

한문고전문헌의 기계번역 평가방안 탐색*

정성훈** · 하지영*** · 김우정****

「차 례」

1. 들어가는 말
2. 기계번역 자동평가
3. 기계번역 수동평가
4. 나가는 말

[국문초록]

이 글은 기계번역을 이용한 한문고전 번역물의 품질평가 방법을 살펴보고, 품질평가의 객관성을 제고하는 동시에 번역품질 향상에 기여할 수 있는 방안을 제안한 것이다. 고립어인 한문 고전문헌은 문체가 다양하고 문법상의 변화도 복잡하다. 또한 기계번역은 평가 기준·평가목적·평가비용·텍스트의 종류 등도 함께 고려하여야 하므로 신뢰성이 높고 간편한 번역 품질 평가모형을 개발하기가 쉽지 않다. 자동평가는 기계번역의 어떤 요소가 번역 품질에 영향을 미치는지는 알 수 없으며, 점수가 가장 높은 기계번역 모델을 보여줄 수는 있지만 기계번역 품질에 대한 타당성을 보장하지는 못한다. 그리고 평가기준도 평가모델에 따라 달라질 수 있고 대량의 데이터를 필요로 하는 경우도 있다. 이런 문제점을 보완하기 위해서는 수동평가가 필요한데, 평가자 각각의 경험이나 수준이 존재하고, 평가기준에 대한 이해가 다를 수 있으며, 평가 환경이나 차수에 따른 차이 등 주관에 치우칠 우려도 불식하기 어렵다. 따라서 자동평가와 수동평가의 장단점을 고려하여 기계번역기의 성격과 목적에 맞는 평가방법을 찾아 적용하되, 기계번역 모델의 성능을 객관적으로 평가할 수 있는 척도를 개발하여야 하며, 궁극적으로 이러한 평가방법이 기계번역 모델의 문제점을 찾아 개

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임
(NRF-2018S1A5A2A03036428).

** 제1저자. 목포대학교 국어국문학과 조교수 / kobe99@mokpo.ac.kr

*** 제1저자. 이화여자대학교 한국문화연구원 연구교수 / zawal@sejong.ac.kr

**** 교신저자. 단국대학교 한문교육과 교수 / rtoran@dankook.ac.kr

선해나가는 데 도움이 될 수 있도록 해야 한다.

주제어: 한문, 기계번역, 자동평가, 수동평가, BLEU, METEOR

1. 들어가는 말

이 글은 신경망 기계번역(Neural machine translation, NMT)의 등장 이후 급격하게 관심이 높아진 기계번역의 번역품질 평가방법들을 살펴보고, 한문 고전문언문 기계번역물의 번역품질 평가에 적합한 방법을 제안하는 것을 목적으로 한다.

한문고전 분야에서는 대체로 원의에 충실하면서도 가독성이 높은 번역문을 좋은 번역문으로 간주한다. 그러나 이는 한문에 대한 이해 수준과 현대 국어에 대한 감각을 함께 요구하는 것이므로, 두 가지 요건을 모두 갖추기란 여간 어려운 일이 아니다. 또한 이 두 가지 조건과 아울러 평가기준·평가목적·평가비용·텍스트의 종류 등도 함께 고려하여야 한다. 이처럼 고려해야 할 요인이 다양하기 때문에 문제가 매우 다양하고 문법상의 변화도 복잡한 한문고전문헌에 적합한 번역 품질 평가모델을 개발하는 것이 쉽지 않다.

그럼에도 불구하고 최근 기계번역 기술이 발전하며 한문고전에 대한 번역품질 평가의 신뢰성, 타당성, 간편성 향상에 대한 기대 수준도 함께 높아지고 있다. 번역의 신뢰성이란 언제, 누가, 어떻게 평가하더라도 번역 품질의 평가 결과가 크게 달라지지 않는 번역을 말한다. 평가자의 배경이나 성실성 등에 따라 평가의 결과가 달라진다면 신뢰성이 낮아질 수밖에 없다. 번역평가모델의 알고리즘이 정해질 경우, 일반적으로 전문가에 의한 수동평가보다는 자동평가가 신뢰성이 높은 것으로 알려져 있다. 번역의 타당성이란 번역문이 정말 좋은 번역이라고 할 만한 충분한 근거를 갖춘 번역을 말한다. 즉, 한문고전에 대한 해박한 지식은 물론 한국어와 한국 문화에 대해서도 풍부한 지식을 갖추어야 하며, 실제 번역 경험에서 나온 번역 판단능력과 기준 등이 명확하고, 번역결과가 원의를 충실히 반영하고 있어야 한다. 일반적인 번역평가에서는 자동평가보다는 전문가에 의한 수동평가가 타당성이 높은 것으로 알려져 있다. 마지막으로 번역의 간편성이란 많은 양의 번역

도 쉽게 평가할 수 있는 평가방식을 말한다. 당연히 많은 양의 번역을 빠른 시간 안에 평가할 수 있는 자동평가가 간편성이 높으나, 자동평가를 위해서는 텍스트의 종류나 평가 목적 등에 적합한 소프트웨어나 프로그램 등을 선정하고 활용하는 능력이 필요하고, 이를 학습하는 데에도 시간과 비용이 소요되기 때문에 누구나 간편하게 자동평가를 진행하기는 어려울 수 있다.

본고에서는 이와 같은 점에 유의하여, 기계번역 평가방식을 자동평가와 수동평가 두 부분으로 나누어 살펴본 후 평가 성능 향상을 위해 지향해야 할 평가방식에 관해 제안하고자 한다.

2. 기계번역의 번역품질 자동평가

2.1. 기계번역 번역품질 자동평가 모델

최근 기계번역 분야에서는 번역모델의 성능 향상은 물론이고, 번역의 품질평가에 대한 중요성이 점차 강조되고 있다. 기계번역 모델의 성능이 비약적으로 향상됨에 따라 기계번역이 다양한 분야에서 사용되기 시작하고 있지만, 기계번역은 예상하지 못한 번역 오류를 포함하고 있으며 동일한 기계번역 모델에서도 투입된 데이터나 병렬 코퍼스 등에 따라 다양한 수준의 번역문들이 생성되는 문제도 나타난다. 즉 기계번역의 품질에 대한 정확한 평가와 확인 과정이 필요하다. 기계번역의 품질에 대한 정확한 평가와 확인 과정은 좋은 기계번역 모델을 만들게 하는 기초 작업이며, 이를 통해 기계번역 모델의 피드백을 주어 해당 기계번역 모델을 적절히 수정하도록 하는 나침반 역할도 한다. 이러한 기계번역의 정확한 품질과 확인 과정에는 일반적으로 자동평가가 수반된다. 자동평가는 기계번역 모델의 성능을 객관적으로 평가할 수 있는 척도를 마련해 줄 수 있고, 개발자에게는 기계번역 모델의 문제점이나 개선 방향을 제공할 수 있다. 또한 기계번역 모델 간 차별성을 부여하여, 목적에 따른 평가기준을 제공하기도 한다.

기계번역의 자동평가는 참조번역(reference translation)과 후보번역(candidate translation)을 고려하여 기계번역의 품질을 평가한다. 자동평가는 명료성

(intelligibility), 정확성(accuracy), 유창성(flucy) 등을 기준으로, 기계가 번역한 일부 텍스트의 품질을 평가하는 방법이다. 명료성은 번역된 문장이 원문 텍스트를 잘 나타내고 있는가를 나타내는 것으로 번역의 타당성과 관련된 기준이고, 정확성은 참조번역의 내용이 기계가 번역한 후보번역에 얼마만큼 잘 유지되어 있는가를 나타내는 것으로, 번역의 신뢰성과 연관된다. 유창성은 기계가 번역한 텍스트가 해당 언어로 얼마나 자연스럽게 번역되었는지를 나타내는 지표이다.

자동평가는 평가에 걸리는 시간과 비용을 절약할 수 있고, 계량적인 지표로 나타내어 기계번역의 품질을 파악할 수 있게 한다. 결과적으로 번역품질에 대한 평가결과를 객관적으로 확보할 수 있다는 장점도 있다. 특히 여러 장르에서 번역된 후보번역의 품질을 한눈에 비교할 수 있기 때문에 기계번역 모델의 성능이 범용인지 특정 장르에 효과적인지를 비교하는 데 도움이 되기도 한다.

그러나 자동평가는 현재 여러 제약도 존재한다. 하나는 번역된 후보번역의 품질에만 관심을 두므로써 기계번역에 필요한 다른 중요한 요소들을 간과할 수 있다는 것이다. 기계번역에 영향을 미칠 수 있는 병렬코퍼스나 기계번역 모델의 알고리즘 등에는 상대적으로 관심을 적게 갖게 할 수도 있다. 동일하지 않은 기계번역 모델을 비교할 때에는 효율성 평가(Evaluation of effectiveness)를 통해 그에 맞는 번역 환경 조성을 해 주어야 더 좋은 번역 결과를 얻을 수 있는데, 자동평가는 번역 환경보다는 번역의 품질향상만을 요구한다.¹⁾ 그러므로 자동평가는 기계번역의 품질만을 평가하는 것이지, 기계번역의 어떤 요소들이 기계번역의 품질에 영향을 미치는지는 알 수 없다. 그리고 기계번역 모델에 대한 평가지표를 비교하여 어느 모델이 가장 높은 점수를 받는 기계번역 모델인지는 보여줄 수 있으나, 기계번역 품질에 대한 완전한 타당성을 부여하지는 못한다. 즉 자동평가의 경우에도, 평가기준이 평가모델에 따라 다소 차이가 나기 때문에 적합하고 타당한 평가결과를 얻기 위해서는 상당히 많은 데이터가 필요하며, 한문고전문헌의 해당 장르의 번역 전문가들의 평가참여도 필수적이다. 자동평가는 한문고전문헌의 원문을 평가하는 것이 아니라 기번역된 참조번역과 기계가 번역한 후보번역을 비교

1) 효율성 평가(Evaluation of effectiveness)는 Van Slype(1982)가 제안한 평가방식으로, 기계번역의 단어 당 비용과 후편집(post-edit)에 드는 비용을 계산하여 기계번역 모델을 평가하는 것이다. 효율성 평가는 기계번역 모델을 전체적 관점에서 평가할 수 있다는 장점이 있다.

하는 것이기 때문이다.

2.2. 기계번역 번역품질 자동평가 지표

현재 기계번역의 품질평가에 사용되는 자동평가 지표는 다양하다. 대표적인 자동평가 지표는 BLEU 점수라고 할 수 있다. BLEU 점수는 주로 범용 목적의 기계번역 품질평가에 사용되며, 기계번역을 거친 후보번역과 전문가가 직접 번역한 참조번역 사이의 유사성을 기반으로 한다. 그 외에도 METEOR, ROUGE, LePOR, BLEUmod., WER 등도 많이 사용된다. 이들 자동평가 지표들은 기계번역의 품질을 평가하는 데에 매우 유용하고 여러 분야에서 활용되고 있지만, 각 지표마다 장단점이 있다.²⁾ 이 연구에서는 일반적으로 널리 사용되는 BLEU 점수의 알고리즘과 여러 연구(Banerjee & Lavie 2005; Denkowski & Lavie 2011; Chung 2020)에서 좋은 결과를 보인 METEOR 점수의 알고리즘을 비교해보고, 한문고전문헌의 기계번역 품질평가에 일반적으로 사용되는 BLEU 점수가 적절한지를 검토해 보기로 한다.

2.2.1. BLEU(BiLingual Evaluation Understudy)

BLEU는 Papineni et al.(2002)가 제안한 기계번역에 대한 자동평가 방식으로, 기계번역의 품질을 측정하는 지표 중 하나이다. 번역에 대한 전문가의 평가가 시간과 비용이 많이 드는 단점을 극복하고, 기계번역의 정밀도(precision)를 간편하게 측정하고자 한 모델이다. BLEU는 전문가의 번역과 유사할수록 기계번역의 품질이 좋을 것이라는 점을 전제로 하여, 전문가의 번역에 대한 유사성(closeness)’을 평가할 정밀도(precision)와 ‘전문가가 번역한 높은 수준의 병렬 코퍼스를 필수적으로 요구한다(Papineni et al., 2002:311).

BLEU의 알고리즘은 일치하는 단어의 수(n)를 계산하는 일종의 n-gram 방식인데, 기계가 번역한 후보번역과 번역 전문가가 미리 번역해 놓은 참조번역을

2) 예를 들어, BLEU 점수는 ‘번역투 효과’에 의해 그 평가결과가 왜곡될 수 있다. ‘번역투 효과’는 BLEU 점수가 매우 높게 나오는 기계번역이나, 어순 등의 이유로 번역된 문장이 매우 어색해지는 현상을 말한다.

1-gram, 2-gram, 3-gram, 4-gram 순으로 비교하여 정밀도를 계산한다.³⁾ 즉 후보번역과 참조번역 간 일치하는 n-gram의 수가 많을수록 기계번역의 품질이 좋은 것으로 평가한다. BLEU 자동평가에서 이론적으로는 5-gram, 6-gram 등도 가능하지만 실제로는 1-gram~4-gram까지의 결과만을 기하평균으로 계산하여 정밀도를 측정하는 경우가 일반적이다. n-gram이 적을수록 참조번역을 얼마나 적절히 옮겼는지에 대한 기계번역의 충분성(adequacy)을 판정할 수 있게 되고 n-gram이 높을수록 기계번역의 자연스러움(fluency)을 평가할 수 있게 된다.

BLEU 점수(BLEU score)의 수식은 다음과 같다.

(1) BLEU 점수

$$BLEU = brevity-penalty \times \exp\left(\sum_{n=1}^N \frac{1}{N} \log P_n\right)$$

기본적으로 정밀도(P_n)와 간결도-벌점(BP)의 곱으로 계산한다. 즉 1-gram, 2-gram, ..., (n-1)-gram, n-gram에 대한 정밀도를 기하평균한 후 가중치에 해당하는 간결도-벌점을 지수승으로 곱한 것이다.

BLEU 점수의 정밀도(P_n)는 다음과 같다.

(2) BLEU 점수의 정밀도(Precision, P_n)

$$P_n = \frac{\text{참조번역과 후보번역이 일치하는 } n\text{-gram}}{\text{후보번역의 총 } n\text{-gram}}$$

일반적으로 정밀도(P_n)는 참조번역 n-gram과 후보번역 n-gram이 일치하는 개수를 후보번역의 총 n-gram 수로 나누어 계산한다. 한편 중복되는 n-gram의 일치를 제거(clipping)하여 수정된 정밀도(P_n)를 측정하기도 하는데, 이 경우 참조번역 n-gram과 후보번역 n-gram이 일치하는 최대 n-gram 수를 후보번역의 총 n-gram 수로 나누어 계산한다.

BLEU 점수의 간결도-벌점(brevity-penalty, BP)은 다음과 같다.

3) n-gram은 번역 텍스트를 n개로 나누는 것을 말한다.

(3) BLEU 점수의 간결도-벌점(BP)

$$\text{brevity-penalty} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-\frac{r}{c})}, & \text{if } c \leq r \end{cases}$$

간결도-벌점(BP)은 참조번역과 후보번역의 번역 길이(n-gram) 간의 차이에 의해 생기는 감점을 말한다. 기계번역에서는 원문에 대해서 누락이 자주 발생하는데, 참조번역의 n-gram과의 비교에서 그 길이가 짧으면 감점을 부여하는 것이다. 즉, 참조번역과 후보번역의 n-gram 길이가 같으면 1, 후보번역이 참조번역보다 짧으면 1보다 작은 수를 가중함으로써 지나치게 짧은 후보번역은 그 점수를 감하는 방식이다.

다음은 BLEU 점수의 알고리즘에 대해 살펴보자. 위에서 기술한 바와 같이, BLEU 점수는 참조번역과 후보번역 사이에 일치하는 n-gram의 수에 대한 비율을 기하평균에 따라 점수를 부여한다. 이때, BP는 단어의 누락 등으로 인해 후보번역이 참조번역보다 짧아질 경우, 기계번역의 품질에 비해서 BLEU 점수가 더 높게 나오는 것을 방지하는 역할을 한다. 보통 BLEU 점수는 결과값에 100을 곱하여 0-100 사이의 척도(scale)로 표현하는데, 범용 목적의 기계번역에서는 30점 이상이면 기계번역의 품질이 좋다고 평가한다.

여기서는 간단한 예제를 통해 BLEU 점수의 알고리즘을 살펴보고 BLEU 점수를 계산해 본다. 조선 중기의 문신 崔嵬의 문집인 『簡易集』에 수록된 「李知事仲薰年兄挽章」의 일부인 ‘公少我二年.’라는 문장으로, 한국어 참조번역과 한국어 후보번역들(candidates)을 만들어 보았다. 후보번역1은 참조번역과 의미는 유사하나, ‘어리다’가 유의어 ‘적다’로 번역이 된 경우, 후보번역2는 ‘어리다’가 반의어 ‘많다’로 잘못 번역된 경우, 후보번역3은 참조번역과 어순을 달리한 경우, 후보번역4는 ‘나보다’가 누락된 경우이다. 이들을 대상으로 1-gram, 2-gram, 3-gram, 4-gram을 측정하고 간단한 BLEU 점수를 계산해 보자.

〈표 1〉 참조번역과 후보번역 간 n-gram(BLEU 점수)

index	0	1	2	3	4	5
한문 원문	公	少	我	二	年	.
참조번역	공은	나보다	두	살	어리다	.
후보번역1	공은	나보다	두	살	적다	.
후보번역2	공은	나보다	두	살	많다	.
후보번역3	나보다	공이	두	살	어리다	.
후보번역4	공은	두	살	어리다	.	.

우선 〈표 1〉처럼 n-gram(1~4)을 통해 참조번역과 후보번역 간에 n-gram이 얼마나 일치하는지 정밀도를 측정한다.

〈표 2〉 참조번역과 후보번역 간 1-gram(BLEU 점수)

	참조번역	후보번역1	일치	후보번역2	일치	후보번역3	일치	후보번역4	일치
	1-gram	공은	공은	1	공은	1	나보다	1	공은
나보다		나보다	1	나보다	1	공은	1	두	1
두		두	1	두	1	두	1	살	1
살		살	1	살	1	살	1	어리다	1
어리다		적다	0	많다	0	어리다	1	.	1
.		.	1	.	1	.	1		
합계			5/6		5/6		6/6		5/5

〈표 2〉는 1-gram의 정밀도를 측정한 것이다. 후보번역1과 후보번역2의 1-gram 정밀도는 5/6=0.833이었고, 후보번역3의 1-gram 정밀도는 6/6=1, 후보번역4의 1-gram 정밀도도 5/5=1이었다.

〈표 3〉 참조번역과 후보번역 간 2-gram(BLEU 점수)

	참조번역	후보번역1	일치	후보번역2	일치	후보번역3	일치	후보번역4	일치
2-gram	BOS 공은	BOS 공은	1	BOS 공은	1	BOS 나보다	0	BOS 공은	1
	공은 나보다	공은 나보다	1	공은 나보다	1	나보다 공은	0	공은 두	0
	나보다 두	나보다 두	1	나보다 두	1	공은 두	0	두 살	1
	두 살	두 살	1	두 살	1	두 살	1	살 어리다	1
	살 어리다	살 적다	0	살 많다	0	살 어리다	1	어리다 .	1
	어리다 .	적다 .	0	많다 .	0	어리다 .	1	. EOS	1
	. EOS	. EOS	1	. EOS	1	. EOS	1		
합계			5/7		5/7		4/7		5/6

〈표 3〉은 2-gram의 정밀도를 측정한 것이다. 후보번역1과 후보번역2의 2-gram 정밀도는 $5/7=0.714$, 후보번역3의 2-gram 정밀도는 $4/7=0.571$, 후보번역4의 2-gram 정밀도는 $5/6=0.833$ 이었다.

〈표 4〉 참조번역과 후보번역 간 3-gram(BLEU 점수)

	참조번역	후보번역1	일치	후보번역2	일치	후보번역3	일치	후보번역4	일치
3-gram	BOS 공은 나보다	BOS 공은 나보다	1	BOS 공은 나보다	1	BOS 나보다 공은	1	BOS 공은 두	0
	공은 나보다 두	공은 나보다 두	1	공은 나보다 두	1	나보다 공은 두	1	공은 두 살	0
	나보다 두 살	나보다 두 살	1	나보다 두 살	1	공은 두 살	1	두 살 어리다	1
	두 살 어리다	두 살 적다	0	두 살 많다	0	두 살 어리다	0	살 어리다 .	1
	살 어리다 .	살 적다 .	0	살 많다 .	0	살 어리다 .	0	어리다 . EOS	1
	어리다 . EOS	적다 . EOS	0	많다 . EOS	0	어리다 . EOS	0		
합계			3/6		3/6		3/6		3/5

〈표 4〉는 3-gram의 정밀도를 측정한 것이다. 후보번역1, 후보번역2, 후보번역3의 3-gram 정밀도는 $3/6=0.500$ 였고, 후보번역4의 3-gram 정밀도는 $3/5=0.600$ 으로 측정되었다.

〈표 5〉 참조번역과 후보번역 간 4-gram(BLEU 점수)

	참조번역	후보번역1	일치	후보번역2	일치	후보번역3	일치	후보번역4	일치
4-gram	BOS 공은 나보다 두	BOS 공은 나보다 두	1	BOS 공은 나보다 두	1	BOS 나보다 공은 두	0	BOS 공은 두 살	0
	공은 나보다 두 살	공은 나보다 두 살	1	공은 나보다 두 살	1	나보다 공은 두 살	0	공은 두 살 어리다	0
	나보다 두 살 어리다	나보다 두 살 적다	0	나보다 두 살 많다	0	공은 두 살 어리다	0	두 살 어리다	1
	두 살 어리다	두 살 적다	0	두 살 많다	0	두 살 어리다	1	살 어리다 . EOS	1
	살 어리다 . EOS	살 적다 . EOS	0	살 많다 . EOS	0	살 어리다 . EOS	1		
합계			2/5		2/5		2/5		2/4

〈표 5〉는 4-gram의 정밀도를 측정한 것이다. 후보번역1, 후보번역2, 후보번역3의 4-gram 정밀도는 2/5=0.400이었고, 후보번역4의 4-gram 정밀도는 2/4=0.500으로 측정되었다.

참조번역과 후보번역 간 1-gram~4-gram의 일치여부를 측정할 값을 이용하여 기하평균을 계산하고 후보번역에 대한 전체 n-gram의 정밀도를 계산해보자. 후보번역1과 후보번역2의 정밀도는 $(\frac{5}{6} \times \frac{5}{7} \times \frac{3}{6} \times \frac{2}{5})^{\frac{1}{4}}$ 이고, 후보번역3의 정밀도는 $(\frac{6}{6} \times \frac{4}{7} \times \frac{3}{6} \times \frac{2}{5})^{\frac{1}{4}}$ 이고, 후보번역4의 정밀도는 $(\frac{5}{5} \times \frac{5}{6} \times \frac{3}{5} \times \frac{2}{4})^{\frac{1}{4}}$ 로 계산된다.

각 후보번역의 간결도-벌점은 간단히 $\min(1, \frac{\text{후보번역 } n\text{-gram의 길이}}{\text{참조번역 } n\text{-gram의 길이}})$ 로 계산할 수 있다. 후보번역1, 후보번역2, 후보번역3의 간결도-벌점은 모두 $\min(1, \frac{6}{6})$ 이므로, 후보번역1의 BLEU 점수는 $1 \times (\frac{5}{6} \times \frac{5}{7} \times \frac{3}{6} \times \frac{2}{5})^{\frac{1}{4}}$, 후보번역2의 BLEU 점수는 $1 \times (\frac{5}{6} \times \frac{5}{7} \times \frac{3}{6} \times \frac{2}{5})^{\frac{1}{4}}$, 후보번역3의 BLEU 점수는 $1 \times (\frac{6}{6} \times \frac{4}{7} \times \frac{3}{6} \times \frac{2}{5})^{\frac{1}{4}}$ 이다. 후보번역4의 간결도-벌점은 $\min(1, \frac{5}{6})$ 이므로 후보번역4의 BLEU 점수는 $\frac{5}{6} \times (\frac{5}{5} \times \frac{5}{6} \times \frac{3}{5} \times \frac{2}{4})^{\frac{1}{4}}$ 가 된다. 이를 정리하면 〈표 6〉과 같다.

〈표 6〉 후보번역의 BLEU 점수

	BP	n-gram의 정밀도	BLEU 점수
후보번역1	1	0.587	0.587
후보번역2	1	0.587	0.587
후보번역3	1	0.581	0.581
후보번역4	0.833	0.707	0.589

기계번역의 정밀도에서는 후보번역4가 가장 높은 평가를 받았고, 간절도-별점이 부여되어도 전체 BLEU 점수는 후보번역4 > 후보번역1 = 후보번역2 > 후보번역3 순으로 높게 나왔다. 후보번역4는 참조번역의 일부가 누락되었지만, 나머지는 참조번역과 일치되는 n-gram이 많은 번역이었다는 점을 감안하면 BLEU 점수는 상대적으로 짧은 번역일 때 점수가 높게 나오는 것을 확인할 수 있었다. 또한 후보번역1과 후보번역2의 BLEU 점수가 같았는데, BLEU 점수는 참조번역의 유의어나 반의어를 유의미하게 구별하지 못하는 것으로 파악된다. 마지막으로 후보번역3의 BLEU 점수가 가장 낮게 나온 것으로 보아, BLEU 점수는 통사적인 측면을 고려하지 않는 기계번역 평가방식으로 보인다.

2.2.2. METEOR 점수(METEOR score)

METEOR(Metric for Evaluation of Translation with Explicit Ordering) 또한 기계번역의 성능을 측정하는 지표 중 하나로서, Banerjee, S. & Lavie, A.(2005)가 제안하였다. METEOR 점수는 1-gram의 정밀도(precision)와 재현율(recall)의 조화평균을 기반으로 측정된다. BLEU 점수에 비해서 재현율에 많은 가중치를 부여한다. METEOR 점수는 BLEU 점수에서 발견된 일부 문제를 해결하고, 단어나 구 수준에서 수동평가와 정(+)의 상관관계를 형성하도록 설계되었다.⁴⁾

BLEU 점수가 참조번역과 후보번역의 정밀도에 대한 지표라면, METEOR 점

4) BLEU 점수는 코퍼스 수준에서 상관관계를 찾는다는 점에서 차이가 있다. Banerjee, S. & Lavie, A.(2005)에 따르면, 동일한 코퍼스 수준에서 수동평가와의 상관관계에서 BLEU의 상관계수 0.817과 비교하여 METEOR은 0.964의 상관계수를 보였다. 문장 수준에서는 수동평가와 최대 상관계수는 0.403이었다.

수는 정밀도뿐만 아니라 재현도의 평균치로 보여주는 지표이다. 정밀도가 후보번역을 기준으로 일치도를 계산한 값으로서, 후보번역의 어느 정도가 참조번역에 가까운지를 보는 유사도 개념이라면, 재현도는 참조번역을 기준으로 일치도를 계산한 값이므로, 참조번역 중에서 기계번역된 후보번역에 단어가 얼마나 나타나는지를 보는 개념이다(Banerjee, S. & Lavie, A., 2005).

METEOR 점수의 평가방식은 1-gram 일치를 판단하는 조화평균과 2-gram 이상의 일치를 보여주는 벌점(penalty)으로 구성된다. BLEU 점수와 마찬가지로, METEOR 점수에서도 1-gram 일치는 주로 '의미'에 맞는 단어를 선택했는지를 평가하는 도구이며, '누락, 첨가'도 작게는 단어 단위로 이루어진다는 점에서 1-gram의 일치로 정밀도를 평가할 수 있다. 또 2-gram 이상의 일치를 보여주는 벌점은 2개 이상의 단어로 이루어진 구나 절을 '문법'과 '논리'에 맞게 번역했는지를 판단한다는 점에서 문법, 논리성, 가독성 등을 평가하는 장치라고 할 수 있다. 이렇듯 METEOR 점수의 평가방식은 수동평가의 방식과 일부 일맥상통하는 점이 있는데, METEOR 점수의 평가방식이 인간의 번역품질을 평가할 수 있는 타당성을 어느 정도 가지고 있음을 알 수 있다.

METEOR 점수(METEOR score)는 다음과 같이, F_{mean} 과 1에서 벌점(penalty)을 감한 수와의 곱으로 표현된다.

(4) METEOR 점수

$$METEOR = F_{mean} \times (1 - penalty)$$

BLEU 점수와 마찬가지로, 기계번역의 평가단위는 n-gram이다. 우선 참조번역과 후보번역 간 1-gram을 사상(mapping)한다. 사상(mapping)은 참조번역과 후보번역의 1-gram을 잇는 하나의 선으로 생각해 볼 수 있는데, 후보번역의 모든 1-gram은 참조번역의 1-gram에 0 또는 1로 사상되어야 한다. 또한 동일한 수의 사상을 가진 경우, 두 사상의 교차가 가장 적은 정렬(alignment)을 선택해야 한다. 최종 1-gram의 정밀도(precision)는 (5)와 같이 계산한다.

(5) METEOR 점수의 정밀도(P)

$$P = \frac{\text{참조번역과 후보번역 사이에 일치하는 1-gram의 수}}{\text{후보번역의 1-gram의 수}}$$

METEOR 점수의 재현도(recall)는 참조번역에 속한 n-gram이 후보번역에 얼마나 사용되었는지로 계산한다.⁵⁾ 즉 참조번역을 기준으로 후보번역과의 1-gram 일치도를 계산한 것이다.

(6) METEOR 점수의 재현도(R)

$$R = \frac{\text{참조번역과 후보번역 사이에 일치하는 1-gram 수}}{\text{참조번역의 1-gram 수}}$$

METEOR 점수의 P와 R은 모두 참조번역과 후보번역 사이에 일치하는 1-gram의 비율이므로, 두 값의 평균을 구할 때는 비율의 평균인 조화평균(harmonic mean)을 사용한다.

(7) 조화평균

$$F = \frac{2 \times PR}{R + P}$$

그런데 Banerjee, S. & Lavie, A.(2005)는 P와 R 중에서 참조번역을 기준으로 하는 R의 중요성을 주장하고 P에 가중치 9, R에 가중치 1을 부여하였다.

(8) METEOR 점수의 조화평균(Fmean)

$$F_{mean} = \frac{10 \times PR}{R + 9P}$$

그런데 METEOR 점수의 P와 R, Fmean은 참조번역과 후보번역 간 1-gram에 대한 일치 여부만을 설명하고 있을 뿐, 2-gram 이상에 대해서는 아무런 설명하고 있지 않는데, METEOR 점수에서는 2-gram 이상의 일치 여부를 별점으로 측정한다. 참조번역과 후보번역 사이에 일치하는 1-gram의 수가 많아지면 별점이

5) 이는 METEOR 점수만의 특징으로, BLEU 점수의 평가 방식은 이를 반영하지 못한다.

작아지고, 참조번역과 후보번역 사이에 일치하는 2-gram 이상의 수가 많아지면 벌점은 커진다. 즉 참조번역과 후보번역에서 인접하지 않은 사상(mapping)이 많을수록 벌점이 커진다. 따라서 참조번역과 후보번역이 완전히 일치하지 않는 이상, 항상 벌점이 존재한다.

(9) METEOR의 벌점

$$penalty = 0.5 \times \left(\frac{\text{chunks의 수}}{\text{일치하는 1-gram의 수}} \right)^3$$

METEOR 점수 알고리즘은 참조번역과 후보번역 사이의 정밀도(P)와 재현도(R)의 조화평균의 가중치에 따라 점수를 부여하는 방식이다. METEOR 점수는 번역의 정밀도 향상을 위해 동의어나 paraphrase 등을 삽입해 다양한 n-gram을 적절한 번역으로 처리하고 있다. 여기서도 BLEU 점수에서 사용한 동일한 예제를 통해 METEOR 점수의 알고리즘을 살펴보고 METEOR 점수를 계산해보자.

다음은 앞서 살펴보았던 중국어 ‘我非常喜欢唱歌.’, 한국어 참조번역과 후보번역들이다. 이들을 대상으로 1-gram을 측정하고 METEOR 점수를 계산해보자.

〈표 7〉 참조번역과 후보번역 간 n-gram(METEOR 점수)

index	0	1	2	3	4	5
한문 원문	公	少	我	二	年	.
참조번역	공은	나보다	두	살	어리다	.
후보번역1	공은	나보다	두	살	적다	.
후보번역2	공은	나보다	두	살	많다	.
후보번역3	나보다	공이	두	살	어리다	.
후보번역4	공은	두	살	어리다	.	.

우선 후보번역을 기준으로 1-gram의 일치여부를 측정하고, 정밀도를 계산한다.

〈표 8〉 참조번역과 후보번역 간 1-gram(정밀도)

	참조번역	후보번역1	일치	후보번역2	일치	후보번역3	일치	후보번역3	일치
	1-gram	공은	공은	1	공은	1	나보다	1	공은
나보다		나보다	1	나보다	1	공은	1	두	1
두		두	1	두	1	두	1	살	1
살		살	1	살	1	살	1	어리다	1
어리다		적다	0	많다	0	어리다	1	.	1
.		.	1	.	1	.	1		
합계			5/6		5/6		6/6		5/5

다음으로 참조번역을 기준으로 하여 1-gram의 일치여부를 측정하고, 재현도를 계산한다.

〈표 9〉 참조번역과 후보번역 간 1-gram(재현도)

	참조번역	후보번역1	일치	후보번역2	일치	후보번역3	일치	후보번역3	일치
	1-gram	공은	공은	1	공은	1	나보다	1	공은
나보다		나보다	1	나보다	1	공은	1	두	1
두		두	1	두	1	두	1	살	1
살		살	1	살	1	살	1	어리다	1
어리다		적다	0	많다	0	어리다	1	.	1
.		.	1	.	1	.	1		
합계			5/6		5/6		6/6		5/6

Fmean은 정밀도(P)와 재현도(R)의 조화평균이다.

$$\text{후보번역1과 후보번역2의 Fmean은 } \frac{10 \times PR}{R + 9P} = \frac{10 \times \frac{5}{6} \times \frac{5}{6}}{\frac{5}{6} + 9 \times \frac{5}{6}} \doteq 0.833 \text{이었고,}$$

$$\text{후보번역3의 Fmean은 } \frac{10 \times \frac{6}{6} \times \frac{6}{6}}{\frac{6}{6} + 9 \times \frac{6}{6}} = 1, \text{ 후보번역4의 Fmean은 } \frac{10 \times \frac{5}{5} \times \frac{5}{6}}{\frac{5}{6} + 9 \times \frac{5}{6}} \doteq$$

0.847이었다.

별점은 참조번역과 후보번역 사이에 일치하는 1-gram과 2-gram 이상 chunk 비율을 말하는데, 후보번역1과 후보번역2의 penalty는

$0.5 \times \left(\frac{\text{chunks의 수}}{\text{일치하는 1-gram의 수}} \right)^3 = 0.5 \times \left(\frac{1}{5} \right)^3 \doteq 0.292$ (공은 나보다 두 살)였고, 후보번역3의 penalty는 $0.5 \times \left(\frac{1}{6} \right)^3 \doteq 0.275$ (두 살 어리다.), 후보번역4의 penalty는 $0.5 \times \left(\frac{1}{5} \right)^3 \doteq 0.292$ 였다(두 살 어리다.).

METEOR 점수는 1에서 벌점을 뺀 값과의 조화평균을 곱해서 계산한다. 후보번역1과 후보번역2의 METEOR 점수는

$Fmean \times (1 - penalty) = 0.833 * (1 - 0.292) \doteq 0.590$ 였고, 후보번역3의 METEOR 점수는 $1 * (1 - 0.275) \doteq 0.725$, 후보번역4의 METEOR 점수는 $0.847 * (1 - 0.292) \doteq 0.600$ 이었다.

이를 정리해 보면 다음과 같다.

〈표 10〉 후보번역의 METEOR 점수

	<i>P</i>	<i>R</i>	<i>Fmean</i>	<i>penalty</i>	METEOR 점수
후보번역1	0.833	0.833	0.833	0.292	0.590
후보번역2	0.833	0.833	0.833	0.292	0.590
후보번역3	1	1	1	0.275	0.725
후보번역4	1	0.833	0.847	0.292	0.600

정밀도(P)에서는 후보번역3과 후보번역4가 가장 높은 평가를 받았고, 재현도(R)에서는 후보번역3이 가장 높은 평가를 받았다. 벌점에서는 후보번역3이 가장 낮게 감점이 되어서 전체 METEOR 점수에서는 후보번역3 > 후보번역1 = 후보번역2 > 후보번역4 순으로 기계번역의 품질이 높게 평가되었다. 이는 BLEU 점수와 비교해 보면 정확하게 역순인 것을 알 수 있다. 물론 절대점수를 고려하면 4개의 후보번역 모두 높은 수준의 기계번역 품질로 평가될 수 있으나, 상대적인 순위로 평가했을 때는 BLEU 점수와 METEOR 점수가 일치하지 않는 모습을 보였다.

2.2.3. BLEU 점수와 METEOR 점수의 한계

우리가 위에서 간략히 살펴본 BLEU 점수와 METEOR 점수는 모두 기계번역의 자동평가 방식으로 기계번역 모델을 평가할 때 자주 사용되는 지표이나, 그

한계도 분명히 존재한다.

BLUE 평가방식은 후보번역 속의 단어나 구(n-gram) 중에서 참조번역에도 사용된 단어나 구(n-gram)가 몇 개인지를 측정하여 후보번역의 n-gram이 얼마나 유사한지를 평가하는 방식이다. BLEU 점수는 짧은 번역에 별점을 부여하는 장치가 있음에도 불구하고 번역문이 짧을수록 상대적으로 높은 점수를 받을 수 있는 것으로 알려져 있다(위의 후보번역4 참조). 또한 BLEU 점수는 대체로 하나의 참조번역만을 반영하는데, 유의어나 반의어(단어 관계), 어순(구의 배열순서)에 따른 적절하고 다양한 참조번역을 반영하지 못하는 결과를 초래할 수도 있다. 극단적인 경우에는, 후보번역1과 2처럼, 원문의 의미와 반대인 후보번역에도 높은 BLEU 점수를 부여하기도 하는 문제점이 존재한다. 최근 위와 같은 BLEU 점수의 한계를 극복하기 위해서 참조번역을 여러 개 사용하여 n-gram의 다양성과 통사구조를 반영하거나 반복적으로 번역되는 동일한 통사구조를 보정하여 n-gram의 정밀도를 수정하는 방식을 사용하기도 한다(Callison-Burch, et al., 2006). 그러나 이 경우에도 참조번역이 많을수록 BLEU 점수가 높아지기는 경향이 존재하고, n-gram의 수를 짧게 하지 않는다고 해서 반드시 문법적으로 적절한 번역을 생성한다는 보장도 없기 때문에 BLEU 점수가 높아졌다고 해서 기계번역의 품질이 꼭 향상되었다고 할 수 없다. 실제 연구에서도 수동평가에서 1등을 한 번역이 BLEU 점수로 6등을 한 경우도 보고되었다(Callison-Burch et al., 2006;2007).

METEOR 평가방식은 조화평균에서 정밀도와 재현도에 부여되는 가중치에 대한 타당성 문제가 존재한다. METEOR 점수의 조화평균은 단순한 조화평균이 아니라 정밀도와 재현도의 비율이 9:1이다. 즉 재현도를 고려하기는 하나, 여전히 정밀도에 높은 가중치를 부여하는 방식이다. 정밀도에 비해 재현도가 기계번역 평가에서 좀 더 의미 있는 지표로 알려져 있어서 재현도를 반영해야 한다는 주장이 설득력이 있으나, 왜 정밀도와 재현도의 비율이 왜 9:1인지에 대한 의문은 존재한다. 최근에는 정밀도나 재현도 모두 각각의 장단점이 있어서 어느 쪽이 기계번역의 품질평가에 더 의미 있는 지표인지 단정하기 어렵다는 연구도 있다. METEOR 점수를 처음으로 제안한 학자 중 하나인 Lavie(2005)도 무조건 재현도에 많은 가중치를 두는 것이 아니라 수동평가와 상관관계를 최대화할 수 있는 방향으로 정밀도와 재현도의 비율을 조정해야 한다는 입장으로 바꾸었다

(Denkowski & Lavie, 2011:87).

최근 기계번역 모델의 성능 향상으로 기계번역의 품질이 크게 향상됐음은 틀림없으나, 일부의 자동평가 모델을 기준으로 기계번역의 품질이 향상되었다고 주장하거나 일반화하는 것이 쉽지 않음을 살펴보았다. 다시 말하면 BLEU 점수 또는 METEOR 점수만으로는 해당 기계번역의 품질이 ‘좋다’ 또는 ‘나쁘다’를 담보할 수 없으며, 한문고전문헌이라고 하더라도 다양한 장르의 텍스트가 존재하기 때문에 일률적인 자동평가 방법을 적용하기도 어려울 것이다. 우리는 한문고전문헌의 기계번역 품질을 평가할 만한 타당한 자동평가 방법을 고안해 내어야 하며, 이를 위해서는 여러 기계번역 자동평가 모델을 고려하고 필요하다면 수동평가의 도움을 받아 기계번역의 성능을 향상시킬 방법을 모색해야 한다.

3. 기계 번역 번역 품질 수동 평가

3.1. 일반적인 기계번역 번역품질 수동평가 방식

이상 살펴본 자동평가는 사람에 의해 이루어지는 수동평가에 비해 시간과 비용이 적게 들고 객관적이라는 장점이 있는 반면, 어휘의 정확성을 평가하는 데에 그칠 뿐 번역 결과의 가독성을 평가하기 어려우며, 오류 분석을 위한 충분한 근거를 제시하기에는 무리가 있다는 한계를 가진다.

수동평가는 시간과 비용이 많이 들고 주관성을 배제할 수 없다는 단점이 있다. 그러나 기계번역의 타당성을 검증하고 시스템의 성능을 개선하는 데 있어서 필요한 피드백을 제공할 수 있는 유일한 수단이기도 하다. 그러므로 신뢰할 수 있는 평가를 위해서는 자동평가 방법과 함께 수동평가가 요청된다(최효은 외, 2017).

기계번역 수동평가 연구는 1966년으로 거슬러 올라간다. Carroll(1966)은 ‘명료성(intelligibility)’과 ‘충실성(fidelity)’이라는 수동평가 기준을 창안하였다. 명료성은 번역 결과가 정상적으로 잘 편집된 문장인지를 평가하는 것이며, 충실성은 원문의 의미를 충실하게 전달했는지 여부를 평가하는 것이다. Hutchins & Somers(1992)는 원문과 동일한 정보의 전달 여부를 살피는 ‘충실성 또는 정확성

(Fidelity or accuracy)', 번역된 내용의 이해 용이성을 살피는 '이해가능성 또는 명확성(Intelligibility or clarity)'과 함께 내용과 의도에 적합한 언어 사용 여부인 '문체(Style)'를 기준으로 제시하였다.

2005년 “The Linguistics Data Consortium”(LDC)에서는 '적절성(adequacy)'과 '유창성(flucency)'을 기준으로 5점 척도로 평가하는 방식을 개발하였다. 이후 이 평가 방식은 수동평가 기준으로 상용되었다. 세부기준은 다음 표와 같다.

〈표 11〉 수동 평가 기준 (The Linguistics Data Consortium, 2005)

fluency: How fluent is the candidate sentence?	5 (flawless)
	4 (good)
	3 (non-native)
	2 (disfluent)
	1 (incomprehensible)
adequacy: How much of the information is present?	5 (all)
	4 (most)
	3 (much)
	2 (little)
	1 (none)

Hampshire, S., & Porta Salvia, C.(2010)가 2010년 제시한 '명료성(clarity)'과 '충실성(Fidelity)' 역시 앞서의 이분화된 기준을 계승한 것이다. 전자는 독자가 번역 결과물을 쉽게 이해하는 정도를, 후자는 원본과 동일한 정보를 포함하는 여부를 평가한다. 전자는 5점 척도로, 후자는 -5점 척도로 평가하였다.

상기 연구들은 평가 항목의 명칭과 평가 방식이 다소 다르기는 하지만 대체로 번역 결과물의 정확성과 가독성을 분리하여 평가한다는 점에서 공통점을 지닌다. 즉, 원문이 의도한 의미를 정확하고도 자연스럽게 전달했는지를 평가 기준으로 삼고 있다. 그럼에도 불구하고 이러한 수동평가 방식에서 여전히 문제가 되는 것은, 별다른 기준이 없이 평가자의 주관에 의해 평가가 이뤄질 수 있다는 점이다.

이에 오류를 정량화하는 평가방식이 제안되기도 하였다(Rossi & Wiggins, 2013)는 수동평가 틀을 개발하여, 평가자가 번역 결과에서 잘못된 부분을 체크하면 자동으로 오류를 계산하도록 하였다. 최근 국내 연구자 이준호도 오류를 단순 오류, 오역, 추가, 누락, 문법, 맞춤법, 구두법으로 세부화하여 각각의 영역을 평가

하는 정량화 방식을 제안하였다. 이와 함께 명확성, 유창성, 실무 적용 가능성 등 다른 척도도 함께 제시, 평가를 병행하였다(이준호, 2019).

이외에도 ‘테스크 기반(task-oriented)’, ‘안전성(suitability)’, ‘수정률(HTER: human translation error rate)’이 수동평가 기준으로 제안된 바 있다. 테스크 기반은 업무 활용 가능 여부와 관련하여 번역 결과를, 적절성은 시스템의 안정성을 평가하는 방식이다. HTER은 포스트 에디팅 단계에서 에디팅 횟수를 측정하여 번역의 품질을 측정하는 방안이다(Han, 2016).

Bazrafshan은 Fluency, Fidelity 기준을 유지하되 평가자에게 번역을 기반으로 한 작업을 수행하는 과정을 통해 평가하는 방식을 제안하였다. 예컨대 평가자에게 대상 번역에서 얻은 정보만 사용하여 원본 문장에 대한 몇 가지 질문에 대답하도록 요청하거나, 일부 단어를 생략하고 평가자에게 해당 단어가 무엇인지 결정하도록 요청하는 방식으로 기계 번역의 유창성과 적절성을 평가하는 방식이다(Bazrafshan, 2014). HTER은 참조 번역과 일치하도록 사람이 수행해야 하는 편집의 양을 측정한다. 삽입, 삭제, 대체 등의 빈도를 측정하여 기계번역기의 성능을 평가한다. 비용이 많이 든다는 단점은 있지만 기존의 유창성, 적절성 두 기준의 평가 방식보다는 세밀하고 객관적인 방법으로 기계 번역기 개선에 기여할 수 있다는 장점을 지닌다(Snover et al., 2006).

3.2. 『승정원일기』 자동번역시스템 수동평가 현황

본고에서는 현재 국내에서 개발된 한문고전문헌 기계번역기인, 『승정원일기』 자동번역시스템을 중심으로 기계번역의 수동 평가 현황을 살펴보고 그 개선 방향을 모색해보았다. 한국고전번역원에서는 2017년 3월부터 2019년 12월까지 3년 동안 진행된 인공지능(클라우드) 기반 고전문헌자동번역시스템 개발을 진행하였으며 2021년 1월 승정원일기 자동번역기를 한국고전번역원 홈페이지에 공개하였다. 승정원일기 자동번역기는 고전 한문을 대상으로 국내에서 최초로 개발된 자동번역시스템이다. 현재 전문 역자와 대중들에게 꽤 많은 평가를 받고 있지만 향후 자동번역기의 성능을 고도화하기 위해서는 평가 시스템의 개선이 필요하다.

한국고전번역원은 그동안 최적화된 번역 모델 선정과 자동번역 모델의 품질

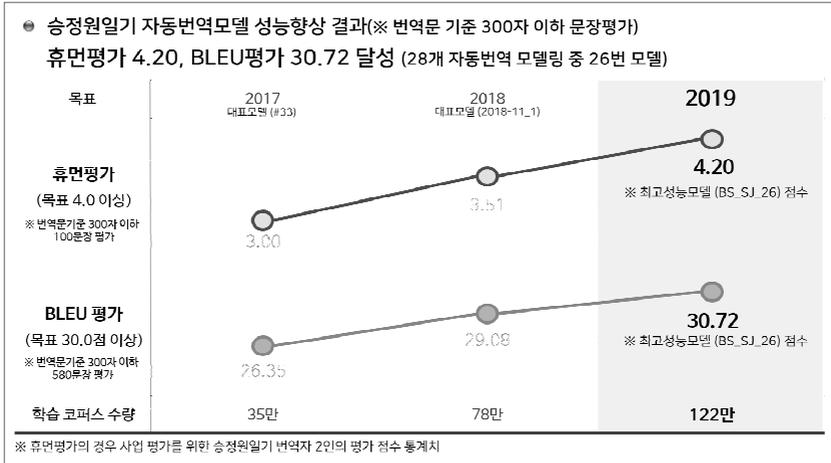
관리 및 성능 향상을 평가하기 위해 전문역자들로 구성된 수동평가를 수행하였다. 2017년 1년차 자동번역 시스템 구축 과정에서는 『승정원일기』 자동번역 모델을 구축하고, 그 학습결과를 대상으로 수동평가를 진행하였으며, 2년차 시스템 고도화 과정에서는 『승정원일기』 자동번역 모델의 고도화 결과와 조선왕조실록 자동번역 모델에 대한 수동평가를 1회 진행하였다. 마지막 3년차 시스템 고도화 과정에서도 『승정원일기』 자동번역모델 고도화 결과와 특수고전(천문분야) 번역 모델 성능에 대한 수동평가를 각각 2회씩 진행하였다.

〈표 12〉 한국고전번역원 인공지능 자동번역모델 수동평가 진행과정 (2017~2019)

	1년차(2017)	2년차(2018)	3년차(2019)
『승정원일기』	성능평가(3회)	고도화 평가1(3회)	고도화 평가2(2회)
조선왕조실록		성능평가(1회)	
특수고전(천문분야)			성능평가(2회)

위와 같은 과정으로 진행된 수동평가를 통해 얻은 정량 데이터를 바탕으로 최적의 자동번역시스템 모델을 선정하였고, 개발된 세 가지 자동번역 모델, 즉 『승정원일기』 자동번역모델, 조선왕조실록 자동번역모델, 특수고전(천문분야) 자동번역모델이 각각 일정 수준 이상의 번역 성능을 보이는 것을 확인하였다.⁶⁾ 특히, 조선왕조실록 및 특수고전-천문분야 자동번역기의 베이스 모델이자 2년간의 고도화 과정을 거친 『승정원일기』 자동번역 모델은 2018년 수동평가에서 3.86점(5점 만점, 2017년 3.0점 대비 128.7% 향상), 2019년 수동평가에서 4.20점(5점 만점, 4점 이상 획득 목표)을 획득하여 번역 인식률이 목표치 이상으로 향상된 것으로 평가되었다.

6) 2017년 개발된 『승정원일기』 자동번역모델은 3.0점(33번 모델 기준, 5점 만점)의 수동평가 점수를 얻었고, 2018년 개발된 조선왕조실록 자동번역모델은 3.18점(2018-16_1과 2018-17_2 두 모델 평균 점수 기준, 5점 만점)을 얻었으며, 2019년 개발된 특수고전(천문분야) 도메인 특화 천문고전 자동번역모델은 4.05점(IDM_AS_19모델 기준, 5점 만점)의 수동평가 점수를 얻었다.



〈그림 1〉 베이스모델 성능향상 추이(2017~2019)

『승정원일기』 자동번역기 번역평가에는 모두 8명의 내외부 전문가가 참여하였으나, 보고서에는 외부 전문가 2명의 평균점만 밝혀져 있다. 수동평가 데이터셋은 구축된 코퍼스를 대표할 수 있도록 코퍼스 전반에서 추출하였다. 자동번역기의 훈련에 반영하지 않은 최소 500문장을 대상으로 번역문 길이가 300자 이내인 자동번역 결과를 갖고 하나의 종결 기호를 갖는 데이터를 선정하였다. 2017년 『승정원일기』 자동번역모델은 54 또는 60문장으로 수동평가 데이터셋을 구성하였으며, 2018년과 2019년에는 단문 60문장⁷⁾과 장문 40문장으로 수동평가 데이터셋을 구성하였다. 2018년과 2019년에 개발된 조선왕조실록과 특수고전(천문분야)도 각각 50문장, 60문장으로 수동평가 데이터셋을 구성하였다.

수동평가는 5점 척도로 외부 전문가 평가위원(번역 전문가)에 의한 수동평가 방식으로 실시되었다. 평가의 기준이 되는 ‘정답문’(즉, 참조번역)과 테스트 모델별 자동번역 결과문을 비교하여 평가하였으며, 참조번역문에 오류가 있을 경우에는 원문을 참고하여 평가위원이 직접 평가하도록 하였다. 이러한 과정을 통해 얻어진 평가위원들의 점수는 그 평균을 계산하여 수동평가 점수로 삼았다. 수동평가에서 평가자에게 제시된 평가기준은 아래와 같다.

7) 2017년 정답문 세트 54문장에 2018년에 구축한 코퍼스 중 6문장을 추가하여 구성.

〈표 13〉 한국고전번역원 수동평가 자체 평가 기준

평가점수	평가기준	보조 설명	오류의 중요성 및 비율
※ 평가요소 : 번역율(누락 여부), 표현 및 문법, 의미 전달의 명확성			
5	의미 전달이 명확하고 문법적 오류가 없는 경우		0%
4	표현 및 문법에 사소한 오류가 있으나 비교적 의미가 이해 가능한 경우	문맥에 큰 영향을 주지 않는 오류	10~20%
3	주요 단어는 번역했으나 의미 전달이 일부 부정확한 경우	번역어휘 선택의 오류, 문법적 오류(비문), 사소한 결역 및 중복번역	30~50%
2	주요 단어를 번역하지 못해 의미 전달이 전반적으로 부정확한 경우	번역 누락 및 중복의 오류(심각한 결역 및 중복번역)	50% 이상
1	번역은 하였으나 이해가 전혀 불가능한 경우	전문 오역(중합 오류)	90%
0	번역 결과가 없는 경우		100%

또한, 2019년 수동평가에서는 2018년 최종모델의 자동번역 결과와 비교 후 번역 품질의 “높아짐(상), 비슷함(중), 낮아짐(하)”을 평가하여 번역기의 성능 향상을 확인할 수 있는 평가 항목을 추가하였다.

앞서 살펴본 주요 수동평가 방식과 『승정원일기』 자동번역시스템 수동평가 방식을 비교해보면 『승정원일기』 자동번역시스템 수동평가 방식에 다음과 같은 문제점이 있다.

첫째, 『승정원일기』 자동번역시스템 수동평가 방식은 일반적인 수동평가 방식, 즉 유창성(가독성)/정확성(적절성)으로 분류하지 않고 번역의 정확성만을 기준으로 하여 5점 척도로 단순화하였다. 이로 인해 번역 품질 평가의 또 다른 주요 요소인 가독성 부분에 대한 평가가 제대로 이루어지지 못했다.

둘째, 각각의 척도에 해당하는 기준이 명확하게 제시되지 않아 평가자의 주관에 따라 평가결과가 크게 달라지는 경우가 나타났다. 평가의 신뢰성을 높이기 위해서는 평가자 훈련과 사전 조율도 필수적인데, 한국고전번역원 수동평가에서는

이 과정이 생략되었다.⁸⁾

셋째, 한국고전번역원의 수동평가는 한국고전번역원의 번역방식에 익숙한 전문역자로만 평가를 진행한 점도 문제다. 전문 번역자가 수동평가를 수행할 경우 지나치게 엄격하게 평가할 우려가 있기 때문에 기계번역에 대한 기본적인 이해를 가지고 있는 인력이 평가를 수행할 필요가 있다(Rossi & Wiggins, 2013). 기계번역은 알고리즘의 선택, 코퍼스의 유형과 가공방법, 학습 절차 등에 따라 상이한 결과가 나올 수 있으므로, 이러한 특성을 이해하고 있는 사람이 평가에 참여하여야 여타 번역기와의 차이도 더 정확히 감별할 수 있고, 기계번역기의 성능을 향상시키는 데에도 활용할 수 있다.

넷째, 평가 결과에 대한 신뢰도 분석을 별도로 수행하지 않은 점도 짚고 넘어갈 일이다. 일반적으로 수동평가를 수행할 때는 평가의 신뢰도를 확보하기 위해 카파 상관계수(Cohen's Kappa Coefficient) 등을 활용하여 평가 일치도를 측정하는 과정을 거친다. 이는 평가 결과가 얼마나 객관적이고 합리적으로 이루어졌는가를 검증하기 위한 것이기도 하지만, 성능 개선에 유용하기 때문이기도 하다.

3.3. 『승정원일기』 자동번역시스템 수동평가 개선 방안

3.3.1. 새로운 평가 척도 제안

본 연구팀은 상기 문제점을 개선하기 위해, 수동평가 지표를 새롭게 모색해 보았다. 평가 지표를 개발하기 위해 검토해야 할 상황은 다음과 같다.

■ 평가지표 개발 :

- 평가자의 지식수준, 환경 등 정량화하기 어려운 변수들도 반영할 수 있는가.
- 평가지표는 번역의 명확성과 유창성을 고루 평가할 수 있도록 구성되었는가.
- 누락되거나 중복된 평가요소는 없는가.

8) 2019년에 수행된 수동평가를 예로 들어 보기로 한다. 이 당시 『승정원일기』 원문 100문장에 대해 테스트 모델 별 자동번역 결과문을 5점 척도로 평가하였는데, 2018년 모델과 비교하였을 때 판단이 엇갈리는 문장, 즉 동일 번역문에 대해 어떤 평가자는 개선되었다고 평가한 반면 비슷하다거나 나아졌다고 본 평가자도 존재하는 사례가 10건이나 보이며, 최고점과 최저점의 차가 3점 이상인 사례도 7건이었다.

- 배점은 평가지표의 특성이나 중요도에 알맞게 부여되었는가.
- 배점의 총합이 번역품질의 차이를 확인할 수 있도록 적절하게 부여되었는가.

■ 평가방식 개선 :

- 기계번역문과 인간번역문 차이를 감별할 수 있는지 평가할 수 있는가. : 자동평가와 수동평가는 단지 평가방식만 다른 것이 아니라 그 목적과 성격도 다르다. 원문, 정답문, 번역문을 모두 제시한 후 평가하는 기존의 평가방식, 즉 기계학습의 결과물임을 인지한 상태에서 진행되는 평가는 ‘원문 VS 번역문’에 대한 평가 외에 ‘정답문 VS 번역문’에 대한 평가도 동시에 이루어지는 셈이어서, 정답문에 의해 왜곡될 가능성이 있다. 한국고전번역원의 수동평가 지침에도 밝혀져 있다시피⁹⁾ 인간이 작성한 정답문이 반드시 정답이라고 할 수는 없다.
- 평가자의 주관에 따른 편차를 줄일 수 있는 절차가 마련되어 있는가.
- 번역품질 평가집단은 합리적이고 객관적으로 구성되었는가.
- 번역기의 활용 가능성(역자들의 번역 지원, 대국민서비스 등)을 판단하는 절차가 포함되었는가.

이상과 같은 사항에 대한 검토를 바탕으로 마련된 새로운 평가안은 다음과 같다.

■ 평가지표 개발 원칙

- 포괄성 : 번역 품질을 전반적으로 평가할 수 있는가.
- 구체성 : 평가요소가 구체적으로 드러나는가.
- 독립성 : 평가요소 간 중복이 없도록 구성되었는가.

■ 평가점수 부여 방식 : 지표의 특성에 맞게 명확성은 감점제, 유창성은 구간별 선택제를 택함

9) “(천문고전) 정답문 항목은 사람이 작성한 것이어서, 정확하지 않을 수 있습니다. 어떤 경우에는 복수의 정답문을 제시한 경우도 있습니다. 원문에 대한 정답문을 평가위원이 직접 정의하고, 기계번역의 결과를 평가하여도 됩니다.”

■ 평가 대상 텍스트 제공 방법

- 원문과 번역문만 제시
- 번역문 중에 전문역자의 번역문을 포함시켜 기계번역과의 차이를 감별할 수 있는지, 동등한 조건 하에서 기계번역 평가결과와 어떤 차이를 지니는지 평가할 수 있도록 함.

■ 평가자 구성

- 전문가 집단과 비전문가 집단으로 나누어 진행
- 전문가 집단은 5년 이상 또는 3종 이상의 전문 번역서 실적이 있는 자로서, 한국고전번역원의 번역기준을 이해하고 있되 일반적인 번역기준까지 함께 고려하여 평가할 수 있는 자 3인 이상으로 구성.
- 비전문가 집단은 한문에 대한 기초적인 이해력을 갖춘 사람을 대상으로 선발.
- 전문가 집단 평가의 목표는 ‘실제 번역 작업에 활용 가능성 진단’, 비전문가 집단 평가의 목표는 ‘대국민 서비스 가능성 진단’

■ 평가방식과 절차의 개선

- 전문가 집단과 비전문가 집단 각각에 적합한 평가방법 적용.
- 전문성을 갖추었으되 한국고전번역원 스타일에 구애받지 않고 평가할 수 있는 평가자가 포함되어야 하며, 전문가 이외의 일반사용자들의 반응을 조사할 수 있도록 비전문가 집단 평가도 병행
- 평가절차 : 평가지표에 대한 충분한 사전 설명 → 예비문항을 통한 상호조율 → 본 평가 → 본 평가 후 평가점수의 편차가 큰(30%) 문항에 대한 의견 조율 → 본 평가 → 평가결과 분석

■ 기타 사항 : 번역 품질의 다각적 검토

- 『승정원일기』 미학습 부분, 실록의 중복(동일 기사)/비중복 부분, 기타 성격이 유사한 문헌자료 등을 평가에 활용
- 『승정원일기』와 문체 및 내용이 상이한 자료(사서, 삼경, 문집류 등)도 평가에 활용
- 바이두 기계번역기와의 비교 평가

이상의 과정을 거쳐 개발한 평가지표는 다음과 같다.

〈표 14〉 정책과제 연구팀 개발 수등평가 평가 기준

평가 영역	평가 지표	평가 요소	배점 (총25점)
명확성	오류 오역	● 어휘의 의미를 바르게 표현하였는가 (인명, 지명, 관명, 서명의 이해, 漢字並記의 적합성, 관용구의 풀이 따위 포함)	3
		● 어구의 연결이 옳은가 (ex: 軍職實職 : ‘군직과 실적’, ‘군직의 실적’)	3
		● 문장의 구조, 구와 구의 관계에 맞게 번역하였는가 (역접/순접, 인과관계 등)	3
		● 문형(평서문/감탄문/의문문 따위), 화자와 청자의 관계, 글의 어조 등에 맞게 번역하였는가	3
	추가 누락 미번역	● 원문에 없는 내용이 추가되었는가	3
		● 원문에 있는 내용이 누락되었는가	
		● 원문을 번역하지 못하고 번역문에 노출하였는가	
유창성	유창성 및 가독성	<ul style="list-style-type: none"> ● (번역의 정확성과 무관하게) 번역문이 자연스러운가 ● 맞춤법, 띄어쓰기의 오류는 없는가 - 10~9점 : 거의 완벽하거나 사소한 오류가 있을 경우 - 8~7점 : 번역투는 있지만 의미는 이해 가능한 경우 - 6~5점 : 맞춤법, 띄어쓰기를 포함한 문법적인 오류, 의미 파악이 어려운 경우 - 4~1점: 내용 이해가 불가능한 경우 	10

명확성 영역은 ‘오류오역’과 ‘추가누락미번역’ 2가지 지표로 나누었다. 이 2가지 지표는 번역 품질평가에서 차지하는 비중이 높음을 감안하여 평가요소를 세분하여 제시하되 각 요소 간 중복이 없도록 예시를 들어 설명하였다. 요소별 배점은 3점으로 하고 결점이 발견될 경우 1점 감점을 원칙으로 하되 복수로 나타나거나 단수로 나타나더라도 심각한 오류오역에 해당할 경우 추가 감점이 가능하도록 하였다. 추가누락미번역은 통합하여 3점을 부여하였다.

유창성 영역은 원문과의 대조로 인한 영향을 배제하기 위해 번역문만 제시하였으며, 맞춤법과 띄어쓰기를 포함하여 평가하도록 하였다. 배점은 10점 만점으로 하되 주관성을 완화하기 위해 구간별로 점수를 부여하도록 하였다.

본 연구팀은 평가 지표의 적절성을 검토하기 위해 전문가 3인으로 평가자를 구성

하여, 상기 지표에 따라 『승정원일기』 자동번역시스템의 평가를 수행하게 하였다.

전문가 평가

- 조사일 : 2020년 11월 14일
- 평가자 구성 : 평가자A(『승정원일기』 역자), 평가자B(번역협동과정 박사학위, 한국고전번역원 외 역자), 평가자C(번역거점연구소 공동연구원, 한국고전번역원 외 역자)
- 평가 결과 (조사용 문장은 [첨부 1] 참조)

유형	Set No.	평가자A			평가자B			평가자C			종합			
		유창성	명확성	합계	유창성	명확성	합계	유창성	명확성	합계	유창성	명확성	합계	
단문	承-人	1	8	15	23	9	14	23	10	14	24	9.00	14.33	23.33
		2	10	15	25	9	14	23	10	15	25	9.67	14.67	24.33
	承-機	3	10	15	25	9	14	23	10	14	24	9.67	14.33	24.00
		4	10	15	25	9	15	24	9	15	23	9.33	15.00	24.00
		5	6	13	19	8	13	21	8	14	22	7.33	13.33	20.67
		6	6	14	20	9	15	24	9	13	22	8.00	14.00	22.00
	其-機	7	8	14	22	9	14	23	10	14	23	9.00	14.00	22.67
		8	8	15	23	8	15	23	10	13	24	8.67	14.33	23.33
		9	10	15	25	8	14	22	8	15	23	8.67	14.67	23.33
		10	8	15	23	10	15	25	9	15	24	9.00	15.00	24.00
		11	6	12	18	6	13	19	8	11	19	6.67	12.00	18.67
		12	10	15	25	10	15	25	8	14	22	9.33	14.67	24.00
전후문장 맥락	承-人	13	10	14	24	9	14	23	10	13	23	9.67	13.67	23.33
		14	6	14	20	6	13	19	5	14	19	5.67	13.67	19.33
	承-機	15	6	11	17	8	13	21	7	13	20	7.00	12.33	19.33
		16	6	14	20	8	13	21	6	13	19	6.67	13.33	20.00
		17	6	12	18	6	14	20	6	12	18	6.00	12.67	18.67
	其-機	18	6	14	20	8	13	21	6	13	19	6.67	13.33	20.00
		19	10	14	24	8	13	21	7	14	21	8.33	13.67	22.00
		20	8	12	20	8	12	20	5	10	15	7.00	11.33	18.33
평균		7.9	13.9	21.8	8.25	13.8	22.05	8.05	13.45	21.45	8.07	13.72	21.77	

* set1, 2, 13은 『승정원일기』 인간번역문(承-人), set3~6, 14~16는 『승정원일기』 기계번역문(承-機), set7~12, 17~20은 기타 자료의 기계번역문(其-機)임.

3.3.2. 한국고전번역원에서 수행한 수동평가 결과와 연구팀 자체 수동평가 결과 분석

본 연구진이 개발한 평가지표의 타당성을 검증하기 위해 앞서 진행한 전문가 평가 결과를 한국고전번역원에서 수행했던 평가결과와 비교해보았다. 비교에 사용된 한국고전번역원 측 자료는 2019년 BS_SJ_26 모델을 대상으로 진행된 결과로, 가장 높은 수동평가 점수(4.20점)를 획득했다. 본 연구진이 수행한 자체평가 대상 원문 및 번역문도 한국고전번역원에서 사용했던 것과 동일한 자료를 사용하여 동등한 비교가 가능하도록 하였다. 일관성을 유지하려면 동일한 평가자가 수행해야 하며, 표본(sample) 수가 충분하지 않다는 한계가 있었지만 몇 가지 의미 있는 결과들을 확인할 수 있었다.(부록 참조)

한국고전번역원 자체 수행 평가자별 점수 평균점은 3.54점부터 4.50점까지 편차가 크게 나는 것에 비해, 연구팀 자체 수행 번역 평가 결과는 명확성(13.45~13.90), 유창성(7.9~8.25)에서 모두 큰 편차가 발생하지 않았다. 이에 한국고전번역원에서 수행한 수동평가 결과에 비해 연구팀이 개발한 수동평가 방식이 『승정원일기』 기계번역기 성능평가에서 더 유의미한 결과를 얻을 수 있다고 판단된다. 표본을 좀 더 늘린다면 기계번역 성능평가에 적용하여도 무방하리라 본다.

3.3.3. 전문 역자 대상 심층 평가 제안

『승정원일기』 번역기는 『승정원일기』 번역 업무의 편의성 및 효율성을 증진시키기 위한 목표를 가지고 있다. 전문 역자들이 해당 번역기를 어떻게 인지하고 활용하는지에 대한 조사는 해당 번역기에 대한 질적 평가에 반영될 수 있으며, 향후 품질 개선에 기여할 수 있으리라 생각한다. 본 설문 조사는 역자들이 승정원일기 자동번역기를 활용한 번역 소요 시간을 측정해 봄으로써 일종의 테스크 기반 평가(task-oriented)를 지향하고 있다. 아울러 자동번역기의 품질, 편리성, 유용성에 대한 의견을 묻는 방식으로 진행하였다.

『승정원일기』 역자 평가

- 조사일 : 2021년 1월 8일~1월 14일

■ 평가자 구성 : 『승정원일기』 번역이나 평가 경험이 있는 역자 9명

평가자	연령대	번역 경력	귀하는 구글, 파파고, 바이두 등 인공지능 번역기를 사용해본 경험이 있습니까?	귀하는 새로운 기술을 사용해 보는 것에 적극적인 편입니까?
A	50대	10년 이상 ~ 15년 미만	사용해 본 적이 있다	그렇지 않다
B	40대	5년 이상 ~ 10년 미만	주 1회 이상 사용	매우 그렇다
C	30대	5년 이상 ~ 10년 미만	사용해 본 적이 전혀 없다	보통이다
D	50대	10년 이상 ~ 15년 미만	사용해 본 적이 있다	보통이다
E	50대	10년 이상 ~ 15년 미만	사용해 본 적이 있다	그렇다
F	30대	5년 미만	가끔 사용한다	보통이다
G	30대	5년 미만	사용해 본 적이 있다	그렇다
H	50대	15년 이상 ~ 20년 미만	주 1회 이상 사용	매우 그렇다
I	40대	10년 이상 ~ 15년 미만	가끔 사용한다	그렇다

설문 조사에 참여한 역자는 총 9명으로 연령과 번역 경력을 고려하여 선정하였다. 이 중 5인(A~E)은 『승정원일기』를 번역한 경험이 있으며, 나머지 4인(F~I)은 『승정원일기』 번역 평가 경험이 있는 역자이다.

본 설문에는 번역기에 대한 평가에 앞서 개인의 혁신성도 함께 조사하였다. 혁신성은 기존의 것과는 다른 ‘새로운 아이디어를 빨리 채택하는 정도’로 규정할 수 있는데,¹⁰⁾ 역자들의 『승정원일기』 번역기 인식에 변인으로 작용할 수 있다고 판단하였다.

평가자 중 『승정원일기』 번역 경험이 있는 5인(A~E)을 대상으로 난이도가 비슷한 원문 텍스트를 각각 2개 제공하여 인간번역과 번역기를 활용한 번역을 비교

10) 천종성(2020), 283면.

하게 하였다. 즉, 원문 1은 평소에 번역하는 방식대로 번역하고, 원문 2는 『승정원 일기』 자동번역기를 활용하여 번역한 뒤 소요된 시간을 반드시 기록하게 하였다.

평가자	소요시간	
	원문 1(인간 번역)	원문 2(번역기 활용 번역)
A	1시간 30분	1시간
B	31분 55초	14분 43초
C	2시간	1시간
D	1시간 30분	1시간
E	1시간 35분	55분

평가자마다 번역에 소요하는 시간은 달랐지만 공통적으로 인간번역보다는 번역기를 활용한 번역의 경우 소요시간이 상당히 절감되었음을 확인할 수 있었다.

다음은 번역기의 번역 품질, 편리성, 유용성에 대한 평가이다.

평가자	품질			편리성	유용성		
	원문의 맥락을 파악하여 번역하고 있는가?	원문의 어휘를 정확하게 번역하고 있는가?	번역문은 가독성이 있는가?		번역 시간을 단축하는 데 도움이 된다고 생각하는가?	승정원일기 자동번역기가 역자들에게 유용하다고 생각하는가?	승정원일기 자동번역기를 사용할 의향이 있는가?
A	보통이다.	보통이다.	보통이다.	그렇다.	보통이다.	보통이다.	보통이다.
B	보통이다.	보통이다.	보통이다.	보통이다.	매우 그렇다.	보통이다.	매우 그렇다.
C	그렇다.	그렇다	보통이다.	그렇지 않다.	그렇다	그렇다.	그렇다.
D	그렇지 않다.	보통이다.	보통이다.	그렇다.	그렇다	매우 그렇다.	매우 그렇다.
E	그렇다.	그렇다	그렇다	매우 그렇다	매우 그렇다.	매우 그렇다.	매우 그렇다.
F	보통이다.	그렇지 않다.	그렇다	-	매우 그렇다.	매우 그렇다.	매우 그렇다.

G	보통이다.	보통이다.	보통이다.	-	그렇다	매우 그렇다.	매우 그렇다.
H	그렇다	그렇다	그렇다	-	매우 그렇다	그렇다	매우 그렇다
I	보통이다.	그렇다.	그렇지 않다.	-	그렇다.	보통이다.	그렇다.

■ 번역 품질

번역기를 활용한 후 번역 품질을 맥락, 어휘, 가독성 부분으로 나누어 평가하게 하였다. 전반적으로 “보통”으로 평가한 의견이 많았다. 세부 의견을 살펴보면 공통적으로 상투적인 구문의 번역 품질은 상당히 뛰어나나 그 밖의 경우에는 번역 품질이 낮다는 의견, 어휘 부분은 고유명사 번역에서 오류가 많다는 의견이 많았다. 또 컨텍스트를 충분히 파악하지 못한다는 점을 한계로 꼽았다.

- 평가자 A : 『승정원일기』의 기사 가운데 반복적으로 나오는 계사의 경우는 자동번역기를 이용하면 번역시간을 단축할 수 있으나 그 밖에 상소문이나 입시 기사의 경우는 자동번역기를 사용하면 번역의 정확성이 그리 높지 않고 가독성도 현저히 떨어지기 때문에 번역 시간은 약간 단축할 수 있겠지만 수정 과정에서 미처 수정하지 못하고 누락하는 곳이 많이 나올 듯함.
- 평가자 B: 자동번역기는 원문의 맥락과 뉘앙스를 충분히 번역하지 못함, 표점이 정확하지 않을 경우 번역 오류가 발생함.
- 평가자 D: 중복하여 번역하는 사례가 있음. 인용 범위 오류, 어휘 풀이 오류, 화자와 청자와의 관계 오류, 문장부호 오류, 번역 안함(예: 병출탑교) 등의 오류가 발견됨.
- 평가자 E: 『승정원일기』에 자주 등장하는 표현이나 반복되는 기사에 대한 번역 품질은 뛰어난 편, 낯선 표현이나 작자 특유의 어휘 사용 등을 마주할 때에는 상대적으로 번역 품질이 낮음. 시제와 낯선 고유명사 처리 등은 주요하게 개선되어야 할 것으로 보임. 이 부분의 오역은 한문을 전혀 모르는 이용자에게 치명적일 것으로 판단됨
- 평가자 I : 『승정원일기』 번역에서 상투적으로 쓰이는 어휘나 상소문의 단어를 자동번역기에서 번역 결과는 우수함. 처음 나오는 단어, 특히 입시기사의 번역의 경우 자동번역기의 한계가 많음. 추후 전통문화연구회에서 번역한 경서라던가 한국고전번

역원에서 번역한 문집 등 더 많은 데이터를 축적하여 개선할 필요가 있음.

■ 편리성

편리성 부분에서는 의견이 갈렸다. 다만 ‘그렇지 않다’고 답한 평가자 C는 기존의 다른 기계 번역기를 사용한 경험이 없기에 승정원일기 번역기에 대한 평가도 긍정적이지 않았던 것으로 보인다.

- 평가자 C : 뒤로 가기(ctrl+z) 기능이 있으면 좋겠음, 번역문과 어휘사전의 표제어가 일치하지 않음, 어휘사전에 해당 어휘에 대한 번호가 부여되면 좋겠음.
- 평가자 E : 번역문이 복사가 안 되는 경우가 있음.

■ 유용성

평가자들은 현재 번역기의 번역 품질에 대해서는 다소 유보적인 태도를 보였으나 업무에 활용 가능성에 대해서는 대체로 긍정적으로 인식하였다. 9명 중 8명이 번역기가 실제 번역 시간을 단축시킬 수 있다고 평가하였다. 9명 중 6명이 업무에 유용하다고 평가하였으며 추후 활용 여부에 있어서도 상당히 긍정적으로 답변하였다. 다만 평가자 A는 전반적으로 『승정원일기』 번역기의 유용성에 대해서 긍정적으로 평가하지 않았는데, 이는 혁신성 부분에서 보수적인 태도가 『승정원일기』 번역기의 평가에 영향을 미쳤을 가능성도 생각할 수 있다.

세부 의견을 살펴보면 평가자들은 번역기의 번역 결과를 초벌 번역 형태로 활용할 수 있는 가능성에 대해 긍정적으로 인식하였다. 특히 특정 어휘나 형식이 반복적으로 나열되는 저맥락 텍스트의 경우 번역 속도를 향상시킬 수 있다고 보았다. 그리고 번역 과정에서 원문 텍스트의 전후 기사의 내용을 번역기를 통해 신속하게 확인할 수 있다는 점, 번역의 규칙을 확인할 수 있다는 점에서 번역기의 활용 가능성을 높게 평가하는 의견이 있었다.

- 평가자 B: 번역위원들이 이를 선택적으로 활용하여, 계사나 상소문과 입시기사의 투식적인 단문에서 활용한다던가, 초벌 번역의 형태로 활용한다면 번역속도를 확실하게 줄여서 상소문이나 입시기사의 중요한 대목에서 시간을 더욱 활용하거나 운문

의 횡수를 늘일 수 있을 것으로 보인다.

- 평가자 D: 대체로 기대 이상의 수준으로 개발되었다고 보여진다. 이 정도의 성과라면 앞으로도 좀더 정교하고 완성도 높은 번역이 가능하리라 생각된다. 실제 번역에 착수하기 전 상소나 경연 기사의 일독을 통해 전체적인 대의나 분위기를 빠르게 파악할 수 있고, 일부 구절이나 문장 연결에 있어 좀더 합당한 번역을 참조할 수 있으며, 심지어 자신의 오역을 수정할 수도 있을 것이다. 또한 한자 병기가 필요한 고유 명사나 어휘, 반복되는 구절이나 문장은 그대로 따옴으로써 입력 시간을 단축할 수도 있고, 이 밖에 다양한 활용을 통해 번역 공기 단축에 도움이 된다는 점에서 충분한 활용 가치가 있어 보인다.
- 평가자 G: 번역 전에 기사 전문을 번역기에 적용시킨 후, 결과물을 한 번 읽어보고 작업에 돌입하면 해당 기사 전반의 분위기나 키워드, 주제 등을 파악하는 데 도움이 된다.
- 평가자 H: 자동번역기와의 협업이 새로운 방식의 번역작업이 되겠지만 적극적으로 활용하는 것이 유용하다고 생각한다. 전통방식에 의한 번역자 양성과정, 발전 기술을 접목하고 응용하는 방식이 병행하는 것이 필요하다.

이상의 설문 조사를 통해 번역기의 번역 품질뿐 아니라 편리성과 효용성 부분에서도 의미 있는 평가 및 제언을 확인할 수 있었다. 『승정원일기』 번역기는 충분히 신뢰할 만한 수준은 아니지만, 전문 역자들의 업무 효율을 높여주는 유용한 도구로 활용될 수 있는 가능성을 지닌다. 이로 볼 때, 『승정원일기』 자동번역 시스템 개발의 애초의 목적은 어느 정도 달성한 것으로 판단된다. 그러나 빈번하게 확인되는 번역 오류를 개선하고 편리성과 효용성을 제고하기 위해, 번역기의 성능을 고도화하는 한편 사용자의 번역 습관·번역기 활용 방식을 고려한 개선이 필요하다는 것을 본 설문 조사를 통해 확인할 수 있다. 이상의 심층 설문 조사는 자동번역 시스템의 구체적인 피드백을 제공하기에, 자동평가 및 수동평가와 함께 수행한다면 향후 『승정원일기』 자동번역 시스템을 고도화하는 데 도움을 주리라 기대한다.

4. 나가는 말

원의를 충실히 살리면서도 가독성이 높아야 좋은 번역문으로 간주된다. 그런데 고립어인 한문 고전문언문은 상이한 문체가 매우 많고 문법상의 변화도 복잡하다. 또한 기계번역은 평가기준·평가목적·평가비용·텍스트의 종류 등도 함께 고려하여야 한다. 이렇듯 고려해야 할 요인이 다양하기 때문에 신뢰성이 높고 타당하면서도 간편한 번역 품질 평가모델을 개발하기가 쉽지 않다.

하지만 기계번역의 품질을 정확히 평가하는 것은 기계번역 모델 개발은 물론 지속적인 성능 개선을 위해서도 반드시 필요하다. 기계번역의 품질을 확인하기 위한 평가방법은 자동평가와 수동평가로 대별되는데, 각각의 특징이 있기 때문에 병행하는 것이 바람직하다.

즉, 자동평가는 기계번역의 품질만을 평가하는 것으로, 기계번역의 어떤 요소들이 번역 품질에 영향을 미치는지는 알 수 없다. 또한 기계번역 모델에 대한 평가지표를 비교하여 어느 모델이 가장 높은 점수를 받는 기계번역 모델인지는 보여줄 수 있지만, 기계번역 품질에 대한 타당성을 보장하지는 못한다. 그리고 평가기준도 평가모델에 따라 달라질 수 있고 대량의 데이터를 필요로 하는 경우도 있다. 이런 문제점을 보완하기 위해서는 수동평가가 필요한데, 평가자 각각의 경험이나 수준이 존재하고, 평가기준에 대한 이해가 다를 수 있으며, 평가 환경이나 차수에 따른 차이 등 주관에 치우칠 우려도 불식하기 어렵다. 따라서 자동평가와 수동평가의 장단점을 고려하여 기계번역기의 성격과 목적에 맞는 평가방법을 찾아 적용하되, 기계번역 모델의 성능을 객관적으로 평가할 수 있는 척도를 개발하여야 하며, 궁극적으로 이러한 평가방법이 기계번역 모델의 문제점을 찾아 개선해나가는 데 도움이 될 수 있도록 해야 한다.

■ 참고문헌

- 김우정 외(2021), 「2020년도 한국고전번역원 기획연구과제 최종보고서: 『승정원일기』 자동번역시스템의 활용방안 연구」, 한국고전번역원.

- 이준호(2019), 「신경망기계번역의 객관적 평가를 위한 예비연구: 자동평가와 수동평가의 균형점」, 『통번역학연구』 23권 5호, 한국외국어대학교 통번역연구소, 171~202.
- 천종성(2020), 「전문번역사들의 기계번역 수용에 관한 연구」, 『한국융합학회논문지』 제11권 제6호, 한국융합학회, 281~288.
- 최효은·이지은(2017), 「특허 기계번역 결과물의 평가 - KIPRIS 의 무료 한영 기계번역을 중심으로」, 『통역과 번역』, 19(1), 139~178.
- 한국고전번역원(2019), 「2019년 클라우드 기반 고문헌 자동번역 확산 서비스 구축: 사업추진결과보고서」.
- Bazrafshan, M.(2014), "*Semantic Features for Statistical Machine Translation*", Unpublished doctoral thesis, University of Rochester, New York, US.
- Banerjee, S. & A. Lavie(2005), "*METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*", Proceedings of the ad workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 65~72.
- Chung, Hye-Yeon(2020), "*Automatische Evaluation der Humanübersetzung: BLEU vs. METEOR*", Lebende Sprachen 65(1), 181~205.
- Denkowski, M, & A. Lavie(2011), "*Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems*", Proceedings of the sixth workshop on statistical machine translation, 85~91.
- Papineni, Kishore, et al(2002), "*Bleu: a method for automatic evaluation of machine translation*", Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 311~318.
- Callison-Burch, C, et al(2006), "*Re-evaluating the role of BLEU in machine translation research*", 11th Conference of the European Chapter of the Association for Computational Linguistics, 249~256
- Callison-Burch, C, et al(2007), "*(Meta-) evaluation of machine translation*", Proceedings of the Second Workshop on Statistical Machine Translation, 136~158.
- S. Hampshire & Porta Salvia, C(2010), "*Translation and the internet: Evaluating the quality of free online machine translators*", Quaderns: Revista de Traduccio, 17, 197~209.
- Han, L(2016), "*Machine translation evaluation resources and methods*". : A survey. arXiv: 1605.04515v8, Cornell University Library.
- J. Hutchins, and S. Harold(1992), *An Introduction to Machine Translation*, London: Academic Press.
- John B. Carroll(1966), "*An experiment in evaluating the quality of translation*", Mechanical Translation and Computational Linguistics, 9(3-4), 67~75.
- Rossi, L., & Wiggins, D(2013), "*Applicability and application of machine translation quality metrics*

- in the patent field*", World Patent Information, 35, 115~125,
- Snover, M, et al(2006), "A study of translation edit rate with targeted human annotation", Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 223~231
- Van Slype, G(1982), "Economic aspects of machine translation", Practical Experience of Machine Translation, Amsterdam: North-Holland, 79-93.

부록

한국고전번역원 자체 수행 번역 평가 결과 분석

① 한국고전번역원 자체 수행 평가자별 점수 분포(전체)

평가자	최소	25%	50%	75%	최대	평균	분산
1	1.00	4.00	5.00	5.00	5.00	4.33	0.911
2	2.00	3.00	4.00	5.00	5.00	4.06	0.814
3	1.00	3.00	4.00	5.00	5.00	3.88	1.121
4	1.00	3.00	3.00	4.00	5.00	3.54	0.947
5	1.00	3.00	4.00	5.00	5.00	3.58	1.273
6	2.00	4.00	5.00	5.00	5.00	4.50	0.792
7	2.00	4.00	5.00	5.00	5.00	4.32	0.873

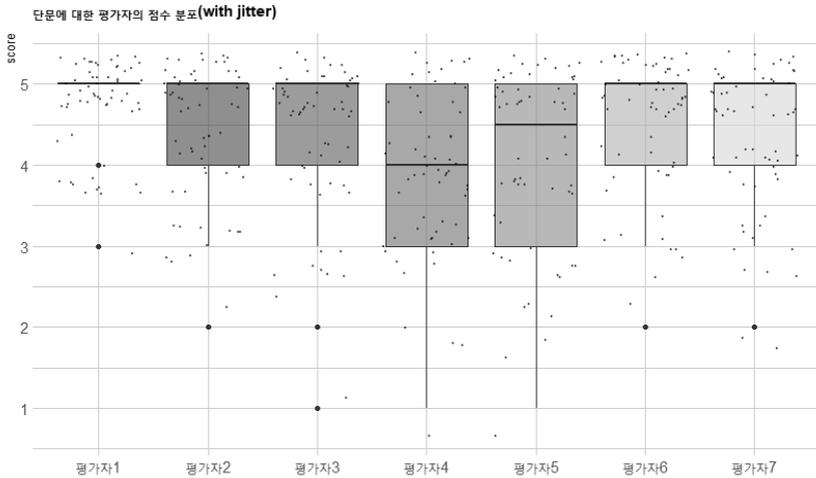
* 점수 분포 구간을 최소, 25%, 50%, 75%, 최대로 나눔.

평가자1~5는 평가 대상 100문장 전체에 참여하였으며, 평가자6, 7은 60문장의 평가에만 참여하였다. 평가점수는 5점 척도로 이루어져 있었는데, 최소점수는 평가자에 따라 1점 또는 2점을 부여하고 있으나 최대점수는 7명 모두 5점 만점까지 부여하고 있음을 볼 수 있으며, 평균점은 3.54점부터 4.50점까지 편차가 크게 남을 확인할 수 있다.

② 한국고전번역원 수행 평가자별 단문 평가 점수 분포

평가자	최소	25%	50%	75%	최대	평균	분산
1	3.00	5.00	5.00	5.00	5.00	4.75	0.474
2	2.00	4.00	5.00	5.00	5.00	4.33	0.816
3	1.00	4.00	5.00	5.00	5.00	4.42	0.889
4	1.00	3.00	4.00	5.00	5.00	3.87	0.947
5	1.00	3.00	4.50	5.00	5.00	4.12	1.075
6	2.00	4.00	5.00	5.00	5.00	4.50	0.792
7	2.00	4.00	5.00	5.00	5.00	4.32	0.873

한국고전번역원 자체 수행 평가자별 점수를 두 개 이상 집단의 평균을 비교하는 통계분석 기법인 分散分析(Analysis of Variance, ANOVA)을 이용하여 검증해본 결과, 평가자간 유의미한 차이가 있음을 확인할 수 있었다. 예를 들어 평가자1은 25%~75% 구간 값이 5점 만점인 반면에 평가자4와 5는 25%~75% 구간 값이 3점~5점에 걸쳐 넓게 분포하고 있으며, 중간값도 평가자1은 5점 만점에 육박함에 비해 평가자4는 4점, 평가자5는 4.5점으로 상당히 차이가 나며, 최소점수도 평가자1은 3점임에 비해 평가자4와 5는 1점으로 많은 확연히 차이가 난다. 이러한 현상은 평가자2, 3, 6을 평가자4와 대조해보아도 동일하게 확인된다. 이는 지터(해당 점수에 있음을 임의로 시각화해서 나타낸 것) 분포를 통해서도 확인할 수 있다.



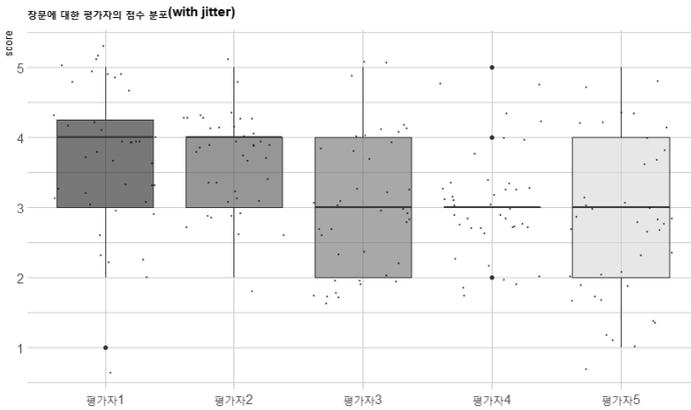
〈그림 1〉 단문에 대한 평가자의 점수(실선은 중간값을 나타냄)

또한 평가자 중 5점의 중간값이 5점 만점이고, 평가자 4와 5도 4점, 4.5점이나 됨을 볼 수 있는데, 이는 5점 척도로는 번역 품질을 온전히 평가하기 어려웠음을 보여준다.

③ 한국고전번역원 수행 평가자별 장문 평가 점수 분포

평가자	최소	25%	50%	75%	최대	평균	분산
1	1,00	3,00	4,00	4,25	5,00	3,70	1,043
2	2,00	3,00	4,00	4,00	5,00	3,65	0,622
3	2,00	2,00	3,00	4,00	5,00	3,08	0,944
4	2,00	3,00	3,00	3,00	5,00	3,05	0,714
5	1,00	2,00	3,00	4,00	5,00	2,78	1,121

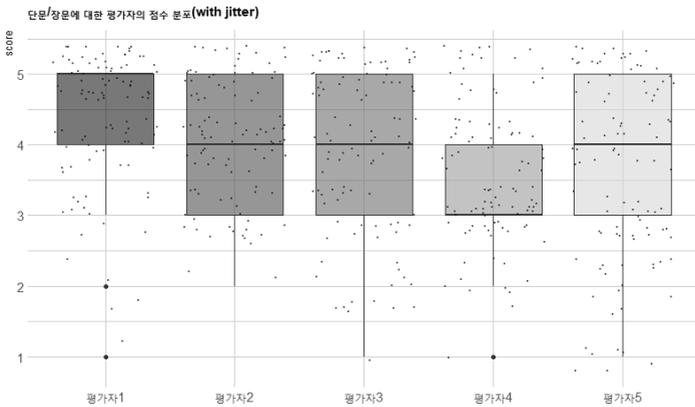
위는 장문 평가에 참여하지 않은 평가자6, 7을 제외한 점수 집계 결과이다. 앞서 살펴본 단문 분석결과와 마찬가지로 장문에서도 평가자 간 유의미한 차이가 확인되었다(평가자1 vs 평가자3, 4, 5 / 평가자2 vs 평가자3, 4, 5).



〈그림 2〉 장문에 대한 평가자의 점수

④ 한국고전번역원 수행 평가자별 단문/장문 전체 점수 분포 : 평가자 간 유의미한 차이(평가자1 vs 평가자3, 4, 5 / 평가자2 vs 평가자4, 5)가 확인되었다.

평가자	최소	25%	50%	75%	최대	평균	분산
1	1.00	4.00	5.00	5.00	5.00	4.33	0.911
2	2.00	3.00	4.00	5.00	5.00	4.06	0.814
3	1.00	3.00	4.00	5.00	5.00	3.88	1.122
4	1.00	3.00	3.00	4.00	5.00	3.54	0.947
5	1.00	3.00	4.00	5.00	5.00	3.58	1.273

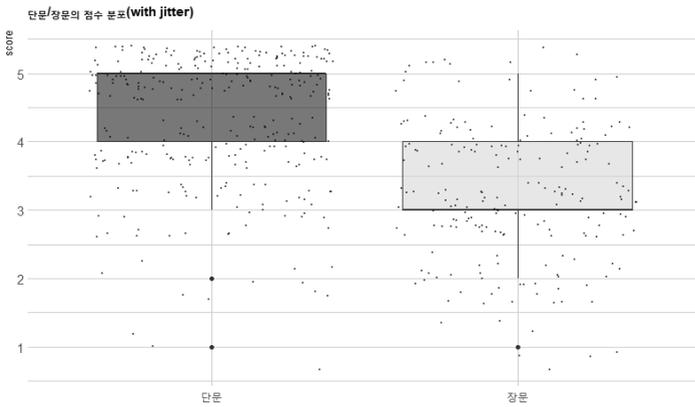


〈그림 3〉 단문/장문에 대한 평가자의 점수

한국고전번역원 수행 평가 결과(단문/장문 비교 분석)

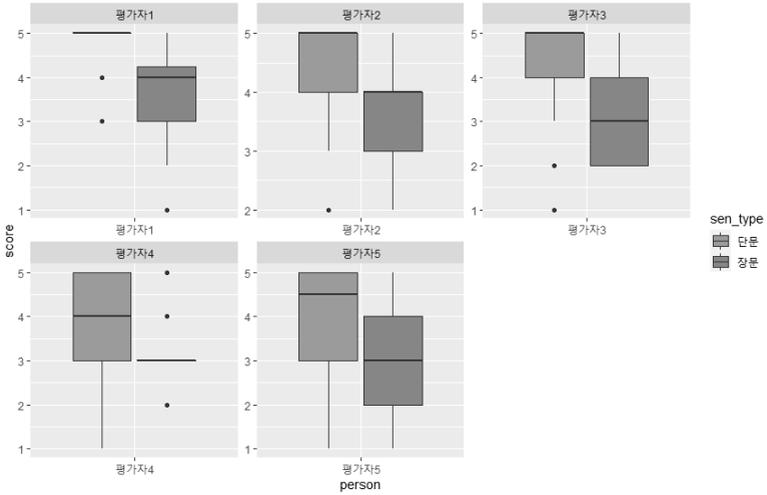
① 장문과 단문 비교 : 문장 종류에 따라 유의미한 차이가 있었다.

문장 종류	최소	25%	50%	75%	최대	평균	분산
단문	1.00	4.00	5.00	5.00	5.00	4.33	0.888
장문	1.00	3.00	3.00	4.00	5.00	3.25	0.971



〈그림 4〉 단문/장문의 점수 분포

② 평가자별 장문과 단문 비교 : 평가자 모두, 장문 점수가 상대적으로 낮았다.

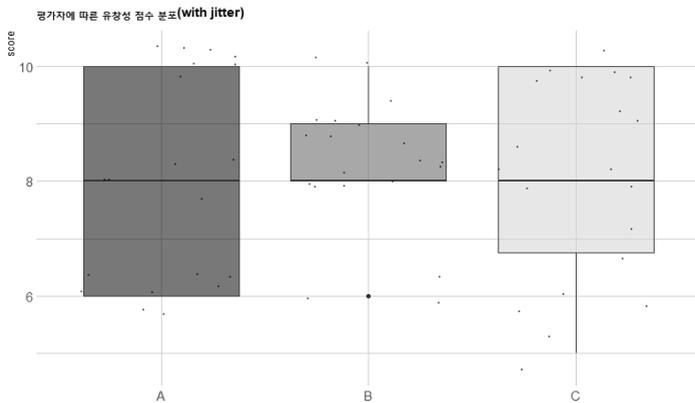


〈그림 5〉 평가자별 장문과 단문 비교

연구팀 자체 수행 번역 평가 결과 분석

① 연구팀 자체 수행 평가자별 유창성 점수 분포 : 평가자 간 유의미한 차이가 없었다.

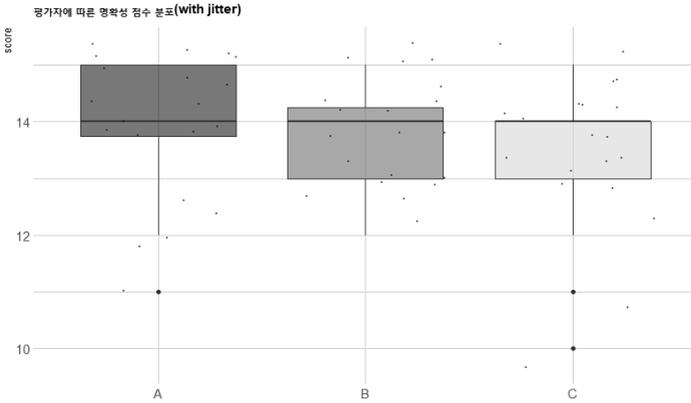
평가자	최소	25%	50%	75%	최대	평균	분산
1	6.00	6.00	8.00	10.00	10.00	7.90	1.774
2	6.00	8.00	8.00	9.00	10.00	8.25	1.164
3	5.00	6.75	8.00	10.00	10.00	8.05	1.761
전체	5.00	6.00	8.00	9.25	10.00	8.07	1.572



〈그림 6〉 평가자에 따른 유창성 점수 분포

② 한국고전번역원 수행 평가자별 명확성 점수 분포 : 평가자 간 유의미한 차이가 없었다.

평가자	최소	25%	50%	75%	최대	평균	분산
1	11.00	13.75	14.00	15.00	15.00	13.90	1.252
2	12.00	13.00	14.00	14.25	15.00	13.80	0.894
3	10.00	13.00	14.00	14.00	15.00	13.45	1.317
전체	10.00	13.00	14.00	15.00	15.00	13.72	1.166

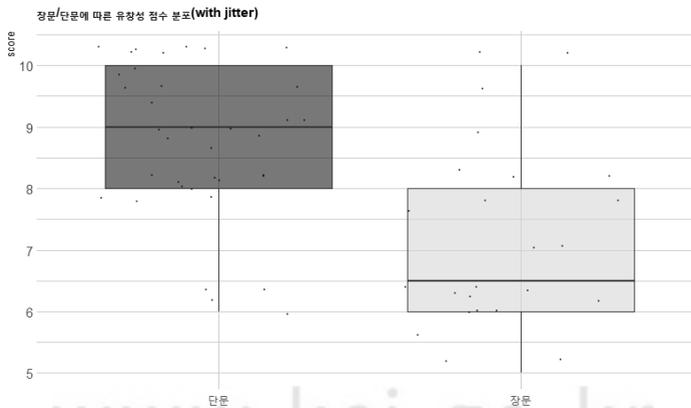


〈그림 7〉 평가자에 따른 명확성 점수 분포

③ 연구팀 자체 수행 장문/단문 유창성 점수 분포 : 장문/단문 간 유의미한 차이가 있었다.

문장종류	최소	25%	50%	75%	최대	평균	분산
단문	6.00	8.00	9.00	10.00	10.00	8.69	1,261
장문	5.00	6.00	6.50	8.00	10.00	7.13	1,541
전체	5.00	6.00	8.00	9.25	10.00	8.07	1,572

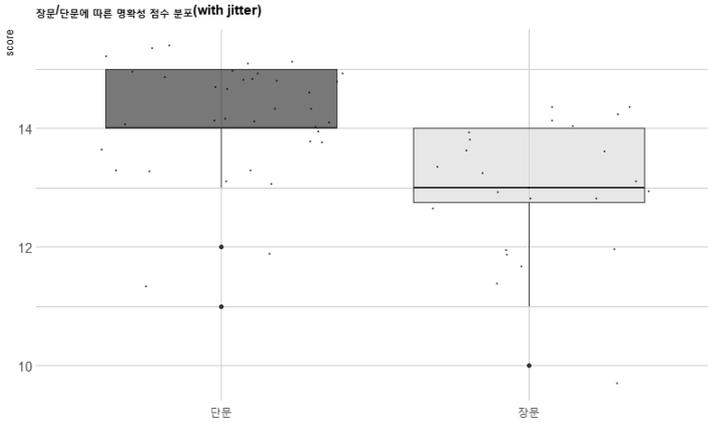
④ 연구팀 자체 수행 장문/단문 명확성 점수 분포 : 장문/단문 간 유의미한 차



〈그림 8〉 장문/단문에 따른 유창성 점수 분포

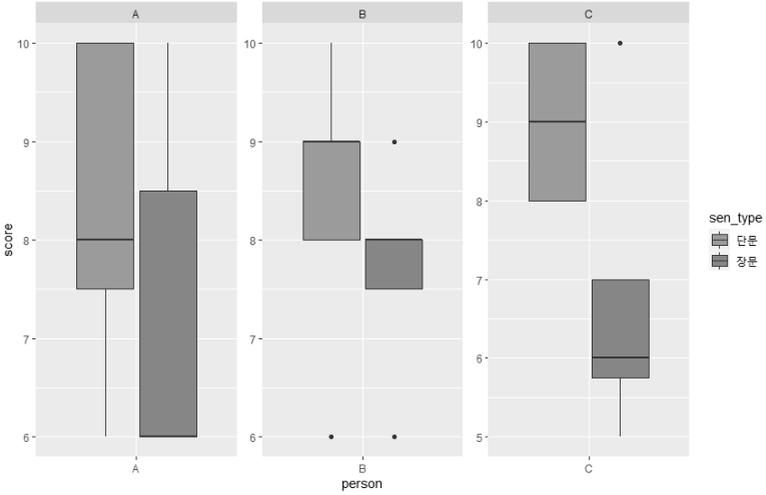
이가 있었다.

문장종류	최소	25%	50%	75%	최대	평균	분산
단문	11.00	14.00	14.00	15.00	15.00	14.19	0.980
장문	10.00	12.75	13.00	14.00	14.00	13.00	1.063
전체	10.00	13.00	14.00	15.00	15.00	13.72	1.166



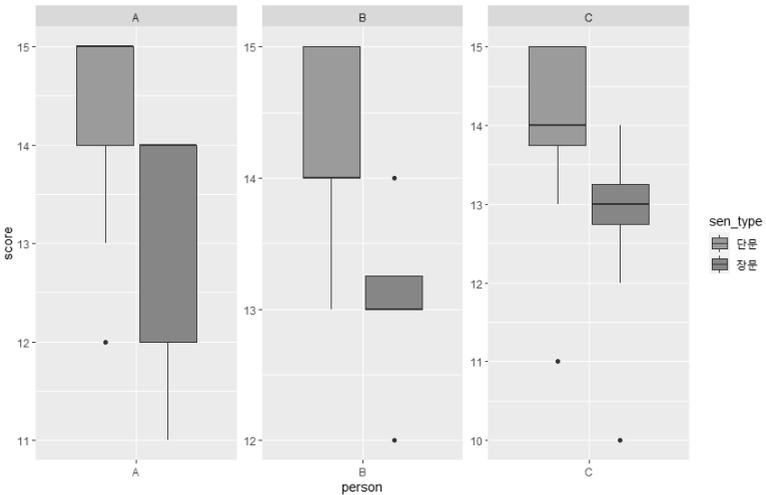
〈그림 9〉 장문/단문에 따른 명확성 점수 분포

⑤ 연구팀 자체 수행 평가자별 유창성 점수 분포(단문/장문)



〈그림 10〉 연구팀 자체 수행 평가자별 유창성 점수 분포(단문/장문)

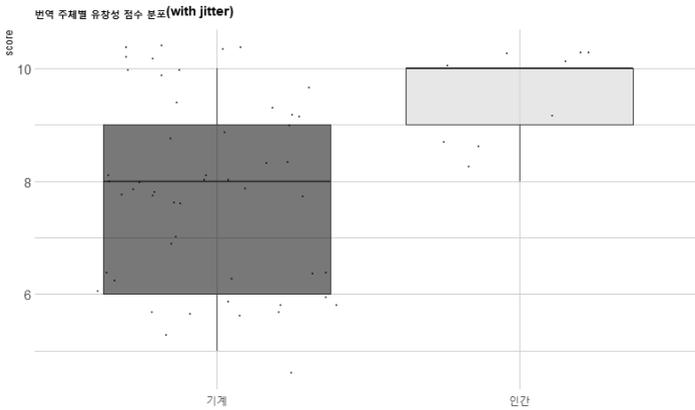
⑥ 연구팀 자체 수행 평가자별 명확성 점수 분포(단문/장문)



〈그림 11〉 연구팀 자체 수행 평가자별 명확성 점수 분포(단문/장문)

⑦ 연구팀 자체 수행 번역 주체별 유창성 점수 분포(기계/인간) : 번역 주체별에 따라 유의미한 차이가 있었다.

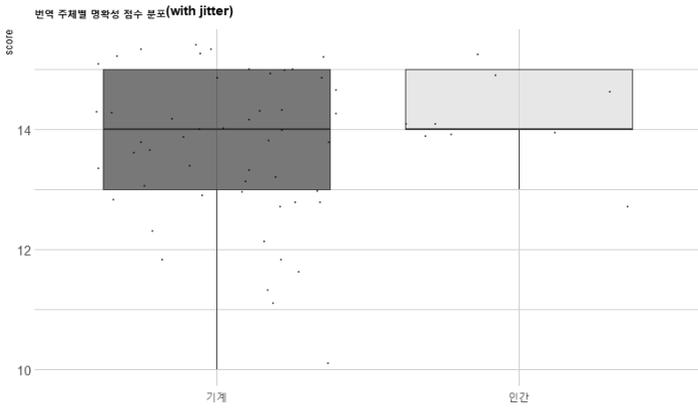
번역주체	최소	25%	50%	75%	최대	평균	분산
기계	5.00	6.00	8.00	9.00	10.00	7.82	1.558
인간	8.00	9.00	10.00	10.00	10.00	9.44	0.726
전체	5.00	6.00	8.00	9.25	10.00	8.07	1.572



〈그림 12〉 번역 주체별 유창성 점수 분포

⑧ 연구팀 자체 수행 번역 주체별 명확성 점수 분포(기계/인간) : 번역 주체별 유의미한 차이가 없었다.

번역주체	최소	25%	50%	75%	최대	평균	분산
기계	10.00	13.00	14.00	15.00	15.00	13.63	1.216
인간	13.00	14.00	14.00	15.00	15.00	14.22	0.667
전체	10.00	13.00	14.00	15.00	15.00	13.72	1.166



〈그림 13〉 번역 주체별 명확성 점수 분포

A Study on the Evaluation Method of Machine Translation in Classical Chinese

Jung, Sunghoon* · Ha, Jiyoung** · Kim, Woojeong***

The purpose of this paper is to examine several methods of translation quality evaluation on the classical chinese using machine translation, and suggest some ways to increase the objectivity of quality evaluation and improve the quality of translation. The classical chinese, an isolated language, have diverse styles and complicated grammatical changes. In addition, it is not easy to develop a reliable and easy translation quality evaluation model because machine translation should also consider evaluation standards, evaluation purposes, evaluation costs, and types of text. Automatic evaluation does not know which elements of machine translation affect translation quality, and although it can show the highest scoring machine translation model, it does not guarantee validity for machine translation quality. In addition, evaluation criteria may vary depending on the evaluation model and may require a large amount of data. To compensate for this problem, manual evaluation is required, which may have different results depending on the experience or level of the appraiser, understanding of the criteria, and the environment or number of evaluations. Therefore, considering the advantages and disadvantages of automatic and manual evaluation, an evaluation method suitable for the purpose of the machine translator shall be found and applied, but a measure shall be developed to objectively evaluate the performance of the machine translation model. And ultimately, these evaluation methods should help identify

* First Author, Assistant Professor, Mokpo National University / kobe99@mokpo.ac.kr

** First Author, Research Professor, Ewha Womans University / zawal@sejong.ac.kr

*** Corresponding Author, Professor, Dankook University / E-mail: rtoran@dankook.ac.kr

and improve the problems of the machine translation model.

Keywords: classical chinese, machine translation, automatic evaluation, manual evaluation, BLEU, METEOR

본 논문은 2021년 9월 9일 투고되어 2021년 10월 21일 심사를 완료하여 2021년 10월 26일에 게재를 확정하였음

www.kci.go.kr

