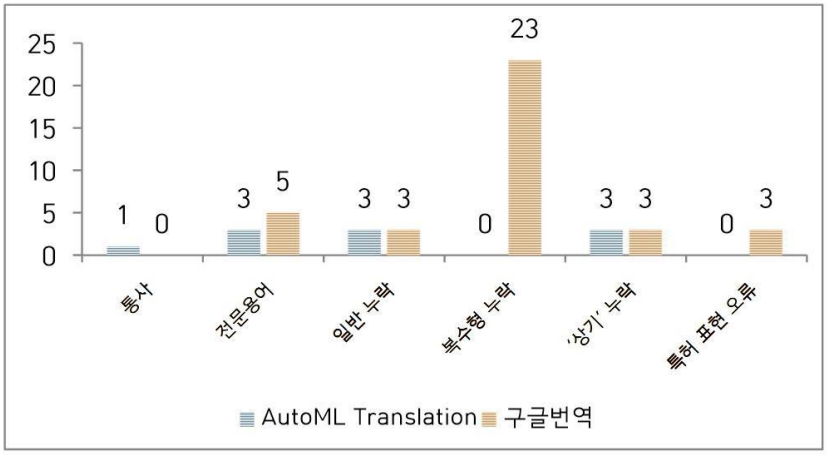


데이터 세트 2의 단문 24개 문장에 대한 AutoML Translation과 구글번역의 오류 차이 또한 데이터 세트 1에서와 마찬가지로 AutoML Translation이 우위였다. 구글번역에서는 오류 없는 문장이 24개 문장 중 5개 문장에 불과하였으며, 이에 반해 AutoML Translation에서는 오류 없는 문장이 14개였다. 한편 데이터 세트 1과 비교하면 데이터 세트 2에서 오류 없는 문장의 비중이 줄었으며, 오류의 총합이 늘어나는 경향을 보였다. <표 6>은 오류 문장과 오류 없는 문장 간 비중 및 전체 오류의 총합을 제시한다. 또한 구체적인 오류 분류별 오류 개수는 아래의 <그림 4>와 같다.

〈표 6〉 데이터 세트 2 단문의 오류 비교

분류	AutoML Translation	구글번역
오류 있음	10문장(41,7%)	19문장(79.1%)
오류 없음	14문장(58.3%)	5문장(20.9%)
오류의 총합	10(문장당 0.41개 오류)	37(문장당 1.54개 오류)

〈그림 4〉 데이터 세트 2 단문의 오류 개수 비교



단문의 번역을 분석한 결과, 데이터 세트 1, 데이터 세트 2 모두 구글번역에서 오류 없는 문장, 즉 정확하게 번역한 문장이 전체 문장에서 차지하는 비율이 20%대를 넘지 못했다. 반면에 AutoML Translation은 데이터 세트 1에서

88.9%의 정확성을 보였으며 데이터 세트 2에서는 조금 저조하기는 하나 구글번역에 비해서는 월등하게 높은 58.3%의 정확성을 보였다.

데이터 세트 1의 경우를 보면, 구글번역에서 '통사', '전문용어'의 오류가 각각 3건과 2건으로 집계되어 문장의 정확성에 영향을 미칠 수 있는 중대한 오류가 포착되는 반면 AutoML Translation에서는 이와 같은 중대한 오류가 단 한 건도 집계되지 않아 데이터 세트 1에 한해서 AutoML Translation은 상당히 정확하게 문장을 번역한 것으로 추론할 수 있다.

데이터 세트 2의 경우를 보면, 구글번역에서 '전문용어'의 오류가 5건으로 집계되었고, AutoML Translation의 경우 '통사'와 '전문용어'가 각각 1건, 3건으로 집계되었다. 데이터 세트 1과 비교해서 데이터 세트 2에서 AutoML Translation의 정확성을 저해할 만한 중대한 오류가 눈에 띄게 증가한 것은 사실이나 '전문용어'의 차원에서는 구글번역에 비해 AutoML Translation의 오류 개수가 적은 것으로 보아 전문용어의 정확성에 있어서 AutoML Translation이 좀 더 우위를 점하고 있는 것으로 보인다.

데이터 세트 1과 데이터 세트 2에서 나타나는 AutoML Translation과 구글번역 간 대표적인 차이점은 구글번역에서 복수표지자 '-들'의 누락이 빈번하다는 점이다('특허 정확성' 중 '복수형 누락'). 즉, 원문의 명사 복수형을 AutoML Translation은 '-들'의 동일한 복수형으로 번역하는 반면 구글번역은 이를 단수형으로 옮기는 경우가 빈번했다. 한편 복수표지자 '-들'은 실무에서 대표 단수의 사용 등으로 빈번하게 누락하고 번역하는 바 번역의 품질에 영향을 미치는 치명적인 오류라고는 보기는 어렵다.

데이터 세트 2에서 구글번역과 AutoML Translation은 공통적으로 원문의 'the'를 따로 번역하지 않고 누락하는 경향을 보였다('특허 정확성' 중 '상기 누락'). 특허번역 실무에서 'the'는 통상 앞서 나온 구성요소를 지칭하는 의미로 'the'를 반드시 '상기(의)'로 번역 및 기입하는 관행을 고려할 때, 구글번역과 AutoML Translation 모두 이와 관련해서 개선의 여지가 있어 보인다.

데이터 세트 2에서 구글번역은 특허 특유의 단어나 구를 관행대로 번역하지 않는 오류를 3건 보였다('특허 정확성' 중 '특허 표현 오류'). 데이터 세트 1과 2 모두에서 '특허 표현 오류'가 단 한 건도 발견되지 않은 AutoML Translation에 비해 구글번역은 특허 특유의 관용어구 번역에 있어서 AutoML

Translation에 비해 덜 정제되어 있을 수 있음을 시사한다.

전체적으로 15단어까지의 비교적 단문에서는 구글번역, AutoML Translation 모두 문장의 정확성을 해칠 수준의 심각한 오류는 비교적 적은 것으로 보이며, 비록 오류 문장들이 존재하나 구글번역, AutoML Translation 모두 상당히 정확한 수준으로 번역되었다는 것을 알 수 있다.

다음으로 장문은 33단어 이상의 문장을 말하며 본 연구에서는 총 36개 문장을 분석하였다. 구체적으로 데이터 세트 1의 23개 문장과 데이터 세트 2의 13개 문장을 분석하였다.

데이터 세트 1의 23개 장문에 대해서, AutoML Translation은 총 10개의 오류를 보였으며 오류가 없이 정확한 문장은 14개로 정확도는 60.9%였다. 반면에 구글번역은 총 39개 오류를 보였으며 오류가 없는 문장은 6개로 26%의 정확도를 보였다. 구체적인 오류 분류는 아래 <표 7>과 같다.

〈표 7〉 데이터 세트 1 장문의 오류 비교

분류		AutoML Translation	구글번역
추가		1개	6개
누락		1개	1개
통사		8개	15개
전문용어		0개	5개
특허 정확성	복수형 누락	0개	9개
	‘상기’ 누락	0개	0개
	특허 표현 오류	0개	0개
오류 없음		14문장(60.9%)	6문장(26%)
오류의 총합		10(문장당 0.43개 오류)	39(문장당 1.7개 오류)

데이터 세트 2의 장문 13개 문장에 대한 AutoML Translation과 구글번역의 오류 차이 또한 데이터 세트 1에서와 마찬가지로 AutoML Translation이 우위였다. 구글번역에서는 오류 없는 문장이 13개 문장 중 1개 문장에 불과하였으며, 이에 반해 AutoML Translation에서는 오류 없는 문장이 6개였다. 구체적인 오류 분류는 아래 <표 8>과 같다.

〈표 8〉 데이터 세트 2 장문의 오류 비교

분류		AutoML Translation	구글번역
추가		0개	0개
누락		0개	1개
통사		4개	6개
전문용어		2개	3개
특허 정확성	복수형 누락	0개	12개
	'상기' 누락	2개	2개
	특허 표현 오류	1개	1개
오류 없음		6문장(46.1%)	1문장(7.7%)
오류의 총합		9(문장당 0.69개 오류)	25(문장당 1.92개 오류)

데이터 세트 1과 데이터 세트 2의 장문을 분석한 결과, 전체적으로 15단어까지의 단문에 비해 정확성은 떨어지고 문장당 오류의 개수는 늘어나는 추세임을 알 수 있다. 한편 데이터 세트 1과 2에서 모두 구글번역의 오류 없는 문장, 즉 정확하게 번역한 문장이 전체 문장에서 차지하는 비율은 최대 26%로 20% 대를 넘지 못하며 특히 데이터 세트 2에서는 7.7%로 정확성이 매우 낮은 편이다. 반면 AutoML Translation에서는 단문에 비해서 그 품질이 저조하기는 하나 구글번역에 비해서 데이터 세트 1과 2 모두에서 월등하게 높은 정확성을 보이고 있음을 알 수 있다.

장문에서는 단문과 비교했을 때, 데이터 세트 1과 데이터 세트 2에서 공통적으로 '추가'와 '누락'의 오류 유형이 추가되었다. 추가와 누락은 번역의 정확성을 가늠하는 주요한 기준으로 이와 같은 오류 유형의 추가는 곧 정확성이 단문에 비해 장문에서 전반적으로 저해되었음을 시사한다. 구글번역과 AutoML Translation을 비교해보면, 데이터 세트 1에서 구글번역의 추가가 6건, 누락이 1건인데 반해 Auto ML은 추가가 1건, 누락이 1건으로, 구글번역이 AutoML Translation에 비해 원문에는 없는 내용을 번역에서 추가하는 경향을 보이며 이로 인해 번역의 정확성이 낮아질 수 있음을 시사한다. 데이터 세트 2에서는 구글번역에서 누락만 1건으로 집계되었으며, AutoML Translation의 경우 본 샘플링 분석에서 사용한 13개 문장에서는 추가 또는 누락의 이슈가 없었다. 데이터 세트 1과 데이터 세트 2 모두에서 추가와 누락의 오류 개수를 정량적으로 비교 분석한 결과, AutoML Translation이 구글번역에 비해 추가와 누락 측면에서 우

위를 점한다.

또한 단문과 비교했을 때 데이터 세트 1과 데이터 세트 2에서 구글번역과 AutoML Translation 모두 ‘통사’의 문제가 빈번하게 나타나는 것으로 드러났다. 구글번역은 단문 데이터 세트 1에서 총 3건, AutoML Translation은 단문 데이터 세트 2에서 총 1건의 오류를 보인 반면, 장문 중 데이터 세트 1에서 구글번역은 총 15건, AutoML Translation은 총 8건의 오류를 보였으며, 데이터 세트 2에서 구글번역은 총 6건, AutoML Translation은 총 4건의 오류를 보였다. 특허에서는 특히 문장이 길어질수록 원문의 수식 관계를 비롯한 전반적인 통사 구조가 매우 복잡해지는데, 구글번역과 AutoML Translation 모두 원문의 이와 같이 복잡한 통사 구조를 정확하게 분석하지 못하는 사례가 빈번해진 것으로 보인다. ‘통사’는 추가, 누락과 같이 번역의 정확성을 가늠하는 주요한 지표 중 하나다. 통사와 관련된 오류가 빈번해지면서 번역의 정확성이 단문에 비해 낮을 수 있음을 알 수 있다. 전반적인 추세는 이러하나 구글번역과 AutoML Translation을 비교해보면, AutoML Translation의 정량적인 오류의 수가 데이터 세트 1과 데이터 세트 2에서 모두 구글번역의 오류 수에 비해 적어 AutoML Translation의 원문 정확도 및 통사 이해도가 구글번역에 비해 좀 더 높은 것으로 판단할 수 있다.

‘전문용어’ 오류의 경우, 단문에서의 오류 개수와 그 수가 크게 차이가 나지는 않는다. 하지만 데이터 세트 1과 데이터 세트 2에서 모두 AutoML Translation이 구글번역에 비해 적은 수의 오류를 기록하여 전문용어 번역의 정확성이 좀 더 높은 것을 확인할 수 있다. 특히 데이터 세트 1에서 Auto ML의 전문용어 오류 개수는 0으로 전문용어 번역의 정확성이 상당한 수준임을 알 수 있다.

장문에서도 구글번역은 단문에서와 마찬가지로 복수표지자 ‘-들’을 누락하고 번역하는 경향을 보인다. 반면 이와 같은 오류가 AutoML Translation에서는 단 한 건도 발견되지 않았다. 이에 반해 원문의 ‘the’를 ‘상기(의)’로 번역하는 문제와 관련해서 데이터 세트 2의 오류를 살펴보면, 구글번역과 AutoML Translation 모두 ‘상기(의)’를 누락하는 경향을 보였다. 따라서 ‘the’의 번역에 한해서 AutoML Translation이 구글번역에 비해 우위에 있다고 보기 어려우며 두 엔진에서 공통적으로 문제가 됨을 알 수 있다.

데이터 세트 1에서 구글번역은 특허 특유의 단어나 구를 관행대로 번역하지 않는 오류를 3건 보였다('특허 정확성' 중 '특허 표현 오류'). 단문에서와 마찬가지로 데이터 세트 1에서 '특허 표현 오류'가 단 한 건도 발견되지 않은 AutoML Translation에 비해 구글번역은 특허 특유의 관용어구 번역에 있어서 AutoML Translation에 비해 덜 정제되어 있을 수 있음을 시사한다. 다만 데이터 세트 2에서 '특허 표현 오류'에 있어 구글번역과 AutoML Translation이 동률인 점은 AutoML Translation이 여전히 개선의 여지가 있을 수 있음을 시사한다고 볼 수 있다.

전체적으로 15단어까지의 단문에 비해 33단어 이상의 장문에서 번역의 정확성이 좀 더 낮은 편이며 두 번역엔진에서 모두 원문의 의미를 정확하게 전달하지 못하는 번역이 긴 문장에서 좀 더 두드러지는 것을 알 수 있다.

무엇보다 구글번역, AutoML Translation 모두 장문에서 '통사' 오류가 문장의 정확성을 저해하는 수준의 심각한 오류로 꼽을 수 있을 것으로 보인다. 통사의 문제만을 따로 떼어 보면, 데이터 세트 1에서 구글번역의 오류 문장이 총 17개인데 통사 오류가 15개, AutoML Translation에서 오류 문장이 총 9개인데 통사 오류가 8개로 두 엔진 모두 오류 문장 대부분이 통사 오류를 가지고 있는 것으로 드러났다. 한편 데이터 세트 2의 경우, 구글번역의 오류 문장이 총 12개인데 통사 오류 6개, AutoML Translation의 오류 문장이 총 7개인데 통사 오류 4개로 데이터 세트 1에 비해서 통사 오류를 가지는 문장의 비중이 낮아지기는 했으나 여전히 상당수의 문장에서 통사 오류를 발견할 수 있다.

또한 단문에서는 찾아볼 수 없었던 추가, 누락의 문제가 장문에서 나오고 있다는 점 즉, 단문과 장문의 오류 양상이 다르다는 점을 주지할 만하다. 장문에서는 공통적으로 통사, 추가, 누락 문제가 두드러지면서 정확성의 문제가 단문에 비해 좀 더 심각한 양상을 띠고 있음을 알 수 있다. 다만 오류 개수와 오류 없는 문장의 개수에서 AutoML Translation이 구글번역에 비해 좀 더 우위를 점하고 있어 미세하게나마 상대적으로 AutoML Translation이 장문에서도 범용엔진에 비해 좀 더 정확한 번역을 수행하고 있음을 알 수 있다.

5. 논의 및 결론

5.1 연구 요약

본 연구에서는 특허에 특화한 AutoML Translation과 범용 엔진인 구글번역에 반도체를 주제로 한 특허 명세서의 영어 문장 200문장에 대한 한국어 번역을 분석 대상으로 선정하여 자동평가와 수동평가를 병행, 특화 엔진과 범용 엔진간 품질의 차이를 비교하여 보았다.

BLEU 스코어 평가 결과, 데이터 세트 1과 데이터 세트 2의 200문장 모두에서 AutoML Translation이 구글번역에 비해 더 높은 BLEU 스코어를 보였으며, 두 번역엔진 간의 스코어 평균은 유의하게 차이가 있음을 t-검정을 통해 살펴보았다. 다만 BLEU 스코어로만 품질을 판단하기에는 한계가 있다는 판단 하에 본 연구진은 두 가지 방법으로 수동평가를 실시하였다.

첫 번째 수동평가는 심각도가 높은 전형적 오류를 찾아서 어떻게 해결했는지를 분석하는 방법으로, 특정 엔진이 상대 우위가 있다면 더 높은 품질을 보일 것으로 추정할 수 있다. 누락, 잘못된 주어 처리, 장문 처리의 취약성, 맥락파악의 취약성 등 전형적이고 심각도가 높은 오류를 로컬라이제이션 표준화 협회의 품질 기준에 따라 중대 오류와 심각한 오류로 분류하여 분석하였다. 분석을 위해서 데이터 세트 1에서는 전반 50문장에서 구글번역 오류를 식별하였으며, 후반 50문장에서는 AutoML Translation의 오류를 식별하였다. 데이터 세트 2는 1과 반대로 오류를 식별하는 방법을 취해 분석을 시행하였다. 분석 결과, 데이터 세트 1과 2 모두에서 구글번역에서 발생한 문제의 상당수가 AutoML Translation에서 해결되었지만, 구글번역은 AutoML Translation에서 발생한 문제 대부분을 해결하지 못했다. 여기에 심각한 오류 2점, 중대한 오류 1점의 가중치를 부여하여, 오류 및 오류 해결을 정량화한 AutoML Translation의 상대적 우위가 더욱 크게 부각되었다.

두 번째 수동평가는 200문장 중 15단어 이내의 단문 42개 문장과 33단어 이상의 36개 문장의 총 78개 문장을 샘플링하여 이 문장들을 대상으로 단문 대비 특허 특유의 복잡한 문장 구조를 지닌 장문에 대해서 특화 엔진과 범용 엔진 간 결과물의 품질에 차이가 있는지 알아보았다. 분석 결과, 단문에서는

AutoML Translation, 구글번역 모두 정확성을 해칠 수준의 심각한 오류는 비교적 적은 것으로 나타났다. 하지만 오류 없는 문장, 즉 완벽한 문장의 비율과 오류의 개수 면에서 AutoML Translation이 구글번역에 비해 완성도가 높은 것으로 드러났다. 장문에서는 두 엔진 모두 단문에 비해 정확성과 관련된 추가와 누락 오류가 더해졌으며 통사 오류 또한 눈에 띄게 많아졌다. 그만큼 장문에서는 아직까지 특화 엔진, 범용 엔진 모두 인간번역을 대체하기에는 어려운 점이 많은 것으로 보인다. 다만 단문에서와 마찬가지로 오류 없는 문장의 비중과 오류의 개수 면에서 AutoML Translation이 구글번역에 비해 완성도가 높았다.

이상과 같이 자동평가와 오류분석 및 샘플링을 활용한 수동평가 결과, AutoML Translation이 구글번역에 비해 전반적으로 영한번역의 품질에 있어서 우위에 있음을 알 수 있었다. 이러한 결과는 특화 엔진을 활용해서 좀 더 후속 작업이 용이한, 정확성과 완성도가 높은 번역물을 도출할 수 있다는 점을 시사한다. 이와 같은 연구 결과로 전문 분야에서 특화 엔진의 더욱 활발한 사용을 기대할 수 있겠다.

5.2 추가 논의

특히 본 연구에 사용된 기계번역 품질 개선 접근법은 두 가지 시사점을 지닌다. 첫째, 본 연구에 사용된 기계번역 학습 방법은 상대적으로 기술적 접근 장벽이 낮은 개방형 도구를 사용했다는 점이다. 달리 말해 기계번역엔진의 데이터와 작동 기전을 개발자들이 독점이 아닌 비개발자와 사용자에게 참여의 기회가 열리는 개념의 전환으로 볼 수 있다. 둘째, 본 연구는 학습데이터의 양도 중요하지만 학습데이터의 질적 개선에 집중하였다. 이를 위해 해당 영역의 전문번역사가 장시간 참여했으며, 이는 번역 전문가와 번역학 연구자들이 수동적 사용자가 아닌 능동적 참여자로 기계번역 품질 개선에 기여할 가능성을 보여주었다는 점에서 의미가 있다. 셋째, 박찬준 외(2020) 등과 같이 기계번역 학습 후 품질 향상을 분석할 때 대부분 BLEU 스코어를 비롯한 자동평가 방법을 활용하고 있지만, 본 연구는 자동평가 외에 오류 분석, 샘플링 분석 등 다양한 수동평가의 방법을 고안 및 적용하였다는 점에서 차별된다. 자동평가뿐만 아니라 다양한 관점의 수동평가를 적용함으로써 번역 결과를 다각도로 분석하여 그 품

질의 차이를 도출하였다는 데 본 연구의 의의가 있다고 할 수 있겠다.

다만 본 연구가 200문장에 한정해서 평가가 진행되었다는 점은 본 연구 결과를 일반화하는 데에는 한계가 있음을 시사한다. 또한 자동평가 결과와 수동평가 결과 모두에서 AutoML Translation이 데이터 세트 1과 데이터 세트 2 간 점수 차이가 있다는 점 또한 본 연구의 한계일 수 있겠다. 여기에 더해 본 연구에서 사용한 데이터는 2022년 7월 기준의 결과물이기 때문에, 본 연구의 분석이 미래 시점에는 유효하지 않을 수 있다.

하지만 본 연구는 데이터 세트에 상관없이 고른 품질의 번역문을 도출하는 데 충분하면서도 현실적인 학습용 데이터의 양에 대해서도 생각해 볼 기회가 된다. 현실적으로 특히 통사 구조가 복잡한 장문에 대해서 특화 엔진, 범용 엔진 모두 정확성 면에서 품질이 아쉽다는 점은 아무리 NMT가 진화했다고 해도 특화 엔진조차도 인간 번역을 대체하는 데에는 무리가 있으며, 포스트에디팅과 같은 기계번역 후 작업의 중요성을 부각한다. 이와 같은 측면에서 특화 엔진에 기반한 번역 생산성 및 생산성과 품질의 상관관계 연구 등도 향후 과제로 진행해 볼 수 있겠다.

참고문헌

김세린, 권혁철 (2022)「도메인 지식에 특화된 신경망 기계번역」,『한국정보과학회 학술발표논문집』678-680.

김정희, 허재무, 김주환, 최희열 (2022)「한국어 자모 단위 구성과 높임말을 반영한 한영 신경 기계번역」,『정보과학회논문지』49(11): 1017-1025.

박찬준, 김경희, 박기남, 임희석 (2020)「Coronavirus Disease-19(COVID-19)에 특화된 인공신경망 기계번역기」,『한국융합학회논문지』11(9): 7-13.

박찬준, 임희석 (2020)「공공 한영 병렬 말뭉치를 이용한 기계번역 성능 향상 연구」,『디지털융복합연구』18(6): 271-277.

이준호 (2022)「법률 특화 번역엔진 성능 평가－한영 계약서 번역을 중심으로」,『T&I Review』12: 169-192.

이지은, 최효은 (2023)「인공신경망 기반 맞춤형 기계번역엔진의 성능 평가: 법

률 및 특허 한영번역 결과물 평가 사례를 중심으로」, 『번역학연구』 24(1): 9-37.

정영준, 박천음, 이창기, 김준석 (2020) 「MASS 와 상대 위치 표현을 이용한 영어－한국어 신경망 기계번역」, 『정보과학회논문지』 47(11): 1038-1043.

최승권 (2007) 「영어 특허문서 자동번역을 위한 특허번역패턴 연구」, 『번역학연구』 8(1): 301-322.

최효은, 이지은 (2017) 「특허 기계번역 결과물의 평가－KIPRIS의 무료 한영 기계번역을 중심으로」, 『통역과 번역』 19(1): 139-178.

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

Etchegoyhen, Thierry, Ana Fernández Torné, Andoni Azpeitia Zaldua, Eva Martínez García and Anna Matamala (2018) 'Evaluating Domain Adaptation in Machine Translation Across Scenarios', *Proceedings of the Eleventh International Conference on Language Resources Evaluation (LREC 2018)*, 6-15.

Guo, Xinze, Chang Liu, Xiaolong Li, Yiran Wang, Guolian Li, Feng Wang, Zhitao Xu, Liuyi Yang, Ma Li and Changliang Li (2019) 'Kingsoft's Neural Machine Translation System for WMT19', *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 196-202.

Kinoshita, Satoshi, Tadaaki Oshio and Tomoharu Mitsuhasi (2017) 'Comparison of SMT and NMT Trained with Large Patent Corpora: Japio at WAT2017', *Proceedings of the 4th Workshop on Asian Translation,* 140-145.

Rossi, Laura and Dion Wiggins (2013) 'Applicability and Application of Machine Translation Quality Metrics in the Patent Field', *World Patent Information* 35: 115-125.

Tsai, Yvonne (2017) 'Linguistic Evaluation of Translation Errors in Chinese–English Machine Translations of Patent Titles', *Forum* 15(1): 142-156.

Wan, Yu, Baosong Yang, Derek Fai Wong, Lidia Sam Choa, Liang Yao, Haibo Zhang and Boxing Chen (2022) ‘Challenges of Neural Machine Translation for Short Texts’, *Computational Linguistics* 48(2): 321-342.

Wang, Xu, Chen Chunyang and Zhenchang Xing (2019) ‘Domain-specific Machine Translation with Recurrent Neural Network for Software Localization’, *Empirical Software Engineering* 24: 3514-3545.

WIPO (2022) Executive Summary: PCT Yearly Review 2022 Available at https://www.wipo.int/edocs/pubdocs/en/wipo-pub-901-2022-exec-summary-en-patent-cooperation-treaty-yearly-review-2022-executive-summary.pdf.

<인터넷 자료>

특허뉴스 (2022) [이슈] 글로벌 IP번역 품질 경쟁 속, 최고의 IP번역 마스터 가려졌다. Available at https://www.e-patentnews.com/8398.

[Abstract]

Evaluation of Patent English-Korean Machine Translations by a Patent-Specific NMT Engine Using AutoML

Hyoeun Choi, Chung-ho Lee & Jun-ho Lee
(Ewha Womans University, Evertran, Chung-Ang University)

This paper compares the quality of English-Korean patent translations by a patent-specific NMT engine trained using AutoML with the general Google Translate. The evaluation was based on both automatic and human evaluations of the Korean translations of 200 English patent sentences excerpted from a number of semiconductor patent gazettes. In automatic evaluation, BLEU scores showed that the patent-specific NMT engine significantly outperformed Google Translate. Human evaluation, carried out by sampling as well as error detection and correction analysis, confirmed the results of automatic evaluation, revealing that patent-specific NMT results were better than Google Translate results. In the error detection and correction analysis, Google Translate had more major errors than patent-specific NMT. Moreover, most errors in Google Translate were addressed in the patent-specific NMT, while errors in the patent-specific NMT still remained in Google Translate. In the sampling analysis, shorter sentences and longer sentences were sampled and analyzed. According to the results, both patent-specific NMT and Google Translate showed better performance in translating shorter sentences. In translating longer sentences, both translation engines exhibited accuracy-related errors and syntactic errors, though patent-specific NMT slightly outperformed Google Translate. Overall, translation results by patent-specific NMT showed better quality than those by Google Translate.

Keywords: machine translation, patent translation, AutoML, BLEU, sampling, MT evaluation
주제어: 기계번역, 특허번역, AutoML, BLEU, 샘플링, 기계번역평가

최효은(1저자)
이화여자대학교 통역번역대학원 초빙교수
hyoeun.choi@ewha.ac.kr
관심 분야: 기계번역, 특허번역, MTPE

이청호(공동저자)
에버트란 대표
john@evertran.com
관심 분야: 기계번역, MTPE, CAT

이준호(교신저자)
중앙대학교 국제대학원 전문통번역학과 조교수
brandon4tni@cau.ac.kr
관심 분야: 기계번역, MTPE, CAT

논문투고일: 2023년 4월 30일
1차 심사 완료: 2023년 6월 5일
2차 심사 완료: 2023년 6월 14일
게재 확정: 2023년 6월 20일