

법률 특화 번역엔진 성능 평가*

– 한영 계약서 번역을 중심으로 –

이준호**

Junho Lee (2021), A Study on Performance Evaluation of an Specialized Machine Translation Engine in the Legal Domain: Focusing on the Korean-to-English translation of legal contracts. *This study aims to examine the possibility to enhance the performance of a machine translation engine in Korean-to-English legal translation. With the rise of neural machine translation, translation scholars quickly responded to investigate the output and limitations of machine translation. However, there has been little research to date on enhancing the quality of raw machine translation itself. As the legal field generates vast demand for machine translation post-editing, any improvement in machine translation engine performance can be of significant contribution. Hence, this paper will introduce an experimental machine translation engine trained with a bilingual legal corpus and verify whether it is possible to create a domain-specific engine that can outperform a freely available general purpose translation service. To that end, the output from the experimental engine and Google Translate will be compared and analyzed in terms of accuracy, fluency, usability, and essential changes. Last but not least, this paper will examine the potential contribution that translators can make and areas of collaboration between translation scholars and translation engine developers. (Hankuk University of Foreign Studies, Korea)*

Keywords: Machine Translation, Legal Translation, Google Translate, MTPE, Domain-specific Translation Engine

주제어: 기계번역, 법률번역, 구글번역, 포스트에디팅, 특화 번역엔진

* 이 논문은 AiLinggo의 기술적 지원을 받아 작성되었으며, 이재욱 대표님과 직원분들께 고마움을 전합니다.

** 한국외국어대학교, 강사

1. 서론

무역과 기업 활동의 글로벌화로 인해 계약, 소송, 중재 등 법률 번역의 수요는 증가하고 있으며, 2019년 기준 국내 법률 번역 시장의 규모는 약 2,000억원 으로 나타났다(법률신문 2019). 또한, 법률문서는 법률적 활동이 이뤄지는 시간의 흐름에 맞춰 적시에 번역되어야 한다는 특성을 가지기에, 정형화된 법률문서는 기계번역을 활용하는 업체들이 국내외에서 모두 등장하였다.

하지만 법률문서는 결코 번역하기 쉬운 텍스트가 아니다. 문체가 복잡하고 삽입절이 다양하고, 서술보다는 명사화가 다수 발생하여 이해하기가 어렵고(Bahatia 1997), 어휘의 부조화, 통사의 복잡성, 텍스트의 특수성, 법률체계와 문화의 차이 등이(Harvey 2002) 존재하기 때문이다. 여기에 더해 법률 번역의 오류가 번역 결과물 사용자들에게 미치는 사회적 여파를 고려한다면, 기계번역 결과물을 그대로 사용하는 것은 합리적이지 않다. 따라서 빠르고 정확한 번역 결과물 생성이라는 요건을 충족하기 위해서 1>양질의 번역을 출력할 수 있는 번역엔진이 필요하고, 2>출력된 결과물을 수정할 수 있는 법률 혹은 번역 전문가의 포스트에디팅이 필요하다.

하지만 양질의 기계번역 결과물 생성은 쉽지 않다. 특히 일반 대중에게 공개된 무료 온라인 기계번역(free on-line machine translation)은 다수의 한계를 노출했으며, 전문 분야에서 사용하기에 적합하지 않은 경우가 많았다. 한국어-영어 번역 성공률이 70~80%에 불과함을 지적한 연구가 존재하며(박찬준 외 2021: 25), 기계번역에 관한 KCI 연구를 통시적으로 분석한 이상빈(2020)에 따르면 다수의 연구가 무료 온라인 기계번역을 대상으로 연구를 진행했으며, 기계번역의 오류가 가장 많이 연구된 영역이라는 점에 주목할 필요가 있다.

물론 무료 온라인 기계번역이 다양한 주제의 텍스트를 번역할 수 있도록 개발된 일반 목적(general purpose) 엔진이라는 점을 고려하면 상기 언급한 한계가 존재하는 것이 자연스러울 수 있다. 즉, 일반적 말뭉치를 활용하여 번역엔진 학습이 이뤄졌기 때문에, 특정 주제 영역의 단어나 글쓰기 스타일 등을 번역 결과물에 반영하기 어려울 수밖에 없다(Wang et al. 2019). 이러한 일반 목적 번역엔진의 단점을 극복하기 위해 해외에서는 90년대 초반부터 번역엔진의 학습 영역을 일반 주제 영역에서 제한된 주제 영역으로 특화하는 다양한 시도가 이뤄졌다(최승권 외 2007: 95).

특정 주제에 대해 학습이 이뤄진 번역엔진은 일반 목적 번역엔진 대비 해당 주제 영역에서 높은 품질의 번역을 제공할 확률이 높으므로, 학술적으로 연구의 가치가 있으며 실무적으로도 많은 장점을 지닌다. 이처럼 더 높은 품질의 기계번역 결과물을 제공하기 위해서는 1>번역엔진의 작동 메커니즘을 조정하거나 2>양질의 병렬 말뭉치를 대량으로 확보해야 하는데, 이 두 가지 요건 중 하나라도 충족하기가 쉽지는 않다. 더욱이 다양한 데이터의 확보가 상대적으로 쉽고 다수의 개발 인력을 지닌 거대 기업들이 일반 번역엔진을 제공하는 상황에서, 소수의 개발자가 개발한 엔진에 양질의 병렬 말뭉치를 적용한다고 해도 일반 번역엔진과 경쟁 가능한 수준의 번역 결과물 생성이 가능할지는 확신이 어렵다.

이에 본고는 법률 번역에 특화된 병렬 말뭉치를 확보하고, 이를 사용하여 학습이 이뤄진 번역엔진이 대기업에서 제공하는 일반 목적 번역엔진 대비 더 우수한 결과물을 생성할 수 있을지 탐구하고자 한다. 이를 위해 정확성, 유창성, 사용 적합성, 포스트에디팅 필수 수정 등의 영역에서 구글번역보다 우수한 결과물을 제공할 수 있는지를 정량적으로 평가하고자 한다. 본 연구를 통해 법률 특화 번역엔진이 더 우수한 결과물을 제공할 수 있다면, 법률 외의 영역에서도 주제 영역에 특화된 번역엔진 연구를 가속하는 하나의 근거가 될 수 있을 것이다. 반면에 특화된 엔진이 더 우수한 기계번역 결과물 제공에 실패한다고 해도, 어떠한 점에서 부족함이 있는지를 파악하고 해당 엔진을 개선하는 계기가 될 수 있을 것이다. 이를 통해 번역학 연구자들과 실무자들이 주제 특화 번역엔진에 대한 이해 및 활용도를 높이는데 작으나마 이바지할 수 있을 것이며, 번역엔진의 개발 및 개선이라는 새롭고 협업적인 연구 주제 발굴에 이바지할 수 있을 것이다.

2. 배경 및 선행연구

2.1. 기계번역의 발전과 번역학의 역할

1950년대 MIT에서 기계번역 연구가 본격화되기 시작한 이후 새로운 기계번역 방식의 개발과 발전은 온전히 공학적 연구의 성과라고 해도 과언이 아니다. 기계번역은 결과물을 구현하는 방식에 따라 규칙기반에서 통계기반 기계번역으로 꾸준히 발전해 왔으며, 최근 널리 사용되고 있는 신경망번역(NMT)에 관한 연구는

2010년대에 들어서 본격적으로 이뤄지기 시작했다. 신경망번역의 구현 방법은 RNN(recurrent neural network)에서 트랜스포머(Transformer) 등으로 지속해서 진화해왔으며 각 방법에 따라 장단점이 있지만, 일부 논문에서는 트랜스포머 기반의 모델이 가장 좋은 성능을 보였다고 보고한 바 있다(Vaswani et al. 2017).

기계번역 활용의 측면에서 보면 이러한 신경망번역의 등장은 희소식이라 할 수 있다. 신경망번역은 기존의 번역 대비 포스트에디팅에 필요한 노력을 낮춰 주며 (Bentivogli et al. 2016), 구글의 신경망번역 서비스는 기존 기계번역 방식대비 더 정확성이 높다는 보고가 있기 때문이다(Wu et al. 2016). 이처럼 신경망번역의 긍정적인 측면과 함께 신경망번역에 관한 관심이 증가했고, 국내에서도 2016년을 기점으로 번역학계에서 기계번역 연구가 본격화되었고 다양한 주제를 다룬 연구가 이미 존재한다(신지선 2020: 90).

하지만 번역학계가 집중한 부분은 기계번역의 한계였다. 인간번역과 기계번역의 비교 그리고 기계번역의 오류 등 기계번역의 단점(서보현과 김순영 2018; 이준호 2019a) 혹은 이를 극복하기 위한 포스트에디팅에 대한 연구(윤미선 외 2018; 이상빈 2018 등)가 번역학계에서 다수 이뤄졌다. 하지만 이러한 번역학계의 연구 흐름이 기계번역 활용에 있어 본질적 문제를 놓친 것은 아닌가 고민이 필요해 보인다.

낮은 품질의 기계번역 결과물에 대해 포스트에디팅을 수행하는 것은 큰 노력이 필요하며 힘든 일이기 때문이다(Koponen 2012: 1). 결국, 번역엔진의 품질 개선은 기계번역 활용에 있어 가장 핵심적인 문제이지만, 주어진 기계번역 결과물의 품질을 분석하고 “좋지 않은” 품질을 효과적으로 활용하는 방안에만 집중하는 것은 아닌가 아쉬움이 남는다.

실제로 기계번역과 관련된 메타 연구를 진행한 이상빈(2020), 현재 국내 연구 기계번역 관련 연구 공백을 제시한 김순미 외(2019), 그리고 해외 연구 동향을 조사한 신지선(2020)의 연구에서는 번역엔진의 성능 개선에 대한 언급을 찾을 수 없었다. 다만 49편의 논문을 분석한 최문선(2019)의 메타 연구는 번역학계의 가장 주된 연구 영역은 기계번역의 생산 보다는 소비 영역임을 지적했으며, 극소수의 번역학 연구만 번역엔진 성능 개선을 위한 “조작”과 관련이 있음을 지적하고 있다.

예를 들어 최승권(2007)은 특허번역문의 구문적 특징을 추출하여 기계번역기에 적용한 이후 번역 결과물 평가를 진행했으며, 최승권과 김영길(2010)은 어휘 번역 패턴을 번역 말뭉치로부터 반자동으로 추출하고 이를 자동 번역 시스템에 활용하

여 번역 결과물을 더욱 자연스럽게 만드는 방법을 제시하였다. 하지만 두 연구는 신경망번역 출시 이전에 출간되었으며, 현재 시도되는 기계번역 품질 개선 방식과는 다소 거리가 있는 접근을 취하고 있다. 결국, 신경망번역 성능 개선은 번역학 연구에 직접적 영향을 주는 분야임에도 불구하고, 국내 번역학계에서는 거의 연구되지 않은 주제 영역이라 할 수 있다.

이에 번역학계의 연구가 기계번역의 소비에만 치중하기보다는 생산에 대해서 이바지할 수 있는 바를 찾을 필요성이 있다 하겠다. 그 시작점으로 본고는 한영 계약서 번역이라는 제한된 분야이지만 신경망번역 품질 개선의 실질적 사례를 분석하고, 해당 결과가 번역학 연구자들에게 가지는 함의를 논의하고자 한다.

2.2. 번역엔진 성능 개선과 평가

번역엔진 개선을 위한 연구는 공학 분야의 주도로 오랫동안 이뤄져 왔다. 대표적으로 기계번역 전문 학술지인 머신 트랜스레이션(Machine Translation)은 1986년에 발행을 시작했으며, 2006년부터 열리기 시작한 컴퓨터 언어학 전문가 컨퍼런스 ACL(Association for Computational Linguistics)은 기계번역 및 엔진 개선과 관련한 다양한 주제를 다루고 있다. 이러한 노력은 신경망번역 출시 이후에도 꾸준히 이어졌다. 구체적으로는 자연어 처리 방식의 개선, 양질의 데이터 사용, 혹은 두 가지를 모두 조합하여 번역엔진의 성능을 높이거나 단점을 극복하기 위한 시도로 요약될 수 있다. 예를 들어 신경망번역의 단점으로 알려진 희귀어(rare word)의 처리를 개선하기 위해 얼라인먼트 기반의 새로운 기법을 제시한 연구가 있다(Luong et al. 2014). 혹은 신경망번역의 문제로 알려진 숫자 번역의 취약점을 해결하기 위하여 최적화된 모델에 숫자 기호화 알고리즘을 적용하여 이 문제를 해결하려 시도한 연구도 존재한다(강청웅 외 2018).

또한, 엔진을 학습시키는 영역을 특화한 시도도 있었다. 기계학습을 활용한 다양한 응용사례의 공통적 문제점은 학습된 데이터와 다른 영역의 업무에서는 성능이 우수하지 못하다는 것이며, 신경망번역 역시 예외는 아니기 때문이다(Müller et al. 2019). 달리 말하면 기계번역 결과물의 품질을 개선하기 위해서는 학습 영역을 좁지만 심화하는 전략이 필요하다.

하지만 특정 주제 영역에 대한 양질의 데이터를 확보하고 엔진을 학습시키는 작업은 쉽지 않기 때문에(Zhang et al. 2019), 다수의 연구가 존재하지는 않는다. 하

지만 박찬준 외(2020)는 코로나19에 특화된 인공신경망 기계번역기 제작을 위해 코로나바이러스에 특화된 말뭉치를 사용하여 번역 모델을 학습시켰으며, 상용화 시스템인 구글 번역기와 비교하여 BLEU 스코어 우위를 보고하였다. 또한, 박찬준과 임희석(2020)은 한국정보화진흥원에서 제공하고 있는 공공 한영 병렬 말뭉치를 사용하여 기계번역 결과물 품질 개선을 시도하였다. 이처럼 모델의 변경도 중요하지만, 데이터의 품질이 더욱 중요할 수 있음을 입증한 연구들이 있기에 효과적으로 양질의 병렬 말뭉치를 생성할 필요성이 높다.

여기에 더해 언어 공학적 연구는 평가 과정에서 자동화 평가만을 사용하고 있어 아쉬움이 남는다. 대표적으로 다수의 연구에서 사용한 BLEU 스코어는 빠른 평가가 가능하다는 장점이 있지만, 일종의 답안 번역과 유사도를 비교하기에 품질을 대변하기는 어렵다는 단점을 가지고 있다(이준호 2019b; 정혜연 2018 등). 따라서 자동화된 평가 방식 외의 평가 방식을 적용해야 더욱 높은 신뢰성을 담보할 수 있으며, 기계번역 결과물 평가와 관련하여 번역학이 기여할 수 있는 부분이 있다 하겠다.

이에 본 연구는 기계번역 결과물 품질 개선 가능성을 검증하기 위해 다음과 같은 실험을 시도하고자 한다. 먼저 효율적 번역엔진 학습 수행을 위해, 학습 영역을 법률 영역으로 한정할 것이다. 이후 계약서를 포함, 법률 영역에서 추출한 양질의 병렬 말뭉치를 엔진 학습에 적용할 것이다(이하 실험엔진). 다음 단계로 실험엔진과 구글번역의 결과물에 대해 다양한 수동 평가를 진행할 것이다. 그리고 마지막 단계로 평가 결과에 기초하여, 양질의 병렬 말뭉치가 기계번역 결과물 품질 개선에 기여할 수 있는지를 논의하고, 구글번역 대비 부족한 점을 식별하여 실험엔진의 추가 개선을 도모하며, 이 과정에서 번역학이 번역엔진의 구축 및 평가에 있어 기여할 수 있는 부분을 제언하고자 한다.

3. 엔진 개발 및 연구의 설계

3.1. 엔진 개발 및 학습 과정

법률 문서의 여러 가지 특징 중 기계번역에 적합할 수 있는 특징이 일부 존재한다. 일반 언어에서 잘 사용하지 않는 용어와 정형화된 문장의 빈도수가 높고,

일상생활에서 사용되는 문장에 비해 문장 구조의 자유도가 높지 않으며, 법률 활동에 사용되는 문장은 구어체 또는 속어의 발생빈도가 낮다. 이에 번역엔진에 법률 문장을 중점적으로 학습시키면 상대적으로 적은 수의 말뭉치를 사용해서 기계 번역 결과물을 개선할 수 있을 것이라 가설을 수립할 수 있다. 실제로 이러한 가설에 기반하여 엔진을 개발한 다수의 업체가 존재하며, 본 연구는 국내에 있는 번역엔진 서비스사의 도움을 받아 진행되었다.

최초의 학습은 오픈 소스 기반인 오픈NMT(seq2seq, Bidirectional RNN)를 사용하였다. PoC를 의뢰한 로펌에서 제공해준 법률 번역 결과물에서 추출한 병렬 말뭉치를 전처리하여 약 20,000쌍의 문장으로 초기 학습을 진행했으며, 수차례의 반복 학습 이후 변호사 및 번역가의 정성적 평가를 통해, 신규 개발 모델이 무료 온라인 번역기 대비 법률 번역에 있어서 우위에 있음을 내부적으로 확인하였다.

하지만 이후 개발 과정에서 트랜스포머 모델이 기존 모델 대비 빠른 학습 속도를 보였고, 내부 정성 평가 및 BLEU 정량 평가에서 기존 모델 대비 우수한 성과를 나타냈다. 이에 대외적으로 검증된 트랜스포머 모델을 제공하는 업체를 찾아, 시스트란(Systran)의 플랫폼을 사용하여 추가 개발을 진행해 왔다. 안정적인 개발 플랫폼을 확보한 이후에는, 학습용 데이터의 질과 전처리 작업에 큰 중점을 두었다. 영한 데이터는 자체 제작한 전처리 모듈을 통해 기계적으로 데이터를 검증하였고, 변호사 및 로스쿨 학생들을 통해 문법적 오류, 오타, 띄어쓰기를 수동으로 검증하는 과정을 진행해 왔다.

이번 실험에서는 계약서 등 엔진 개발사 자체 보유데이터 30만, AI 허브에서 제공한 20만 등 약 50여만 쌍의 병렬 말뭉치를 사용했으며, 민감 정보(이름, 금액, 날짜 등)를 제거하고 문서 소유자에게 데이터 사용 및 연구에 대한 동의를 득하였다. 내부 테스트 결과 BLEU 스코어가 온라인 무료 번역기 대비 25점이 높은 79점을 기록했으며, 추가 수동 평가 연구를 통해 실험엔진의 미래 활용 가능성을 탐색하고자 한다.

3.2. 연구의 설계

실험에 사용될 텍스트 선정에 있어 법률 문서 중 가장 번역 의의가 많은 계약서를 텍스트를 선정하였다. 이후 연구 신뢰도 확보를 위해 번역엔진 학습에 계약서 텍스트가 사용되었음을 확인했으며, 분석 대상 텍스트가 엔진 학습에 사용되

지 않았음을 보증하기 위해 영문 번역본이 존재하지 않는 계약서 텍스트만을 분석 대상으로 선정하였다. 이후 대표성을 지니는 계약서 텍스트 선별을 위해 대한상사중재원에서 제공하는 일반 계약서 샘플 52건을 검토하여 연구 텍스트를 구성하였다.

본 연구에서는 “기계번역의 품질”이란 전형적이고 평범한 문장에서 인간번역 수준의 결과물을 생산할 수 있으며, 난제가 있는 문장에서도 적은 오류가 발생해야 함으로 정의하였다. 따라서 전형적인 계약서 표준 문구 30문장(총 558단어)을 연구자가 선별하였다(이하 “일반문장”). 다음으로는 기계번역이 처리하기 어려운 속성을 가지고 있는 30문장을(총 892단어) 선별했으며, 선별 과정에서 고유명사 등 저빈도 어휘, 복잡한 주술 관계를 포함한 장문 위주로 선정하였다(이하 “난제문장”).

최종적으로 가장 널리 사용 중인 무료 온라인 번역 서비스 중 하나인 구글 번역을 실험엔진과 비교 대상으로 선정하였다. 실험 데이터 구성에 있어 출발언어 문장별로 기계번역 결과물을 무작위로 배치하여 평가자의 주관 개입을 최소화하도록 노력했으며, 각 평가 영역당 1명의 전문가를 할당하였다.

4. 결과 분석

이번 장에서는 정확성, 유창성, 법률적 사용 적합성, 그리고 포스트에디팅 필수적 수정이란 네 가지 영역에서 진행한 평가를 정량적으로 요약하고, 평가 결과가 가지는 의미를 4.5에서 논의한다.

4.1. 정확성 분석

정확성 평가를 위해 출발언어 텍스트의 의미를 도착언어 텍스트에서 오류 없이 전달하고 있음을 정확성으로 정의했다. 국내 법무법인의 상근 번역사로 5년 이상 근무한 인원에게 하기 기준에 따라 평가를 의뢰했으며, 정확성만 평가하고 유창성 등 여타 요소는 평가에서 배제할 것을 요청했다.

10점: 출발언어 텍스트의 의미가 완전하게 전달됨

- 1점 차감: 명사/대명사 등에서 사소한 오류가 1~2개 있다.
- 1점 차감: 동사나 형용사 등 서술어에서 사소한 오류가 1개
- 2점 차감: 명사/대명사 등에서 중대한 오류가 1~2개 있다.
- 2점 차감: 동사나 형용사 서술어에서 중대한 오류가 1개

정확성 평가 결과를 분석하기 위해 먼저 전체 문장에 대한 점수 평균을 비교했으며, 다음으로 문장 단위별로 어떤 엔진에서 출력된 번역이 더 높은 점수를 득했는가(이하 상대 순위) 빈도를 비교하였다.

표 1. 구글번역과 실험엔진 정확성 비교 (일반문장)

엔진	평균	문장별 상대 순위
구글번역	8.9	6
실험엔진	9.0	9

일반문장에 대한 정확성 평균은 실험엔진이 구글번역 대비 미세하나마 높지만, 이를 근거로 실험엔진과 구글번역의 성능을 비교하기는 어렵다. 하지만 문장별 상대 순위에서는 구글번역이 6회, 실험엔진 9회의 우위를 보였으며, 15회의 동률이 관찰되었다. 실제로 아래 예시에서 볼 수 있듯 두 엔진 모두 높은 정확성을 보인 사례가 다수 관찰되었다.

[ST]

갑과 을은 본 계약으로 지득한 상대방의 업무상 및 기술상의 기밀을 상대방의 승락이 없는 한 제3자에게 누설하여서는 아니된다.

[구글번역]

A and B shall not disclose the business and technical secrets of the other party acquired through this contract to a third party without the consent of the other party. (평가자 정확성: 10점)

[실험엔진]

Party A and Party B shall not disclose to a third party any confidential information of the other party's business and technology which they have learned in this Agreement, unless the other party has consented thereto. (평가자 정확성: 10점)

따라서 두 엔진의 정확성 차이는 크지 않으나, 상대 우위 빈도를 고려하면 실험 엔진의 결과물이 미세한 정확성 우위를 보였다 해석할 수 있다.

반면 난제문장은 기계번역 오류가 발생할 가능성이 큰 문장을 선별했기 때문에 평가자가 일반문장 대비 낮은 점수를 줄 것이라 가정하는 것이 합리적이다. 실제 평가 결과에서도 두 엔진 모두 일반문장 대비 다소 낮은 평균을 기록하였다.

표 2. 구글번역과 실험엔진 정확성 비교 (난제문장)

엔진	평균	문장별 상대 우위
구글번역	8.33	3
실험엔진	8.66	12

일반문장의 경우 두 엔진의 평균 차이가 0.1에 불과했지만, 난제문장에서는 두 엔진의 평균 차이가 0.33으로 증가했다. 또한, 실험엔진이 8.66으로 더 높은 평균을 득하였다. 여기에 더해 문장별 상대 우위에서도 실험엔진이 12번의 우위를 보인 반면 구글번역은 3번의 우위만을 보였다. 따라서 난제 문장에 대해서는 실험엔진의 결과물이 정확성 우위를 보였다고 해석할 수 있다.

4.2. 유창성 분석

유창성 평가를 위해 법률적 교육 배경을 가진 원어민 관점에서 도착언어가 법률 영어의 읽기 및 쓰기 관습에 부합함을 유창성으로 정의하였다. 미국 변호사이며, 통번역학과에 강의 중인 원어민에게 다음의 기준에 따라 평가를 의뢰하였다.

- 10점: 어색함이 없으며, 원어민의 글쓰기와 구분이 어려움
 - 9점: 번역임을 인지할 수 있으나, 어색하지 않음
 - 8점: 번역임을 인지할 수 있으며, 사소한 어색함이 존재
 - 7점: 번역임을 인지할 수 있으며, 사소한 어색함이 3개 이상 혹은 중차대한 어색함이 존재
 - 6점: 2개 이상의 중차대한 어색함이 존재
 - 5점: 다수의 중차대한 어색함이 존재
- (연구자의 번역)

일반문장에 대한 평균을 비교하면 실험엔진이 구글번역 대비 통계적으로 유의미하지는 않지만 미세하나마 높은 점수를 기록하였다. 반면 문장 단위별 상대 순위에서는 동물 11회, 구글번역 우위 8회, 실험엔진 우위 11회로 나타났다.

표 3. 구글번역과 실험엔진 유창성 비교 (일반문장)

엔진	평균	문장별 상대 순위
구글번역	7.36	8
실험엔진	7.46	11

하지만 난제문장에 대해서는 일반문장 대비 큰 폭으로 평균이 하락했으며, 구글번역은 1.7 그리고 실험엔진은 1.66의 하락을 보였다. 물론 이러한 평균 하락에도 불구하고 실험엔진은 난제문장에서도 여전히 구글번역 대비 높은 평균을 보였으며, 문장별 상대 순위에서 실험엔진이 12번의 우위를 보인 반면 구글번역은 10번의 우위를 보였다.

표 4. 구글번역과 실험엔진 유창성 비교 (난제문장)

엔진	평균	문장별 상대 순위
구글번역	5.76	10
실험엔진	5.80	12

따라서 일반문장에서는 실험엔진 결과물이 구글번역 결과물 대비 미세한 유창성 우위가 있다고 볼 수 있다. 하지만 난제 문장에서는 두 엔진 모두 5점대의 매우 낮은 점수 평균을 기록했으며, 아래 예시에서 볼 수 있듯 수용성이 매우 낮은 경우도 다수 있었다.

[ST]

기성부분에 대한 검사가 끝난 경우에는 그 부분에 상당하는 금액을 하도급대금에서 공제한 금액을 기준으로 지체상금을 계산한다. 다만 기성부분의 인수는 그 성질상 분할할 수 있는 제조 등의 위탁에 대한 완성부분으로 인수하는 것에 한한다.

[구글번역]

When the inspection of the ready-made part is completed, the compensation for delay is calculated based on the amount equivalent to the part deducted from the subcontract price. However, the acceptance of the ready-made part is limited to the acceptance of the completed part for the consignment of manufacturing, etc. that can be divided due to its nature. (평가자 수용성: 4점)

[실험엔진]

Where the inspection of the completed part is completed, the penalty for delay shall be calculated based on the amount obtained by deducting the amount equivalent to such part from the subcontract price. However, the acquisition of the completed part is limited to the acquisition of the completed part for consignment of manufacturing, etc., which can be divided due to its nature. (평가자 수용성: 3점)

따라서 난제문장의 수용성 평가 결과에 기반하여 구글번역과 실험엔진의 수용성 우위를 비교하는 것은 큰 의미가 없어 보인다.

4.3. 텍스트 사용 적합성 분석

텍스트 사용 적합성 분석을 위해 사용 적합성이란 용어는 법률 문서에서 의도하는 행위가 그대로 이뤄질 수 있도록 번역이 구성되었는가로 정의했으며, 국내에서 통번역대학원을 졸업하고, 미국에서 법학 박사를 취득한 인원에게 다음의 평가 기준에 따라 평가를 요청하였다.

10점 만점으로 평가를 하며 다음의 조건에서 점수를 차감합니다.

1점 차감: 텍스트가 의도하는 법률적 행위의 이해를 사소하게 저해하는 번역이 있다 (간단한 수식이나 잘못된 명사의 사용 등)

2점 차감: 텍스트가 의도하는 법률적 행위의 이해를 크게 저해하는 번역이 있다 (잘못된 서술어나 주체가 사용 등)

일반문장에 대한 평균을 비교해보면 구글번역이 실험엔진 대비 높은 점수를 보였다. 또한, 문장 상대 우위 빈도에 있어서도 구글번역 11회, 실험엔진은 6회의 우위, 동물 13회로 나타났다.

표 5. 구글번역과 실험엔진 사용 적합성 비교 (일반문장)

엔진	평균	문장별 상대 순위
구글번역	8.6	11
실험엔진	8.13	6

텍스트 사용 적합성 역시 난제문장에 대해서는 일반문장 대비 큰 폭으로 평균이 하락하였다. 구글번역은 2.67 그리고 실험엔진은 1.9점의 평균 감소가 있었다. 하지만 전체 평균에서 실험엔진이 더 높은 점수를 보였으며, 문장별 상대 순위에서도 실험엔진이 13번의 우위를 보인 반면 구글번역은 7번의 우위를 보였다.

표 6. 구글번역과 실험엔진 사용 적합성 비교 (난제문장)

엔진	평균	문장별 상대 순위
구글번역	5.93	7
실험엔진	6.23	13

따라서 일반문장에서는 구글번역이 사용 적합성에 있어 약간의 우위가 있다고 볼 수 있다. 하지만 난제 문장에서는 두 엔진 모두 매우 낮은 점수 평균을 기록하였다는 점을 감안하면 두 엔진의 우위를 논의하는 것이 큰 의미가 없지만, 실험엔진이 상대적 사용 적합성 우위를 보였다고 요약할 수 있다.

4.4. 필수 수정의 정량적 비교

4.1에서 4.3까지 실시한 평가는 연구자가 제공한 평가 기준을 사용하긴 했지만, 평가자가 주관적 판단으로 점수를 부여했을 가능성을 완전히 배제할 수는 없다. 따라서 평가의 신뢰도를 높이기 위해 기계번역 결과물에 필수적으로 수정이 필요한 부분을 정량적으로 평가하였다. 필수적 수정에 대해 드 알메이다(de Almeida 2013: 100)는 다음과 같이 정의하고 있다.

- 1> 문법적으로 어긋남 (표준 문법책에 내용에 명백하게 반하는 내용)
 - 2> 출발언어 텍스트와 비교하여 정확하지 않음 (출발언어 텍스트에 있는 정보가 없거나, 반대로 출발언어 텍스트에 없는 정보가 추가된 경우)
- (연구자의 번역)

상기 기준에 근거하여 연구자는 맹검 상태에서 모든 문장을 분석하였으며 문장 별로 필수적 수정이 필요한 횟수를 정량화하여 번역 결과물 우위를 평가하였다.

표 7. 구글번역과 실험엔진 필수 수정에 따른 우위 비교 (일반문장)

엔진	문장별 상대 우위
구글번역	8
실험엔진	6
판단 어려움 (긍정)*	14
판단 어려움 (부정)**	2

*심각한 오류가 있는 것은 아니며 사소한 오류가 있지만, 우위 판단이 어려움

**두 번역 모두 오류가 많으며 우위 판단이 어려움

일반문장에서는 8:6으로 구글번역 결과물이 상대적으로 필수적 수정이 적게 필요한 것으로 분석되었다. 특히 법률 텍스트 외에도 일상적으로 사용될 법한 문장에서는 구글번역이 더 적은 오류를 보인 경우가 있었으며, 구글번역이 더욱 방대한 데이터 세트와 상세한 언어처리 방식을 적용했기 때문으로 추정된다.

구체적으로는 숫자의 처리는 신경망번역의 단점으로 알려져 있는데, 실험엔진의 경우 단순한 숫자 처리에는 오류가 없었으나, 아래 예시에서 볼 수 있듯 상대적으로 복잡한 숫자 처리에서는 오류를 범한 사례가 있었다.

[ST]

하도급대금의 **1000분의 1.5**를 곱한 금액을 원칙으로 한다.

[구글번역]

multiplied by **1.5/1000** of the subcontract price.

[실험엔진]

1.5% of the subcontract price.

다음으로는 단어 등가 차원의 명백한 오역이 실험엔진에서 발생한 경우가 더 많았다. 일례로 국문 계약서에 “갑”과 “을”이라는 단어가 사용되는 것이 일반적인 예도 불구하고 실험엔진은 “을”을 “lessee(임차인)”로 오역하였다. 또한 “갑”이라는 단어도 “lessor(임대인)”와 “lessee(임차인)”로 일관되지 않게 오역을 하기도 하였다.

[ST]

“을”은 생산한 제품에 대하여 “갑”이나 “갑”이 지정하는 자 또는 기관의 검사를
필하여야 하며

[구글번역]

B shall complete the inspection of the manufactured product by "A" or a person or
organization designated by “A”

[실험엔진]

**Lessee shall inspect the products produced by the person or institution designated by
the Lessor or the Lessee**

마지막으로 출발언어 텍스트의 내용과 무관하게 명백한 문법적 및 통사적 오류
가 실험엔진에서 더 많이 발견되었다. 일례로 통사적으로 명백한 오류가 있는 동
어 반복이 구글번역에서는 발견되지 않았으나, 실험엔진에서는 아래 예시에서 볼
수 있듯 동어 반복이 발견되었다.

[ST]

갑과 을은 상호 존중 및 신의와 성실의 원칙에 따라 계약을 이행하여야 한다.

[구글번역]

Party A and Party B shall perform the contract in accordance with the principle of
mutual respect, good faith, and sincerity.

[실험엔진]

A and B shall perform the contract in accordance with the principles of mutual respect
and **good faith and good faith.**

이상의 필수적 수정 분석 결과, 일반문장에 대해서는 구글번역이 번역의 정확
도와는 별개로 문장을 구성하는 능력에 있어 더 완성도가 높은 것으로 보인다. 다
만 번역의 핵심적인 의미 오류에 대해서는 구글번역과 실험엔진 모두 오류를 보
였지만 그 빈도가 실험엔진 결과물에서 더 많이 관찰되었다.

하지만 난제문장에 대해서는 일반문장과 다른 분석 결과가 나왔다. 우선 일반
문장의 경우 전반적으로 필수적 수정이 필요한 경우가 많지 않았다. 하지만 난제

문장은 필수적 수정이 필요한 경우가 많았으며, 두 개의 기계번역 결과물을 비교 하기가 쉽지 않은 경우도 있었다. 그럼에도 불구하고 더 적은 필수적 수정이 필요한 문장을 우위로 선정했으며 그 결과는 다음과 같다.

표 8. 구글번역과 실험엔진 필수 수정에 따른 우위 비교 (난제문장)

엔진	문장별 상대 우위
구글번역	4
실험엔진	13
판단 어려움 (긍정)*	4
판단 어려움 (부정)**	9

우위 빈도에 있어 실험엔진이 구글번역을 압도하고 있지만, 공통적으로 문제를 보인 지점도 다양하게 나타났다. 대표적으로 문장의 길이가 길고 복잡한 주술 관계를 지닌 문장의 처리에서 다수의 오류가 있었다. 그 외에도 미등재어(out of vocabulary) 처리에 있어 두 엔진 모두 취약점을 보였다. 대표적으로 고유명사인 사람의 이름을 처리하는 과정에서는 두 엔진 모두 오류가 있었다. 여기에 더해 아래 예시에서 볼 수 있듯 두 엔진 모두 “도급인”과 “수급인”을 동일한 단어로 인식 하여 오역이 발생한 사례도 있었다.

[ST]

본 계약은 도급인과 수급인간의 광고물 제작 설치에 관한 기본사항을 정한 것으로서 이와 달리 정하는 별도의 서면약정이 없는 한 도급인과 수급인은 이를 준수하여야 한다.

[구글번역]

This contract sets out the basic matters for the production and installation of advertisements **between the contractor and the contractor**, and unless there is a separate written agreement stipulated otherwise, **the contractor and the contractor** must comply with them

[실험엔진]

This Agreement stipulates the basic matters concerning the production and installation of advertisements **by the Contractor and the Contractor**, and unless otherwise specified in a written agreement, **the Contractor and the Contractor** shall comply with it.

구글번역에서 “도급인과 수급인”만 따로 입력해도 “contractor and contractor”를 출력하고, 비교적 간단한 문장인 “이 계약은 건설 프로젝트의 도급인과 수급인의 관계를 정의한다.”를 입력해도 “This contract defines the relationship between the contractor and the contractor for the construction project.”로 오역이 이뤄진다. 따라서 “도급인”과 “수급인”에 대해 명확한 의미 학습이 이뤄지지 않은 것으로 추정된다.

하지만 이러한 미등재어 문제에도 불구하고 실험엔진은 구글번역 대비 법률과 관련된 일반명사 및 고유명사 번역에 있어 우위를 보였다. 특히 국내 법령에 대한 명칭이 법제연구원 법령번역센터의 공식 번역과 일치하는 모습을 자주 보여주었다.

[ST]

갑과 을은 본 계약의 이행에 있어서 **하도급거래공정화에관한법률, 독점규제및공정거래에관한법률** 및 관련 법령의 제규정을 준수하여야 한다.

[구글번역]

In the implementation of this contract, A and B shall comply with the provisions of the Subcontract Fair Trade Act, Monopoly Regulation and Fair Trade Act, and related laws and regulations.

[실험엔진]

arty A and Party B shall comply with the provisions of the **Fair Transactions in Subcontracting Act, the Monopoly Regulation and Fair Trade Act**, and related Acts and subordinate statutes in the performance of this Agreement.

두 엔진 모두 여타 미등재어 처리에 있어 취약한 모습을 보였기에, 실험엔진이 법령명 처리에 있어 우위를 보인 것은, 실험엔진 학습에 있어 법령 관련 데이터가 다수 사용되었기 때문으로 풀이된다. 여기에 더해 양질의 데이터가 미등재어 문제 극복에 이바지할 수 있다는 자명한 원리를 다시금 보여준 사례라 하겠다.

또 한 가지 실험엔진이 우위를 보인 부분은 신경망번역의 단점으로 잘 알려진 누락(omission) 이었다. 총 30문장 중 실험엔진은 누락이 있는 기계번역 결과물이 없었지만, 구글번역은 세 개의 문장에서 누락이 나타났다.

[ST]

13세 이상 15세 미만인 자에 대해서는 고용노동부장관으로부터 취직인허증을 교부받아야 하며, 이 계약에 정함이 없는 사항은 근로기준법령에 의함

[구글번역]

Persons between the ages of 13 and 15 must obtain a work permit from the Minister of Employment and Labor. <이하 누락>

[실험엔진]

With respect to a person aged between 13 and under 15 years of age, a certificate of employment permit shall be issued by the Minister of Employment and Labor, and matters not provided for in this contract shall be governed by the Labor Standards Act and subordinate statutes

하지만 누락의 원인이 특정 문장의 특성에 기인하여 나타난 것인지, 아니면 구글번역과 시스트란 플랫폼에서 만들어진 실험엔진의 언어처리 방식 차이 때문인지는 알기 어려우며, 이 부분은 추가적인 연구가 필요한 영역이라 하겠다.

여기에 더해 실험엔진은 출발언어 텍스트의 의도를 완전히 다르게 번역한 경우가 전혀 없었지만, 구글번역은 출발언어 텍스트의 의도를 완전히 다르게 번역한 예도 있었다.

[ST]

“을” 및 “정”은 “갑”의 승인을 얻은 경우를 제외하고 본 차용금채무의 담보로 제공할 것을 약속한 자산을 제3자에게 양도·임대·담보로 제공하거나 또는 담보로 제공할 것을 약속하고 제3자를 위하여 그 자산 상에 지상권을 설정하는 등 “갑”에게 손해를 주거나 또는 그러한 우려가 있는 행위를 하지 못한다.

[구글번역]

B and “Jeong” **shall transfer**, lease, or provide the assets promised to be provided as collateral for this loan to a third party as collateral, or provide them as collateral, except in cases where the approval of “A” has been obtained. You must not commit any act that causes damage to “A” or is likely to do so by making a promise and setting superfcies on the assets for a third party.

[실험엔진]

Except for the cases for which the approval of Party A has been obtained, the Party B and the Party C **shall not transfer**, lease, provide as security the assets promised to be offered as security for this loan obligation to a third party, or shall promise to provide as security the assets, and shall establish superfcies on the assets for the third party, thereby

inflicting damage on the Party A or performing any act which may cause such damage.

상기 문장은 양도 금지 조항으로 타인에게 양도, 임대, 담보 제공 행위를 금하고 있다. 하지만 구글번역은 “shall transfer”로 번역하여 출발언어 텍스트가 의도한 양도 금지를 정반대로 번역하였다.

이상의 포스트에디팅 작업 용이성 관점에서 분석한 필수적 수정의 정량적 평가 결과를 요약하자면, 일반문장에서는 두 번역엔진 모두 필수적 수정의 빈도가 높지는 않았다. 하지만 일반문장에 대해서는 구글번역 결과물이 상대적으로 적은 필수적 수정이 필요했으며 그 원인은 숫자, 구두법 등 기본적인 번역 결과물 산출 규칙과 관련이 있어 보인다. 반면 난제문장에서는 두 번역엔진 모두 필수적 수정의 빈도가 높았다. 하지만 실험엔진 결과물이 상대적으로 적은 필수적 수정이 필요했으며 그 원인은 법률과 관련된 용어의 처리, 누락 등에서 구글번역보다 일부 우세했기 때문으로 풀이된다.

4.5. 평가 결과에 대한 논의

특정 번역엔진의 성능을 평가하는 것은 어려운 일이며, 특히 다른 엔진과 성능 비교는 신중하고 보수적으로 접근할 필요성이 있다. 그렇기에 단 하나의 평가요소로 엔진의 성능을 단정하는 것은 합리적이지 않다. 이에 본고는 제한적인 텍스트이긴 하지만, 네 가지 다른 방법으로 평가를 진행했으며 다음과 같이 평가 결과를 해석하고자 한다.

표 9. 일반문장에 대한 평가 결과 비교

엔진	정확성 평균	정확성 상대 우위	유창성 평균	유창성 상대 우위	사용 적합성 평균	사용 적합성 상대 우위	필수 수정 상대 우위
구글번역	8.9	6	7.36	8	8.6	11	10
실험엔진	9.0	9	7.46	11	8.13	6	4

일반문장에서 실험엔진이 정확성과 유창성에 있어 구글번역 대비 상대적 우위를 보였다. 하지만 구글번역이 사용 적합성과 필수적 수정에 있어 실험엔진 대비

상대적 우위를 보였다. 따라서 두 엔진 중 어떤 엔진이 우위가 있는지는 판단하기 어렵다. 다만 정확성과 사용 적합성에서 8점 이상 그리고 유창성에서 7점 이상의 평균을 득했다는 점, 그리고 필수 수정을 해야 하는 오류가 많지 않았다는 점을 감안하면 두 엔진 모두 일반문장에 대해서 준수한 성능을 보여주었다고 평가해야 할 것이다. 또한, 한 엔진이 특정 문장에서 다른 엔진보다 더 높은 점수를 받은 상대 우위 빈도를 보아도 실험엔진의 성능이 구글번역 대비 열등하다고 평가하기는 어려워 보인다. 오히려 상기 데이터는 두 개의 비슷한 성능의 엔진이, 특정 문장에서는 상대 엔진보다 더 높은 점수를 득하지만 다른 문장에서는 그렇지 않을 수 있음을 보여준 사례라 해석하는 것이 합리적일 것이다.

다음으로 난제문장에 대한 평가 결과를 요약하기 전에, 모든 영역에서 평가자들이 난제문장에 대해서 더 낮은 점수를 주었다는 점은 주목할 가치가 있다. 그림1에서 확인할 수 있듯 모든 영역에서 고르게 평균 점수가 낮아졌다는 점은, 난제문장의 기계번역 결과물이 일반문장의 기계번역 결과물 대비 품질이 낮음을 시사한다.

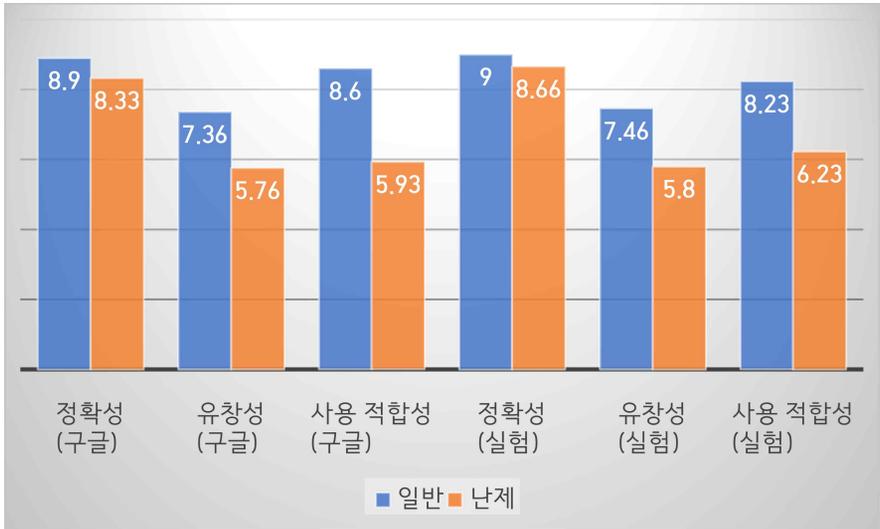


그림 1. 일반문장에서 난제문장으로 전환에 따른 평균 감소

또한, 일반문장과 난제문장 사이의 평균 감소 폭을 비교해보면, 실험엔진이 정확성(0.23)과 사용 적합성에(0.67)서 더 적은 평균 점수의 감소가 있었으며, 유창성에서 있어서는 구글번역이 미세하게 적은 평균 점수 감소가 있었다(0.06). 이는

실험엔진이 구글번역 대비 난제문장에서 품질 악화가 상대적으로 적었음을 보여 준다. 여기에 더해 4.1에서 4.4에서 실시한 모든 평가 항목에서 실험엔진은 구글번역 대비 상대적인 우위를 보였다.

표 10. 난제문장에 대한 평가 결과 비교

엔진	정확성 평균	정확성 상대 우위	유창성 평균	유창성 상대 우위	사용 적합성 평균	사용 적합성 상대 우위	필수 수정 상대 우위
구글번역	8.33	3	5.76	10	5.93	7	4
실험엔진	8.66	12	5.80	12	6.23	13	13

따라서 난제문장에서는 실험엔진과 구글번역이 모두 상대적으로 낮은 품질을 보였지만, 실험엔진이 구글번역 대비 우수한 품질의 결과물을 출력했다고 해석할 수 있다.

5. 결론

본 연구는 법률 번역에 특화된 한영 기계번역의 품질 개선 가능성을 모색하기 위해 설계되었다. 번역엔진의 품질 개선을 위해 양질의 병렬 말뭉치를 법률이라는 주제에 국한하여 수집한 이후, 상용으로 제공하는 번역엔진 플랫폼에서 학습 과정을 거쳤다. 그 결과 일반적인 계약서 문구에서는 무료 온라인 상용 번역 시스템인 구글번역 대비 비열등성을 보였다. 반면 난제가 있는 계약서 문장에서는 실험엔진이 정확도, 유창성, 사용 적합성, 필수적 수정의 필요성 등 모든 항목에서 구글번역 대비 우위를 보였다.

따라서 본 연구는 양질의 데이터가 있으면 번역엔진의 결과물 품질을 높일 수 있다는 가능성을 번역학의 관점에서 검증했다는 점에서 그 의미가 크다. 특히 난제 문장에서는 실험엔진의 우수성이 부분적으로 입증되었다. 이는 번역 실무자 및 연구자들이 양질의 말뭉치 생성 및 검증에 더욱 적극적으로 참여한다면 기계번역 엔진을 통해 더 양질의 결과물을 생성할 수 있음을 시사한다.

하지만 다양한 평가 방식을 통해 객관화를 위해 노력했음에도 불구하고 제한적

인 샘플 데이터에 기반한 결과이기에, 특정 엔진의 우위를 주장하는 것은 합리적이지 않으며, 이는 본 연구의 목적도 아니다. 본 연구가 주목한 바는 특정 엔진이 어떠한 강점이 있는지를 파악하고, 품질 개선의 방안과 가능성을 모색하는 것이었음을 다시 한번 강조하고자 한다.

본 연구 결과에서 알 수 있듯 한 엔진이 모든 문장에서 다른 엔진의 결과물 품질을 압도하는 것은 아니다. 따라서 실무적인 관점에서는 여러 번역엔진의 결과물 장단점을 번역학 전문가들이 우선적으로 평가하고 해당 텍스트에 가장 적합한 번역엔진의 결과물 2-3개를 작업자의 번역 도구 내에서 제시하여 작업자가 포스트에디팅 과정에서 선택할 수 있도록 사용자 인터페이스를 구성해 주는 것이 가장 현실적인 대안이라 판단된다.

마지막으로 번역학 전문가의 관점에서 파악한 실험엔진의 단점은 실험엔진의 개발자들에게 유용한 추후 개발 과제를 제시할 수 있었다. 이는 본 연구를 넘어서 엔진 개발자와 번역학 전문가의 협업 분야를 새롭게 제시했다는 점에서 의미가 있다.

이러한 기여에도 불구하고 제한적인 데이터를 사용했다는 사실, 그리고 시간의 감에 따라 엔진에서 출력되는 결과물이 달라질 수 있다는 점은 본 연구의 약점이라 할 수 있다. 이에 추후 연구에서는 본 연구에서 발견된 사항에 기초하여 엔진의 언어처리 방식을 개선하고, 여러 차례의 성능 평가를 더 많은 데이터를 사용하여 진행할 필요가 있다. 또한, 텍스트의 학습과 성능과의 연관 관계 파악을 위해 텍스트의 종류 역시 더욱 세분화하여 연구를 진행할 필요가 있다. 마지막으로 평가 과정에서 있어 자동 평가와 수동 평가를 적절하게 조합하여 더욱 객관적인 엔진 발전 방향성을 제시해야 할 것이다.

참고문헌

- 강청웅·노영한·김지수·최희열. (2018). 「숫자 기호화를 통한 신경기계번역 성능 향상」. 『한국 디지털콘텐츠학회 논문지』 19(6): 1161-1167.
- 김순미·신호섭·이준호. (2019). 「번역학계와 언어서비스업체 (LSP) 간 산학협력연구: 포스트에디팅 생산성과 기계번역엔진 성능 비교」. 『번역학연구』 20(1): 41-76.
- 박찬준·박기남·문현석·어수경·임희석. (2021). 「인공신경망 기계번역에서 말뭉치 간의 균형성을 고려한 성능 향상 연구」. 『한국융합학회논문지』 12(5): 23-29.
- 박찬준·김경화·박기남·임희석. (2020). 「Coronavirus Disease-19 (COVID-19) 에 특화된 인공신경망 기계번역기」. 『한국융합학회논문지』 1(9): 7-13.
- 박찬준·임희석. (2020). 「공공 한영 병렬 말뭉치를 이용한 기계번역 성능 향상 연구」. 『디지털융복합연구』 18(6): 271-277.
- 서보현·김순영. (2018). 「기계번역 결과물의 오류유형 고찰」. 『번역학연구』 19(1): 99-117.
- 신지선. (2020). 「기계번역 포스트에디팅에 관한 해외 연구 동향」. 『번역학연구』 21(4): 87-114.
- 윤미선·김택민·임진주·홍승연. (2018). 「영어-한국어 언어쌍에 적합한 포스트에디팅 가이드라인: 미래 포스트에디터를 위한 가이드라인과 그 사례」. 『번역학연구』 19(5): 43-76.
- 이상빈. (2018). 「국내학부 번역전공자의 구글 기계번역 포스트에디팅 과정 (process) 및 행위 연구」. 『번역학연구』 19(3): 259-286.
- 이상빈. (2020). 「기계번역에 관한 KCI 연구논문 리뷰: 인문학 저널 논문 (2011~ 2020년 초) 의 논의내용과 연구방법을 중심으로」. 『통역과 번역』 22(2): 75-104.
- 이준호. (2019a). 「문학번역 적용을 위한 기계번역의 한주소」. 『통번역학연구』 23(1): 143-167.
- 이준호. (2019b). 「신경망기계번역의 객관적 평가를 위한 예비연구: 자동평가와 수동평가의 균형점」. 『통번역학연구』 23(3): 171-202.
- 정혜연. (2018). 「번역의 자동평가: 기계번역 평가를 인간번역 평가에 적용해보기」. 『통번역학연구』 22(4): 265-287.
- 최문선. (2019). 「국내 번역학 기계번역 연구 동향: 내용 분석과 키워드 분석을 중심으로」. 『언어학연구』 24(1): 275-297.
- 최승권. (2007). 「영어 특허문서 자동번역을 위한 특허번역패턴 연구」. 『번역학연구』 8(1): 301-322.
- 최승권·김영길. (2010). 「번역 말뭉치로부터 추출한 어휘 번역 패턴의 의미 분류와 자동번역 시스템에의 활용」. 『번역학연구』 11(3): 277-301.
- 최승권·박은진·김영길. (2007, 5월). 제한된 도메인에 특화된 기계번역 기술 개발-특허 전문 영한 번역기를 중심으로. 한국정보처리학회 학술대회 발표논문.
- Bhatia, V. K. (1997). Translating legal genres. In Trosborg, A. (ed.), *Benjamins Translation Library* 26. Amsterdam/Philadelphia: John Benjamins, 203-216.
- Bentivogli, L., Bisazza, A., Cettolo, M. and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. The 2016 Conference on Empirical Methods in Natural Language Processing Proceedings.
- de Almeida, G. (2013). *Translating the Post-editor: An Investigation of Post-editing Changes and Correlations with Professional Experience across Two Romance Languages*. Unpublished

- PhD dissertation, Dublin City University.
- Harvey, M. (2002). What's so special about legal translation?. *Meta: Translators' Journal* 47(2): 177-185.
- Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. The 7th Workshop on Statistical Machine Translation Proceedings.
- Luong, M. T., Sutskever, I., Le, Q. V., Vinyals, O. and Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing Proceedings.
- Müller, M., Rios, A. and Sennrich, R. (2019). Domain robustness in neural machine translation. The 14th Conference of the Association for Machine Translation in the Americas Proceedings.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. and Polosukhin, I. (2017). Attention is all you need. The 31st International Conference on Neural Information Processing Systems Proceedings.
- Wang, X., Chen, C. and Xing, Z. (2019). Domain-specific machine translation with recurrent neural network for software localization. *Empirical Software Engineering* 24(6): 3514-3545.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W. and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Zhang, X., Shapiro, P., Kumar, G., McNamee, P., Carpuat, M. and Duh, K. (2019). Curriculum learning for domain adaptation in neural machine translation. The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Proceedings.

[인터넷 자료]

법률뉴스 (2019). 연 매출 2000~3000억 규모... '법률번역 시장' 선점 경쟁. [https://m.lawtimes.co.kr/Content/Article?serial=156645\(2022.03.19.검색\)](https://m.lawtimes.co.kr/Content/Article?serial=156645(2022.03.19.검색)).

This paper was received on 15 April 2022; revised on 10 June 2022; and accepted on 20 June 2022.

Author's email address

cuefit@gmail.com

About the author

Junho Lee is a Lecturer at Hankuk University of Foreign Studies. His research interests include machine translation and post-editing.