

## 인공신경망 특허 기계번역 성능에 관한 연구 - Patent Translate와 WIPO Translate 한영 번역 결과물의 누락과 통사 오류 분석을 중심으로 -

이지은\* · 최효은\*\*

*Jeun Lee and Hyeoun Choi (2022). A study on the quality of patent neural machine translation: A comparison of omission and syntactic errors in the Korean-English translations by patent-specialized Patent Translate and WIPO Translate. This paper aims to evaluate the quality of patent translations produced by the two patent-specialized machine translation engines, EPO's Patent Translate and WIPO's WIPO Translate. For manual evaluation, four experienced patent translators or patent translation service managers evaluated the quality of 106 English sentences from the translations of 30 Korean patent abstracts by the two MT engines. In the automatic evaluation, Patent Translate slightly outperformed WIPO Translate, whereas in the manual evaluation WIPO Translate outperformed Patent Translate. According to the error annotations provided by the evaluators, WIPO Translate produced more omission errors than Patent Translate but handled the complex syntax of the source text better, while Patent Translate produced more syntactic errors than WIPO Translate. The results indicate that in automatic evaluation, MT outputs with fewer omissions were rated higher, while in manual evaluation, comprehensible and accurate syntactic structures appeared to determine the overall quality evaluation. (Ewha Womans University, Korea)*

---

\* 이화여자대학교 통역번역대학원, 교수, 교신저자

\*\* 이화여자대학교 통역번역대학원, 초빙교수, 공동저자

**Keywords: Korean-English patent machine translation, patent-specialized neural machine translation engine, machine translation quality evaluation, Patent Translate, WIPO Translate**

**주제어: 한영 특허 기계번역, 특허전문 인공지능망 번역엔진, 기계번역 품질 평가, Patent Translate, WIPO Translate**

## 1. 서론

2016년 구글이 인공지능망 기계번역(NMT)을 세상에 내놓으면서 문법기반, 통계기반에 의존하던 기계번역 기술의 패러다임 전환이 이루어졌고, 이후 인공지능망 기술을 적용한 파파고, 카카오 등 무료 범용 기계번역 엔진이 속속 등장하였다. 가장 대표적인 ‘Google Translate(이하 구글번역)’은 100개 이상 언어로 연간 30조 문장 이상을 번역할 정도로 많은 사람들이 이용하고 있다(Kuczmarski 2018). 특허 문건은 길고 복잡한 통사구조와 기술용어 비중이 높아 전문성이 요구됨에도 불구하고(Rossi and Wiggins 2013: 115; Ying et al. 2021: 13) 기계번역이 활발하게 활용되고 있다. 그 이유는 특허 번역 수요 물량이 많고, 해외 특허 데이터베이스에서 선행 기술을 검색하고 해당 특허청에서 제공하는 기계번역을 통해 발명의 요지를 파악하기 위한 수요가 높기 때문이다(Olohan 2015: 171; 최효은과 이지은 2017: 142). 일례로 미국 특허청(PTO)은 거절 심사용으로 기계번역을 활용한다(USPTO 2018, §1207.02).

특허 기계번역의 주된 목적 중 하나는 해당 기술 분야에 대해 통상의 지식을 가지고 있는 사람들이 외국어로 작성된 특허 문서를 읽고 내용의 관련성을 판단할 수 있도록 요지를 제공하는 것이다(Nurminen 2020: 100; Olohan 2015: 167; 최효은과 이지은 2017: 143). 특허 기계번역의 품질에 대한 연구는 여전히 소수에 불과하며, 기존 연구에 의하면 특허 기계번역 품질은 그다지 높지 않았다(Rossi and Wiggins 2013; Tsai 2017; 최효은 2016; 최효은과 이지은 2017). 지금까지 특허 기계번역의 품질을 다룬 연구 상당수가 인공지능망 기술이 본격적으로 도입되기 전 대부분 통계기반의 엔진(SMT)을 대상으로 했다는 점을 고려할 때, 주요한 변곡점인 2016년 인공지능망 도입 후 특허 기계번역의 품질에 변화가 있는지 살펴보는 것이 유의미하다고 여겨진다.

www.kci.go.kr

본고는 특허 분야에 특화된 번역 엔진인 ‘Patent Translate’와 ‘WIPO Translate’의 한영 특허초록 기계번역 품질을 비교 평가해보고자 한다. 유럽특허청(EPO)이 제공하는 ‘Patent Translate’는 유럽특허청과 구글이 개발하여 한국어를 비롯한 총 32개 언어에 대해 기계번역 서비스를 제공한다.<sup>1)</sup> 구글이 기계번역에 인공지능경망 방식을 도입하면서 품질이 비교적 양호한 것으로 평가 받는다(Olohan 2015: 166; 최효은과 이지은 2017: 142). 다만 ‘Patent Translate’와 구글번역 간의 차이는 명확하지 않다.<sup>2)</sup> ‘WIPO Translate’는 특허협력조약(PCT)에 따른 해외 출원을 관장하는 UN 산하 지식재산권기구(WIPO)가 제공하는 인공지능경망 기계번역 서비스다. WIPO는 2009년부터 구문을 기반으로 한 통계기반 기계번역 시스템인 ‘TAPTA’<sup>3)</sup>를 사용하다가(Olohan 2015: 126), 한국어를 포함한 특허협력조약(PCT) 10개 공식 언어의 번역이 가능한 인공지능경망 기반 ‘WIPO Translate’를 2017년부터 제공하고 있다.<sup>4)</sup> 이 두 인공지능경망 기계번역 엔진은 특허 분야에서 활용되고 있는 대표적인 특허 전문 무료 기계번역 엔진으로 한국어를 출발어 및 도착어로 하는 방향의 번역을 지원하여 국내 특허업계 종사자들의 활용 빈도가 높다는 점에서 특허 전문 기계번역 엔진으로서 연구 가치가 있다고 판단되었다.

인공지능경망 도입 전후 규칙기반 기계번역과 구문을 바탕으로 한 통계기반 기계번역(SMT)과 인공지능경망 기계번역(NMT)의 품질을 비교 평가한 연구들이 비교적 활발하게 이루어지고 있다(Castilho et al. 2017; Costa et al. 2015; Kinoshita et al. 2017 등). 국내에서는 2016년 이후 구글번역과 파파고로 대별되는 범용 기계번역의 품질에 대한 연구가 주를 이룬 점을 고려할 때(강병규와 이지은 2018; 김동미 2018; 서보현과 김순영 2018; 이준호 2019; 한현희 2020 등), 특허 전문 번역엔진인 ‘Patent Translate’나 ‘WIPO Translate’의 결과물 품질을 살펴보는 것은 국내 관련 연구의 범위를 확장하는 의미가 있다고 판단되었다.

본고의 구성을 간략히 소개하자면 다음 장에서는 특허 기계번역과 기계번역 평가 관련 선행연구를 살펴보고, 이어서 본 연구의 연구문제와 연구방법, 그리고 연

- 
- 1) EPO 공식 홈페이지. <https://www.epo.org/news-events/press/releases/archive/2013/20131217.html>
  - 2) 2020년 11월 2일자 공개된 KPA 중 특허의 출원번호 및 발명의 명칭은 각각 1020200022186 캠핑카 오수처리장치, 1020190050822 원적외선 스마트 난방기, 1020200001985 악취제거 신발이다. 본 연구에 앞서 한국특허초록(KPA) 등 3건을 선정하여 한국어로 작성된 발명의 요약 부분의 범용 구글번역과 ‘Patent Translate’의 영어번역 결과물을 비교한 결과 차이가 없었다.
  - 3) WIPO Translate의 전신이라 할 수 있는 TAPTA는 ‘Moses’를 바탕으로 WIPO가 1억 8천만 단어의 자체 특허문서로 학습시킨 특허 전문 기계번역 엔진이다(Poliquen 2015: 2-3; 최효은 외 2017: 143).
  - 4) WIPO 공식 홈페이지. <https://patentscope.wipo.int/translate/wtapta-user-manual-en.pdf>

구결과를 차례로 소개한다. 마지막 결론에서는 연구결과를 요약하고 본 연구의 한계점과 함의를 논하며 글을 맺고자 한다.

## 2. 주요 선행 연구

### 2.1. 기계번역 품질평가 방식과 기준

선행 연구에 나타난 특히 기계번역의 품질 평가 방식과 기준을 간략히 살펴보면 크게 자동평가와 수동평가의 두 가지 방식이 있다. 사람에 의한 수동평가에 비해 시간과 비용이 적게 드는 자동평가는 기계번역 결과물과 사람 번역과의 유사성을 기준으로 평가한다(Rossi and Wiggins 2013: 118). 자동평가는 객관적이고 효율적이라는 장점이 있고 어휘 정확성 등을 평가하는 데에 효과적이거나 가독성 평가와 오류 분석 면에서 한계가 있다(Rossi and Wiggins 2013: 118; 최효은과 이지은 2017; 이준호 2019; 정혜연 2018; 정혜연 외 2020). 반면 수동평가는 세밀하고 유용한 정보를 제공할 수 있어 자동평가에 비해 선호되는 방식이기는 하나 평가자의 주관적인 판단에 의존하는 것과 시간과 비용의 소모가 심하다는 단점이 있다(Castilho et al. 2018: 10; Coughlin 2003: 63; Dorr et al. 2011; Popovic 2018: 154; 최효은과 이지은 2017: 147). 이러한 한계를 보완하고 기계번역의 품질을 평가하는 데에 유용한 정보를 얻기 위해서는 자동평가와 수동평가가 병행되어야 한다(Chatzikoumi 2020: 158; 이준호 2019; 최효은과 이지은 2017: 147).

일반적으로 기계번역의 수동평가에서 충실성(fidelity)<sup>5)</sup>과 가독성(intelligibility or readability)<sup>6)</sup>의 두 가지가 널리 사용되는 기준이다(Bazrafshan 2014: 17; Brkic et al. 2013: 314; Castilho et al. 2018: 18; Chatzikoumi 2020: 138). 충실성은 기계번역 결과물이 원문 또는 사람의 번역과 비교했을 때 기계번역 결과물이 얼마나 원문이 담고 있는 의미와 정보를 충실하게 전달하고 있는지의 여부를 가리키며 원문 대비 내용 전달의 정확성을 평가하는 기준이다(Bazrafshan 2014: 17; Lavie 2013: 7; 최효은과 이지은 2017: 147). 한편 가독성은 기계번역 결과물이 목표 언어 규범을

5) 충실성은 정확성(accuracy) 또는 적절성(adequacy)으로도 불린다.

6) 가독성은 유창성(flouency)으로도 불린다.

따르고 자연스러운지 언어적 측면을 판단하며 원문 없이 평가 가능하다(Castilho et al. 2018: 18; Lavie 2013: 7; 최효은과 이지은 2017: 147). 가독성에는 문법, 적절한 단어의 선택 여부, 텍스트의 스타일 등이 포함될 수 있으며(Bazrafshan 2014: 17; 최효은과 이지은 2017: 147), 문법 오류뿐 아니라 오역과 미번역도 가독성에 영향을 미칠 수 있다(Castilho et al. 2018: 18).

두 가지 평가 기준의 상호 영향을 완전히 배제할 수 없기 때문에 충실성과 가독성 두 가지 평가 기준을 하나로 합쳐서 용인성(acceptability)이라는 단일 평가 기준을 사용하기도 한다(Coughlin 2003: 64; 박경리 외 2013: 207; 최효은과 이지은 2017: 147). 번역품질 요인분석 연구(한승희 2021: 165)에 따르면 목표 언어 관습에서 자연스러운 언어적 표현에 해당하는 유창성이 의미의 적합성에 포함되어 의미와 표현을 하나의 측정항목으로 보는 것이 타당하다는 결과도 확인된다.

수동평가의 방식으로는 척도 평가와 서열 평가(ranking)가 대표적이다. 척도 평가에서는 보통 4점 또는 5점 척도를 사용하여 문장 또는 세그먼트(segment)<sup>7)</sup> 단위로 평가한다. 이로 인해 맥락에 대한 고려와 텍스트 차원에서 문장간 응집성에 대한 평가가 부실하다는 점 등이 문제로 지적된다(Castilho et al. 2018: 18; Chatzikoumi 2020: 147; Läubli et al. 2018). 서열 평가는 동일한 원천 텍스트에 대한 복수의 기계번역 결과물을 도출하여 결과물 간의 비교 평가를 위해 사용한다. 평가자는 구체적인 평가 기준 또는 가독성과 같이 비교적 일반적인 평가 기준에 따라 무작위 또는 익명화 처리한 목표텍스트 문장의 순위를 매긴다. 최대 세 가지 중에서 최상을 선택하거나 최상부터 최하까지 순위를 매기는 방식으로 간단하고 신속하며, 효율적이라는 평가를 받는다(Castilho et al. 2018: 21). 평가 연구에서는 연구의 타당성을 확보하기 위해서는 평가자 정보와 평가자 훈련 및 전문성, 평가 가이드라인, 조작적 정의 요건 등이 중요하다(Chatzikoumi 2020: 146; Doherty 2017). 정확한 오류 분석을 위해서는 오류 유형의 일반화 가능성은 물론이고 분석가(annotator) 변수, 유형분석의 일관성, 오류의 정의, 오류 항목의 수 등을 세심하게 규정할 필요가 있다(이상빈 2020: 86). 오류 항목이 많을수록 상세한 정보를 제공해줄 수 있지만 분석 작업의 인지부하를 높이고 일관성 및 분석가 간 일치도를 저하시킬 수 있다는 점을 유의해야 한다(Popovic 2018: 139).

특히 문서의 기계번역 결과를 평가할 때는 다른 기계번역에서와 마찬가지로 번역이 어떤 목적을 띠며, 어떠한 맥락에서 사용될 것인지를 반드시 고려해야 한다

7) 완벽한 문장은 아니나 평가의 단위가 되는 구 또는 절을 말한다.

(Doherty 2019). 특히 기계번역 수동평가 기준에는 일반 기계번역 평가에서 많이 사용되는 충실성과 가독성 외에 검색 가능성이 특히 기계번역의 품질을 판단하는 기준으로 꼽힌다. 로시와 위긴스(Rossi and Wiggins 2013: 116)는 특히 분야에서 기계번역 결과물이 발명의 요건인 신규성이나 진보성을 검증하거나 침해 여부를 판단하기 위한 도구로 주로 사용되고 있다는 점을 고려하여 검색 가능성과 가독성을 핵심 평가 기준으로 꼽았다. 검색 가능성을 위해서는 용어 완결성과 정확성이 필수요건이며, 가독성은 문장에서 구의 순서와 논리 연결의 정확성에 의해 결정된다(Rossi and Wiggins 2013: 116). 이들은 특히 기계번역의 경우 요지 파악이 주목적이기 때문에 다른 분야의 기계번역 평가에서와 달리 형태와 통사 오류는 일반적인 요지 파악이나 검색에 상대적으로 미치는 영향이 크지 않으므로 평가에서 배제 가능하다는 견해를 보였다(Rossi and Wiggins 2013: 116). 그렇지만 본 연구에서는 가독성을 논할 때 통사 오류는 번역 품질에서 중요한 요소로 인식하고 분석 대상으로 삼았다.

## 2.2. 특허 기계번역 평가 관련 연구

특허 기계번역에 대한 연구가 많지 않지만 여기에서는 본 연구에 영향을 미친 주요 연구 몇 가지만 소개하기로 한다. 로시와 위긴스(2013)는 특허 전문 번역엔진 LexisNexis를 대상으로 일본어 특허 명세서에서 길이와 통사구조가 다양한 1,000개 문장을 추출하여 영어 기계번역 결과물에 대해 자동평가와 수동평가를 실시하였다. 자동평가는 BLEU 35점을 기준으로 그 이상은 ‘수용 가능’, 그 미만은 ‘수용 불가능’으로 정했다. 이때 수동평가는 2명의 전문가가 문장별로 평가하게 하였다. 용어, 정보 누락, 추가, 단어 순서를 평가 기준으로 채택하였다. 이 중 처음 세 가지는 검색 가능성의 측면에서 적절성을 평가하는 기준이며, 마지막 네 번째 단어 순서는 가독성에 대한 평가 기준이다(Rossi and Wiggins 2013: 120). 자동평가의 단점을 수동평가로 보완할 수 있도록 하였으며 자동평가와 수동평가 결과 모두 ‘수용 가능’인 경우 최종적으로 수용 가능한 것으로 정하였다(Rossi and Wiggins 2013: 124). 평가결과 검색가능성 측면에서는 94.3%라는 높은 수준의 수용 가능성을 보인 데 비해 가독성에 대한 수용 가능성은 45.04%로 낮은 수준을 보였다(Rossi and Wiggins 2013: 124).

카스틸호 외(Castilho et al. 2017)는 중영 특허 명세서에 대한 NMT 엔진과 SMT

엔진의 번역 결과물 품질을 자동평가와 수동평가로 비교하였다. BLEU를 사용한 자동평가 결과에 의하면, 발명의 명칭 번역에서 NMT가 좀 더 우수하고 요약 번역에서는 SMT가 조금 우세한 것으로 나타나 NMT가 SMT에 비해 반드시 우수한 번역 결과물을 산출한 것은 아닌 것으로 확인되었다(Castilho et al. 2017: 114). 두 명의 평가자에 의한 수동평가에서는 SMT가 NMT보다 결과물 품질이 우세한 것으로 나타났는데 문장 길이에 따라 결과의 차이가 있었다. 한편 NMT 번역 오류 중 누락이 가장 많았고 SMT에서는 문장 구조 오류 비중이 더 높았다. 오류 없이 완벽한 번역문은 SMT의 경우 25%, NMT의 경우 2%를 차지하였다. 자동평가에서는 특히 발명의 명칭 번역에 대해 SMT에 비해 NMT가 우세한 것으로 보였으나 수동평가는 SMT를 높게 평가하였다.

프레몰리 외(Premoli et al. 2019)는 자체적으로 구축한 인공지능망 기계번역엔진의 영어-이탈리아어 특허 기계번역에 대한 수동평가 연구다. 평가자 2명이 4점 척도를 사용하여 평가한 결과, 엔진이 학습한 전문 분야인 기계 분야의 특허 번역에 대해서 이해가 어려운 세그먼트는 10개, 이해가 아예 불가능한 세그먼트는 6개에 불과했으며, 번역 문제는 주로 문법 오류, 지나친 직역, 생경한 어휘에 대한 추측 때문으로 확인되었고, 그 외에 문장부호 오류 등은 매우 미미하게 나타났다(Premoli et al. 2019: 37). 이 연구자들은 특허 기계번역 품질이 대체로 용인 가능한 수준이라고 평가한 한편 용어 번역의 비일관성은 특허문서라는 장르 상 매우 치명적인 품질 저하 요인으로 간주하였다(Premoli et al. 2019: 37).

최효은과 이지은(2017: 151)은 한국 특허청의 특허검색사이트인 키프리스에서 무료로 제공하는 SMT인 K2E-PAT를 통해 얻은 반도체 분야의 특허 요약서 총 38건에서 추출한 100 문장을 참조번역인 KPA 100 문장과 각각 비교하였다. 자동평가를 위해서는 BLEU 점수를 사용하고, 수동평가를 위해서는 전문 번역사 2명이 충실성과 가독성의 두 가지 평가 기준을 적용하고 5점 척도를 사용하여 충실성과 가독성 점수를 각각 도출하였다(최효은과 이지은 2017: 165). 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 SMT 중영 번역 결과물 중 발명의 명칭 473개에 대한 수동평가 연구다. 대만특허청 소속 번역사들의 수동평가 결과를 토대로 차이(2017: 147)는 철자법 오류(구두점, 대소문자 오류 포함), 형태적 오류(동사, 명사 등), 어휘(어휘 추가 및 누락), 의미 오류(다의어, 동음이의어, 표현 오류)와

통사적 오류(접속사, 전치사, 관사, 문장구조) 등 크게 다섯 가지로 오류를 분류하였다. 발명의 명칭에서 발견된 오류는 총 692개로 발명의 명칭마다 약 1.5개의 오류가 있는 것으로 드러나 품질이 낮은 것으로 평가받았다(Tsai 2017: 149). 오류 유형상 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).

한편 본 연구의 대상 중 하나인 ‘WIPO Translate’의 특허 문건 번역 결과와 범용 엔진인 구글번역의 BLEU 점수를 비교한 폴리퀸(Poliquen 2017)에 따르면 ‘WIPO Translate’가 구글번역에 비해 대체로 높은 점수를 받았다. 영한 번역의 경우 ‘WIPO Translate’의 BLEU 점수는 39.20, 구글번역의 점수는 32.65였다. 하지만 그의 연구는 자동평가에만 기반하였으며, 아시아 언어를 분석하기는 하였으나 한영 특허기계번역에 대한 분석은 이루어지지 않았다.

### 3. 연구 문제 및 방법

본 연구는 특허 전문 기계번역 엔진의 한영 특허 초록 번역 결과물의 품질을 자동평가와 수동평가를 통해 비교 분석하고, 번역 결과물의 차이 및 특징을 오류 분석을 중심으로 살펴보고자 한다. 이를 위해 반도체 분야의 한국어 특허 공보 30개를 임의로 추출하여 공보의 내용 중 요약 부분에 대해 ‘Patent Translate’와 ‘WIPO Translate’에 의한 영어 번역 결과물의 품질을 비교 분석하였다.

자동평가로는 기계번역의 대표적인 자동평가 방식인 BLEU(Papineni et al. 2002)와 METEOR(Lavie and Agarwal 2007) 두 가지를 활용하였다. 참조번역으로는 인간번역인 KPA(Korean Patent Abstracts)과 비교하여 자동평가 결과를 도출하였다.

수동평가를 위해서 4명의 특허 번역 전문가(E1~E4)를 섭외하여 한영 특허 기계번역 결과물을 평가하도록 하였다. 평가의 공정성을 위해 본 연구에서는 가급적 전문적인 특허 번역 경험이 풍부한 전문가를 위주로 평가자를 구성하고자 하였으며, 평가자들은 모두 특허 번역 경력 10년 이상으로 전문 특허번역사와 특허 번역 관리자로 구성하였다. 평가자와 관련하여 구체적으로 E1은 인하우스 특허번역사, E2는 프리랜서 특허번역사, E3은 인하우스 특허 번역 매니저, E4는 프리랜서 특허번역사로 모두 10년 이상의 특허 번역 경력을 갖추었다. 평가의 공정성을 위해 번역 엔진의 종류를 평가자들에게 공개하지 않았고, 번역 결과물의 순서 또

한 무작위로 섞어서 제시하였다.

현업에 종사하는 평가자들의 바쁜 일정을 고려하여 충분한 시간 여유를 가지고 평가할 수 있도록 8주 기간 내에 30개 요약의 53개 문장에 대해 각 기계번역 엔진에서 추출한 번역 결과물(총 106개 문장, 8,065단어 분량)을 평가해 줄 것을 요청하였다. 평가자들은 문장 단위로 충실성과 가독성의 전반적인 품질을 바탕으로 5점 척도(품질 매우 불량(0점) ~ 품질 매우 양호(4점))에 따라 점수를 매겼다. 평가자들이 평가 대상 텍스트가 어느 기계번역 엔진에서 나온 것인지 알 수 없는 상태에서 평가하도록 텍스트를 혼합하여 제공하였다. 평가자들은 MQM-DQF(Lommel 2018)과 로시와 위긴스(2013)의 기계번역 평가 기준을 차용한 추가, 누락, 통사, 용어, 기타 등 다섯 가지 평가 기준에 따라 평가하고 오류 분석을 하였다. 해당 문장에서 각 평가 기준에 해당하는 오류가 있을 시 평가자들이 오류에 대해 구체적인 내용을 상술하게 한 것이다. 최효은과 이지은(2017)보다 세분화한 평가 기준을 적용하고 평가자들이 오류 코멘트를 제시하도록 하여 평가자의 눈으로 본 오류 분석을 시도한 차이가 있다. 평가자에게 평가 방식에 대한 설명을 충분히 제공하고, 평가 작업을 위해 평가자들에게 아래 표 1과 같은 평가표를 이용하여 문장 별로 평가하도록 하였다.

표 1. 수동평가를 위한 평가표

(예시) 요약 번역 1(텍스트 #1)					
종류	문장 1				
점수	0	1	2	3	4
평가 의견	추가	1. 번역의 A는 원문에 없는 내용임. 2. 번역의 B는 원문에 없는 내용임.			
	누락	1. 원문의 C가 번역에서 누락됨. 2. 원문의 D가 번역에서 누락됨.			
	통사	1. The present invention to a circuit.은 완전하지 않은 통사구조를 지님. 2. 원문에 의하면 A, B, C, D가 발명의 구성요소로 대등하게 나열되어야 하나 번역에서는 A, B, C만 대등하게 나열되고, D는 별도의 문장으로 번역됨.			
	용어	1. (a) 부정확한 용어의 사용으로 원문의 A는 A'로 번역되어야 하나 번역에서 B'로 번역됨. 2. (b) 용어 일관성의 문제로 원문에서 A가 2회 제시되는데, 번역에서 각각 B와 C로 일관되지 않게 번역됨.			
	기타	1. A는 오타임.			

## 4. 자료 분석

여기에는 전반적인 자동평가 및 수동평가 결과를 간단히 소개하고, 평가 결과에서 두드러지게 나타난 두 엔진 간 차이인 누락과 통사 관련 오류에 대해 좀 더 구체적으로 분석하여 두 엔진의 평가 결과에 우열 요인을 중심으로 살펴보겠다. 본 장에서는 편의상 ‘Patent Translate’를 ‘PT’, ‘WIPO Translate’를 ‘WT’로 칭한다.

### 4.1. 평가 결과

본 연구에서 활용한 자동평가 기준인 BLEU와 METEOR 점수는 아래 표 2와 같다.

표 2. 자동평가 결과

번역 엔진	BLEU	METEOR
Patent Translate	35.42	35.7
WIPO Translate	33.80	31.3

표 2의 결과를 보면 BLEU와 METEOR 모두 PT가 보다 높은 점수를 기록하였음을 알 수 있다. 자동평가가 참조번역과의 유사성 정도에 따라 점수를 매긴다는 점을 고려할 때, 이는 PT가 KPA와 좀 더 유사한 특징을 가지고 있음을 뜻한다. 한편 이 수치는 이전에 SMT 기반의 K2E-PAT가 기록한 BLEU 22.90점(최효은과 이지은 2017: 165)과 비교했을 때, NMT 기반의 PT와 WT가 좀 더 나은 수준의 번역 결과를 보이는 것으로 판단할 수 있다. 이는 본 연구보다 5년 앞선 폴리퀸(Poliquen 2017)의 영한 특허기계번역 BLEU 수치에 비해 낮은 편이다. 당시에는 WT가 구글 번역을 앞섰는데, 본 연구에서는 WT에 비해 PT의 BLEU 수치가 조금 높다.

아래의 그림 1과 그림 2는 각각 PT와 WT에 대한 BLEU 결과이다.

BLEU:	<b>35.42</b>
Precision x brevity:	35.42 x 100.00
Type	1-gram 2-gram 3-gram 4-gram
Individual	66.40 42.43 28.66 19.49
Cumulative	66.40 53.08 43.22 35.42
Export data	<a href="#">CSV</a>

그림 1. Patent Translate에 대한 상세 BLEU 결과

www.kci.go.kr

BLEU:	<b>33.80</b>
Precision x brevity:	40.75 x 82.93
Type	1-gram 2-gram 3-gram 4-gram
Individual	72.13 47.67 33.48 23.96
Cumulative	59.82 48.63 40.34 33.80
Export data	<a href="#">CSV</a>

그림 2. WIPO Translate에 대한 상세 BLEU 결과

수동평가 결과를 살펴보면, 우선 0~4점의 5점 척도를 이용한 평가자들의 개별 평가 점수와 번역 엔진별 평균 평가 점수는 표 3과 같다.

표 3. 5점 척도에 따른 번역 엔진별 수동평가 결과

구분	Patent Translate	WIPO Translate
E1	1.92	1.94
E2	2.13	2.19
E3	1.98	2.19
E4	2.62	2.56
전체	2.16 <sup>8)</sup>	2.22 <sup>9)</sup>

표 3의 수동평가 결과를 보면, 자동평가 결과와 달리 E4를 제외한 나머지 평가자 3인은 정도의 차이는 있으나 근소하게나마 WT를 PT보다 높게 평가하였다. 평가자들의 평가 점수를 합산하여 평균을 계산한 전체 점수 또한 WT가 2.22점으로 2.16점인 PT를 조금 앞섰다. 2점 초반대라는 점에서 전반적으로 품질은 낮은 것으로 평가되었다.

평가자 4인이 제공한 평가 의견을 살펴보면, PT 결과물 53문장에 대한 오류 코멘트는 총 442건, WT 결과물에 대한 오류 코멘트는 총 416건으로 집계되었다. PT에 대한 오류 코멘트가 WT에 비해 근소하게 많아 WT의 품질이 조금 앞선 것으로 해석할 수 있다. 두 번역 엔진에 대한 평가 의견은 총 858건이었는데 이를 평가 기준에 따라 분류한 결과는 아래의 표 4와 같다.

- 8) 평가자 4인이 시행한 평가 결과에 대한 급내 상관 계수는 0.954로 신뢰 수준이 매우 높다.  
 9) 평가자 4인이 시행한 평가 결과에 대한 급내 상관 계수는 0.918로 신뢰 수준이 매우 높다.

표 4. 평가 기준별 평가 의견

평가 기준	추가	누락	통사	용어	기타	전체
코멘트 수 (%)	32 (3.7)	154 (18.0)	323 (37.6)	210 (24.5)	139 (16.2)	858 (100)

표 4에 의하면, 평가자들이 지적한 문제는 통사(37.6%)가 가장 많고 그 다음으로 용어(24.5%)와 누락(18.0%) 순이다. 번역에서 통사의 문제가 가장 두드러진 이유는 특히 문서 특유의 고도로 정형화된 장르 특성인 ‘특히 특유의 문체’ (Nurminen 2020: 106)에 기인한다고 할 수 있다. 일반적인 문장에 비해 지나치게 길고 복잡한 통사 구조를 가지는 요약의 문장을 기계가 정확하게 분석하여 이해하고 번역하는 데에는 아직 무리가 있는 현실에 대한 방증이기도 하다.

한편 여기서 기타(16.2%) 또한 적지 않은 비중을 차지하는데, 기타는 추가, 누락, 통사, 용어에 해당하지는 않으나 번역 품질에 영향을 미칠 수 있는 요소들로 대소문자 문제, 부정관사와 정관사의 사용 문제, 단복수형 사용 문제, 오타 문제, 문장부호 사용 문제가 대부분이었다. 일례로 ‘형식: 불필요한 대문자 사용(예, of the Facility and the Device)’, ‘번역 ‘and more particularly, to’에서 and와 to 앞의 콤마는 불필요함.’과 같은 코멘트가 여기에 해당한다.

이에 대해 PT와 WT를 비교한 결과는 아래 그림 3과 같다.

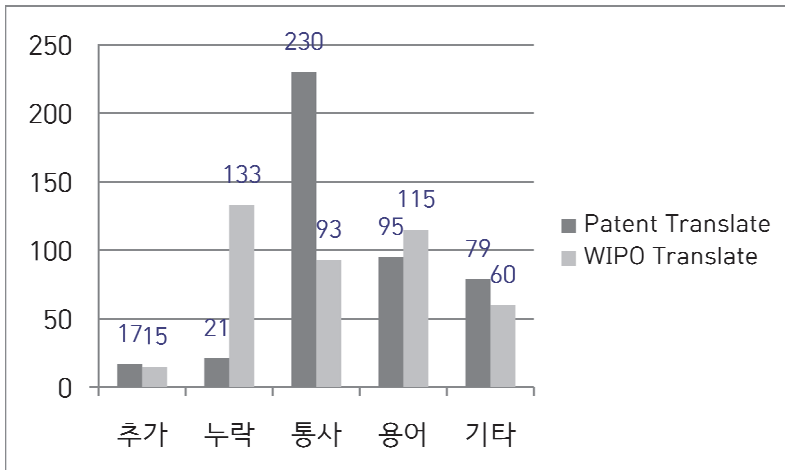


그림 3. 번역 엔진 간 오류 코멘트 수 비교

그림 3을 보면, 눈에 띄는 차이가 누락과 통사이다. 평가자들이 지적한 누락을 먼저 살펴보면, 두 번역 엔진에 대한 총 누락 154건 중 133건이 WT에서 발생했음을 알 수 있다. 상대적으로 누락이 거의 없는 PT가 수동평가에서와 달리 자동평가에서 전체적으로 높은 점수를 받았다는 사실을 볼 때, WT의 빈번한 누락이 자동평가에서의 점수 차이를 가를 수 있는 결정적인 요소일 수 있음을 추론해볼 수 있다.

반대로 통사의 경우, 전체 323건 중 230건이 PT에서 나타났다. WT의 통사의 경우는 불과 93건으로 PT에 대한 통사 지적의 절반에도 미치지 않는 적은 수인 것으로 나타났다. 이는 PT의 문장 구조가 품질에 영향을 미칠 만큼 심각한 오류가 다수 있음을 나타내며, WT 대비 수동평가에서 낮은 점수를 얻은 결정적인 요인일 수 있음을 시사한다. 즉, 평가자들은 누락보다는 통사 문제에 좀 더 가중치를 두어 각 문장을 평가했을 가능성이 크다는 점을 유추할 수 있다.

한편 그 외의 추가, 용어, 기타 문제는 두 번역 엔진 간 큰 차이 없이 비슷한 빈도로 나타나 두 엔진 간 품질 차이에 영향을 미친 결정적인 요소는 아닐 수 있음을 알 수 있다.

## 4.2. 누락과 통사의 오류 양상 비교

이제 평가자들이 가장 많이 지적한 주요 오류인 누락과 통사를 중심으로 두 기계번역 결과물에 나타난 오류를 보다 세부적으로 분석하겠다. 이 두 가지 오류는 실제 오류 개수와는 다르고 평가자별 차이가 있겠지만 PT와 WT에 대한 평가 결과 차이가 커서 두 개의 엔진 성능을 비교 분석하기 좋은 오류 유형이라 판단했다.

우선 누락과 관련해서 살펴보면 PT에서 21건, WT에서 133건이나 언급되었다. 엔진별로 개별 평가자가 지적한 누락 오류의 수는 아래 표 5와 같다.

표 5. 평가자별 각 엔진에 대해 지적한 누락 오류의 개수

구분	Patent Translate	WIPO Translate
E1	0	20
E2	12	37
E3	2	35
E4	7	41
총합	21	133

표 5에서 평가자별로 지적한 누락 오류의 개수를 살펴보면, Patent Translate와 WT에서 공통적으로 평가자 E2와 평가자 E4가 지적한 오류 개수가 4명 중 1, 2위로 가장 많다. 특히 PT에서는 E2와 E4가 지적한 누락 오류의 합이 전체 21개 오류 중 19건으로 전체의 90%를 차지할 정도로 대다수를 차지한다. 반면 E1과 E3이 지적한 누락 오류의 개수는 E2와 E4에 비해 모두 적은 편이다.

다음으로 각 누락 부분에 해당하는 원문 단어를 5개 단위별로 분류한 결과는 아래 표 6, 표 7과 같다.

표 6. Patent Translate의 누락 오류에 대한 길이 분류

원문 기준 누락 단어 수	0-5	6-10	11-20	전체
오류 개수 (%)	17 (81)	3 (14.2)	1 (4.8)	21 (100)

표 7. WIPO Translate의 누락 오류에 대한 길이 분류

원문 기준 누락 단어 수	0-5	6-10	11-20	21-30	31-40	41-50	51 이상	전체
오류 개수 (%)	53 (40)	21 (15.8)	20 (15)	13 (9.8)	10 (7.4)	10 (7.4)	6 (4.6)	133 (100)

표 6과 표 7을 비교해 보면, 앞서 여러 차례 언급한 바와 같이 PT이 WT에 비해 누락 오류의 수가 훨씬 적다. 그런데 여기에 더해 원문을 기준으로 한 누락 단어 수를 살펴보면, PT의 경우 전체 누락 오류의 절대 다수인 81%에 해당하는 17건의 오류가 모두 0-5단어 사이의 사소한 누락에 해당한다. 더욱이 11-20단어 사이의 상대적으로 긴 누락은 전체의 4.8%에 불과한 단 한 건에서만 찾아볼 수 있었다. 사소한 누락인 ‘0-5단어’에 해당하는 PT의 사례는 아래 [예시 1]과 같다.

[예시 1] Patent Translate의 누락 양상

ST: 본 발명에 따른 기관 지지 유닛은 기관의 배면(背面)의 중심 영역을 지지하도록 설치된 서셉터 및 서셉터의 둘레에 설치되어, 일면에 상기 기관 배면의 가장자리 영역을 지지하며, 상기 기관 배면의 가장자리가 지지되는 일면으로부터 기관 외측 방향으로 마련되고, 상기 기관 배면의 가장자리가 지지되는 일면에 비해 표면 높이가 높은 단턱부를 가지는 프레임을 포함한다.

TT: The substrate support unit according to the present invention includes a susceptor installed to support a central region of a rear surface of a substrate and a susceptor installed around the susceptor to support an edge region of the rear surface of the substrate on one surface, and the edge of the rear surface of the substrate is and a frame provided in an outward direction from one supported surface of the substrate, and having a stepped portion having a higher surface height than the one on which the edge of the rear surface of the substrate is supported.

상기 [예시 1]과 관련해서 평가자 E2는 “원문 ‘상기 기판 배면의 가장자리가 지지되는 일면에 비해’가 번역에서 ‘than the one on which the edge of the rear surface of the substrate is supported’로 번역되어 문맥에서 밑줄 친 일면이 ‘one’로만 번역되어 ‘면’이란 번역어가 누락됨.”으로 오류를 지적하였다. 즉, 원문의 ‘일면’에서 ‘면’이 ‘side’와 같은 표현으로 구체화되지 않으면서 원문의 의미가 누락된 것이다. 이 경우 원문 기준 한 단어의 누락이다.

이에 반해 WT의 경우를 살펴보면 누락 오류의 수가 PT에 비해 절대적으로 많은 것에 더해 0-5단어 사이의 비교적 사소한 누락은 전체 누락 오류의 40%에 그쳐 PT의 80%에 비해 사소한 누락의 비율이 훨씬 적은 것으로 드러났다. 그뿐만 아니라 PT의 최대 누락 단어수가 원문 기준 11-20단어 사이인데 반해 WT의 경우 20단어 이상의 누락도 빈번하게 발생하며 원문 단어수 기준 60단어 이상 누락되는 경우도 총 4.6%의 6건으로 비교적 그 수가 적기는 하지만 대량 누락도 존재한다는 사실을 확인할 수 있었다. 아래 [예시 2]는 WT의 번역 중 51단어 이상 누락에 해당하는 경우다.

#### [예시 2] WIPO Translate의 누락 양상

ST: 본 발명은 반도체 설비 장치의 상태 보고 및 제어 시스템에 있어서, 특정 반도체 설비 장치의 I/O(Input/output) 신호의 상태정보를 알기 위해 SVID 리스트 정보를 요청하고, 상기 요청에 대한 응답으로 SVID 리스트 정보를 전달받으며, 상기 전달받은 SVID를 상태정보 요청메시지에 실어 보내고 상태정보 응답 메시지를 통해 SV 정보를 받아 특정 I/O 신호의 현재 상태를 파악하는 상위 제어기, 특정 반도체 설비 장치와 연결되어 상기 연결된 반도체 설비 장치의 I/O 신호 상태파악을 위한 SVID 정보를 저장, 관리하는 설비 제어기 및, 상기 설비 제어기와 연결되지 않은 반도체 설비 장치와 연결하여 상기 연결된 반도체 설비 장치의 I/O 신호 상태정보 파악을 위해 상기 설비 제어기가 가지고 있는 SVID와

중복되지 않는 SVID를 할당해서 생성하고, 상기 상위 제어기의 SVID 리스트 정보 요청시 상기 설비 제어기가 가지고 있는 SVID 정보와 상기 생성된 SVID 정보를 더해서 상기 상위 제어기로 보고하는 VID 생성기를 포함하여 이루어진 반도체 설비 장치의 상태 보고 및 제어 시스템에 관한 것으로, 상위 제어기가 기존 설비 제어기에 연결된 I/O에 대해서뿐만 아니라, 설비 사용자가 임의로 장착한 설비(Facility)와 디바이스 그리고, 기존 설비 제어기가 SVID로 제공하지 못하는 I/O에 대해서도 현재 상태 파악을 할 수 있게 되어 제어를 할 수 있게 된다.

TT: In a state report and control system of a semiconductor equipment device, SVID list information is requested to know state information of an I/O (Input/Output) signal of a specific semiconductor equipment device, SVID list information is received in response to the request, The present invention relates to a state report and control system of a semiconductor equipment device comprising a VID generator for allocating and generating SVID which does not overlap with an SVID having the facility controller in order to grasp O signal state information, and a VID generator for adding the SVID information of the facility controller and the generated SVID information to the upper controller when the SVID list information of the upper controller is requested.

상기 [예시 2]에 대해 평가자 E4는 “원문의 ‘상기 전달받은 SVID를 상태정보 요청메시지에 실어 보내고 상태정보 응답 메시지를 통해 SV 정보를 받아 특정 I/O 신호의 현재 상태를 파악하는 상위 제어기, 특정 반도체 설비 장치와 연결되어 상기 연결된 반도체 설비 장치의 I/O 신호 상태파악을 위한 SVID 정보를 저장, 관리하는 설비 제어기 및, 상기 설비 제어기와 연결되지 않은 반도체 설비 장치와 연결하여 상기 연결된 반도체 설비 장치의’가 번역문에서 누락됨.”으로 평가하였으며, E4가 지적한 해당 부분은 원문 기준 57단어로 ‘51단어 이상’에 해당한다. 이와 같이 원문의 내용이 대량 누락되는 경우는 [예 1]과 같이 사소한 누락에 비해 정확성 차원에서 심각한 문제를 초래할 수 있다. 또한 기계번역 결과물에 의존해야 하는 독자의 경우 제대로 된 정보를 얻는 데 어려움이 있을 수 밖에 없다.

이와 같은 누락의 정도 차이와 평가 결과를 연계해서 보면, BLEU, METEOR의 자동평가 결과 PT가 WT에 비해 비교적 높은 점수를 받았는데, 그 이유가 WT의 이와 같은 대량 누락 문제 때문임을 유추할 수 있다. 자동평가의 기준이 되는 참조번역의 경우 이와 같은 대량 누락 사태는 찾아보기 어렵기 때문에 자동평가에서 WT의 번역 결과물이 참조번역과의 유사도가 낮을 수 밖에 없을 것이다. 한편 전문가에 의한 수동평가의 결과는 그 반대로 WT의 점수가 PT에 비해 전반적으

로 높은 편이었는데, 이는 아마도 기계번역에만 의존하는 일반 독자들과 달리 본 연구의 평가자들이 특히 번역 전문가들로 원문과 번역문을 비교 분석하여 읽는 능력을 갖추고 있으며, 따라서 WT에서 누락된 부분들을 단순히 길이와 양으로만 평가하는 것이 아니라 누락된 부분이 전체 문장에서 차지하는 중요성을 감안하여 평가한 것으로 유추해 볼 수 있다.

이어서 통사 오류 문제에 대해 살펴보도록 하겠다. 통사 오류는 PT에서 총 230건, WT에서 총 93건이 지적되었고, 누락과 달리 PT가 월등히 높은 통사 오류 수를 기록하였다. 엔진마다 평가자들이 지적한 통사 오류의 수를 집계해 본 결과는 아래 표 8과 같다.

**표 8. 평가자별 각 엔진에 대해 지적한 통사 오류의 개수**

구분	Patent Translate	WIPO Translate
E1	47	23
E2	55	19
E3	69	29
E4	59	22
총합	230	93

PT와 WT에서 공통적으로 평가자 E3이 지적한 오류의 개수가 두 번역 엔진에서 모두 가장 많다.

한편 평가자 4인이 제시한 통사 오류를 연구자들이 다섯 가지로 세부 분류하였다. 첫 번째, ‘구성요소 오류’는 원문에 제시된 발명의 여러 구성요소가 통사 문제로 인해 번역에서 반영되지 않거나 잘못 반영된 오류를 가리킨다. 두 번째, ‘수식어구 오류’는 통사 문제로 피수식어와 수식어구의 관계가 원문과 달라진 오류를, 세 번째, ‘불명확한 번역 오류’는 문장의 통사구조 자체에 문제가 없고 번역이 앞선 두 오류와 같이 원문과 완전히 다른 의미를 전달하는 것은 아니나 번역이 전달하는 의미가 불분명하여 개선의 여지가 있는 경우를, 네 번째, ‘불완전한 문장 오류’는 번역 문장의 통사구조 자체에 문제가 있어 완전한 문장이 이루어지지 못한 비문의 오류를, 마지막으로 ‘문장 나눔 오류’는 원문의 문장을 불필요하게 번역에서 분절하여 나누어 번역함으로써 발생하는 오류를 말한다. 각 유형별 오류 건수와 비중은 아래의 <표 9>와 같다.

표 9. 각 유형별 통사 오류 분류별 건수와 비중

구분	Patent Translate	WIPO Translate
구성요소 오류	54개(23.5%)	17개(18.2%)
수식어구 오류	45개(19.6%)	13개(14%)
불명확한 번역 오류	92개(40%)	53개(57%)
불완전한 문장 오류	24개(10.4%)	6개(6.5%)
문장 나눔 오류	15개(6.5%)	4개(4.3%)
총합	230개(100%)	93개(100%)

표 9에서 PT와 WT의 통사 오류 비중을 비교해 보면, PT에서 ‘(통사로 인한) 불명확한 번역 오류’를 제외한 ‘구성요소 오류’, ‘수식어구 오류’, ‘불완전한 문장 오류’, ‘문장 나눔 오류’의 비중이 WT에 비해 모두 높다. 한편 WT의 경우, ‘불명확한 번역 오류’가 전체 통사 오류의 57%로 절반 가량을 차지하며 PT에 비해 유일하게 높은 비중을 보인다. 여기서 ‘불명확한 번역 오류’에 해당하는 사례는 아래 [예시 3]과 같다.

[예시 3] WIPO Translate의 불명확한 번역 오류

ST: 본 발명에 의하면, 검사장치의 구조가 간단하여 제조과정을 단순화할 수 있어 비용을 절감할 수 있고 효과가 있다.

TT: According to the present invention, the structure of the inspection device is simple to simplify the manufacturing process, thereby reducing costs.

위의 [예시 3]과 관련해서 평가자 E1은 “이유(의)의 부사절과 효과의 주절로 기재하여야 함. Be simple to simplify ~ → since ~ be simple, it is possible to simplify ~”으로 오류를 지적하여, 문장의 전체 구조에 문제가 있는 것은 아니며 독자가 이해하는 데에도 큰 문제는 없으나 다만 원문의 이유와 효과의 차이가 번역에서 좀 더 분명하게 드러나게 할 필요가 있다고 언급하였다. ‘불명확한 번역 오류’는 PT와 WIPO Translate에서 모두 이와 같이 문장 자체에 문제가 있는 것은 아니나 구조 개선 시 원문을 좀 더 명확하게 반영할 수 있는 번역 문제를 가리킨다. 즉, 다른 통사 오류들에 비해 정확성 및 가독성에 있어서 상대적으로 미치는 영향이 작은 오류라 할 수 있다.

이에 반해 PT에서 두드러지는 통사 오류의 분류 중 ‘(발명의) 구성요소 오류’는 해당 발명을 구성하는 요소들을 제대로 반영하지 못하여 잘못된 번역한 경우로 발명을 이해하는 데 있어 치명적인 수준의 오류라 할 수 있다. ‘구성요소 오류’에 해

당하는 예는 아래 [예시 4]와 같다.

[예시 4] Patent Translate의 불명확한 번역 오류

ST: 본 발명에 따른 검사장치는, 판상 콘택트가 슬릿에 끼워져 지지되어 있는 콘택트 지지부, 상기 검사회로 상에 배치되며, 상기 콘택트 지지부의 하부를 수용하여 판상 콘택트의 하단이 상기 검사회로의 피검사접점에 접촉하는 하부 소켓 및 상기 하부 소켓 상에 배치되며, 상기 콘택트 지지부의 상부를 수용하여 판상 콘택트의 상단이 상기 피검사체의 검사접점에 접촉하는 상부 소켓을 포함하여 구성될 수 있다.

TT: The inspection apparatus according to the present invention includes a contact support part on which a plate-shaped contact is fitted and supported by a slit, and is disposed on the inspection circuit, and accommodates the lower part of the contact support part so that the lower end of the plate-shaped contact contacts the inspection target point of the inspection circuit. and a lower socket disposed on the lower socket, the upper socket accommodating the upper part of the contact support part so that the upper end of the plate-shaped contact contacts the test contact point of the subject.

상기 [예시 4]의 ‘검사장치’를 구성하는 요소는 ‘콘택트 지지부’, ‘하부 소켓’, ‘상부 소켓’의 세 가지이다. 그런데 PT에서는 ‘콘택트 지지부(a contact support part)’만 구성요소로 제대로 표현되었을 뿐 나머지 ‘하부 소켓’과 ‘상부 소켓’은 문장의 구조가 꼬이면서 발명의 구성요소로 표현되지 못하였다. 이에 대해 평가자 E1은 “검사장치는 콘택트 지지부, 하부소켓, 및 상부소켓을 구성요소로 구비한다는 문장구조가 전혀 반영되지 않음”으로 오류를 기술하여 문장구조의 문제로 발명이 구비하는 구성요소가 제대로 표현되지 않았음을 지적하였다. 이와 같은 오류는 앞서 [예시 3]에서 설명한 ‘불명확한 번역 오류’에 비해 그 심각성이 좀 더 클 수 있음을 알 수 있다.

완전한 문장이 만들어지지 않고 불완전한 문장이 생산된 ‘불완전한 문장 오류’ 또한 번역 품질에 직접적인 영향을 미치는 오류일 것이다. 아래의 [예시 5]는 PT에 의한 ‘불완전한 문장 오류’의 사례이다.

[예시 5] Patent Translate의 불완전한 문장 오류

ST: 상기 박막층의 결함 제거 방법은, 기판 상에 시드층을 형성하고, 상기 시드층 상에 무기 박막층을 형성하고, 상기 무기 박막층을 형성하는 과정에서 발생한 결함을 제거하는 것을 포함하되, 상기 결함을 제거하는 것은, 자기조립 단분

www.kci.go.kr

자막(Self-Assembled Monolayers)을 형성하여 제거한다.

TT: The method for removing defects in the thin film layer includes forming a seed layer on a substrate, forming an inorganic thin film layer on the seed layer, and removing defects generated in the process of forming the inorganic thin film layer, wherein the defect is removed It is removed by forming Self-Assembled Monolayers.

상기 [예시 5] 역시 PT의 사례로 TT의 밑줄 그은 부분을 보면 wherein 이하에는 주어와 동사가 각각 하나씩 나와 완전한 절이 구성되어야 한다. 이에 대해 PT는 wherein 다음에 주어로 ‘the defect’를, 동사로 ‘is removed’를 사용한 뒤 바로 이어서 다시 주어로 ‘It’을, 그리고 동사로 ‘is removed’를 반복한다. 이로 인해 wherein 이하에는 주어와 동사가 각각 두 개가 되면서 통사적으로 완전하지 않은 구문이 형성되었다. 이와 같은 불완전 문장에 대해 평가자 E4는 “‘the defect is removed It is removed’는 비문법적인 표현임. 원문에 ‘제거하는’, ‘제거된다’가 반복됨으로 인해 발생한 오류로 판단됨.”으로 지적한 바 있다. 이와 같은 오류 또한 ‘구성요소 오류’ 만큼은 아니더라도 문장이 완전하지 않아 독자의 이해를 저해한다는 점에서 가독성에 치명적인 영향을 미치는 것으로 판단되고 따라서 앞서 제시한 ‘불명확한 번역 오류’에 비해 그 심각성이 높은 것으로 여겨진다.

이 외에 일반적으로 특히 번역에서는 특별한 이유 없이 원문의 문장을 분절하는 것에 대해서 인간 번역사들은 소극적인 데 반해 PT는 WT에 비해 원문의 문장을 좀 더 빈번하게 나누는 편이며 이로 인해 발생하는 오류 또한 빈번한 것으로 드러났다.

## 5. 결론

본고는 한국어 특허 공보의 요약 30건(총 53문장)에 대해 무료 특허 전문 기계 번역 엔진인 ‘Patent Translate’와 ‘WIPO Translate’로 영문 번역을 추출한 후 이에 대한 자동평가와 수동평가를 실시하여 2종의 전문 번역 엔진 간 품질의 차이를 살펴보았다.

BLEU와 METEOR 지수를 살펴보면 ‘Patent Translate’가 ‘WIPO Translate’에 비해 조금 높은 점수를 기록하였다. 수동평가에서는 자동평가와 달리 ‘WIPO Translate’가 ‘Patent Translate’를 근소한 차로 앞섰다. 즉, 특허번역 전문가들은

www.kci.go.kr

WIPO Translate의 품질이 나은 것으로 평가한 것이다. SMT를 대상으로 한 최효은과 이지은(2017)의 결과와 비교해 볼 때 NMT의 품질이 개선되었다고 할 수 있지만 여전히 요지번역 이상의 기능을 기대하기는 어려운 품질 수준이라 할 수 있다.

오류 분석 결과에 의하면 전체적으로 가장 빈도 높게 지적된 오류는 통사였으며, 그 뒤를 누락과 용어 문제가 따랐고, 추가와 기타는 큰 비중을 차지하지 않았다. 두 엔진 간 빈도 차이가 가장 심한 오류 유형이 누락과 통사였다. 누락은 ‘Patent Translate’에 비해 ‘WIPO Translate’에서 압도적으로 많았고, 통사 오류는 ‘Patent Translate’에서 훨씬 많았다. 본 연구를 통해 드러난 엔진 간의 차이를 요약하자면, ‘Patent Translator’가 누락 없이 원문을 좀 더 충실히 옮기는 반면 ‘WIPO Translate’는 복잡하고 긴 문장을 좀 더 정확하게 이해하여 이에 맞는 문장 구조로 번역하는 특징이 있다고 할 수 있다.

실제 누락 오류들을 살펴본 결과 누락 양상의 차이가 있었다. ‘Patent Translate’의 경우 누락은 5단어 이하의 사소한 누락이 절대 다수를 차지하는 반면 ‘WIPO Translate’의 누락은 5단어 이상의 누락 비중이 적은 데 비해 무려 50단어 이상까지의 누락도 나타나는 것으로 드러났다. 이러한 누락의 정도 차이가 자동평가에서 ‘Patent Translate’가 더 높은 점수를 받을 수 있는 주요한 요인이 되었을 것이라는 유추가 가능하다.

통사 오류로는 주로 구성요소 오류, 수식어구 오류, 불명확한 번역 오류, 불완전한 문장 오류, 문장 나눔의 오류가 대부분이었다. 통사 오류 수가 적은 ‘WIPO Translate’의 경우, 이 중 가장 품질에 미치는 영향이 미미할 것으로 여겨지는 불명확한 번역 오류의 비중이 가장 높은 반면, 통사 오류가 빈번한 ‘Patent Translate’의 경우 문장의 정확성과 가독성에 절대적으로 영향을 미치는 구성요소 오류, 불완전한 문장 오류 등의 비중이 월등히 높았다. 또한 ‘Patent Translate’에서는 ‘WIPO Translate’에 비해 문장 나눔으로 인한 통사 오류 또한 더 높은 비중으로 발생하는 것을 확인하였다. 이러한 통사 오류의 양적 및 질적 차이는 결국 수동평가 결과에 영향을 미쳐 ‘WIPO Translate’가 조금 더 높은 점수를 받을 수 있었던 것으로 판단된다.

통사 오류 문제와 평가 결과를 연계해서 보면, 수동평가와 달리 자동평가에서는 ‘Patent Translate’가 상대적으로 높은 점수를 받았는데, 그 이유가 자동평가의 경우 참조번역과의 유사도만을 판단하기 때문에 ‘Patent Translate’ 결과물이 원문의 내용을 충실하게 반영하지 못하고 번역 결과물이 불완전한 문장이거나 통사적

문제가 있는 문장일지라도 높은 점수를 받을 수 있는 가능성을 시사한다. 이러한 자동평가와 수동평가 결과의 괴리에 대한 지적은 선행연구(한현희 2020 등)에서 확인된 바 있으며, 이 때문에 기계번역 결과물에 대한 정확한 평가를 위해서는 수동평가가 필요함을 재확인할 수 있었다. 본 연구에서 수동평가 결과는 자동평가와 반대로 ‘WIPO Translate’의 점수가 전반적으로 높은 편이었는데, 이는 특히 번역 전문가인 평가자들의 시각에서는 원문과 비교하여 번역문을 보았을 때 번역문이 원문의 구조와 발명의 구성요소, 그리고 이 구성요소들을 수식하는 내용이 정확하게 번역된 경우에 제대로 번역이 되었다고 판단하는 경향이 있음을 유추해 볼 수 있다. 평가자들의 시각에서 앞서 언급한 바와 같이 ‘불명확한 번역 오류’ 정도는 평가에 크게 영향을 미치지 않을 수 있는 통사 오류이나 ‘구성요소 오류’, ‘불완전한 문장 오류’ 등은 아마도 결과물의 평가에 영향을 미칠 수 있는 통사 오류일 것이다. 따라서 이러한 주요 통사 오류의 비중이 높은 ‘Patent Translate’의 평가 점수가 상대적으로 주요 통사 오류의 비중이 낮은 ‘WIPO Translate’에 비해 낮았으리라 추측해 볼 수 있다. 자동평가에서는 기계번역 결과물을 평가할 때 통사보다는 누락 문제에 좀 더 가중치를 두어 평가를 할 가능성이 높고, 수동평가에서는 누락 문제보다는 통사 문제에 좀 더 가중치를 두어 문장을 평가할 가능성이 높음을 시사한다.

본 연구는 적은 데이터를 기반으로 인공지능망 한영 특히 기계번역 평가를 진행한 한계가 있는 터라 연구 결과를 일반화하기는 어렵다. 다만 특정 전문 분야에 특화된 기계번역 엔진이 동일한 NMT 방식이라도 훈련을 위해 사용된 데이터의 차이 등으로 인해 번역 품질의 차이가 있을 수 있음을 확인하였으며, 그간 NMT 번역 품질이 주로 범용 기계번역 엔진을 대상으로 한 것에 비해 다국어 특히 기계번역 엔진의 성능을 살펴볼 수 있었다는 점에서는 의의가 있다고 할 수 있겠다. 한편 평가 결과는 평가자별로 누락이나 통사 오류에 대한 민감성 내지 번역 오류 수용 가능성이 다를 수 있음을 시사하였다. 평가자 차이나 인식 차이는 본 연구의 범위는 아니지만 번역 평가자들의 평가 성향이나 품질 인식의 차이에 관한 연구가 기계번역의 수동평가에서도 흥미로운 연구 주제가 될 수 있다고 판단하여 후속연구로 제안한다.

## 참고문헌

- 강병규·이지은. (2018). 「신경망 기계번역의 작동 원리와 번역의 정확률」. 『중어중문학』 73: 253-295.
- 곽중철·한승희. (2018). 「포스트에디팅 측정지표를 통한 기계번역 오류 유형화 연구」. 『통번역학 연구』 22(1): 1-25.
- 김동미. (2018). 「의약품 설명서의 인공지능 번역에 대한 전략연구」. 『인문사회21』 9(6): 533-544.
- 김보영·김연주·서승희·송신애·이진현·전경아·최지수·홍승빈·정혜연·허탁성. (2020). 「번역자동 평가에서 풀리지 않은 과제」. 『번역학연구』 21(1): 9-29.
- 박건영. (2021). 「정보성 텍스트의 한영 기계번역 포스트에디팅 가이드라인 제시 - 신경망 기계 번역(NMT)을 사용한 뉴스 기사문 번역의 사례」. 『번역학연구』 22(1): 109-137.
- 박경리·조수연·전종섭. (2013). 「웹기반 영한 번역을 위한 중간 언어의 효율성 연구」. 『통번역학 연구』 17(3): 201-229.
- 서보현·김순영. (2018). 「기계번역 결과물의 오류유형 고찰」. 『번역학연구』 19(1): 99-117.
- 이상빈. (2020). 「기계번역에 관한 KCI 연구논문 리뷰: 인문학 저널 논문(2011-2020년 초)의 논 의내용과 연구방법을 중심으로」. 『통역과 번역』 22(2): 75-104.
- 이준호. (2019). 「신경망기계번역의 객관적 평가를 위한 예비연구: 자동평가와 수동평가의 균형 점」. 『통번역학연구』 23(3): 171-202
- 정혜연. (2018). 「번역의 자동평가: 기계번역 평가를 인간번역 평가에 적용해보기」. 『통번역학연구』 22(4): 265-287.
- 최효은. (2016). 한영 특허 번역 품질 평가 연구. 이화여자대학교 통역번역대학원 통역번역학과 박사학위논문.
- 최효은·이지은. (2017). 「특허 기계번역 결과물의 평가 - KIPRIS 의 무료 한영기계번역을 중심으로」. 『통역과 번역』 19(1): 139-178.
- 한승희. (2021). 「기계번역 품질에 대한 요인분석」. 『통번역학연구』. 25(2): 147-170.
- 한현희. (2020). 「한-노 기계 번역, 어디까지 왔나?: Google과 Papago 번역 성능 비교를 기반으로」. 『노어노문학』 32(3): 63-93.
- Bazrafshan, M. (2014). *Semantic Features for Statistical Machine Translation. Unpublished Doctoral Thesis*, University of Rochester, New York, US.
- Brkić, M., Seljan, S., and Vičić, T. (2013). Automatic and human evaluation on English-Croatian legislative test set. *Computer Science* 7816: 311-378.
- Castilho, S., Moorkens, J., Gaspari F., Calixto, I., Tinsley, J., and Waya, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108: 109-120.
- Castilho, S., Doherty, S., Gaspari, F., and Moorkens, J. (2018). Approaches to human and machine

- translation quality assessment. In Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.), *Translation Quality Assessment: From Principles to Practice*. New York: Springer, 9-38.
- Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering* 26: 137-161.
- Costa A., Ling W., Luís T., Correia R., and Coheur L. (2015). A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation* 29: 127-161
- Coughlin, D. (2003, September). Correlating automated and human assessments of machine translation quality. Paper presented at MT Summit IX. New Orleans, US.
- Doherty, S. (2017). Issues in human and automatic translation quality assessment. In Kenny, D. (ed.), *Human Issues in Translation Technology*. London & New York: Routledge, 131-148.
- Doherty, S. (2019). Translation technology evaluation research. In O'Hagan, M. (ed.), *The Routledge Handbook of translation and technology*. London: Routledge, 339-353.
- Dorr, B., Snober, M., and Madnami, N. (2011). Machine translation evaluation and optimization. In Olive, J., Christianson, C., and McCary, J. (eds.), *Handbook of Natural Language Processing and Machine Translation*. New York: Springer, 745-843.
- González, M. and Giménez, J. (2014). An open toolkit for automatic machine translation (meta-)evaluation. *Technical Manual*, version 3.0. TALP Research Center, LSI Department, Universitat Politècnica de Catalunya.
- Giménez, J and Márquez, L. (2010). Linguistic measures for automatic machine translation evaluation. *Machine Translation* 24: 209-240
- Klubička, F., Toral A., and Sánchez-Cartagena, V. M.(2018). Quantitative fine-grained human evaluation of machine translation systems: A case study on English to Croatian. *Machine Translation* 32: 195-215
- Kuczmariski, J. (2018). A new look for Google Translate on the web. Retrieved from <https://www.blog.google/products/translate/new-look-google-translate-web/> on 20 May 2022.
- Lavie, A. (2013). Automated metrics for mt evaluation. *Machine Translation*, 11: 731.
- Lavie, A. and Argawal, A. (2007 June). Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. Paper presented at the Third Workshop on Statistical Machine Translation. Ohio, USA.
- Läubli, S., Sennrich, R. and Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. Paper presented at the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium.
- Lommel, A. (2018). Metrics for translation quality assessment: A case for standardising error typologies. In Castilho, S., Doherty, S., Gaspari, F., and Moorkens, J. (eds.), *Translation Quality Assessment: From Principles to Practice*. New York: Springer, 109-128.
- Nurminen, M. (2020). Raw machine translation use by patent professionals: A case of distributed

- cognition. *Translation, Cognition & Behavior* 3(1): 100-121.
- Olohan, M. (2015). *Scientific and Technical Translation*. London: Routledge.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. Paper presented at the 40th Annual Meeting of the Association for Computational Linguistics. Pennsylvania, USA.
- Poliquen, B. (2015). Full-text patent translation at WIPO: Scalability, quality and usability. Paper presented at MT Summit XV: The Sixth Workshop on Patent and Scientific Literature Translation (PSLT6). Miami and Florida, US.
- Poliquen, B. (2017). WIPO Translate: Patent neural machine translation publicly available in 10 languages. Paper presented at the Seventh Workshop on Patent and Scientific Literature Translation. Nagoya, Japan.
- Popovic, M. (2018). Error classification and analysis for machine translation quality assessment. In Castilho, S., Doherty, S., Gaspari, F., and Moorkens, J. (eds.), *Translation Quality Assessment: From Principles to Practice*. New York: Springer, 129-158.
- Premoli, V., Murgolo, E., and Cresceri, D. (2019 August). MTPE in patents: A successful business story. Paper presented at the MT Summit XVII. Dublin, Ireland.
- Rossi, L. and Wiggins, D. (2013). Applicability and application of machine translation quality metrics in the patent field. *World Patent Information* 35: 115-125.
- Tsai, Y. (2017). Linguistic evaluation of translation errors in Chinese-English machine translations of patent titles. *Forum* 15(1): 142-156.
- USPTO. (2018). Manual of patent examining procedure. USPTO United States Patent and Trademark. Retrieved from <https://mpep.uspto.gov/RDMS/MPEP/current#/current/d0e122292.html> on 20 May 2022.
- Wu, Y., Schuster, M., Chen, Z., Quoc, V. Le, Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv: 1609.08144. Retrieved from <https://arxiv.org/pdf/1609.08144.pdf>. on 20 May 2022.
- Ying, C., Shuyu, Y., Jing, L., Lin, D., and Qi, Q. (2021). Errors of machine translation of terminology in the patent text from English into Chinese. *ASP Transactions on Computers* 1(1): 12-17.

This paper was received on 31 October 2022; revised on 16 December 2022; and accepted on 20 December 2022.

---

***Authors' email addresses***

jieun.lee@ewha.ac.kr

cutedinojr@naver.com

***About the authors***

Jieun Lee (lead author) is a Professor at GSTI of Ewha Womans University. Her research interests include legal interpreting and translation, community interpreting, machine translation, interpreter and translator education.

Hyoeyun Choi (co-author) is an Adjunct Professor at GSTI of Ewha Womans University. Her research interests include legal translation, patent translation, machine translation, interpreter and translator education.

www.kci.go.kr