

## 법률 도메인 특화 기계번역의 번역 품질 개선 분석 - 계약서 중한 번역을 실례로 -

최승혁\* · 강병규\*\*

*Seunghyuk Choi and Byeongkwu Kang (2025). Analyzing Quality Improvements in Legal Domain-Adaptive Machine Translation: A Case Study of Chinese-Korean Contract Translation. This study evaluates domain-adaptive machine translation for legal texts by developing a “contract-adaptive machine translation model”(contract-adaptive model). Using a 53,334-character corpus of Chinese contracts, we fine-tuned a neural translation system and assessed outputs using two semantic-oriented automatic metrics: BERTScore and BLEURT. The evaluation results demonstrate that the contract-adaptive model consistently achieved higher scores across both metrics. In histograms visualizing BERTScore and BLEURT, the contract-adaptive model was concentrated in the high-score regions, and its median values were higher in the corresponding box plots. Analysis of actual translation examples showed that the contract-adaptive model delivered improved translation quality tailored to contractual usage - in terms of syntactic and semantic features. (Chung-Ang University, Sogang University, Korea)*

**Keywords:** Legal translation, domain-adaptive machine translation, TQA, BERTScore, BLEURT

**주제어:** 법률 번역, 도메인 특화 기계번역, 번역품질평가(TQA), BERTScore, BLEURT

---

\* 중앙대학교 국제대학원 전문통번역학과 한중과 강사, 제1저자

\*\* 서강대학교 중국문화학과 교수, 교신저자

# 1. 들어가며

새로운 기술은 일정 수준 보급되고 난 후에 차별화가 이루어진다. 즉, 사용자 맞춤형 전략을 통해 개별 수요를 충족시키는 방향으로 발전하는 것이다. 기계번역도 예외는 아니다. 신경망 기계번역의 사용이 보편화되고 성능도 꾸준히 향상되면서 이제는 IT, 금융, 의료 등 다양한 분야별로 맞춤화하는 추세가 나타나고 있다. 대표적인 분야는 법률이다. 현재 법률 번역은 시장 규모가 16조 원 정도로 추정될 만큼 수요가 거대하고, 법률 이외에도 금융, 임상, 조선 등 여러 전문 분야로 확장 중이다.<sup>1)</sup>

법률은 우리 생활과 밀접하다. 일상에서 쉽게 접할 수 있는 법률 관련 문서로 계약서를 꼽을 수 있다. 온라인 회원가입 약관부터 크고 작은 계약에 관련된 규범적 텍스트가 모두 계약서 범주에 포함된다(Sarcevic 1997: 11). 계약서에는 수많은 전문용어와 관습화된 표현 등 계약서 특유의 사용역(register)이 존재하기 때문에 번역할 때 주의를 기울여야 한다. 또한 계약서 번역은 단순한 언어 변환이 아닌 정확한 법률적 개념을 전달해야 하기 때문에 해당 국가의 법률 체계와 문화적 배경을 이해해야 한다. 하지만 인간 번역사가 이러한 법률 관련 지식을 모두 갖추는 것은 현실적으로 불가능하다.

기계번역을 활용해 법률 문서를 번역하면 이런 단점을 장점으로 바꿀 수 있다. 법률은 하나의 어휘가 문맥에 따라 다의(多義)로 해석되는 것을 지양하기 때문에 단의성(單義性)을 가진다(李振宇 2006: 81-83). 즉, 법률 전문용어와 관습화된 표현은 형식과 의미가 1:1로 대응하기 때문에 문학작품과 같이 독자에 따라 다르게 해석되지 않는다. 또한 기계는 방대한 법률 지식을 학습해 일관성 있는 번역 결과물을 생성할 수 있다. 정리하면 법률은 번역 수요의 증가, 텍스트 유형적 특징, 신경망 기계번역 기술의 빠른 발전 등으로 인해 도메인 특화의 효용을 기대할 수 있는 중요한 분야이다(최승혁 2023a: 352).

본 연구는 ‘계약서의 언어적 특징을 기계번역에 학습시키면 계약서 중한 번역의 품질이 향상될 수 있을까?’라는 질문에서 출발하였다. 즉, 계약서 중심의 법률 도메인에 특화된 기계번역 모델이 번역 품질 측면에서 어떤 효용이 있을지에 관한 궁금증이었다. 이에 본 연구는 계약서 도메인 특화 기계번역 모델의 구축과 그 번역 품질에 대한 자동평가를 통해 법률 도메인 특화 기계번역의 번역 품질 개선에

---

1) <https://news.mt.co.kr/mtview.php?no=2024072317155448271>(2025.04.05. 검색).

대해 논하고자 한다.

먼저 2장에서 법률 도메인 특화 기계번역에 관한 선행 연구를 통해 연구 현황을 이해한다. 3장에서 계약서 특화 모델 구축, 4장에서 의미 기반 자동평가 모델 BERTScore, BLEURT에 대해 소개한다. 5장에서는 계약서 특화 모델의 번역 결과를 분석해 번역 품질 개선도를 구체적으로 살펴본다.

## 2. 선행 연구

현재 국내에서 법률 도메인 특화 기계번역에 관한 연구는 매우 소수이며, 주로 특허의 영한 번역에 집중되어 있다.

먼저, 이진철(2022)은 특허에 특화된 NMT-PT(Neural Machine Translation - Patent Translation) 모델을 구축하고 특화 전후 모델과 범용 기계번역(Google Translate, Naver Papago) 모델의 성능을 비교했다. 특정 도메인에서의 방대한 데이터를 추가 학습(fine-tuning)하는 것이 번역 성능 개선에 효과적임을 검증하였다. 이준호(2022)는 계약서 병렬 말뭉치 50여만 쌍을 이용해 법률 특화 번역 엔진을 구축하여 한영 계약서의 기계번역 결과를 분석했다. 국내에서 도메인 특화 엔진을 직접 사용해 분석했다는 점에서 의미가 매우 크다. 이지은과 최효은(2023)은 법률과 특허에 특화된 NMT를 구축한 후 한영 번역 결과물에 대해 수동평가와 자동평가를 실시했다. 법률 도메인의 평가 결과는 모두 양호한 편이었지만 특허 도메인은 비교적 낮은 평가를 받았다. 특히, 전문용어 사용의 정확성 향상의 필요성을 강조했다. 이 연구는 동일한 법률 도메인이더라도 세부 도메인의 사용역에 따라 번역 품질에 차이가 존재하고 도메인 특화에서 전문용어 학습이 중요하다는 사실에 있어 중요한 시사점을 주었다. 최승혁(2024)은 법률 도메인에 특화된 기계번역 모델을 구축한 후에 특화 모델에서 번역의 정확성 개선 여부를 어휘, 통사, 텍스트 층위에서 분석함으로써 도메인 특화 기계번역의 효과를 입증하였다. 무엇보다 법언어학, 번역학, 기계번역 등 학제적 관점에서 법률 텍스트를 가공하고 이를 기계번역에 학습시킴으로써 기존 연구와 차별성을 두었다. 이 외에 법률 도메인 기계번역에 관한 연구로 특허 기계번역에서 발생하는 전문용어 오류에 관한 연구(최효은 2022), 법령 기계번역 포스트에디팅(MTPE)에 관한 연구(김혜림 2021; 이지은과 최효은 2024) 등이 있다.

이상의 선행 연구는 그 수가 제한적이며 연구 범위도 다소 편향된 양상을 보인다. 분석 대상이 특허 문서에 집중되어 있으며, 언어 쌍도 영-한 번역에 치우쳐 있다. 연구 방법은 대부분 기계번역 엔진의 성능 평가를 중심으로 번역 결과물의 오류 유형의 분석, 평가에 초점이 맞춰져 있다. 이와 같이 일부 연구자들이 법률 도메인 특화 기계번역에 관심을 갖고 지속적인 연구를 이어오고 있지만, 양질의 학습 데이터 구축의 어려움과 기계번역 모델 구축에 소요되는 기술적, 시간적, 비용적 제약으로 인해 관련 연구가 활발히 진행되지 못하는 상황이다(이준호 2022: 173).

### 3. 계약서 특화 모델의 구축<sup>2)</sup>

도메인 특화 기계번역이란 특정 도메인의 번역 품질 개선을 위해 특정 도메인의 데이터를 범용 기계번역에 학습시키는 것을 말한다. 계약서에서 자주 사용되는 ‘续租’의 경우, 네이버 중국어 사전에서 찾아보면 ‘임대 기한이 끝난 후 계속 임대함’이라는 해석을 제공한다.<sup>3)</sup> 하지만 실제 계약서에서는 ‘임대 연장’, ‘계약 갱신’, ‘재계약’ 등으로 표현하는 것이 계약서 사용역(register)에 부합하고, 이러한 전문용어나 정형화된 표현을 정확하게 사용해야 번역 품질이 좋다는 평가를 받는다. 본 연구의 계약서 특화 모델은 계약서를 중심으로 그 언어적 특징을 학습시켜 계약서 번역 시 더 높은 품질의 번역 결과물을 얻는 것으로 이해할 수 있다.

분석을 위해 계약서 특화 모델이 필요하다. 그 구축 과정은 크게 ① 범용 기계번역 모델과 계약서 원천 데이터 확보 ② 계약서 원천 데이터를 학습 데이터로 가공(중-한 번역 쌍 구축) ③ 미세조정(fine-tuning) 세 단계로 나누어 이해할 수 있다.

첫째, 범용 기계번역 모델과 계약서 원천 데이터 확보 단계이다. 계약서 도메인에 특화하려면 계약서 데이터를 학습하지 않은 범용 기계번역 모델이 필요하다. 파파고, 구글 번역 등 기존의 범용 기계번역 모델은 연구자 개인이 미세조정을 할 수 없다. 따라서 본 연구를 위해 AI 법률번역 엔진 오토란(Otran)의 범용 모델을 제공받았다.<sup>4)</sup> 계약서는 외부에 비공개하는 민감 정보를 담고 있어 폐쇄성이 크다.

---

2) 본 연구는 대규모 다국어 코퍼스로 사전학습한 기계번역 모델을 ‘범용 모델’, 계약서 데이터로 학습시킨 법률 도메인 특화 기계번역 모델을 ‘계약서 특화 모델’로 부른다.

3) <https://zh.dict.naver.com/#!/search?query=%E7%BB%AD%E7%A7%9F>(2025.03.05. 검색).

법령이나 협정문과 같이 온라인에 공개되어 있지 않아 데이터 확보에 어려움이 있다. 이러한 현실적인 어려움 때문에 필자 및 다른 번역사가 직접 번역했던 중국어 계약서 원문과 한국어 번역문을 원천 데이터로 확보했다. 다른 번역사로부터 계약서 데이터를 제공받은 이유는 데이터의 편향을 줄이기 위함이다. 해당 계약서에서 민감 정보와 불용어 등을 제외하는 전처리를 통해 총 53,334자 규모의 데이터를 얻었다.

둘째, 계약서 원천 데이터를 학습 데이터로 가공(중-한 번역 쌍 구축)하는 단계이다. 미세조정을 위해서는 앞에서 얻은 데이터를 문장 단위로 분절해 중국어 원문과 한국어 번역문을 대응시켜 표로 작성해야 한다. 이 과정에서 단순한 대응 작업 외에 원문과 번역문의 번역이 정확한지 대조하고, 한국어 계약서에서 사용되는 전문용어, 정형화된 표현, 문체 등을 파악해 수정 작업을 함께 진행해야 한다. 수정 시 공정거래위원회에서 제공하는 임대주택 표준 임대차 계약서<sup>5)</sup>와 법무부에서 제공하는 주택 임대차 표준계약서(2023 수정)<sup>6)</sup>를 참고하였다. 이렇게 가공된 중-한 번역 쌍을 엑셀 파일에 총 1,357개 셀로 정리하였으며, 이는 미세조정 시 범용 기계번역에 계약서 도메인 지식을 전수하는 학습 데이터로 활용된다. 아래 <그림 1>은 위에서 구축한 학습 데이터 중-한 번역 쌍의 실제 예시이다.

- 
- 4) 오탁란(Otran)은 AILingGo에서 제공하는 AI 법률번역 엔진으로 법률에 특화된 신경망 기계번역 플랫폼과 함께 사용자 맞춤형 학습을 통한 고품질의 번역 서비스를 제공한다.
  - 5) [https://www.ftc.go.kr/www/selectBbsNttView.do?pageUnit=10&pageIndex=1&searchCnd=all&searchKwrd=%EC%9E%84%EB%8C%80%EC%B0%A8&key=202&bordCd=201&nttSn=11098\(2025.05.05.검색\)](https://www.ftc.go.kr/www/selectBbsNttView.do?pageUnit=10&pageIndex=1&searchCnd=all&searchKwrd=%EC%9E%84%EB%8C%80%EC%B0%A8&key=202&bordCd=201&nttSn=11098(2025.05.05.검색)).
  - 6) [https://www.immigration.go.kr/bbs/moj/118/569781/artclView.do\(2025.05.05.검색\)](https://www.immigration.go.kr/bbs/moj/118/569781/artclView.do(2025.05.05.검색)).

| No. | 중국어 원문   | No. | 한국어 번역문   |
|-----|--|-----|---|
| 31- | 乙方权利和义务  | 31+ | 윤의 권리와 의무   |
| 32- | 乙方不得超越代理权限, 必须按条款1中约定的组展企业范围及条款2中约定的组展价格组织相应的展商参展。           | 32+ | 윤은 대행권한을 초과할 수 없고 반드시 조항1에서 약정한 전시대행 업체의 범위 및 조항2에서 약정한 전시대행 가격에 따라 관련된 참가업체를 조직해야 한다.                      |
| 33- | 乙方在组展过程中应以甲方代理机构身份开展工作, 不得以洽谈组委会名义进行宣传。                      | 33+ | 윤은 참가대행 과정에서 갑의 대행기관 자격으로 업무를 수행하고, 투자상담회 조직위원회 명의로 홍보해서는 아니 된다.  |
| 34- | 组展通用招展邀请函由甲方提供(中英文版)。  | 34+ | 참가대행에 통용되는 초청전시회 초대장은 갑이 제공한다(중·영문 버전).   |
| 35- | 乙方若因组展需要需自行印制组展邀请函, 文字内容须经甲方审核, 设计、印刷由乙方负责。                  | 35+ | 윤이 만일 참가대행의 필요로 인해 자체적으로 참가대행 초청서를 인쇄하고 제작해야 하는 경우, 글 내용은 반드시 갑의 확인을 거쳐야 하고, 디자인과 인쇄는 윤이 책임진다.              |
| 36- | 乙方组织的参展企业需遵守展会规定, 禁止现场售卖展品。                                  | 36+ | 윤이 조직한 참가업체는 상담회 규정을 준수하며 현장에서 전시품의 매매를 금지한다.   |
| 37- | 组委会谢绝中药材、玉器、珠宝等商品及假冒伪劣商品参展。                                  | 37+ | 조직위원회는 중국약재, 옥기, 보석 등 상품 및 위조상품의 전시를 사절한다.  |
| 38- | 乙方对所组织的展商须实行全程服务。  | 38+ | 윤은 조직된 참가업체에 대해 전반적인 서비스를 제공한다.   |
| 39- | 企业在缴交展位费后, 乙方应收集、整理参展企业信息, 在规定时间内交给甲方制作名录、楣牌等, 协助展商上网申办参展证件。 | 39+ | 업체가 부스비용을 지불하면 윤은 참가업체 정보를 수집하고 정리한 후 정해진 시간 안에 이를 갑에게 제공하여 명단과 업체간판 등을 제작하고, 참가업체가 온라인으로 참가서류를 수속하도록 협조한다. |
| 40- | 布展期间负责所组企业报到、协调。   | 40+ | 전시기간 동안 조직된 업체의 등록과 조율을 책임진다.   |

그림 1. 계약서 특화 모델 구축을 위한 학습 데이터 중-한 번역 쌍

셋째, 미세조정(fine-tuning)으로 바로 특화 단계이다. 미세조정은 대규모 다국어 코퍼스나 범용 데이터로 사전학습된 기계번역 모델 위에서 전이학습(transfer learning)을 통해 이루어진다. 미세조정은 전문용어나 특수한 문체에 맞는 번역문 제공, 적은 양의 데이터로도 효과적인 성능 향상 가능, 범용 모델의 일반 지식과 특정 도메인 지식 제공 등의 장점이 있다.

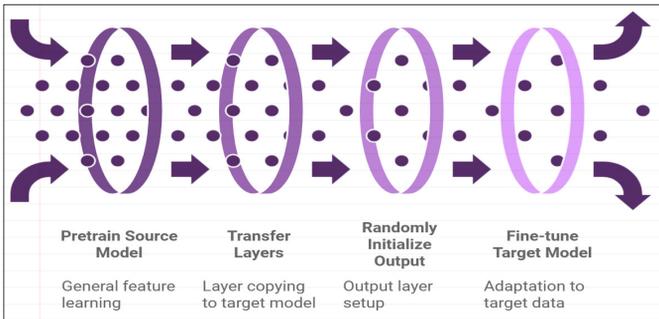


그림 2. 미세조정(fine-tuning) 과정

7) 전이학습은 사전학습(pre-train)과 미세조정(fine-tuning)으로 나뉜다. 사전학습은 방대한 데이터로부터 범용 도메인에 대한 지식을 학습하는 것이고, 미세조정은 사전학습 한 데이터를 기반으로 특정 도메인의 특수성에 맞추어 가공하는 것이다.

<그림 2>는 미세조정을 포함한 전이학습의 전반적 과정을 보여준다. 먼저 Pretrain Source Model은 본 연구에서 사용한 오토란(Otran) 범용 모델에 해당하며, 일반 도메인 데이터(Source data set)로 사전학습된 범용 기계번역 모델이다. 이 범용 기계번역 모델은 Transfer Layers에서 범용 모델의 출력층(Output layer)을 제외한 모든 파라미터가 Target model로 복사된다. 그런 다음 Random initialization을 통해 Target model에 출력층을 추가하고 가중치를 무작위 초기화<sup>8)</sup>한다. 마지막으로 계약서 중-한 번역 쌍으로 Target model의 모든 층(layer)을 미세조정(fine-tuning)하면 계약서 특화 모델이 구축된다. 훈련 데이터를 한번 모두 학습하는 주기인 에포크(epochs)를 18-22, 신경망에서 비이상적 정확률을 나타내는 과적합(overfit)을 방지하는 드롭아웃(dropout)은 0.2-0.3으로 설정하였다.

#### 4. 의미 기반의 자동평가 모델 - BERTScore, BLEURT

기계번역의 번역 품질 평가는 크게 자동평가와 수동평가로 나눈다. 수동평가는 인간 번역사가 직접 평가하는 것으로 여전히 가장 신뢰 받는 방법이다. 하지만 커다란 시간과 비용 소모, 신뢰도를 갖춘 평가자 확보의 어려움, 평가자 간 일관성 유지의 어려움, 동일한 조건에서 동일한 결과를 얻는 반복성 및 재현성 불가 등 여러 한계가 존재한다(Wenbo Wang 2023: 108). 반면 자동평가는 인간 평가자 간 편차의 영향을 받지 않아 객관성과 일관성을 확보할 수 있다. 무엇보다 수치에 기반한 점수 산출 방식에 따라 t-검정과 같은 통계 분석에 용이하다.

지금까지 기계번역 품질 평가를 위해 BLEU, Meteor, ChrF, SacreBLEU, COMET, BERTScore, BLEURT 등과 같은 다양한 자동평가 모델이 개발되었다. 기계번역 품질 평가의 관건은 정답 번역과 후보 번역의 쌍을 비교하는 것으로 자동평가 모델은 크게 형태 기반의 유사도 평가 방식과 사전학습 모델 기반의 평가 방식으로 나눌 수 있다.

형태 기반의 자동평가는 문자나 어휘 위주의 N-gram 계산, 즉 문장 표면에 드러나는 형태적 유사성에 의존한다. BLEU, Meteor, ChrF 등이 대표적이다. 이러한 자

---

8) 레오나르도 스카비니 외(Leonardo Scabini et al. 2022: 4)에 따르면 무작위 초기화(random initialization)는 정규 분포 등에서 값을 무작위(random)로 생성해 초기화하는 것으로 미세조정 시작 시점의 가중치(weight)에 변화를 주어 네트워크의 대칭성 문제를 해결하는 역할을 한다.

동평가 모델은 N-gram 일치율, 즉 단어 일치 여부에 지나치게 의존하기 때문에 문장의 의미적 유사성이나 문맥을 반영하지 못하고, 어순에 민감하여 창의적 표현을 평가하기 어렵다는 단점이 있다. 이러한 단점을 극복하기 위해 신경망 기술을 활용해 사전학습 모델을 기반으로 한 BERTScore, BLEURT, COMET 등의 자동평가 모델이 등장했다. 이 모델들은 의미 기반의 토큰 임베딩을 사용해 의미적 유사도를 측정하기 때문에 동의어나 의역(paraphrase)을 보다 정확하게 평가하고 먼 거리 의존성과 어순 정보를 효과적으로 학습하도록 설계되어 있어 형태 기반의 평가 모델 방법들보다 더 우수한 성능을 보인다(Tianyi Zhang et al. 2020: 1-3).

계약서와 같은 법률 문서의 번역은 전문용어 사용의 정확성뿐만 아니라 문맥에 따라 정확한 의미를 전달하는 것이 중요하다. 만일 단순히 어휘나 문장 일치와 같은 형태적 유사성에만 의존해 평가할 경우, 의미 정확성이 높아도 감점을 받아 낮은 품질의 문장으로 평가될 수 있다. 이에 보고는 법률 문서인 계약서의 특성을 고려해 형식과 의미를 모두 평가할 수 있는 사전학습 모델이 번역 품질에 더 적합하다고 판단한다. 또한 상술한 수동평가의 한계를 피하고자 번역 품질 개선 분석을 위한 번역 품질 평가는 딥러닝 기반의 자동평가로만 진행한다.

자동평가 모델로는 BERT 기반의 BERTScore와 BLEURT를 사용한다. BERT는 언어 이해의 단위가 정교하고, 문맥을 고려하도록 훈련되었으며, 통사구조를 반영할 수 있는 장점이 있으며(정혜연과 서수연 2022: 122-123), 수동평가와의 상관성이 높게 나타나(Tianyi Zhang et al. 2020; Yiming Yan et al. 2023) 인간 평가 결과를 예측할 수 있는 근거로 삼을 수 있다. 특히, AI를 기반으로 한 자동평가는 기계 번역 시스템의 성능 평가에 있어 수동평가보다 신속성, 객관성, 효율성을 보장해 줄 뿐 아니라, 여러 자동평가 지표를 결합하면 보다 신뢰성 있는 결과를 도출할 수 있다(임형재와 리번켈빈 2024: 209).

#### 4.1. BERTScore

BERTScore는 BERT와 같은 대규모 사전학습 모델을 기반으로 후보 문장과 참조 문장 간의 유사성을 평가하는 자동평가 모델로 토큰(Token)과 문장 층위에서 평가를 실시한다. 먼저 토큰 층위에서 정확성(Precision), 재현율(Recall), 정확성과 재현율의 조화평균(F1-score)을 활용해<sup>9)</sup> 후보 문장과 참조 문장 간의 유사성을 계산한다.

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j, \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j, \quad F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

그런 다음 각 토큰의 F1-score를 평균화하여 문장 층위 점수를 계산한다. 0에서 1 사이의 점수로 평가하며 값이 클수록 두 문장 간 의미적 유사성이 높다. 문맥적 의미를 반영하는 BERTScore는 인간의 주관적 평가와 상관관계가 높아 이 지표를 활용하면 문장 간의 의미적 유사성을 효과적으로 평가할 수 있다(Tianyi Zhang et al. 2020; 정혜연과 서수연 2022).

### 표 1. BLEU Score와 BERTScore 점수 비교

| 원문   | 기계번역  | BLEU Score | BERT Score |
|--|---|------------|------------|
| 因不可抗力致使合同无法继续履行的, 应当通知对方当事人, 本合同自通知到达对方当事人时解除。 | 불가항력으로 인하여 계약의 계속 이행이 불가능한 경우, 상대방 당사자에게 통지하고 본 계약은 상대방 당사자에게 통지할 때 해지한다. | 0.028      | 0.868      |

<표 1>에서 기계번역 번역문은 ‘不可抗力’, ‘当事人’, ‘解除’ 등 중국어 계약서 원문의 전문용어뿐 아니라 의미도 정확하게 번역했다. 이에 대해 BLEU Score는 N-gram 일치율에만 의존해서 중국어와 한국어의 어순 차이를 반영하지 못하고 0.028점으로 품질을 낮게 평가했다. 반면 BERTScore는 불가항력이 발생한 가정적 상황과 그에 따른 계약의 해지에 관한 의미적 유사성을 반영해 0.868점으로 높게 평가했다.

## 4.2. BLEURT

BLEURT 역시 BERT 모델을 기반으로 문장의 의미를 평가한다. BLEURT는 단

- 
- 9) 정확성(Precision)은 후보 문장을 기준으로 참조 문장과의 유사성을 나타낸다. 재현율(Recall)은 참조 문장을 기준으로 후보 문장과의 유사성을 보여준다. F1-score는 정확성과 재현율의 조화평균으로 두 문장 간의 유사성을 균형 있게 평가한다.

어의 고정된 의미에서 벗어나 텍스트 맥락에 따른 의미를 반영하기 위해 문맥 임베딩(contextual embedding) 방식을 사용한다.<sup>10)</sup> 또한 수동평가와 차이를 최소화하기 위해 회귀 모델을 이용해 최종 점수를 0에서 1 사이로 산출하는데 값이 클수록 참조 문장과 번역 문장 간 유사도가 높다.

$$\hat{y} = f(x, \tilde{x}) = Wv_{\text{BERT}} + b$$

(y: 최종 점수, W: 가중치, b: 편향)

BLEURT 역시 사전에 학습된 언어 모델을 활용해 문장의 의미적 유사성을 평가하기 때문에 단어 일치율과 같은 형태적 유사성에 기반한 지표보다 성능이 우수하고 수동평가와 높은 상관관계를 보인다(Yiming Yan et al. 2023).

**표 2. BLEU Score와 BLEURT 점수 비교**

| 원문                                   | 기계번역  | BLEUScore | BLEURT |
|--------------------------------------|---|-----------|--------|
| 对本合同中未约定或约定不明的内容, 双方可根据具体情况签订书面补充协议。 | 본 계약에서 약정되지 않았거나 약정이 불분명한 내용에 대해서는 양 당사자가 구체적인 상황에 따라 서면 보완 계약을 체결할 수 있습니다. | 0.228     | 0.679  |

<표 2>의 예시에서 기계번역은 ‘双方’을 ‘양 당사자’로, ‘补充协议’을 ‘보완 계약’으로 번역했다. 이를 쌍방 혹은 추가 계약으로 번역하는 것이 적합하지만 전반적인 의미를 이해하는 데 큰 지장을 주지는 않는다. 이에 대해 BLEU Score는 0.228로 낮은 점수를 주었고, BLEURT는 그보다 3배 정도 높은 0.679를 주었다. BLEU Score만 본다면 이 번역문은 활용 가치가 낮은 문장으로 판단할 수 있지만, BLEURT는 약 0.7점으로 1점에 가까운 점수를 얻었기 때문에 참조 문장에 근접하고 수용도가 높다.

10) 단어를 고정된 차원의 연속적인 벡터 공간에 위치시키는 것으로 의미적 유사성이 높은 단어들이 벡터 공간에서 가까운 위치에 놓인다. 단어 간의 의미적, 문법적 관계의 수치적 파악이 가능하다.

## 5. 번역 품질 개선 분석

범용 모델과 계약서 특화 모델의 번역 품질 개선을 비교 분석하기 위해 아래의 테스트를 실시하고 BERTScore, BLEURT를 통해 번역 품질 개선도를 살펴보았다.

### 5.1. 범용 모델과 계약서 특화 모델의 번역 품질 테스트

번역 품질 테스트를 위한 기계번역 모델, 즉 범용 모델과 계약서 특화 모델은 앞에서 준비를 마쳤다. 두 모델의 번역 품질 테스트를 위해서는 테스트 문장이 필요하다. 테스트 문장의 중국어 원문은 ‘全国合同示范文本库’<sup>11)</sup>에서 주택 임대차 계약서(房屋租赁合同(示范文本), 河北省房屋租赁合同(示范文本))에서 추출하였으며, 개별 계약서의 글자 수가 많지 않아 동일 유형의 계약서 두 종류를 선정하였다. 계약서 원문에서 ‘[ ], :’ 등 문장부호, 각 조항의 순서와 표제(예: ‘第十一条 合同解除’), 계약 당사자의 선택사항(예: [是][否], [有][无] 등) 및 서명이나 날인처럼 단순 어휘로 이루어진 부분 등 테스트에 불필요한 요소 및 두 계약서에서 중복된 부분을 제거하였다. 그런 다음 계약서 조항을 문장 단위로 엑셀 파일에 150개 셀로 정리하였다. 이 150개 테스트 문장을 범용 기계번역 모델과 계약서 특화 모델에서 번역하고, 각각의 번역 문장을 BERTScore, BLEURT로 자동평가하였다.<sup>12)</sup>

```
73 # BERTScore
74 P, R, F1 = bert_score(translated, reference, lang="ko")
75 df_results["BERTScore"] = F1.tolist()
76
77 # BLEURT (Hugging Face Transformers)
78 def compute_bleurt(hypothesis, reference):
79     inputs = bleurt_tokenizer(reference, hypothesis, return_tensors="pt", padding=True, truncation=True)
80     with torch.no_grad():
81         scores = bleurt_model(**inputs).logits.squeeze().tolist()
82     return scores
83
84 df_results["BLEURT"] = [
85     compute_bleurt(hyp, ref)
86     for ref, hyp in zip(reference, translated)
87 ]
```

그림 3. BERTScore, BLEURT 자동평가 화면

11) [https://htsfwb.samr.gov.cn/\(2025.03.15.검색\)](https://htsfwb.samr.gov.cn/(2025.03.15.검색)).

12) git hub 참고문헌 참조.

## 5.2. 범용 모델과 계약서 특화 모델의 번역 품질 개선 비교

### 5.2.1. 범용 모델과 계약서 특화 모델의 BERTScore 비교

<그림 4>는 범용 모델과 계약서 특화 모델의 BERTScore 점수를 히스토그램과 박스플롯으로 시각화한 그래프이다.

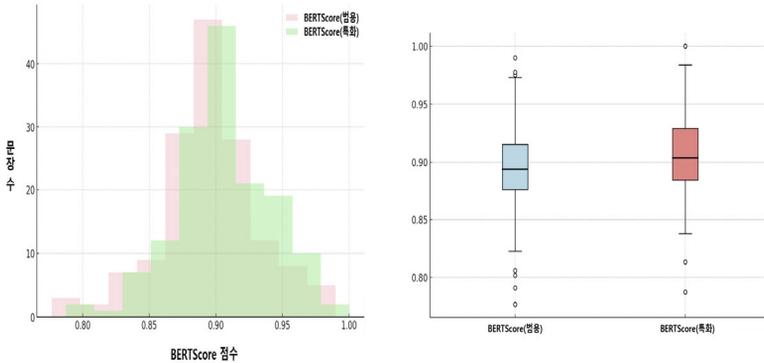


그림 4. 범용 모델과 계약서 특화 모델의 BERTScore 점수 히스토그램(좌)과 박스플롯(우)

왼쪽의 히스토그램은 범용 모델과 계약서 특화 모델의 BERTScore 분포를 보여 준다. 히스토그램에서 두 모델은 모두 높은 점수에 집중되어 있다. 그중 특화 모델의 BERTScore가 조금 더 오른쪽으로 치우쳐 범용 모델보다 더 높은 점수 영역(9.2~9.5)에 집중해 있다. 이는 특화 모델의 BERTScore가 전반적으로 더 높아 번역 품질이 상대적으로 우수함을 보여준다. 평균과 표준편차를 살펴보면, 범용 모델의 평균은 0.8941, 표준편차는 0.0374였다. 계약서 특화 모델의 평균은 0.9054, 표준편차는 0.0362였다. 특화 모델이 평균도 더 높고 표준편차도 더 작아 일관된 성능을 나타냈다. 동일한 테스트 문장에 대한 범용 모델과 특화 모델 간 BERTScore의 상관관계를 살펴보기 위해 t-검정을 실시하였다. 그 결과, t-검정 값 -4.0005, p값 0.000099로<sup>13)</sup> 유의미한 차이를 보였으며, 이는 전체적으로 특화 모

13) 일반적으로 p값은 유의수준 0.05를 기준으로 한다.  $p < 0.05$ 이면 우연히 발생할 가능성이 없어 통계적으로 유의하다고 판단하며,  $p \geq 0.05$ 이면 우연히 발생했을 가능성이 있어 통계적으로 유의미하지 않음을 나타낸다.

델의 번역 품질과 일관성이 높다는 점을 시사한다.

오른쪽의 BERTScore 분포 박스플롯에서 두 모델 모두 중앙값이 약 0.90으로 번역 품질이 평균적으로 유사하게 나타났다. 하지만 계약서 특화 모델의 중앙값이 범용 모델보다 높았고, 이는 전체 번역 문장 중 50% 이상이 참고 문장과 높은 의미적 유사성이 있음을 보여준다. 이상치의 경우, 범용 모델이 계약서 특화 모델보다 분포가 더 넓고, 하위 이상치의 개수도 많다. 이는 범용 모델이 특정 문장에서 상대적으로 낮은 의미 유사도를 보였음을 의미한다.

분석 결과를 정리하면 BERTScore는 문장 수준에서 의미적 일치를 측정하는 지표이므로 중앙값에서 우위를 갖는다는 사실을 통해 계약서 특화 모델이 우수한 품질의 번역을 일관성 있게 생성했음을 확인했다. 이는 특정 도메인에 특화된 모델이 의미적 정확성 향상에 긍정적 영향을 미쳤다는 근거로 해석할 수 있다.

<표 3>은 전체 테스트 문장 150개의 BERTScore 점수를 5개 구간으로 나눈 것이다. 계약서 특화 모델의 점수가 더 높게 나타난 구간의 예문을 통해 어떤 부분에서 번역 품질 개선이 이루어졌는지 살펴보자.

**표 3. BERTScore 점수 구간별 문장 수**

| 점수 구간 |             | 문장 수          |               |
|-------|-------------|---------------|---------------|
|       |             | BERTScore(범용) | BERTScore(특화) |
| 1     | 0.777~0.820 | 5             | 3             |
| 2     | 0.821~0.865 | 18            | 15            |
| 3     | 0.866~0.910 | 85            | 71            |
| 4     | 0.911~0.954 | 32            | 48            |
| 5     | 0.955~1.000 | 10            | 13            |

<표 3>에서 고득점 구간인 4, 5번 구간의 계약서 특화 모델 문장 수가 더 많다. 계약서 특화 모델이 고품질의 번역 문장을 더 많이 생성하였음을 명확히 알 수 있다.

[ST]

2. 违反本合同约定, 擅自将房屋转租或者出借给他人的 ;

[TT1] BERTScore(범용), 점수: 0.874

2. 본 계약의 약정을 위반하여 제멋대로 주택을 전대하거나 타인에게 대여하였다.

[TT2] BERTScore(특화), 점수: 0.906

2. 본 계약의 약정을 위반하여 무단으로 부동산을 타인에게 전대 또는 대여한 경우

위 예문을 어휘 층위에서 살펴보면, [ST]의 ‘擅自’를 [TT1]은 ‘제멋대로’로, [TT2]는 ‘무단으로’라고 번역해 후자가 계약서 사용역에 부합하는 어휘를 사용하였다. 통사 층위에서 보면, 원문의 ‘的’구문을 [TT1]이 시제와 가정의 용법에서 오류를 보인 것과 달리, [TT2]는 가정의 의미를 정확히 살려 번역하였다. 계약서 등 중국 법률 문서에서 ‘的’구문은 특별한 통사적 구조로 [XP+的, ] 형식으로 사용하면 가정의 의미를 갖는다.<sup>14)</sup> 그리고 이 [XP+的, ] 구조를 기계번역에 가정적 의미로 일정하게 번역하도록 학습시키면 해당 구조에 대한 의미의 정확성을 확보할 수 있다(최승혁 2023: 117). 위 예문에서 계약서 특화 모델은 미세조정을 통해 계약서 데이터의 [XP+的, ] 구조를 학습했기 때문에 가정적 의미를 정확하게 생성해 냈다. 즉, 미세조정 결과의 번역문의 어휘적, 의미적 층위에 정확히 반영되어 번역 품질이 개선되었음을 확인할 수 있다.

### 5.2.2. 범용 모델과 계약서 특화 모델의 BLEURT 비교

<그림 5>는 범용 모델과 계약서 특화 모델의 BLEURT 점수를 히스토그램과 박스플롯으로 시각화한 그래프이다.

---

14) 왕제(王洁 1996), 류홍잉(刘红婴 2003), 동슈광(董秀芳 2003), 천쑤(陈炯 2004), 순이화(孙懿华 2006) 등 중국 법언어학자는 법률 텍스트에서 구조조사 ‘的’가 문미에 오는 명사구 또는 그러한 구조를 ‘的’구문으로 정의하고 특별한 용법으로 보았다. 특히 이러한 구문은 법률 텍스트에서 가정적 조건 관계를 표시하는 동시에 고유의 문체적 특징까지 드러낸다고 보았다.

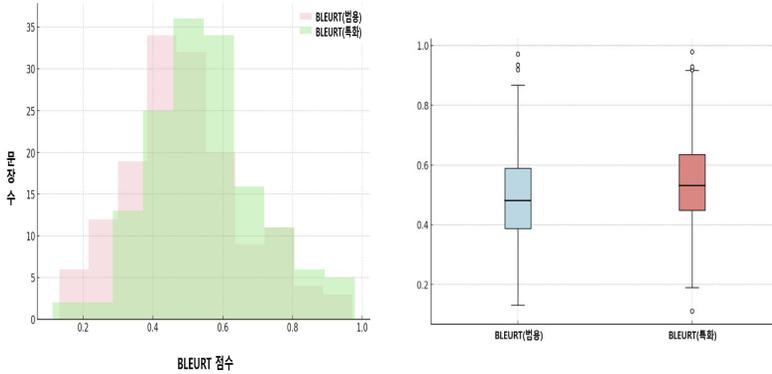


그림 5. 범용 모델과 계약서 특화 모델의 BLEURT 점수 히스토그램(좌)과 박스플롯(우)

왼쪽의 히스토그램은 범용 모델과 계약서 특화 모델의 BLEURT 점수 분포를 보여준다. 두 모델 모두 비교적 넓은 분포를 보였지만, 고득점 구간에서 더 많은 분포를 보인 것은 계약서 특화 모델이다. 평균과 표준편차를 살펴보면, 범용 모델의 평균은 0.4831, 표준편차는 0.2097이었다. 계약서 특화 모델의 평균은 0.5378, 표준편차는 0.2022이었다. 계약서 특화 모델의 평균이 더 높고 표준편차도 더 작아 성능의 일관성을 확인할 수 있었다. 동일한 테스트 문장에 대한 범용 모델과 계약서 특화 모델 간 BLEURT의 상관관계를 살펴보기 위해 t-검정을 실시했다. 그 결과, t-검정 값 -3.9669, p값 0.000113으로 유의미한 차이를 보였다. 히스토그램 분석 결과, 계약서 특화 모델의 번역 품질의 일관성이 높았다.

BLEURT는 문장의 의미적 유사성을 정량적으로 평가하는 척도로 번역 문장과 참조 문장과의 의미적 일치성이 높을수록 값이 크다. 오른쪽의 BLEURT 박스플롯을 살펴보면, 두 모델의 BLEURT 점수는 0.4~0.7 범위에 분포했다. 계약서 특화 모델은 중앙값이 약 0.52로 범용 모델의 약 0.45 대비 다소 상승해 계약서 특화 모델이 참조 문장에 가까운 번역 문장을 생성하였다. 이상치 분석에서 주목할 만한 점은 상한을 초과하는 고득점 이상치가 두 모델에서 각각 3건씩 발생했는데 이는 번역문의 의미적 일치도가 매우 높음을 의미한다. 해당 원문의 길이나 특정 어휘의 연어 관계 등 문장의 특성을 파악해 어떤 유형의 문장에서 번역 품질이 특히 높은지 파악할 수 있는 단서로 삼을 수 있다.

분석 결과를 정리하면 BLEURT 점수를 통해 계약서 특화 모델에서 생성한 번

역 문장이 계약서의 언어적 특성을 잘 반영하였을 뿐 아니라 의미의 일치성이 높아 계약서 특화 모델에서 번역 품질이 개선되었음을 확인할 수 있다.

<표 4>는 전체 테스트 문장의 150개 BLEURT 점수를 5개 구간으로 나눈 것이다. 계약서 특화 모델의 점수가 더 높게 나타난 구간의 예문을 통해 어떤 부분에서 번역 품질 개선이 이루어졌는지 살펴보자.

**표 4. BLEURT 점수 구간별 문장 수**

| 점수 구간 |             | 문장 수       |            |
|-------|-------------|------------|------------|
|       |             | BLEURT(범용) | BLEURT(특화) |
| 1     | 0.111~0.284 | 15         | 4          |
| 2     | 0.285~0.457 | 49         | 38         |
| 3     | 0.458~0.631 | 59         | 70         |
| 4     | 0.632~0.804 | 21         | 27         |
| 5     | 0.805~1.000 | 6          | 11         |

<표 4>를 보면 고득점 구간인 4, 5번 구간의 계약서 특화 모델 문장 수가 더 많다. 계약서 특화 모델이 고품질의 번역 문장을 더 많이 생성하였음을 명확히 알 수 있다.

[ST]

承租人有第十一条第(五)项约定情形的, 应按年租金的 %向出租人支付违约金, 或者出租人要求承租人将房屋恢复原状或赔偿相应损失。

[TT1] BLEURT(범용), 점수: 0.395

임차인이 제11조 제5항 약정이 있는 경우에는 임대인에게 연 임대료의 %에 해당하는 위약금을 지급하거나 임대인이 임대인에게 주택의 원상복구 또는 그에 따른 손해배상을 요구하여야 한다.

[TT2] BLEURT(특화), 점수: 0.607

임차인이 제11조 제5항에서 약정한 상황이 있을 경우, 연간 임대료의 %에 따라 임대인에게 위약금을 지불하거나, 임대인은 임차인에게 부동산의 원상회복 또는

이에 따른 손해배상을 요구하여야 한다.

위 예문의 번역 품질 차이를 어휘 층위에서 살펴보자. [ST]의 ‘支付’를 [TT1]은 ‘지급’으로, [TT2]는 ‘지불’로 다르게 번역하였다. 모두 상대방에게 돈을 준다는 뜻이다. 한국법제연구원에서 제공하는 법령 용어에 따르면 ‘지급’은 주로 봉급, 수당, 여비 등 급여적 성질을 가진 것 및 보험급여 등을 주는 것이고<sup>15)</sup>, ‘지불’은 금전채권의 이행으로서 금전을 주는 것을 말한다.<sup>16)</sup> 따라서 ‘지불’이 계약서 사용역에 적합하다.

계약서는 주체와 대상의 관계가 중요하다. 계약서 등 법률 문서는 최소한 두 명의 주체가 필요하고, 이들은 서로 대척점에 서서 동작이나 행위에 방향성을 발생시키기 때문에 이를 정확하게 번역해야 한다(최승혁 2023b: 524). 예문에서 원문의 ‘承租人’을 [TT1]은 ‘임대인’으로, [TT2]는 ‘임차인’으로 번역해 후자의 정확성이 높았다.

분석의 편의상 문장 단위로 분절했지만 [ST]의 ‘有第十一条第(五)项约定情形的’에는 6가지 하위 항목이 포함되어 있다. 이를 [TT1]은 제5항 자체에 한 가지 상황만 있는 것으로 번역한 반면 [TT2]는 하위 항목이 있음을 추측할 수 있도록 원문의 의미를 보다 세밀하고 정확하게 전달했다.

BLEURT 점수에서도 계약서 특화 모델이 어휘적, 의미적 층위에서 계약서 사용역에 더욱 정확하게 부합해 높은 번역 품질을 보여주었다.

## 6. 나오며

본고는 중국어 계약서 데이터를 미세조정하여 구축한 계약서 특화 모델을 통해 중한 계약서 번역 품질의 개선을 살펴보았다. 번역 품질은 의미 기반의 평가가 가능하고 인간 평가와의 상관성이 높은 BERTScore, BLEURT를 사용했다. 평가 결과, 계약서 특화 모델의 BERTScore, BLEURT 점수가 전체적으로 높게 나타나 번역 품질 개선을 확인했을 뿐 아니라 도메인 특화의 효용도 확인하였다.

BERTScore, BLEURT 점수를 시각화한 히스토그램에서 계약서 특화 모델의 점

15) [https://www.klri.re.kr/kor/business/bizLawDicKeyword.do\(2025.05.05.검색\)](https://www.klri.re.kr/kor/business/bizLawDicKeyword.do(2025.05.05.검색)).

16) [https://www.klri.re.kr/kor/business/bizLawDicKeyword.do\(2025.05.05.검색\)](https://www.klri.re.kr/kor/business/bizLawDicKeyword.do(2025.05.05.검색)).

수가 고득점 구간에 분포하였으며 박스플롯에서도 중앙값이 더 높았다. 구체적으로 계약서 특화 모델이 생성한 번역문은 전문용어, [XP+的, ] 구조, 주체와 대상 등 법률 문서인 계약서의 사용역을 어휘, 통사, 의미 층위에서 정확하게 번역해 범용 모델에 비해 개선된 번역 품질을 보여주었고, 이는 BERTScore, BLEURT 점수로도 증명되었다. BERTScore, BLEURT 점수는 인간 평가와 상관성이 높아 계약서 특화 모델의 번역문은 수용성이 높을 것으로 예상할 수 있다.

법률 도메인 기계번역에 관한 연구가 부족하고 영어와 특허에 연구가 집중된 상황에서 중국어 계약서를 대상으로 한 본 연구는 연구 외연의 확장과 도메인 특화 기계번역의 효용 증명의 관점에서 의미를 찾을 수 있다.

하지만 아직 여러 한계점과 함께 후속 연구의 필요성도 존재한다. 첫째, 데이터 규모의 한계이다. 계약서는 매매, 임대차, 상속 등 매우 구체적인 하위 분류가 있고, 분류별 사용역이 다양하다. 본 연구는 임대차 계약서를 중심으로 분석을 진행했지만, 향후 다양하고 방대한 계약서 데이터를 구축해 도메인 특화의 효용성을 극대화할 필요가 있다. 둘째, 본 연구에서 사용한 BERTScore와 BLEURT 자동평가 지표는 빠르고 일관된 평가에는 용이하지만, 수동평가만큼 완벽하지 못하다. 따라서 후속 연구에서는 수동평가와의 상관성 분석이나 보완적 측면에서 수동평가 추가 등 보다 정교한 평가 체계를 고민할 필요가 있다. 셋째, 번역 품질 개선 항목의 구체적인 유형을 어휘, 통사, 화용 등 언어학적 관점에서 체계적으로 귀납, 정리하고, 이를 다시 미세조정 데이터 가공 과정에 반영해 양질의 학습 데이터를 생성하는 데 효과적으로 활용하는 선순환을 형성해야 한다.

도메인 특화 기계번역은 법률 분야 이외에도 의료, 금융, IT 등 확장 가능성과 수요가 높은 분야이다. 현재 연구는 소수에 불과하지만, 본 연구의 여러 제한점을 극복하는 후속 연구가 활발히 이루어지기 기대한다.

## 참고문헌

- 김혜림. (2021). 「중한 법령 기계번역 포스트에디팅 교육을 위한 예비 연구」. 『번역학연구』 22: 65-99.
- 이건철. (2022). 특허데이터로 학습한 트랜스포머 기반의 인공지능법 번역모델. 국민대학교 소프트웨어융합대학원 석사학위논문.
- 이준호. (2022). 「법률 특화 번역엔진 성능 평가—한영 계약서 번역을 중심으로」. 『T&I review』 12: 169-192.
- 이지은·최효은. (2023). 「인공신경망 기반 맞춤형 기계번역엔진의 성능 평가: 법률 및 특허 한영번역 결과물 평가 사례를 중심으로」. 『번역학연구』 24: 9-37.
- 이지은·최효은. (2024). 「인간번역과 기계번역 및 포스트에디팅의 오류 비교: 한영 법률 번역 사례를 중심으로」. 『번역학연구』 25: 11-39.
- 임형재·리변철킨. (2024). 「한국어 번역 교육을 위한 인공지능 기반 평가모델 연구 - 다중 자동평가 지표 활용을 중심으로」. 『한국언어문화학』 21: 183-210.
- 정혜연·서수영. (2022). 「버트(BERT)를 활용한 인간번역의 자동평가 : 여러 모델의 성능 비교 및 활용 가능성」. 『통번역학연구』 26: 117-137.
- 최승혁. (2023a). 「법률 특화 기계번역을 위한 중국 법률 텍스트의 어휘적 특징 분석: 중화인민공화국 민법전을 예로」. 『중어중문학』 92: 349-380.
- 최승혁. (2023b). 「법률 도메인 특화 번역을 위한 중한 기계번역 오류의 귀납적 분석: 중국 민법전의 파과고 번역을 중심으로」. 『번역학연구』 24: 507-537.
- 최승혁. (2024). 법률 도메인 특화 기계번역 모델 연구: 중국 민법전을 실례로. 서강대학교 일반대학원 중국문화학과 박사학위논문.
- 최효은. (2022). 「기계번역의 특허 전문용어 한영번역에서 나타나는 오류 양상 분석—구글 번역 결과물을 중심으로」. 『언어학연구』 27: 217-244.
- Scabini, L., De Baets, B., and Bruno, O. M. (2022). Improving Deep Neural Network Random Initialization Through Neuronal Rewiring. arXiv preprint arXiv:2207.08148.
- Sarcevic, S. (1997). *New Approach to Legal Translation*. Hague: Kluwer Law International.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y. (2020). *BERTScore: Evaluating text generation with BERT*. In *International Conference on Learning Representations (ICLR)*. arXiv:1904.09675.
- Wenbo, Wang. (2023). A review of machine translation quality assessment. *Frontiers in Computing and Intelligent Systems* 5: 108-110.
- Yan, Y., Wang, T., Zhao, C., Huang, S., Chen, J., Wang, M. (2023). BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* 1: 5428-5443.
- 陈炯. (1998). 『法律语言学概论』. 西安: 陕西人民教育出版社.
- 董秀芳. (2003). 「“的”字短语做后置关系小句的用法—兼评法律文献中“的”字短语的用法」. 『语言文字应用』 4: 120-126.
- 李振宇. (2006). 『法律语言学新说』. 北京: 中国检察出版社.

刘红婴. (2003). 『法律语言学』. 北京: 北京大学出版社.  
孙懿华. (2006). 『法律语言学』. 长沙: 湖南人民出版社.  
王洁主编. (1996). 『法律语言学教程』. 北京: 法律出版社.

[인터넷 자료]

머니투데이. (2024). 대형로펌 뛰쳐나온 변호사, 챗GPT보다 똑똑한 ‘AI 번역’ 해냈다  
<https://news.mt.co.kr/mtview.php?no=2024072317155448271>(2025.04.05.검색).  
네이버 중국어 사전. <https://zh.dict.naver.com/#/search?query=%E7%BB%AD%E7%A7%9F>  
(2025.05.05.검색).  
법무부. <https://zh.dict.naver.com/#/search?query=%E7%BB%AD%E7%A7%9F>(2025.03.05.검색).  
한국법제연구원. <https://www.klri.re.kr/kor/business/bizLawDicKeyword.do>(2025.05.05.검색).  
GitHub. [https://github.com/Tiiiger/bert\\_score](https://github.com/Tiiiger/bert_score)(2025.02.20.검색).  
GitHub. <https://github.com/google-research/bleurt>(2025.02.20.검색).  
国家市场监督管理总局. <https://htsfwb.samr.gov.cn/>(2025.03.15.검색).

This paper was received on 9 May 2025; revised on 9 June 2025; and accepted on 16 June 2025.

---

***Authors' email addresses***

choi00joy@hanmail.net

kgb43@sogang.ac.kr

***About the authors***

Sunghyuk Choi (first author) is a Lecturer in the Korean-Chinese Interpretation and Translation Program at the Graduate School of International Studies, Chung-Ang University. His main research interests include machine translation and AI-based translation.

Byeongkwu Kang (corresponding author) is a Professor in the Department of Chinese Culture at Sogang University. He received his Ph.D. from the Institute of Computational Linguistics at Peking University. His main research interests include modern Chinese grammar, linguistic typology, corpus linguistics, and natural language processing.