http://dx.doi.org/10.15749/jts.2023.24.1.001

번역학연구
24권 1호
(2023.3.31.)

인공신경망 기반 맞춤형 기계번역엔진의 성능 평가: 법률 및 특허 한영번역 결과물 평가 사례를 중심으로

이 지 은 · 최 효 은
(이화여대)

1. 서론

번역 주체에 따라 인간 개입을 배제한 기계번역 즉, 자동번역(full MT)에서부터 인간의 개입 정도에 따른 인간보조번역(HAMT)과 컴퓨터보조번역(MAHT/CAT)으로 구분 가능하며, 이와 대비되는 개념으로 기계번역을 배제한 인간번역(HT)으로 구분할 수 있다(Hutchins and Somers 1992: 148). 그러나 최근 번역기술의 발전으로 번역메모리와 기계번역의 경계가 불분명해졌다(Moorkens 2017: 471). 이하 본고에서 기계번역은 자동번역을 가리킨다.

인공지능(AI) 기술이 발전함에 따라 인공신경망 기계번역(NMT)에 대한 관심이 높아졌고, 최근 기계번역 품질은 상당히 개선되었지만 품질 면에서 인간번역과 격차가 존재한다(Hassan et al. 2018; Ragni and Nunes Vieira 2022). 품질 관리를 위해 프리에디팅과 포스트에디팅과 같은 작업을 필요로 함에도 불구하고 언론 보도는 기계번역의 성능에 관한 기계번역 업체의 홍보 자료를 그대로 싣기도 하는 등 긍정적인 측면만 부각하는 경향이 있다(Nunes Vieira 2020: 113-114). 당연한 것일 수 있겠지만 기계번역 기술 개발자들과 업체들은 자체

개발한 자사의 기계번역 품질이 상당히 우수한 것으로 평가한다. 마이크로소프트와 구글을 비롯하여 기계번역 개발자가 참여하여 발표한 논문은 대부분 인간 번역과의 유사성만 계산하는 자동평가에만 의존하는 등 신뢰성이 부족한 번역 품질 평가방식을 사용하거나 상호심사를 거치지 않은 한계가 있다(Carmo 2022). 과거 통계기반 기계번역(SMT)이 주를 이룰 때에 비해 NMT에 대한 연구는 크게 늘었고 기계번역 성능평가 연구도 많지만 이 가운데 연구방법상 취약한 연구도 적지 않아 신뢰할 만한 평가방식의 중요성과 번역학자가 참여한 기계번역 성능평가 연구가 적기에 이에 대한 관심을 환기시킨 바 있다(Rivera-Trigueros 2022: 612).

이제는 기계번역에 대한 번역사의 수용 태도나 인식 문제를 넘어 번역사 관점에서 NMT의 성능과 유용성에 대한 연구는 학문적으로나 실무적 차원에서도 필요하다(Ragni and Nunes Vieira 2022: 152). 국내 번역업체들도 앞다퉈 기계번역 기술을 도입하여 서비스를 제공하고 있으며, 산학협력을 통한 기계번역 및 포스트에디팅 교육 관련 연구도 시도되었다(김순미 외 2019; 이준호와 김순미 2022). 기계번역 서비스 업체의 예를 보면, 한컴인텔리전스가 번역업체인 에버트란과 손잡고 AI기반 번역서비스 '나루트랜스랩'을 출시하였으며, 솔트룩스 파트너스는 특허, 법률 서비스를 위한 기업용 자동번역시스템을 개발하였다(백지영 2019; 방은주 2019). 이 분야에서 스타트업의 진출도 눈에 띈다. XL8와 같이 구글보다 규모가 작지만 정제된 데이터를 기초로 학습시킨 엔진을 사용하여 기계번역 같지 않다는 점을 홍보하는 NMT 업체가 있는가 하면(장형태 2022), 전문 분야에 특화된 기계번역 서비스 스타트업도 등장하였다. 베링랩은 법률, 특허 분야에 특화한 번역엔진을 개발하였고, 리용은 AI법률번역 서비스 알파 버전을 공개하였다(임영신과 우수민 2021). AILingo의 '오트란'은 법률 특화 번역엔진이지만 텍스트 유형 선택 메뉴에 법률, 비법률(non-legal), 특허, 회계 분야를 제공하고 있다(안경애 2021).

이같이 유료 기계번역 서비스가 날로 확대되고 있는 상황에서 그간 연구가 미진했던 전문 분야에 해당하는 법률과 특허 문건의 기계번역 품질을 살펴보기 위해 본 연구는 연구자 접근이 허용된 맞춤형 번역엔진 '오트란'[1]의 번역결과

1) 오트란은 AILingo(https://ailinggo.ai/)에서 개발한 기계번역 서비스로 시스트란

물에 대한 평가를 다룬다. 맞춤형 기계번역엔진은 전문분야에 특화된 커스텀 번역엔진을 뜻한다. 본 연구를 소개하기에 앞서 다음 장에서는 법률 및 특허 기계번역을 포함한 기계번역 평가에 관한 선행연구를 간략히 살펴보겠다.

2. 기계번역 평가연구

2.1 NMT 품질 평가

NMT 연구 중 가장 큰 비중을 차지하는 것이 품질 평가에 대한 연구다(이상빈 2019; Ragni and Nuns Vieira 2022). 기계번역 평가는 참조번역인 인간번역과 얼마나 유사한가를 측량하는 자동평가와 인간 평가자에 의한 수동평가 방식이 있다(최효은과 이지은 2017). BLEU는 자동평가에 대표적으로 많이 사용되는 평가방식으로 0에 가까울수록 인간번역과 거리가 멀고, 1에 가까울수록 유사하다는 것을 뜻한다. 자동평가에 사용되는 측정법은 BLEU 수치 외에도 METEOR, NIST 등이 있다(이지은과 최효은 2022a). 자동평가는 실제 기계번역의 품질을 알려주는 지수로 부족하고, 성능에 대해 수동평가와 상이한 결과를 보이는 경우도 있어 수동평가와 상관관계가 확인되지 않은 연구가 많다(한현희 2020).[2] 따라서 자동평가와 수동평가를 병행하여 신뢰할 만한 평가 결과를 도출하는 것이 바람직하다(이준호 2019; 최효은과 이지은 2017: 147; Chatzikoumi 2020: 158).

한편 수동평가는 평가 기준을 적용하여 점수나 등수를 매기는 방법으로 수행되며, 오류 분석도 함께 이루어져 보다 정확한 번역 품질 평가가 가능하다(이지은과 최효은 2022a). 한국어 조합 언어쌍에 대한 기계번역 결과물 품질 평가 연구에서 제시된 오류 유형 분류는 연구자별 차이가 있기는 하지만 누락, 어휘, 단어 순서, 문장구조, 문법, 오역, 오탈자와 그 외 기타 등 상당 부분 겹친다(박

(Systran) 플랫폼을 기반으로 하여 품질이 검증된 법률과 특허 데이터로 학습시킨 모델이다. 이재욱 대표에 의하면 학습 데이터는 한국어 기준 5천만 문장 이상이다. 다음의 링크에서 번역 서비스를 제공한다. https://otran.io/

2) 한편 높은 상관관계를 보여주는 연구도 있다(김보영 외 2020).

옥수 2017; 기유미 2018; 서보현과 김순영 2018; 한현희 2019, 2020).

수동평가는 사용적합성(usability)과 같이 단일 평가 기준을 사용하기도 하지만 충실성, 유창성 등과 같은 평가 기준을 사용하여 기준별 점수를 매기게 하거나 기준을 활용하여 세그먼트별 평가 점수를 제시하도록 한다(이지은과 최효은 2022a; 최효은과 이지은 2017). 한편 문장 단위의 평가가 문맥을 고려할 수 없는 한계가 있음을 인식하고 최근에는 텍스트 층위의 기계번역 품질 평가가 이루어지기도 한다.

한국어와 영어, 중국어, 러시아어 등 주요 외국어 간의 기계번역 품질을 평가한 연구는 주로 범용 NMT인 구글 번역, 네이버 파파고, 카카오*i* 등의 기계번역 결과물의 오류를 분석하거나 품질 평가를 수행하였다(이상빈 2019; 이준호 2019, 2020; 최문선 2020). 다양한 유형의 텍스트를 이용한 범용 NMT 연구에 의하면 엔진 성능이 조금씩 다르지만 대체로 오류가 많고 NMT 품질이 미흡한 결과를 보였다. 언어쌍, 언어 방향, 텍스트 유형에 따라 번역엔진 성능에 대한 평가의 차이가 있어 어떤 엔진이 우세하다고 단정하기는 어렵다(강병규와 이지은 2018; 기유미 2018; 이준호 2020; 한현희 2020). NMT 연구 동향 분석에 따르면 충실성보다 유창성 향상이 두드러졌다(Ragni and Nuns Vieira 2022: 142).

앞서 언급한 바와 같이 기계번역 평가 관련 선행연구들은 자동평가 또는 수동평가 한 가지에만 의존하거나 원문과 대조하지 않고 번역문만 평가하여 유창성이 높게 평가된 경우(Koehn and Knowles 2017: 6; Pym 2019), 그리고 평가 및 평가자에 대한 정보가 불명확하거나 연구자가 직접 평가하여 연구자의 주관성을 배제할 수 없는 등의 문제점이 있었다. 이러한 이유로 연구자들은 평가자 정보와 평가자 훈련 및 전문성 확보, 평가 가이드라인 제시 등이 중요하다고 강조한다(Doherty 2017; Chatzikoumi 2020: 146). 또한 정확한 오류 분석을 위해서는 오류 유형의 일반화 가능성은 물론이고 분석가/평가자 변수, 유형 분석의 일관성, 오류의 정의, 오류 항목의 수 등을 세심하게 규정할 필요가 있다(이상빈 2020: 86). 신뢰할 만한 평가 결과를 도출하기 위해서는 이러한 유의사항을 고려하여 기계번역 평가 연구를 설계해야 할 것이다.

2.2 법률 기계번역 연구

법률 분야에서 번역 수요가 높기 때문에 기계번역 활용 가능성에 대한 논의가 있다. 국내에서도 형사소송절차에서 인공지능 통번역 애플리케이션 활용과 법령번역서비스를 통한 법 정보화에 관심이 높아지고 있다(김한균 2021; 이정민과 후쿠바 히카루 2021). 그렇지만 법률 기계번역의 오류는 법률적으로 심각한 결과를 초래할 수 있기 때문에 기계번역의 특성에 대한 정확한 이해와 경각심을 요한다. 눈스 비에이라, 오해이건과 오설리번(Nunes Vieira, O'Hagan and O'Sullivan 2021)와 김한균(2021)이 인용한 미국 판례를 보면 형사 및 난민 사건과 같은 법률 분야에서 법률 기계번역에 대한 문제의식 부족으로 인해 어떠한 문제가 발생했는지 알 수 있다. 법률종사자를 중심으로 번역 수요가 워낙 많다 보니 대강의 의미를 파악하고자 할 때 그리고 중요한 소송 문건이 아닐 경우 법률 기계번역을 일부 활용할 수 있다는 긍정적 시각을 확인할 수 있다(Giordano 2013: 467; Roberts 2022).

하지만 이와 같은 긍정적인 시각에도 불구하고 아직 법률번역에서 번역 메모리를 비롯한 기계번역 활용도나 법률 기계번역 품질에 대한 최신 정보는 부족하다. 이는 국내외 MT 연구를 통틀어 법률 기계번역에 대한 연구는 소수에 불과하기 때문이기도 하다. NMT 개발 이전과 이후에 이루어진 연구로 구분되는데 전자에 해당하는 연구에서는 대체로 법률 기계번역의 품질이 낮다. 그나마 최근 일부 긍정적인 평가 결과가 보이기는 하나 요지 파악을 위한 기능을 할 수 있을지 여전히 품질에 대한 의문이 있다. NMT 개발 이전에 이뤄진 몇 가지 관련 선행연구를 소개하자면, 멕시코와 독일의 민법 조문을 영어로 번역한 바벨피쉬 결과물을 분석한 예이츠(Yates 2006)는 심각한 오류로 요지 파악을 위해 기계번역을 사용하기에 미흡하다는 결론을 내렸다. 키트와 옹(Kit and Wong 2008)은 바벨피쉬와 구글, 시스트란, 윌드링고 등 6종의 기계번역엔진의 영어 법률번역 결과물을 자동평가한 결과 BLEU와 NIST 수치는 대체로 유럽 언어쌍에 비해 아시아 언어의 기계번역 품질이 낮은 것으로 나타났다(Kit and Wong 2008: 315). 법률 용어 번역에 기계번역을 활용할 수 있을지 스페인어-영어 용어 기계번역 결과를 살펴본 킬먼(Killman 2014)은 연구자료의 64%에 해당하는 스페인어 용어가 영어로 적절히 번역된 것으로 평가하며 법률 기계번역

활용 가능성을 긍정적으로 평가하였다. 한편 브키치, 셀잔과 비치치(Brkić, Seljan and Vičić 2014)는 영어-크로아티아어 간의 법률 텍스트 두 건에 대한 번역 결과물을 자동평가와 수동평가 방식을 모두 사용하여 평가하였다. BLEU 수치는 33.70과 31.11을 기록하였으며 수동평가와 자동평가 간의 상관관계는 확인되지 않았다(Brkić, Seljan and Vičić 2014: 5). 두 가지 텍스트의 수동평가 결과 유창성은 4점 기준 3.03, 3.30을 각각 기록했고, 충실성은 3.04, 3.67을 기록하였다(Brkić, Seljan and Vičić 2014: 4). 상기 연구에서는 형태적 오류가 가장 많았고, 그다음으로 어휘 오류, 비번역, 통사 오류 순이었다(Brkić, Seljan and Vičić 2014: 4).

NMT 연구에 해당하는 드파우 외(Defauw et al. 2019)는 자체 개발한 기계번역 엔진과 구글 번역으로 일반 텍스트와 법률 텍스트 두 가지 텍스트 유형을 영어-아일랜드어 양방향으로 번역하게 하고 자동평가를 실시하였다. BLEU 수치에 의하면 양방향에서 모두 구글 번역보다 자체 개발 엔진의 성능이 나은 것으로 평가되었다(Defauw et al. 2019: 35). 또한 NMT 엔진 DeepL Translator와 CAT 툴인 MateCat를 이용하여 이탈리아어 법률 텍스트를 독일어로 기계번역한 결과물을 연구자 자신이 직접 평가한 비스만(Wiesmann 2019)에서는 충실성과 유창성 모두 점수가 낮았고, 충실성은 유창성보다 더 품질이 낮았다.

법률 기계번역과 관련된 국내 번역학 연구 논문은 김혜림(2021), 이지은과 최효은(2022b), 이준호(2022)가 전부이다. 하지만 김혜림(2021)은 중한 법령 기계번역의 포스트에디팅 교육과 관련된 연구이고, 이지은과 최효은(2022b)은 법조문의 이중주어구문에 대한 구글번역과 파파고의 한영 번역 양상을 분석한 것으로 NMT 성능 평가 연구는 아니다. 법률 기계번역 평가를 다룬 것은 한국어 계약문건의 영어 번역 결과물을 오트란과 구글번역 엔진을 이용하여 추출하고 비교 평가한 이준호(2022)가 유일하다. 이준호(2022)는 오트란이 구글번역보다 우수한 품질임을 확인하였지만 60문장만을 대상으로 하여 연구 규모가 작고, 평가 영역 당 단 한 명의 평가자의 평가에 의존하는 등 연구결과를 일반화하기에는 한계가 있어 후속 연구가 필요하다.

2.3 특허 기계번역 연구

법률 분야보다 특허 분야에서 기계번역이 훨씬 먼저 활용되기 시작했다. 특허 기계번역은 사용자가 알지 못하는 언어로 작성된 특허 문서의 요지를 파악하기 위해 사용된다(최효은과 이지은 2017: 143; Kinoshita et al. 2017; Nurminen 2020: 100; Olohan 2015: 167). 그렇지만 특허 기계번역 엔진의 번역 품질에 대한 선행연구를 살펴보면 요지 번역으로서 기능하기에 부족한 심각한 오류가 나타났다. 특허 번역 수요가 급증하면서 기계번역이 많이 활용되는 경향이 있는데 기계번역의 품질 문제로 인한 출원 지연 등의 문제가 발생하기도 한다(Smyth et al. 2015: 154).

기계번역 연구가 폭발적으로 증가하였지만 특허 기계번역의 품질에 대한 연구는 여전히 소수에 불과하다(Castilho et al. 2017; Kinoshita et al. 2017; Poliquen 2015, 2017; Tsai 2017 등). 이중 본 연구에 참고할 만한 몇 가지 주요 선행연구를 살펴보겠다. 먼저 로시와 위긴스(Rossi and Wiggins 2013)는 일본어 특허 명세서 중 길이와 통사구조가 다양한 1,000개 문장을 추출하여 SMT 특허 전문 기계번역 엔진 LexisNexis의 영어 기계번역 결과물을 평가하였다. 자동평가의 경우 애초 번역 엔진을 설계할 때 기준인 BLEU 35점을 최소 기준으로 삼아서 '수용 가능'과 '수용 불가능'으로 평가하였다(Rossi and Wiggins 2013: 121). 수동평가에서는 용어, 정보 누락, 정보 추가, 단어 순서의 네 가지를 평가 기준으로 적용하였으며, 전문가 2인이 문장별로 평가하도록 하였다. 오류 분류에 따라 1점(수용불가능)에서 4점(이상적)까지의 점수를 매기도록 하였다. 오류 분류별 점수, 단어 비중에 따른 점수, 단어 순서에 따른 점수 중 최하점을 해당 문장의 최종 점수로 확정하였으며, 최종 점수 3점 미만은 '수용불가능'으로, 3점 이상은 '수용 가능'으로 분류하였다.

자동평가와 수동평가 결과가 모두 '수용 가능'인 경우 최종 평가를 '수용 가능'으로 하고 자동평가가 '수용 가능'인데 반해 수동평가가 '수용 불가능'인 경우 역시 최종 평가를 '수용 불가능'으로 확정하였다. 한편 자동평가가 '수용 불가능'인데 반해 수동평가가 '수용 가능'인 경우, BLEU 점수와 원인 분석 등을 거쳐 선택적으로 '수용 가능'한 문장과 '수용 불가능'한 문장으로 나눔으로써 자동평가의 단점을 수동평가로 보완할 수 있는 평가 틀을 제시했다는 평가

를 받는다(Rossi and Wiggins 2013: 124).

로시와 위긴스(2013)에서는 검색가능성의 측면에서는 높은 수준의 수용 가능성(94.36%)을 보였지만 가독성에 대한 수용 가능성은 45.04%로 낮은 편이었는데, 이와 관련해서 특히 단어 순서가 가독성에 미치는 영향이 클 수 있으며, 실제로 문장 길이가 길어질수록 단어 순서에 대한 점수가 낮아짐을 확인했다(Rossi and Wiggins 2013: 124).

최효은과 이지은(2017)은 한국 특허청의 특허검색사이트인 키프리스에서 무료로 제공하는 K2E-PAT로 생산한 한영 특허 요약서에서 추출한 100 문장을 인간번역인 KPA 100문장과 비교하여 BLEU 점수를 사용하여 자동평가하고, 특허 전문 번역사 2명이 충실성과 가독성 점수를 각각 도출하였다(최효은과 이지은 2017: 165). 한영 특허 기계번역 결과물은 자동평가에서 BLEU 22.90점으로 낮은 점수를 기록했고, 수동평가에서도 충실성, 가독성 모두 평균 3점 이하의 낮은 점수를 받았다(최효은과 이지은 2017: 165).

차이(Tsai 2017)는 대만특허청의 SMT에 의한 발명의 명칭 중영 번역 결과물에 대해 대만특허청 소속 번역사들이 수동평가를 수행한 연구다. 오류를 철자법 오류, 형태적 오류, 어휘, 의미 오류와 통사적 오류 등 크게 다섯 가지로 분류하였는데 연구 결과, 473개의 발명의 명칭에서 692개의 오류가 발견되었다. 발명의 명칭당 대략 1.5개의 오류가 있는 셈으로 통사적 오류의 비중이 53.76%로 가장 높았고 어휘 오류(18.64%), 의미 오류(14.60%)가 그 뒤를 이었으며 형태 오류가 3.18%로 가장 낮았다(Tsai 2017: 149).

NMT 특허 번역연구도 찾아볼 수 있다. 카스틸호 외(Castilho et al. 2017)는 특허 명세서 중 발명의 명칭과 요약의 중국어-영어 번역에 대해서 NMT과 SMT 번역 품질을 자동평가와 수동평가를 실시하여 비교하였다. BLEU를 사용한 자동평가 결과에 의하면 발명의 명칭 번역에서 NMT가 조금 더 우수한 것으로, 요약 번역에서는 SMT가 조금 우세한 것으로 확인되었다(Castilho et al. 2017: 114). 평가자 2인이 수행한 수동평가에서는 SMT의 품질이 우세한 것으로 나타났는데 문장 길이에 따라 결과가 상이하게 나타났다. 두 종류의 번역엔진 결과물에서 번역 오류 중 누락이 가장 많았고, SMT에서는 문장 구조 오류 비중이 더 높았다. 오류 없이 완벽한 번역문은 SMT의 25%, NMT의 2%를 차지하여 SMT의 품질이 우세한 것으로 판단되었다.

한편 프레몰리 외(Premoli et al. 2019)는 영어-이탈리아어 특허 기계번역에 대한 수동평가 연구로 자체적으로 구축한 특허 번역에 특화된 NMT 번역 결과물의 품질을 분석하였다. 번역사 2인이 4점 척도를 사용하여 평가한 결과, 엔진이 학습한 기계 분야의 특허번역에 대해서 이해가 어려운 문장은 10개, 이해 불가한 문장은 6개에 불과했으며 대체로 품질이 용인 가능한 수준으로 평가되었다(Premoli et al. 2019: 37). 그럼에도 불구하고 용어 번역의 비일관성 문제는 치명적인 품질 저하 요인으로 간주되었다(Premoli et al. 2019: 37).

세계지적재산권기구(WIPO)의 인공신경망 기계번역 서비스 WIPO Translate의 특허 문건 번역 결과물과 범용 엔진인 구글번역의 BLEU 점수를 비교한 폴리켄(Poliquen 2017: 24-25)에 의하면, 대체로 WIPO Translate의 성능이 나은 것으로 나타났다. 폴리켄은 영일 언어쌍을 비롯하여 여러 언어 번역결과물을 평가하였는데, 영한 번역의 경우 WIPO Translate가 39.20, 구글번역이 32.65였다(Poliquen 2017: 24-25).

특허 전문 기계번역 엔진 2종 Patent Translate와 WIPO Translate을 평가한 이지은과 최효은(2022a)에 따르면, BLEU와 METEOR 수치로 측정한 자동평가에서는 Patent Translate가 우세한 것으로 나왔지만 수동평가는 WIPO Translate가 조금 높은 점수를 받았다. BLEU는 Patent Translate와 WIPO Translate 결과물이 각기 35.42, 33.80, METEOR는 35.7, 35.3이었으며 특허 전문 번역사 4인의 수동평가 점수는 2.16, 2.22였다(이지은과 최효은 2022a: 114-115). 두 엔진 간 수동평가 결과의 차이가 큰 오류는 누락과 통사였다. WIPO Translate는 Patent Translate보다 통사 오류가 훨씬 적어 유창성이 뛰어난 것으로 평가되었으며 누락 빈도는 높았지만 5단어 이하의 사소한 누락이 압도적으로 많아 상대적으로 양호한 것으로 평가받았다(이지은과 최효은 2022a: 125). 이 같은 결과를 보더라도 자동평가와 수동평가는 반드시 결과가 일치하지 않아 정확한 성능평가를 위해서는 자동평가와 수동평가 결과를 모두 고려할 필요가 있다는 것을 알 수 있다.

3. 연구 방법

본 연구는 NMT 엔진인 오트란의 성능을 분석하기 위해 국문 법령 2종에서 발췌한 180개 세그먼트와 특허공보 30건에서 발명의 명칭과 요약을 발췌한 83개 세그먼트의 영어 번역텍스트를 대상으로 자동평가와 수동평가를 수행하였다. 자동평가는 BLEU[3]와 METEOR[4] 수치를 이용하였고, 수동평가는 분야별로 전문성을 갖춘 2인의 평가자를 활용하였다. 법률번역은 번역을 전공한 석사 이상의 학력과 5년 이상의 경력자, 특허번역은 번역 석사 전공을 요구하지 않는 대신에[5] 10년 이상의 경력자를 평가자로 섭외하였다. 본 연구에 참여한 법률번역 평가자는 번역을 전공한 석사 이상의 학력 소지자로 각각 6년과 7년의 법률번역 경력자이고, 특허번역 평가자는 각각 20년, 21년의 특허번역 경력자다 (<표 1> 참고).

〈표 1〉 원천 텍스트와 평가 텍스트

	법률	특허
텍스트	물관리기술 발전 및 물산업 진흥에 관한 법률 & 외국인투자촉진법 발췌 - ST 기준 2,725 단어 - TT 기준 5,252 단어 - 총 180세그먼트	전기전자 관련 특허공보 30건에서 발명의 명칭과 요약 발췌 - ST 기준 2,792 단어 - TT 기준 3,704 단어 - 총 83세그먼트
평가자	번역 전공 석사 이상 학력자로 법률 번역 실무 경력 5년 이상 보유 평가자 2인(E1, E2)	특허번역 실무 경력 20년 이상 보유 평가자 2인(E3, E4)

3) Tilde Custom Machine Translation에서 제공하는 BLEU 계산기를 사용하였다. https://www.letsmt.eu/Bleu.aspx

4) CMU Language Technologies Institute에서 제공하는 METEOR 계산 로직을 활용하였다. https://www.cs.cmu.edu/~alavie/METEOR/

5) 특허번역은 번역을 전공한 전문가 이외에 공학이나 자연과학을 전공한 전문가들도 다수 참여하고 있어 번역을 전공하지 않은 번역 전문가들이 많으므로 번역 석사 제한을 두지 않았다.

수동평가의 방식에 대해서는 법률번역과 특허번역 모두 동일하게 평가자들에게 세그먼트별로 5단계 척도에 해당하는 0 ~ 4 중 가장 근접한 점수를 매기도록 하였다. 평가자 가이드라인을 통해 0에서 4까지 대략적인 수준에 대해 기술하여 평가자들이 참고할 수 있도록 하였다. 구체적으로 4는 원문의 의미를 완전히 정확하고 자연스럽게 전달하는 매우 양호한 수준의 번역을, 반대로 0은 원문과 비교하였을 때 오류가 번역으로 간주하기 어려울 만큼 심각한 수준이며 문장 구조 또한 이해 불가한 매우 불량한 수준의 번역을 의미한다.

척도 평가 점수와 함께 평가자들에게 연구자들이 제시한 오류 기준에 따라 각 세그먼트에서 해당하는 오류가 있을 경우 평가 의견을 기재해줄 것을 요청하여 평가자들이 부여한 평가 점수에 대한 근거를 제시할 수 있도록 하였다. 연구자들이 제시한 오류 기준은 아래 <표 2>와 같다. 아래의 기준은 기계번역의 수동평가에서 주로 사용하는 기준들을 추려서 구성하였으며, 법률번역과 특허번역에 있어서 특수하게 문제가 될 수 있는 콜론, 세미콜론의 사용 등 스타일 문제 등을 기타의 분류에 넣었다(Castilho et al. 2017: 114; Chatzikoumi 2020: 151; Rossi and Wiggins 2013: 121 등).[6]

〈표 2〉 법률번역과 특허번역의 수동평가를 위한 오류 기준

대분류	소분류
정확성	추가
	누락
유창성	통사
	문법
용어	용어
기타	오타, 스타일 등

6) 법률과 특허 기계번역 연구 자료 수집 시기가 다르며 시기적으로 나중에 실시한 법률번역 평가는 특허번역 평가에 비해 좀 더 상세한 평가 기준을 적용하였지만 본고에서는 비교를 위해 공통적인 분류 기준을 정리하여 제시하였다.

4. 연구 결과

4.1 법률 기계번역 평가

4.1.1 자동평가

오트란 법령 기계번역에 대한 BLEU 수치는 58.5점, METEOR는 44.5점을 기록했다. 이와 같은 수치는 로시와 위긴스(2013)에서 번역 엔진을 설계할 때 최소 품질 기준인 BLEU 35점보다 높아 '수용 가능'한 수준의 기계번역 결과물이라고 할 수 있다. 영어-크로아티아어 법률 텍스트 두 건에 대해 기계번역 결과물을 자동평가한 브키치, 셀잔과 비치치(2014)에서 BLEU 점수가 각각 33.70점과 31.11점이었던 점을 고려했을 때, 오트란의 법률 기계번역 결과물에 대한 자동평가 결과는 수용 가능한 수준이라고 할 수 있겠다.

4.1.2 수동평가

수동평가 결과, 아래 <표 3>과 같이 180개 세그먼트에 대해 4점 만점 기준으로 E1은 평균 3.41점을, E2는 평균 3.16점을 부여하였고, 두 점수의 평균은 3.28점이다. E1이 E2에 비해 전반적으로 높은 점수를 주었으나 평가자 두 사람 모두 3점 대의 점수 평균을 기록했다는 점에서 둘 간의 큰 편차는 보이지 않은 것으로 보인다. 평가자 2인 모두 3점 대의 점수를 주었다는 것은 오트란의 번역 결과가 완벽하지는 않으나 전반적으로 양호하여 이해가 가능한 수준임을 가늠할 수 있다.

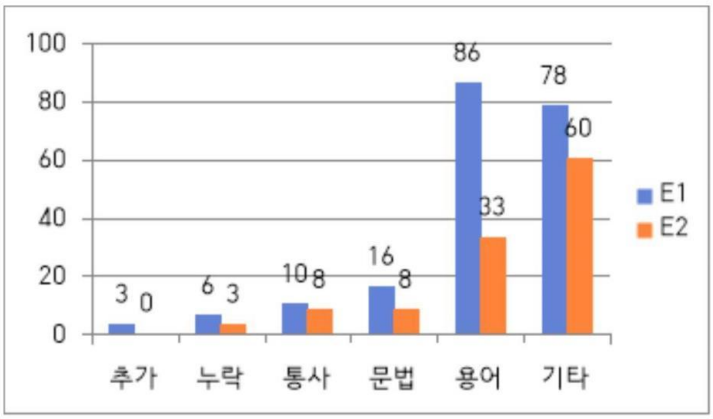
〈표 3〉 법률 기계번역에 대한 수동평가 평균 점수

	평가자 E1	평가자 E2	E1과 E2의 평균
평균 점수	3.41점	3.16점	3.28점

다음으로 세그먼트 별 오류 기준에 따른 각 평가자의 코멘트를 살펴보면 E1은 총 199개의 코멘트를, E2는 총 112개의 코멘트를 제시하여 E1은 하나의 세그먼트에 평균 1개 정도의 코멘트를, E2는 세그먼트 별로 평균 0.6개 정도의 코멘트를 남긴 것으로 볼 수 있다.

앞서 제시한 평가표의 소분류에 따른 평가자 E1과 E2의 코멘트 수는 아래 <그림 1>과 같다.

〈그림 1〉 법령 기계번역 결과물에 대한 평가자 2인의 오류 기준별 코멘트 수



법률 기계번역에 대한 평가자 E1과 E2의 세그먼트 별 오류 기준에 따른 코멘트 양상을 살펴보면, 오트란의 법률 기계번역 결과물에서 정확성에 해당하는 추가나 누락의 문제는 상당히 적은 비중을 차지하는 것을 알 수 있다. 또한 유창성에 해당하는 통사와 문법도 마찬가지로 큰 비중을 차지하지 않아 문장 자체 및 구조의 완성도 또한 상당히 높다는 사실을 알 수 있다.

한편 평가자들은 용어와 기타 문제 지적이 많았고 건수는 평가자별 차이를 보인다. 용어 사용의 정확성과 일관성을 묻는 용어 문제에 대해서는 E1이 지적한 오류의 수가 전체 분류 중 가장 많았다.

용어와 관련된 코멘트로 E1과 E2가 공통적으로 지적한 사례를 들어보면, 아래의 <예 1>의 '신청인'에 대해서 E1은 '신청인 claimant는 부자연스러운 표현'으로, E2는 '신청인을 claimant로 번역한 것이 적절하지 않음'으로 코멘트를 제시했다. 실제로 'claimant'는 권리를 주장하는 당사자를 칭하는 용어이며, 이에 반해 아래 <예 1>의 '신청인'은 허가를 신청한 당사자이므로 'applicant'가 아닌 'claimant'는 용어를 부적절하게 사용한 경우다. 이와 같이 오트란에서 생성된 용어와 관련된 오류는 전문용어를 사용하되 미세하게 그 의미가 다른 용

어를 사용하여 정확성이 저해되는 경우가 대부분이었다.

<예 1> 외국인투자촉진법 제6조제2항
ST: ② 산업통상자원부장관은 제1항에 따른 허가신청을 받으면 대통령령으로 정하는 기간에 그 허가 여부를 결정하고 신청인에게 알려야 한다.
TT: (2) Upon receipt of an application for permission filed under paragraph (1), the Minister of Trade, Industry and Energy shall determine whether to grant such permission within a period prescribed by Presidential Decree and notify the claimant of his/her determination.

스타일 문제와 오타 등 번역 결과물의 품질에 있어서 문제가 될 수 있는 기타 오류는 다른 오류에 비해 E1과 E2가 동일하게 많은 코멘트를 제시하였다. E1과 E2가 공통적으로 지적한 기타 오류의 예를 살펴보면, 아래 <예 2>의 마침표에 대해 E1은 '마지막에 : 와야 함'으로, E2는 '나열의 시작 부분이므로 : 필요함(의미 이해에는 그다지 지장 없음).'으로 의견을 제시했다. 실제로 아래 <예 2>의 제2조2호 이하에는 여러 목을 통해 사업을 구체적으로 나열하여 기술하고 있으며, 따라서 ST와 달리 TT에서는 앞으로 나열이 있을 것을 암시하는 ':(콜론)'을 써 주는 것이 법령번역의 스타일에 부합한다. 다만 E2가 언급한 바와 같이 TT의 말미에 콜론 대신 마침표를 썼다고 해서 의미 이해에 큰 지장이 있는 것은 아니다.

<예 2> 물관리기술발전 및 물산업 진흥에 관한 법률 제2조제2호
ST: 2. "물산업"이란 다음 각 목의 어느 하나에 해당하는 사업을 말한다.
TT: 2. The term "water industry" means any of the following businesses.

이와 같은 평가 결과를 종합하면 전반적인 문장의 이해 및 정확성에 영향을 미치는 치명적인 오류가 비교적 적고, 이해에는 큰 지장이 없는 스타일, 오타 등의 오류가 주류를 이루고 있으므로 전반적인 번역 수준은 양호하다고 볼 수 있을 것이다. 다만, 전문용어의 비중이 큰 법률번역의 특성상 용어의 번역이 정확하지 않은 점은 개선이 필요하다고 볼 수 있겠다.

4.2 특허 기계번역 평가

4.2.1 자동평가

오트란 특허 기계번역에 대한 BLEU 수치는 38.7점, METEOR는 36.7점을 기록했다. 특허 기계번역 결과물은 법률 기계번역 결과물의 자동평가 결과와 비교했을 때, BLEU, METEOR 모두 좀 더 낮은 편에 속했다. 로시와 위긴스(2013)의 최소 기준을 근소하게 넘었지만 법률 기계번역 결과물과 마찬가지로 '수용 가능'한 수준임을 알 수 있다.

하지만 특허 기계번역 선행연구인 최효은과 이지은(2017)에서 NMT 이전 세대인 K2E-PAT의 BLEU 점수 22.90점에 비해 성능이 우월하다. 언어 방향은 다르지만 범용 NMT의 대표격인 구글번역을 대상으로 영어-한국어 특허 문건의 번역에 대해 자동평가를 실시한 폴리켄(2017)에서 구글번역의 BLEU 점수가 32.6점으로 나온 결과와 비교하였을 때에도 오트란이 K2E-PAT뿐만 아니라 범용 NMT인 구글번역에 비해 우수한 성능임을 확인할 수 있다. 또한 이지은과 최효은(2022a)에서 오트란과 유사한 맞춤형 특허 전문 번역엔진인 Patent Translate(BLEU 35.42, METEOR 35.7), WIPO Translate(BLEU 33.80, METEOR 31.3)의 자동평가 결과를 살펴보았을 때, 오트란의 BLEU와 METEOR 수치가 근소하나마 높다는 점을 알 수 있다.

4.2.2 수동평가

83개 세그먼트에 대한 수동평가한 결과를 보면 E3은 평균 2.77점을, E4는 평균 2.84점을 부여하였고, 둘의 평균은 2.80점이다(<표 4> 참고). E4가 E3에 비해 평균 점수가 약간 높으나 평가자 두 사람이 모두 2점 대 후반의 점수 평균을 기록했다는 점에서 둘 간의 큰 편차가 없다. 법률 기계번역 결과물이 3점대의 평균을 기록한 것에 비해서 특허 기계번역은 2점 대 후반으로 법률 기계번역에 비해 그 품질이 조금 아쉬울 수 있음을 시사한다. 이와 같은 수동평가의 점수 흐름은 자동평가의 점수 흐름과도 동일하여, 자동평가와 수동평가 간 어느 정도 관련이 있을 수 있음을 시사한다.

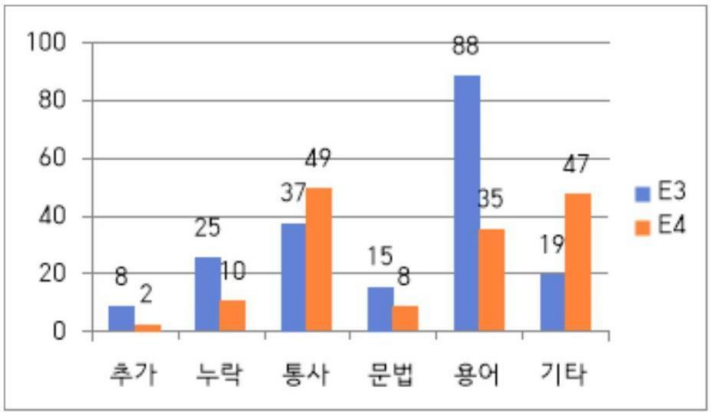
〈표 4〉 법률 기계번역에 대한 수동평가 평균 점수

	평가자 E3	평가자 E4	E3과 E4의 평균
평균 점수	2.77점	2.84점	2.80점

다음으로 평가자들의 오류 코멘트를 살펴보면 E3은 총 192개의 코멘트를, E4는 총 151개의 코멘트를 제시하여 E3은 하나의 세그먼트에 평균 2.3개 수준의 코멘트를, E4는 세그먼트별로 평균 1.8개 수준의 코멘트를 남긴 것으로 볼 수 있다. 법률 기계번역에서 E1과 E2가 제시한 세그먼트별 코멘트 수에 비해 특허 기계번역에서 E3과 E4가 제시한 세그먼트별 코멘트 수가 더 많다. 이와 같은 오류 코멘트 수의 많고 적음이 평균 점수와도 연관이 있음을 알 수 있다. 즉, 법률 기계번역과 특허 기계번역을 비교한 결과, 오류 코멘트 수가 많을수록 평균 점수가 낮은 경향을 보였다.

앞서 제시한 평가표의 소분류에 따른 평가자 E3과 E4의 코멘트 수는 아래 <그림 2>와 같다.

〈그림 2〉 특허 기계번역 결과물에 대한 평가자 2인의 오류 기준별 코멘트 수



특허 기계번역에 대한 평가자 E3과 E4의 세그먼트별 오류 기준에 따른 코멘트 양상을 살펴보면, 오트란의 특허 기계번역 결과물에서는 법률 기계번역에서와 달리 정확성에 해당하는 누락의 문제가 조금 더 빈번히 일어나는 것을 알 수 있다. 아래 <예 3>과 관련해서 평가자 E3은 '원문의 '제조과정을 단순화할

수 있어'에서 '제조과정을'이 번역에서 누락됨. 또한 원문의 '효과가 있다'가 번역에서 누락됨.'으로 코멘트를 기재하여, 오트란 결과물에서 여러 구절의 누락 문제를 지적하였다. 특히 문서임을 고려할 때 'thus can be simplified' 부분은 '제조과정을'의 누락으로 인해 '검사장치'인 'the test apparatus' 자체가 '단순화되었다'라고 잘못 해석될 가능성이 높다.

<예 3> 출원번호 1020130063633의 요약 일부
ST: 본 발명에 의하면, 검사장치의 구조가 간단하여 제조과정을 단순화할 수 있어 비용을 절감할 수 있고 효과가 있다.
TT: According to the present invention, the test apparatus has a simple structure and thus can be simplified, thereby reducing costs.

법률 기계번역 결과물과 달리 유창성에 해당하는 통사의 오류가 차지하는 비중 또한 상당히 높다. 아래 <예 4>와 관련해서 평가자 E4는 통사의 오류를 지적하면서 '구성요소 파악이 잘못됨(예, 촬상 유닛 내에 촬상부와 촬상 헤드가 구비되어 있으나 번역은 이를 개별 구성요소 번역함. 이외에도 도포 유닛도 파악이 잘못됨)'으로 기재하였다. 실제로 아래 <예 4>의 원문을 보면 '본 발명에 따른 도포 장치'는 '촬상 유닛', '도포 유닛', '도포 수평 이동부'로 구성되어 있다. 이 중 '촬상 유닛'이 다시 '촬상부'와 '촬상 헤드'를 구비하며, '도포 유닛'이 다시 '도포부'와 '도포 헤드'를 구비하는 것이다. 한편 오트란에 의한 번역을 보면 '본 발명에 따른 도포 장치'는 'an imaging unit(촬상 유닛)', 3개의 'a coating unit(도포 유닛)', 그리고 'a coating horizontal moving unit(도포 수평 이동부)'로 구성되어 원문의 발명을 구성하는 구성요소의 구조를 제대로 파악하지 못해 완전히 잘못 번역하였음을 알 수 있다. 특허번역의 경우, 법률번역과 달리 발명의 구성요소를 설명하는 매우 길고 복잡한 문장이 빈번하게 출현하는데 발명을 설명하는 핵심 내용으로 매우 중요하다. 법률 기계번역에 비해 특허 기계번역 결과물에서 통사와 관련된 오류가 더 빈번하게 나타난 대표적인 원인 중 하나가 이와 같이 길고 복잡한 구조로 발명의 구성요소를 나열 및 설명하는 문장을 기계번역에서 제대로 소화하지 못하고 엉뚱한 구성요소를 갖춘 것으로 번역했기 때문이다. 또한 이러한 문제로 인해 특허 기계번역 결과물의 수동평

가 점수가 법률 결과물에 비해 낮아졌을 수 있다.

<예 4> 출원번호 1020140047591의 요약 일부

ST: 본 발명에 따른 도포 장치는 일 방향으로 나열되어 이격 적재된 복수의 피처리물의 이미지를 촬상하여 획득하는 촬상부 및 일측에 상기 촬상부가 지지 설치되며, 촬상부가 일 방향으로 적재된 상기 복수의 피처리물 각각에 대응 위치하도록 이동시키는 촬상 헤드를 구비하는 촬상 유닛과, 촬상부의 일측 및 타측 중 어느 하나에 위치하며, 상기 복수의 피처리물 각각의 도포면에 접합제를 도포하는 도포부 및 일측에 상기 도포부가 지지 설치되며, 도포부를 일 방향으로 적재된 상기 복수의 피처리물 각각에 대응 위치하도록 이동시키고, 상기 도포부가 각 피처리물의 도포면의 연장 방향을 따라 수평 이동하면서, 상기 피처리물이 기울어진 경로를 따라 이동하여, 상기 피처리물 도포면의 두께 방향의 중심을 따라 수평 이동하도록 두께 방향의 위치를 조절하는 도포 헤드를 구비하는 도포 유닛과, 일측에 상기 도포 헤드 및 촬상 헤드 각각이 설치되어, 상기 도포 유닛 및 촬상 유닛을 각각이 상기 피처리물 도포면의 연장 방향으로 수평이동하도록 가이딩하는 도포 수평 이동부를 포함한다.

TT: The coating apparatus according to the present invention comprises: an imaging unit which captures and obtains images of a plurality of objects arranged in one direction and stacked to be spaced apart from each other; a coating unit which has the imaging unit supported on one side thereof, and has an imaging head for moving the imaging unit to be positioned to correspond to each of the plurality of objects stacked in one direction; a coating unit which is positioned on one side or the other side of the imaging unit, and coats an adhesive on a coating surface of each of the plurality of objects; a coating unit which has the coating unit supported on one side thereof, moves the coating unit to be positioned to correspond to each of the plurality of objects stacked in one direction, and has the coating head for adjusting a position in a thickness direction to horizontally move the objects along an inclined path to horizontally move the objects along a center in a thickness direction of the coating surface of the objects; and a coating horizontal moving unit which has the coating head and the imaging head on one side thereof, respectively, and guides the coating unit and the imaging unit to horizontally move in an extension direction of the coating surface of the objects.

용어와 관련해서는 두 평가자 모두 다수의 오류를 지적했는데, 특히 E3의 용어 오류 지적이 눈에 띄게 많은 편이었다. 아래의 <예 5>와 관련해서 평가자 E3은 '원문 패시베이션막은 'passivation film'이나, 번역은 passivation layer(패시베이션층)으로 번역됨.'으로 오류를 지적하였다. 실제로 '막'과 '층'은 완전히 다른 개념으로 구분해서 사용되어야 한다. 또한 <예 6>과 관련해서 평가자 E3은 '원문의 리볼링은 reballing이나, 번역에서는 resolding으로 번역됨.'으로 오류를 지적하여 음차한 표현인데도 불구하고 전혀 다른 표현으로 번역되었음을 지적하였다. 이와 같이 특허 기계번역의 경우, 용어가 터무니없이 다른 표현으로 번역되는 경우가 종종 있으며, 이로 인해 수동평가의 점수 결과가 낮을 수 있음을 시사한다.

<예 5> 출원번호 1020130094477의 발명의 명칭
ST: 패시베이션막 형성방법 및 이를 포함하는 AlGaN HFET의 제조방법
TT: Method of forming passivation layer and method of manufacturing AlGaN/GaN HFET including the same

<예 6> 출원번호 1020130049575의 발명의 명칭
ST: 반도체 리볼링 장치
TT: Semiconductor resolding apparatus

법률 기계번역에서와 마찬가지로 특허 기계번역에서도 기타에 해당하는 오류 또한 상당수 존재했다. 한편 특허 기계번역에서의 기타 오류 역시 법률 기계번역과 마찬가지로 문장부호, 대소문자의 지적이 대부분이었다. 아래의 <예 7>과 관련해서 E4는 '문장 중 불필요한 대문자 사용(Provided is)'으로 오류 코멘트를 기재하였다. 소문자를 사용해야 할 곳에 대문자를 사용한 경우 번역 결과물의 품질에 영향을 미칠 수 있다. 다만 법률 기계번역에서 언급한 바와 같이 이와 같은 오류가 의미 전달 등에 치명적인 영향을 주는 것은 아니다.

<예 7> 출원번호 1020130046223의 요약 일부
ST: 본 발명에 따르면, 일측에 세척물질이 유입 가능하도록 유입구(32)가 형성된 노즐본체(31)와; 상기 노즐본체(31)에 내장되어 초음파에 의해 세척물질을 가진시키도록 된 진동소자(35)를 포함한 초음파 진동자(34)와;

> 상기 초음파 진동자(34)에 의해 가진된 세척물질을 배출하도록 일정 길이로 연장된 도파관(37)을 포함하여 이루어진 미세세정 노즐장치가 제공된다.
>
> TT: According to the present invention, the nozzle body (31) comprises: an inlet (32) formed at one side such that a washing material can flow thereinto; an ultrasonic vibrator (34) including a vibration element (35) embedded in the nozzle body (31) to excite a cleaning material by ultrasonic waves; Provided is a micro-cleaning nozzle apparatus including a waveguide (37) extended to a predetermined length to discharge a cleaning material vibrated by the ultrasonic vibrator (34).

평가 결과를 종합하면 법률번역에 비해 특허번역은 누락이 많았으며, 유창성에 해당하는 통사 문제 또한 발명의 구성요소를 설명하는 문장에서 상당한 오해를 불러일으킬 수 있는데 그 수가 적지 않아 품질이 좀 더 낮게 평가되었다. 연구자 간의 코멘트 차이는 있었지만 용어 문제 또한 상당수 확인되었다. 기타 오류 또한 적지 않게 존재했는데, 법률 기계번역 결과물에서와 마찬가지로 주로 대소문자의 사용 문제나 콜론, 세미콜론의 사용 등 스타일 문제에 국한된 오류가 주를 이루었다.

5. 논의 및 결론

기존의 범용 NMT 중심의 기계번역 평가 연구에서 탈피하여 본 연구에서는 법률번역과 특허번역에 특화된 NMT 기반의 자동번역엔진인 오트란의 한영 기계번역 결과물 2종을 토대로 번역 품질을 평가하였다. 법률 기계번역 결과물 평가를 위해 법조문 총 180개 세그먼트를 취합하였고, 특허 기계번역 결과물 평가를 위해 특허공보 건에서 발명의 명칭과 요약을 발췌하여 총 83개 세그먼트를 취합하였다. 평가 대상 국문 텍스트를 오트란을 통해 영어로 번역한 뒤 번역 결과물에 대해 자동평가와 수동평가를 진행하였다.

자동평가는 잘 알려진 지표인 BLEU와 METEOR 두 가지를 활용하였으며, 수동평가는 법률과 특허 각각 법률번역 전문가와 특허번역 전문가 2인에게 의

뢰하여 세그먼트 별 점수를 기재하고, 추가, 누락, 통사, 문법, 용어, 기타의 오류 분류 기준에 따라 오류 코멘트를 제시하였다.

자동평가 결과, 법령 기계번역에 대한 BLEU 점수는 0.585, METEOR는 0.445로 번역 결과물이 수용 가능하며 상당히 양호한 수준임을 보여주었다. 한편 특허 기계번역에 대한 BLEU 점수는 0.387, METEOR는 0.367로, BLEU와 METEOR 모두 법령 기계번역 결과물에 비해 낮은 수치를 보였다. 하지만 오트란의 특허 기계번역 결과물은 BLEU를 기준으로 수용 가능한 수준이며, 기존의 연구들에서 제시한 NMT 이전의 기계번역엔진이나 범용 NMT의 대표격인 구글번역에 비해 높은 수준의 번역 결과물을 산출하였다.

수동평가 결과, 법령 기계번역에 대한 평가 점수 평균은 4점 만점에 3.28점으로 높은 편이었다. 평가자 코멘트를 오류 기준별로 분류한 결과, 정확성에 해당하는 추가와 누락의 오류는 거의 없었으며, 유창성에 해당하는 통사와 문법의 오류도 미미한 편이었다. 다만 용어와 기타의 오류에 코멘트가 집중되는 경향을 보였는데, 용어의 경우 전문용어 내에서 한국어 대응어가 같을 수 있는 난해한 경우의 오류가 상당수 있었고, 기타 오류는 대부분 콜론, 세미콜론의 사용과 같은 스타일과 관련된 오류가 대부분이었다. 따라서 전반적인 오류 코멘트를 기반으로 살펴보았을 때, 법령 기계번역 결과물은 정확성과 유창성이 크게 저해되지 않아 비교적 정확하고 이해가 가능한 수준의 번역 결과물이 산출되었음을 알 수 있었다.

한편 특허 기계번역에 대한 수동평가 점수 평균은 2.80점으로 법령 기계번역 결과물 평가 점수에 비해 낮은 편이었으며 이는 특허 문건의 길고 복잡한 문장 특성에 기인한 것으로 보인다. 평가자 코멘트를 오류 기준별로 분류한 결과, 정확성 중 추가에 해당하는 오류는 거의 없었으나 누락에 해당하는 오류는 법령 기계번역에 비해 빈번히 발생하는 편이었다. 또한 유창성에 해당하는 통사의 오류도 빈번하여 특히 발명의 구성요소를 설명하는 길고 복잡한 문장의 경우 제대로 번역되지 않는 경우가 다수 있었다. 용어의 문제 또한 빈번하게 발생하였으며, 법령 기계번역에 비해 특허 기계번역에서 용어의 문제가 좀 더 일차원적이며 광범위했다. 기타 오류는 대소문자의 사용, 콜론, 세미콜론 등 문장부호의 사용과 같이 의미 전달에는 영향을 미치지 않으나 번역 품질에는 영향을 줄 수 있는 오류들이 대부분이었다.

이와 같은 분석 결과는 맞춤형 번역엔진이라고 하더라도 전문 분야에 따라 그 결과물에 대한 평가 결과가 확연하게 다를 수 있으며, 따라서 맞춤형 번역엔진 구축 시 전문 분야의 특징을 잘 살려서 번역엔진을 학습시킬 필요가 있음을 시사한다.

무엇보다 전문용어의 비중이 높고 중요한 법률번역과 특허번역의 특징을 고려할 때, 법률과 특허 기계번역 결과물 모두에서 용어에 대한 오류 지적이 많았다는 점 또한 눈여겨볼 만하다. 정확한 용어의 사용 문제뿐만 아니라 일관된 용어의 사용 문제가 두드러진다는 번역 전문가들의 오류 지적은 앞으로 오트란이 용어의 차원에서 좀 더 개선의 여지가 필요함을 시사한다.

본 연구는 범용 NMT와 차별화되는 맞춤형 NMT의 품질을 해당 분야 전문 번역사들의 수동평가와 자동평가를 병행하여 살펴보았다는 데 그 의의가 있다. 특히 BLEU와 METEOR로 살펴본 자동평가 결과와 각 분야의 전문가들이 분석한 수동평가 결과의 추이가 일치하는 경향을 보며 본고의 분석에 한해 수동평가와 자동평가 간 연관성이 있다고 볼 수 있겠다.

본고에서 살펴본 법률 기계번역과 특허 기계번역 간 성능의 차이가 존재하기는 하나 NMT 이전 세대의 번역엔진 및 범용 NMT에 관한 기존 연구 결과에 비해서 맞춤형 NMT의 성능이 비교 우위에 있다는 점을 확인하였다. 비록 제한된 데이터를 기반으로 한 사례연구로서 연구결과를 일반화할 수 없는 한계가 있지만 장르별 기계번역 결과물을 비교함으로써 성능과 오류를 이해하는 데 도움이 되었고, 시간의 경과에 따른 NMT 엔진의 성능 개선을 가늠해볼 수 있었다. 본 연구에서 충분히 다루지 않은 수동평가 결과 및 오류 등에 대해서는 후속 연구로 미룬다.

참고문헌

강병규, 이지은 (2018)「신경망 기계번역의 작동 원리와 번역의 정확률」,『중어중문학』73(2): 253-295.

기유미 (2018)「한중 기계번역 오류의 문형별 비교분석: 네이버 파파고 번역기와 구글 번역기를 중심으로」,『중국연구』74: 3-32.

김보영, 김연주, 서승희, 송신애, 이진현, 전경아, 최지수, 홍승빈, 정혜연 (2020) 「번역자동평가에서 풀리지 않은 과제」, 『번역학연구』 21(1): 9-29.

김순미, 신호섭, 이준호 (2019) 「번역학계와 언어서비스업체(LSP)간 산학협력연구」, 『번역학연구』 20(1): 41-76.

김한균 (2021) 「형사절차상 인공지능기반 통번역 애플리케이션 활용과 적법절차」, 『형사소송』 13(3): 147-174.

김혜림 (2021) 「중한 법령 기계번역 포스트에디팅 교육을 위한 예비 연구」, 『번역학연구』 23(3): 65-98.

박옥수 (2017) 「한영 기계번역에서 ST의 유형적 특징에 따른 번역 오류 분석」, 『동아인문학』 44: 151-171.

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

이상빈 (2020) 「기계번역에 관한 KCI 연구논문 리뷰: 인문학 저널 논문(2011~2020년 초)의 논의내용과 연구방법을 중심으로」, 『통역과 번역』 22(2): 75-104.

이준호 (2019) 「신경망기계번역의 객관적 평가를 위한 예비연구: 자동평가와 수동평가의 균형점」, 『통번역학연구』 23(3): 171-202.

이준호 (2022). 「법률 특화 번역엔진 성능 평가: 한영 계약서 번역을 중심으로」, 『T&I Review』 12(1): 169-192.

이준호, 김순미 (2022) 「풀 포스트에디팅에 대한 고찰: 풀 포스트에디팅 생산성에 영향을 주는 요소를 중심으로」, 『번역학연구』 23(5): 119-144.

이정민, 후쿠바 히카루 (2020) 「정보기술 발달에 따른 사법통역의 현재와 미래」, 『4차산업혁명 법과 정책』 2: 171-202.

이지은, 최효은 (2022a) 「인공신경망 특허 기계번역 성능에 관한 연구: Patent Translate와 WIPO Translate 한영 번역 결과물의 누락과 통사 오류 분석을 중심으로」, 『T&I Review』 12(2): 105-130.

이지은, 최효은 (2022b) 「기계번역에서 이중주어 구문의 한영 번역 양상: 구글 번역과 네이버 파파고의 법조문 번역 사례 비교」, 『T&I Review』 12(1): 211-240.

최문선 (2020) 「국내 번역학 기계번역 연구 동향: 내용 분석과 키워드 분석을

중심으로」,『언어학연구』24(1): 275-297.

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

한현희 (2019)「한노 기계번역의 오류 유형화 및 품질 개선을 위한 프리에디팅(pre-editing) 규칙 제안」,『통번역학연구』23(3): 291-327.

한현희 (2020)「한-노 기계번역, 어디까지 왔나?: Google과 Papago 번역 성능 비교를 기반으로」,『노어노문학』32(3): 63-93.

Brkić, Marija, Sanja Selijan and Tomislav Vičić (2013) 'Automatic and Human Evaluation on English-Croatian Legislative Test Set', *Computer Science* 7816: 311-378.

Carmo, Félix do (2022) Félix do Carmo 2022 'Debunking 'No Language Left Behind', 'Human Parity' and Other Machine Translation Myths', University of Surrey Centre for Translation Studies LITHME Seminar on 14 October 2022)

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

Chatzikoumi, Eirini (2020) 'How to Evaluate Machine Translation: A Review of Automated and Human Metrics', *Natural Language Engineering* 26: 137-161.

Defauw, Arne, Sara Szoc, Tom Vanallemeersch, Anna Bardadym, Joris Brabers, Frederic Everaert, Kim Scholte, Koen Van Winckel, Joachim Van den Bogaert (2019) 'Developing a Neural Machine Translation System for Irish', in Arne Defauw, Sara Szoc, Tom Vanallemeersch, Anna Bardadym, Joris Brabers, Frederic Everart, Kim Scholte, Koen Van Winckel and Joachim Van den Bogart (eds) *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, 32-38.

Doherty, Stephen (2017) 'Issues in Human and Automatic Translation Quality Assessment', in Dorothy Kenny (ed.) *Human Issues in Translation Technology*, London: Routledge, 131-148.

Giordano, Stella Szantova (2013) 'It's All Greek to Me: Are Attorneys Who Engage in or Procure Legal Translation for Their Clients at Risk of Committing an Ethical Violation?', *Quinnipiac Law Review* 31(2): 447-487.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou (2018) 'Achieving Human Parity on Automatic Chinese to English News Translation', ArXiv manuscript. Available at https://arxiv.org/abs/1803.05567.

Hutchins, W. John and Harold L. Somers (1992) *An Introduction to Machine Translation*, London: Academic Press.

Killman, Jeffrey (2014) 'Vocabulary Accuracy of Statistical Machine Translation in the Legal Context', in Sharon O'Brien, Michel Simard and Lucia Specia (eds) *Third Workshop on Post Editing Technology and Practice* (Proceedings of the 11th Conference of the Association for Machine Translation in the Americas), 85-98.

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

Kit, Chunyu and Tak, Wong Ming (2008) 'Comparative Evaluation of Online Machine Translation Systems with Legal Texts', *Law Library Journal* 100(2): 299-321.

Koehn, Phillip and Rebecca Knowles (2017) 'Six Challenges for Neural Machine Translation', in Thang Luong, Alexandra Birch, Graham Neubig and Andrew Finch (eds) *Proceedings of the First Workshop on Neural Machine Translation Association for Computational Linguistics*, 28-39.

Moorkens, Joss (2017). ‘Under Pressure: Translation in Times of Austerity’, *Perspectives* 25(3): 464-477.

Nurminen, Mary (2020) ‘Raw Machine Translation Use by Patent Professionals: A Case of Distributed Cognition’, *Translation, Cognition & Behavior* 3(1): 100-121.

Olohan, Maeve (2015) *Scientific and Technical Translation*, London: Routledge.

Poliquen, Bruno (2015) ‘Full-text Patent Translation at WIPO: Scalability, Quality and Usability’, *Proceedings of the 6th Workshop on Patent and Scientific Literature Translation*, Available at https://aclanthology.org /2015.mtsummit-wpslt.1.

Poliquen, Bruno (2017) ‘WIPO Translate: Patent Neural Machine Translation Publicly Available in 10 Languages’, Patent and Scientific Literature Translation Workshop at MT Summit Conference, Nagoya, Japan.

Premoli, Valeria, Elena Murgulo and Diego Cresceri (2019) ‘MTPE in Patents: A Successful Business Story’, *Proceedings of MT Summit XVII*, volume 2, 36-41.

Pym, Anthony (2019) ‘How Automation through Neural Machine Translation Might Change the Skill Sets of Translators’, Draft article written as part of the project ‘Language Competence and Work’. Available at https://www.academia.edu/40200406/How_automation_through_neural_machine_translation_might_change_the_skill_sets_of_translators.

Rivera-Trigueros, Irene (2022) ‘Machine Translation Systems and Quality Assessment: A Systemic Review’, *Language Resources & Evaluation* 56: 593-619.

Ragni, Valentina and Lucas Nunes Vieira (2022) ‘What Has Changed with Neural Machine Translation? A Critical Review of Human Factors’, *Perspectives* 30(1): 137-158.

Roberts, Ben (2022) ‘Machine vs. Human Translation: When to Use Which for Legal Translation’. Available at https://www.attorneyatwork.com/machine-translation-vs-human-translation-when-to-use-which-for-legal-translation/.

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

Somers, Harold (2007) 'The Use of Machine Translation by Law Librarians — A Reply to Yates', *Law Library Journal* 99(3): 611-620.

Smyth, Darren, Robert Barker and Timothy Belcher (2015) 'Rage against the (Translation) Machine', *Journal of Intellectual Property Law & Practice* 10(3): 153-154.

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

Vieira, Lucas Nunes (2020) 'Machine Translation in the News: A Framing Analysis the Written Press', *Translation Spaces* 9(1): 98-122.

Vieira, Lucas Nunes, Minako O'Hagan and Carol O'Sullivan (2021) 'Understanding the Societal Impacts of Machine Translation: A Critical Review of the Literature on Medical and Legal Use Cases', *Information, Communication & Society* 24(11): 1515-1531.

Wiesmann, Eva (2019) 'Machine Translation in the Field of Law: A Study of the Translation of Italian Legal Texts into German', *Comparative Legilinguistics* 37: 117-153.

Yates, Sarah (2006) 'Scaling the Tower of Babel Fish: An Analysis of the Machine Translation of Legal Information', *Law Library Journal* 98(3): 481-502.

[Abstract]

A Case Study of the Evaluation of Legal and Patent Korean-English Machine Translations by a Domain-Specific NMT Engine

Jieun Lee & Hyoeun Choi
(Ewha Womans University)

This paper addresses the quality of Korean-English legal and patent translation outputs by a commercial neural machine translation engine customized for legal and patent translations. The current research is based on both automatic and human evaluations of Otran's English translations of Korean statutes and Korean titles of invention and abstracts extracted from patent gazettes. In automatic evaluation, both BLEU and METEOR scores revealed that legal translation outperformed patent translation. Human evaluation results confirmed the automatic evaluation results, showing Otran's legal translation receiving better evaluation than its patent translation. According to the error comments provided by evaluators, terminology and other errors, mostly stylistic issues, were the most prevalent error types in the legal translation, while terminology and syntax errors were the most frequent in the patent translation. In the legal translation, accuracy and fluency errors were far scarcer than in the patent translation. The results suggest that the domain-specific NMT engine needs improvement in handling terminology in both legal and patent translation, and its legal translation output proved to be good enough for gisting. The findings from this case study cannot be generalized and thus call for further research.

Keywords: NMT, legal MT, patent MT, automatic translation, MT evaluation
주제어: NMT, 법률기계번역, 특허기계번역, 자동번역, 기계번역 평가

이지은(1저자)
이화여자대학교 통역번역대학원 교수
jieun.lee@ewha.ac.kr
관심 분야: 통번역교육, 법률통번역, 번역기술

최효은(공동저자)
이화여자대학교 통역번역대학원 초빙교수
hyoeun.choi@ewha.ac.kr
관심 분야: 번역교육, 법률번역, 번역기술

논문 투고: 2023년 1월 30일
1차 심사 완료: 2023년 3월 8일
2차 심사 완료: 2023년 3월 15일
게재 확정: 2023년 3월 23일