http://dx.doi.org/10.15749/jts.2020.21.1.001

번역학연구
2020 봄호 제21권1호

번역자동평가에서 풀리지 않은 과제*

김보영· 김연주· 서승희· 송신애· 이진현· 전경아· 최지수· 홍승빈· 정혜연** (한국외대)
허탁성 (한림대)

1. 번역평가의 어려움

좋은 번역평가란 무엇인가. 이 문제를 놓고 오랫동안 많은 연구가 이루어졌다. 이 중에는 번역품질에 대한 이론적 접근도 있었고 (House 1997: Jakobsen 2011 등), 구체적 평가방법을 논하는 실질적 접근도 있었다 (PACTE 2008 등). 나아가 실제 적용 가능한 번역평가 모델이 개발되기도 하였다 (호주의 NAATI, 미국의 J2450, 캐나다의 Sical, CTTIC 등). 하지만 아직까지 세계적으로, 혹은 국내에도 널리 통용되는 '하나의' 번역평가 방법은 없다. 번역평가 방식이 이렇게 통일되지 않는 이유에는 여러 가지가 있겠지만 무엇보다 번역평가 결과에 영향을 미칠 수 있는 변수가 많다는 것이 중요한 원인일 것이다. 번역평가자의

* 이 연구는 (2019학년도) 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것임
** 주저자, 교신저자

배경, 평가모델의 기준, 평가자의 성실성, 텍스트 종류, 언어, 언어조합 등 그러한 변수로 이에 따라 번역평가 결과가 달라질 수 있기 때문이다. 이렇듯 변수를 통제하기 어렵다보니 신뢰성 높은 번역평가 모델을 개발하기가 쉽지 않다.

더 큰 문제는 이런 변수를 잘 통제할 수 있는 번역평가 방법이 개발된다 하더라도 그 방법을 적용하기 쉽지 않다는 점이다. 변수 통제를 위해 정교한 방식을 적용해야 할 텐데 이렇듯 정교한 평가에는 보통 많은 노력이 들기 때문이다.

객관적 평가는 이렇듯 어려운데 평가의 객관성을 요구하는 목소리는 점점 높아지고 있다. 사회 전반적으로 공정성과 투명성을 요구하는 목소리가 높아지면서 번역평가의 공정성과 과정, 결과의 투명성에 대한 관심이 높아진 것이다. 그러다보니 평가자의 전문성, 경험을 뛰어넘는 보다 객관적인 평가 근거가 필요해졌다. 여기서 객관적 평가란 타당성과 신뢰성을 갖춘 평가를 말한다.

본고에서는 비교적 적은 노력으로 보다 객관적인 평가를 할 수 있는 한 가지 방법으로 기계에 의한 자동평가를 제안하고자 한다. 그 중에서도 실효성을 인정받아 널리 사용되고 있는 BLEU와 METEOR를 연구대상으로 삼는다. 두 시스템은 기계에 의한 자동평가이기 때문에 (정답번역이 바뀌지 않는 한) 언제 누가 평가하더라도 같은 결과가 나온다. 신뢰성(reliability)이 어느 정도 보장된다는 의미이다. 따라서 본 연구에서는 이 시스템의 타당성(validity)에 더 무게를 두고자 한다. 그리고 자동평가 시스템의 타당성, 신뢰성을 향상시키기 위한 방법이 있는지를 알아보는 것이 본고의 궁극적 목적이다. 그 과정에서 두 시스템이 영어뿐 아니라 아랍어, 일본어 등의 다양한 출발어에 대해 비슷한 수준의 타당성, 신뢰성을 갖는지도 확인한다. 이를 위해 먼저 두 시스템의 원리를 간단히 설명하고 이 과정에서 두 시스템이 가지고 있는 미결과제를 질문 형식으로 소개한다. 실험 부분에서는 번역평가 데이터 분석을 통해 이 질문에 대한 답변을 시도한다.

2. 정량평가 vs 정성평가

실제 번역평가에 사용되고 있는 기존 모델을 살펴보면 종류는 다양하지만 평가방법에 따라 크게 정량적(quantative) 방법과 정성적(qualitive) 방법으로 구

분되는 것을 볼 수 있다 (Williams 2001). 이 방식은 평가모델에 따라 오류분석 기반(error analysis-based), 전체적(holistic) 방식이라는 이름으로 불리기도 한다 (Waddington 2001b).

이 중 정량적 방식은 평가항목을 정해놓고, 각 항목마다 얼마큼의 점수를 더하거나 감할까를 결정하는 방식으로 평가를 하는 것이다. 즉, 어떤 종류의 항목을 정하고 항목마다 점수 폭을 어떻게 잡느냐에 따라 평가가 달라진다고 할 수 있다. 주로 번역 텍스트의 길이가 길지 않은 교육현장에서 많이 사용하는 방식이다 (Waddington 2001b). 정량평가는 평가자마다 오류를 정의하는 시각은 조금씩 다르지만 전체적으로 보았을 때 평가자 간의 점수 차이는 크지 않다 (Waddington 2001a; Lai 2011; Kunilovskaya 2015). 이렇듯 정량평가는 누가, 언제 하더라도 크게 결과가 변하지 않는다는 점이 장점이지만 오류 감점 방식이다 보니 번역을 바라보는 시각이 미시적이다. 이런 미시적 시각으로는 가독성이나 논리성과 같은 번역 전체 품질을 평가하기 어렵다. 또 오류를 찾아 일일이 점수를 감해야 하기 때문에 시간과 노력도 많이 들어가는 것도 단점이다.

반면, 정성평가의 경우, 평가자의 거시적 관점에서 번역을 평가하므로 가독성, 논리성을 평가하기 좋고 정량적 방식에 비해 평가가 복잡하지 않다는 장점이 있다. 하지만 평가자의 직관이 작용하기 때문에 평가자 간의 점수 차이가 날 수 있다는 단점도 있다. 번역 길이가 길고 단순한 의미 오류만으로 평가할 수 없는 가치를 판단해야 하는 문학번역 등에서 자주 쓰이는 방식이다 (박혜주 2007; 이영훈 2010; 조성원 2007).

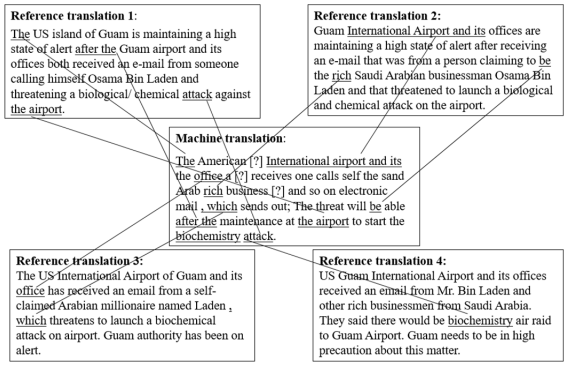
얼핏 생각할 때 두 가지 평가방식의 장점을 결합할 수 있다면 이상적인 평가를 할 수 있을 것으로 보인다. 그러나 관련 연구결과를 살펴보면 반드시 그렇지 않다. 대표적으로 정량평가와 정성평가를 비교한 Waddington (2001b)에서는 순수한 정량적 평가방법이 순수한 정성적 평가방법이나 정량, 정성을 70:30으로 섞은 방법에 비해 번역능력과 가장 높은 상관관계를 보였다 (순수 정량평가: 630** / 순수 정성평가: 578** / 정량, 정성평가 70:30 혼합방식: 623**). 그것도 특히 특정 오류에 가중치를 두지 않은[1] 단순한 오류 감점방식이 가장 상관관계가 높았다. 정량평가와 정성평가를 결합한 방식이 가장 높은 상관관계를 보일 것이라는 예상이 빗나갔던 것이다. 정성평가는 더 간편한 방식이지만 전문가의 직관에 의거하는 만큼 평가의 신뢰성을 보장할 수 없다. 같은 이유로 타당성의 근거도 분명하다 할 수 없다 (내용타당도만 확보). 이를 볼 때 평가의 신뢰성을 높이기 위해서는 정량평가의 비율을 70% 이상으로 확대하는 것이 안전해 보인다. 다만 이렇게 신뢰성 좋은 정량평가를 간편하게 하는 방법은 없는 것일까?

1) Waddington (2001b)은 순수한 정량적 방법을 크게 두 가지로 나누었다. 하나는 모든 오류를 똑같이 취급하는 것이고, 또 하나는 순수한 언어오류보다 번역오류에 더 큰 가중치를 두어 감점하도록 하는 방식이다. 후자에서는 번역오류가 번역품질에 미치는 부정적 영향을 점수화하여(-2~-12) 반영하도록 했다.

3. 정량평가 도구로서의 기계

오류의 개수를 세고 그에 따라 점수를 매기는 일은 인간에게 번거로운 일이지만 기계에게는 간단한 과제일 수 있다. 매칭 여부의 확인과 정량화는 잘 알려진 기계의 장점이기 때문이다. 문제는 기계가 어떻게 오류가 오류인 것을 알아보고 이를 수치화하느냐이다. 기계가 오류를 판단하는 방식은 '정답'과의 매칭을 통해 이루어진다. 미리 마련된 정답('참조번역')과 평가대상인 번역('후보번역')을 비교하고 두 번역이 얼마나 일치하는지를 계산하여 이를 수치로 나타내는 것이다.

〈그림 1〉 참조번역과 후보번역 (BLEU의 예)[2]



2) 이미지 출처: https://dalpo0814.tistory.com/46

자동평가 시스템은 기계번역을 평가하기 위해 개발되었기 때문에 보통 참조번역은 인간이 한 번역, 후보번역은 기계번역을 말한다. 그러나 본고는 인간번역을 평가하므로 참조번역은 예를 들어 교사(전문가)의 번역, 후보번역은 학생번역이 되겠다.

자동평가 시스템은 정답과의 일치도를 판단하는 기준(어절 일치, 통사 일치, 의미의 벡터값 일치, 오류수정 개수)과 계산하는 방식에 따라 여러 종류로 나뉜다(BLEU, METEOR, LePOR, BLEUmod. WER 등). 본고에서는 그 중에서도 대표적으로 널리 사용되는 BLEU와 여러 연구(Banerjee & Lavie 2005; Denkowski & Lavie 2011; Chung 2020)에서 그보다 좋은 결과를 보인 METEOR를 소개하고자 한다.

3.1 BLEU

BLEU는 크게 두 요소로 구성되어 있다. 하나는 참조번역과 후보번역 간의 1~4그램 일치도이고, 나머지는 후보번역이 참조번역보다 짧을 때[3] 가해지는 벌점이다. 직관적으로 말하자면 두 번역 간에 단어와 구, 절, 나아가 텍스트 전체가 얼마나 일치하는지를 계산하는 것이다. 일치도가 높으면 높은 점수를 주고 후보번역이 너무 짧으면(텍스트 전체 기준) 감점을 한다.

$$\text{BLEU}^{4)} = \underbrace{(\text{벌점(Brevity Penalty)})}_{\text{번역이 짧으면 감점}} \quad \underbrace{(\text{1~4그램 일치도(Precision)})^{1/4}}_{\text{정확도}}$$

1그램의 일치는 한 단어의 일치인 셈이니 의미, 표현 등의 정확성을 본다고 할 수 있다. 2그램~4그램의 일치는 두 단어~네 단어의 일치이니 구나 절 같은 통사단위의 정확성을 판단하는 도구가 될 수 있다. 그 밖에도 2그램~4그램 일치는 가독성 판단하는 데에 쓰일 수도 있겠지만 언어의 형태에만 기초한 판단이니 매우 조악한 방법이라 할 수 있다.[5] 한편, 짧은 번역 벌점은 두 번역의 길이가 같으면 1, 후보번역이 짧으면 그보다 작은 수를 곱하도록 함으로써 지나치게 짧은 번역에서 점수를 감하는 장치이다.

3) 기계번역에서는 누락이 자주 발생하기 때문에 생긴 벌점이다.

4) 위는 계산식의 의미를 이해하기 쉽게 풀어쓴 것으로 실제 계산법은 아니다. BLEU와 METEOR의 실제 계산법과 예시는 부록 및 김유섭 (2017) 참조.

5) 1~4그램의 일치도는 정보인출(information retrieval)의 정확도(precision) 를 계산한 것이다. 4그램이라는 수, 가중치 1/4는 Papineni et al. (2002)이 정한 것이다.

3.2 METEOR

METEOR 역시 두 요소로 구성되어 있다. 하나는 1그램 일치도이고 나머지는 벌점이다. 얼핏 BLEU와 비슷해 보이지만 일치도, 벌점의 계산법이 다르다.

$$\text{METEOR} = \underbrace{(1 - \text{벌점(Penalty)})}_{\text{정답과 완벽 일치할 때 가장 낮음}} \quad \underbrace{(\text{1그램 일치도(Fmean)})}_{\text{정확도와 재현도의 평균}}$$

BLEU에서 일치도가 정확도(precision, P)만을 의미했다면 METEOR의 일치도는 정확도, 그리고 재현도(recall, R)란 값의 평균치로 나타난다. 여기서 P는 후보번역을 기준으로 일치도를 계산한 것이고, R은 참조번역을 기준으로 일치도를 계산한 것이다(Banerjee & Lavie 2005, 아래 참조). 이해하기 쉽게 설명하자면 P는 학생번역 중 얼마만큼이 정답인지를 보는 것이고, R은 정답번역 중 학생번역에서 나타난 단어가 얼마만큼이나 있는지를 보는 것이다.

$$\text{정확도 (P)} = \frac{\text{참조번역과 후보번역 간의 일치 1그램 수}}{\text{후보번역 1그램 수}}$$

$$\text{재현도 (R)} = \frac{\text{참조번역과 후보번역 간의 일치 1그램 수}}{\text{참조번역 1그램 수}}$$

P값과 R값은 모두 비율이므로 두 값의 평균을 구할 때 비율의 평균인 조화평균(harmonic mean)을[6] 사용한다(Sasaki 2007). METEOR를 개발한 Banerjee & Lavie (2005)는 두 값 중, 정답을 기준으로 하는 R을 훨씬 중요하게 보고 R에 9라는 큰 가중치를 두었다.

6) 일반적으로 알려진 (a+b+...n) ÷ n 와 같은 계산식은 산술평균이라 한다.

〈표 1〉 정확도(P)와 재현도(R)의 비율

P와 R의 조화평균	R에 9의 가중치를 둔 조화평균
$F = \frac{2PR}{R+P}$	$Fmean = \frac{10\,PR}{R+9P}$

한편, 벌점은 대략적으로 말해 1그램 일치와 청크(2그램 이상의 모든 그램, 아래 참조) 일치의 비율을 의미한다. 청크 수가 많아지면 벌점도 높아지고, 1그램 일치가 많아지면 벌점은 낮아진다(Banerjee & Lavie 2005: 68). 참조번역 전체의 어순이 정답번역 전체의 어순과 완벽하게 일치하면 청크 수가 1이 되는데, 이 때 벌점이 가장 낮아진다. 즉, 참조번역과 후보번역의 단어 배치가 완전히 일치하지 않는 한, METEOR값은 벌점에 의해 낮아지게 되어 있다.

참조번역: China is the main diplomatic ally and trading partner for North Korea
후보번역 China is the most important ally and trading partner for North Korea
청크의 수: 2개 (China is the / ally and trading partner for North Korea)

이상을 종합해 METEOR의 평가 원리와 인간의 평가 원리를 비교해보면 다음과 같다. BLEU와 마찬가지로 METEOR에서도 1그램 일치(Fmean)는 주로 '의미'에 맞는 단어, 즉 '표현'을 선택했는지를 평가하는 도구이다 (예: important: 중요한). '누락, 첨가'도 작게는 단어 단위로 이루어진다는 점에서 1그램 일치로 평가할 수 있다. 또 2그램 이상의 일치를 보여주는 벌점은 두 개 이상의 단어, 즉 구, 절을 '문법'과 '논리'에 맞게 번역했는지를 판단한다는 점에서 문법, 논리성, 나아가 가독성을 평가하는 장치라고 할 수 있다 (예: the most important ally: 가장 중요한 동맹국).

아래에서는 BLEU와 METEOR의 계산법을 인간평가 기준과 비교해 <표 2>로 정리해 보았다. BLEU에서도 1그램은 의미와 표현을, 2그램은 문법을, 그 이상을 평가하는 벌점은 논리성과 가독성을 평가하는 장치라고 보았다. 이렇듯 보듯 자동평가의 계산법은 인간평가의 평가방법과 일맥상통하는 점이 있는데,

이것을 보아도 자동평가의 계산법이 인간번역의 품질을 평가할 수 있는 타당성을 어느 정도 가지고 있음을 알 수 있다.

〈표 2〉 자동평가의 타당도 (인간평가 기준과의 비교) (주저자 정리)

	일치도		벌점
BLEU	1그램	2~4그램	짧은 번역 벌점
인간평가	의미, 표현, 누락/첨가	문법	논리성, 가독성
METEOR	Fmean		벌점
인간평가	의미, 표현, 누락/첨가		문법, 논리성, 가독성

4. 자동평가에서 풀리지 않은 과제

위에서 BLEU와 METEOR의 원리에 대해 살펴보았고 두 시스템의 원리에 차이가 있음을 확인하였다. 그렇다면 두 시스템 중 번역평가에 더 적합한 시스템은 어떤 것일까? Banerjee & Lavie (2005), Chung (2020)에서는 METEOR가 더 좋은 성적을 거둔 바 있다. Banerjee & Lavie (2005)의 경우, 중국어와 아랍어의 두 가지 언어를 사용하는 등 데이터의 다양성을 추구했으나 번역데이터가 텍스트가 아닌 문장이었던 점, 평가자가 번역사가 아닌 일반인이었던 점, 평가 방식에 있어 적합성(Adequacy)과 유창성(Fluency)에 대해 1~5점까지 주도록 한 직관적 방식을 사용한 점이 부족해 보인다. Chung (2020)에서는 전문번역사를 평가자로 하고 보다 정교한 평가모델을 사용하였으나, 데이터 크기가 2596 단어로 작았고, 언어를 한 가지만 사용한 게 단점이었다.

한편, 두 시스템은 기계번역을 평가하기 위해 개발되었기 때문에 그 계산식이 인간번역을 평가하는 데에도 적합한 것인지, 적합하지 않다면 어떠한 점을 수정 보완할 수 있는지의 여부를 다각적으로 검토할 필요가 있어 보인다. 계산식의 변화와 관련해서는 다음과 같은 문제를 제기해볼 수 있다.

첫째, 의미의 정확성을 단어 형태의 일치 여부로 판단하는 것이 대표적 문제점이다. 이 문제를 개선하기 위해 이미 여러 해결책이 사용되고 있는데, 동의어나 패러프레이즈를 삽입해 여러 개의 단어를 정답으로 처리하도록 하는 방식

(EBLEU, ParaEval)이 대표적 방식이다. METEOR도 그 방식을 사용하고 있다. 반면 BLEU의 경우, 같은 효과를 위해 여러 개의 참조번역을 정답으로 사용하고 있다.

둘째, 번역이 짧으면 무조건 벌점을 주는 방식은 인간번역에 적합하지 않을 수 있다. BLEU에서 벌점은 기계번역의 특징인 불완전하고 짧은 번역을 벌하기 위해 고안된 장치이다 (각주 3 참조). 하지만 인간번역은 기계번역과 달리, 문법을 완전히 틀리는 경우나 미완성으로 끝나는 경우가 드물다. 오히려 번역의 길이가 짧으면 문체가 군더더기 없이 간결할 가능성이 높아서 더 좋은 번역일 가능성이 크다. 따라서 벌점을 주더라도 짧은 번역 벌점(Brevity Penalty)대신, 길이 벌점(Length penalty)를 사용하는 것이 더 타당해 보인다. 길이 벌점은 짧은 번역뿐 아니라 지나치게 긴 번역 역시도 감점하는 벌점으로 긴 번역의 감점도 짧은 번역의 감점과 계산법이 같다. 정답 번역의 길이와 달라진 폭만큼 '벌'을 받는 것이다. 이 방법은 Han et al. (2012: 443)이 제안한 바 있다.

셋째, METEOR 수식의 P값과 R값의 비율이 적절한지도 의문시되고 있다. 앞에서 설명했듯이 정답번역을 기준으로 하는 R값이 번역평가에 좀 더 의미 있는 값인 것으로 알려졌다. 하지만 실질적으로는 P값과 R값은 각각의 장단점을 가지고 있어 어느 쪽이 번역평가에 더 의미 있는 값인지 단정하기 쉽지 않다. 2004년 논문에서 R의 가치를 강조했던 METEOR 개발자 Lavie도 Denkowski & Lavie (2011)에서는 한 발짝 물러선 입장을 취했다. 무조건 R에 많은 가중치를 두는 대신, 인간평가와의 상관관계를 최대화할 수 있는 방향으로 P값과 R값을 조정하도록 입장을 바꾼 것이다 (Denkowski & Lavie 2011: 87).

넷째, 점수와 등수 평가의 우위도 토론 대상이 될 수 있다. 인간평가에도 절대평가와 상대평가가 있다. 절대평가는 점수로 상대평가는 등수로 표현되는데 자동평가의 점수와 등수는 인간평가의 점수, 등수와 각각 얼마나 높은 상관관계를 가질까? 점수와 등수 중 일관적으로 더 높은 상관관계를 갖는 것이 있다면 그에 따라 자동평가를 절대평가에 활용하는 것이 좋은지, 상대평가에 활용하는 것이 좋은지를 판단할 수 있을 것이다.

이상의 질문을 정리해보면 다음과 같다. (1) BLEU, METEOR 중 무엇이 더 좋은 평가도구인가? (2) 의미의 정확성을 측정할 수 있는 더 좋은 도구는 없을까? (3) BLEU 계산식에서 짧은 번역 벌점 대신 길이 벌점 (Han et al. 2012)

을 주면? (4) METEOR 계산식에서 P값과 R값의 비율은 적절한가? (5) 점수와 등수 중 어느 쪽이 더 좋은가? 이 중 (2), (3)번은 후속 연구를 위해 남겨두고 다음 장에서는 우선 (1), (4), (5)번에 대한 답을 실험을 통해 구해보고자 한다.

5. 실험

본 실험은 4장에서 언급한 (1), (4), (5) 문제에 대한 답을 구하려는 목적으로 진행되었다. 자동평가는 도착어 형태를 기반으로 하기 때문에 출발어의 영향은 미미할 것으로 예상했지만 혹시 언어별 특징이 관찰되는지도 살펴보았다.

5.1 실험 자료 및 방법

5.1.1. 인간평가

출발어는 독일어, 스페인어, 아랍어, 영어, 일본어 5개 언어이고 언어별로 4개의 텍스트가 선정되었다. 텍스트 유형은 경제와 기술에 대한 신문기사였다. 번역은 H대학 통번역대학원 재학생이 맡았는데 언어별로 5명의 학생이 4개의 신문기사를 한국어로 번역했다 (언어별 총 20개 도착텍스트). 평가는 각 언어당 2명의 번역사가 맡았으며 한국외대 번역평가인증 연구팀 (2016)의 평가모델에 의해 평가했다. 정량평가 항목은 '의미', '표현', '문법', '누락/첨가', '텍스트 형식'이었고 정성평가 항목은 '논리성'과 '기능성'이었다. 본 연구에서는 기존 모델을 수정해 사용했는데, 구체적으로는 정량평가의 비율을 높였다. 이는 순수 정량평가가 실제 번역평가에 더 좋은 결과를 가져왔다는 Waddington (2001b)에 의거한 것이었다. 이렇게 미시적 평가 비율을 높이는 것이 평가자 간의 신뢰도를 높이는 데에 도움을 줄 것으로 기대했다. 다만, Waddington (2001b)과 달리, 정성평가도 포함하기로 했는데 그 비율은 같은 연구에 의거해 30%보다 낮게 잡기로 했다. 그리하여 최종적으로 정량평가 95%, 정성평가 5%의 비율로 점수를 산출했다. 평가자는 평가모델에 의해 각각 평가를 하고 그 결과를 평균내고, 두 평가 간의 상관관계를 구하였다.

5.1.2. 자동평가

자동평가는 H대학 융합소프트웨어학과에서 담당했다. 후보번역은 피험자의 번역 20개로, 이를 형태소 처리하여 사용하였고, 참조번역은 언어별 평가자가 제공했다. 그렇게 구한 인간평가와 자동평가를 비교하여 위 질문에 대한 답을 구하고자 하였다. 먼저 (1)번 질문에 대한 답을 구하기 위해 BLEU와 METEOR의 평가 결과와 인간평가의 평균과의 상관관계를 구하였다. 상관관계가 높은 쪽이 더 인간평가에 가까운 좋은 시스템으로 볼 수 있겠다. 그리고 (4)번 질문에 대한 답을 구하기 위해 P값-인간평가 평균, R값-인간평가 평균 간의 상관관계를 구했다. 여기서도 마찬가지로 인간평가와 높은 상관관계를 보이는 값이 평가에 더 중요한 값이라 할 수 있겠다. 여기에 추가로 R값에 9배 가중치를 둔 Fmean을 그렇지 않은 F1과도 비교했다. 마지막으로 (5)번 질문에 답을 구하기 위해 점수와 등수의 상관관계도 각각 비교하였다.

5.2 실험 결과 및 분석

5.2.1. BLEU와 METEOR

5개 언어 별로 BLEU-인간평가, METEOR-인간평가 간의 상관관계[7]를 구한 결과, 대부분의 경우 METEOR의 결과가 더 좋은 것으로 나타났다. 점수의 경우는 4개 언어에서 (독일어 예외), 등수의 경우는 5개 언어 모두에서 METEOR와 인간평가 간의 상관관계가 더 높았다.

METEOR에는 R값이 크게 반영되어 있고, BLEU은 P값만을 반영하는 만큼 METEOR와 R값, BLEU와 P값은 같은 방향으로 변화하는, 즉 상관관계가 더 높은 성향이 관찰되었다.

7) 자동평가-인간평가와의 상관관계에서는 인간평가자 2인의 평균을 사용했다. 인간평가자 간의 상관관계는 아래와 같이 대부분 높은 수준이었다 (스페인어는 평가자 상관관계가 낮아 1인 평가 결과만 반영). 그렇지 않은 언어의 경우, 자동평가-인간평가의 상관관계도 감안해서 볼 필요가 있다. 독일어: 점수 .941* / 등수 .9*, 아랍어: 점수 .942* / 등수 .8, 영어: 점수 .807 / 등수 .9*, 일어: 점수 .55 / 등수 .7

〈표 3〉 BLEU-인간평가 상관관계

독일어		스페인어		아랍어		영어		일본어	
점수	등수	점수	등수	점수	등수	점수	등수	점수	등수
.728	.9*	.386	.6	.073	.4	.674	.7	-533	-.3

* p<.05 ** p<.01 *** p<.001

〈표 4〉 METEOR-인간평가 상관관계

독일어		스페인어		아랍어		영어		일본어	
점수	등수	점수	등수	점수	등수	점수	등수	점수	등수
.440	.9*	.626	.8	.583	.8	.89*	.9*	-.384	-.1

* p<.05 ** p<.01 *** p<.001

5.2.2. 정확도(P)와 재현도(R)

점수에서는 3개 언어에서 (독일어, 일본어 예외), 등수에서는 5개 언어에서 R이 더 높은 상관관계를 보였다. 이는 P보다 R을 더 중요시 해 METEOR값에 반영한 Banerjee & Lavie (2005)와도 부합한다. METEOR는 R값에 9라는 큰 가중치를 두었기 때문에 METEOR와 인간 상관관계가 높았다는 것(5.2.1)은 R값과 인간평가의 상관관계가 높다는 의미이기도 하다. 다만, R에 무조건 큰 가중치를 두는 게 좋은지는 의문이다. R에 9의 가중치를 둔 Fmean보다 2의 가중치를 둔 F1이 더 좋은 결과를 냈기 때문이다. 점수의 경우, 3개 언어에서 (스페인어, 영어 예외), 등수의 경우, 4개 언어에서 (일본어 예외) F1이 같거나 더 높은 상관관계를 보였다. 이는 F1이 더 좋은 결과를 보인 Chung (2020)과 같은 결과이다.

이를 종합해보면, R값-인간평가의 상관관계가 높은 편이고 따라서 R값을 포함한 METEOR의 평가가 더 타당성이 높지만, 그렇다고 R값에 절대적 비중을 두는 것은 조심해야 할 필요가 있어 보인다 (Denkowski & Lavie 2011).

〈표 5〉 정확도(P)-인간평가 상관관계

독일어		스페인어		아랍어		영어		일본어	
점수	등수	점수	등수	점수	등수	점수	등수	점수	등수
.775	.7	.218	.4	.563	.4	.710	.3	-.402	-.3

* p<.05 ** p<.01 *** p<.001

〈표 6〉 재현도(R)-인간평가 상관관계

독일어		스페인어		아랍어		영어		일본어	
점수	등수	점수	등수	점수	등수	점수	등수	점수	등수
.325	.7	.576	.8	.580	.5	.85	.7	-.422	-.1

* p<.05 ** p<.01 *** p<.001

〈표 7〉 F1-인간평가 상관관계

독일어		스페인어		아랍어		영어		일본어	
점수	등수	점수	등수	점수	등수	점수	등수	점수	등수
.683	.7	.440	.5	.723	.9*	.834	.9*	-.435	-.4

* p<.05 ** p<.01 *** p<.001

〈표 8〉 Fmean-인간평가 상관관계

독일어		스페인어		아랍어		영어		일본어	
점수	등수	점수	등수	점수	등수	점수	등수	점수	등수
.410	.7	.555	.5	.699	.8	.854	.7	-.436	-.1

* p<.05 ** p<.01 *** p<.001

5.2.3. 점수와 등수

5개 언어 중 4개 언어가 등수에서 더 높은 상관관계를 보였다 (영어 예외). 등수가 점수에 비해 수의 편차가 크지 않기 때문이 아닐까 하는 추측을 해볼 수 있다. 점수의 경우, BLEU, METEOR는 1~100 (BLEU에서 0값은 제외시킴) 사이에서, P, R, F1, Fmean은 0~1 사이에서 수많은 값을 가질 수 있지만, 등수의 경우, 1, 2, 3, 4, 5 중 하나의 값이기 때문이다. 그만큼 인간평가와 자동평가가 차이가 크지 않을 가능성이 높다. 이는 많은 학생을 대상으로 상대평가를 해야 하는 번역수업의 경우, 희소식이라 할 수 있다. 1인당 번역물의 수가 충분히 많다는 전제 하에 자동평가의 도움으로 등수를 보다 쉽게 매길 수 있는 가능성이 열렸기 때문이다. 등수 매기기를 전적으로 자동평가에 의존할 수는 없겠지만 번역물만 충분하다면 등수 결정에 자동평가 값을 참조하는 것은 가능하다.

한편, 예상했던 대로 출발어에 따른 특징은 발견되지 않았다. BLEU, METEOR, P, R, F1, Fmean 비교에서 모든 언어가 비슷한 양상을 보였고, 예외도 거의 모든 언어에서 골고루 나왔다. 이는 BLEU와 METEOR를 위시해 본 연구에 사용된 모든 계산법이 기본적으로 도착어의 형태에 기반 하기 때문이라

생각된다. 따라서 도착어에 따라서는 상관관계에 영향이 있을 가능성이 있다. 다른 언어에 비해 상관관계가 유의미하게 높게 나온 언어의 경우, 언어의 영향보다는 평가자와 평가 방법의 적용의 영향이 컸던 것으로 추정된다.

6. 요약 및 결론

본 연구에서는 기계번역 평가에 널리 사용되는 BLEU, METEOR의 평가방법이 타당한지, 인간번역 평가에 활용할 수 있는지를 알아보고, 또 활용할 수 있다면 개선해야할 점이 무엇인지를 알아보았다. 구체적으로는 (1) BLEU, METEOR 중 무엇이 더 좋은 평가도구인가? (2) METEOR 계산식에서 P값과 R값의 비율은 적절한가? (3) 점수와 등수 중 어느 쪽이 더 좋은가 라는 질문을 던졌고 실험을 통해 이에 대한 답을 구하고자 하였다. 실험 결과는 다음과 같았다. 첫째, BLEU와 METEOR 평가를 인간평가와 비교한 결과, BLEU보다는 METEOR가 인간평가와 더 높은 상관관계를 보였다. 둘째, R값이 인간평가와 더 높은 상관관계를 보였다. 하지만 R값의 가중치가 일방적으로 높은 경우, 결과가 부정적이었다. 셋째, 점수보다는 등수가 인간평가와 더 비슷한 것으로 나타났다.

이 결과의 의미를 실질적으로 해석하자면 다음과 같다. 인간번역 평가에 자동평가를 사용하기에는 아직 부족한 점이 많다(아래 한계점 참조). 다만 인간평가 중에서도 정량평가에서는 자동평가 수치를 참조해 볼 수는 있겠는데 이 목적으로는 METEOR가 보다 타당성 높은 평가 시스템으로 보인다. 단, 인간평가에 사용하기 위해서는 METEOR 수식 내 R의 가중치를 조정해 사용해야 할 가능성도 있다. 그리고 절대평가보다는 상대평가에 활용하는 것이 더 바람직해 보인다.

이 연구는 여러 한계점을 안고 있다. 무엇보다 데이터 양이 많지 않은 점이 가장 큰 한계였다. 전체로 보면 100개의 텍스트였지만 평가는 언어별로 이루어져 사실상 상관관계는 20개의 텍스트를 대상으로 구한 것이었다. 이렇듯 데이터 양이 적으면 본 연구에서 보듯 상관관계가 높더라도 그 결과가 유의미하기 어렵다. 그럼에도 불구하고 이 결과를 제시한 것은 데이터 양이 100개 이상으

로 많았을 때 METEOR과 인간평가의 비교에서 상관관계 뿐 아니라 유의미성도 높게 나왔던 선행연구가 있기 때문이다 (Chung 2020). 본 연구도 양질의 데이터가 더 확보될 경우, 더 유의미한 결과를 얻을 수 있을 것으로 기대한다.

그리고 평가의 신뢰도라는 변수도 평가자 훈련 및 신뢰도 검사를 통해 사전에 통제할 수 있다면 보다 분명한 결과를 얻을 수 있을 것으로 보인다. 본 실험에서도 평가자의 평가 방법과 평가자 간의 신뢰도가 인간평가와 자동평가의 상관관계에 큰 영향을 미쳤기 때문이다. 향후 연구에서는 평가 경험이 많은 평가자를 선정하여 이들에게 동일한 평가방법을 익히도록 한 후, 이를 사용하여 평가를 하도록 한다면 평가자 간의 상관관계를 높일 수 있을 것이다.

또 도착어가 아니라 출발어를 다양하게 한 것도 본 연구의 한계라 할 수 있다. 평가가 도착어 형태에 기반 해 이루어지기 때문에 출발어보다는 도착어가 다양했을 때 자동평가의 결과가 달라질 가능성이 크다. 향후에 도착어의 영향을 알아보는 것도 흥미로울 것이라 생각한다.

마지막으로 본 연구의 결과를 실질적으로 활용하는 데에는 여러 가지 전제조건이 있음을 언급하고자 한다. 우선 후보 1인당 많은 번역물을 확보해야 한다. 12개의 텍스트를 번역하도록 해 총 120개의 텍스트에 대해 자동평가-인간평가를 비교 연구한 선행연구에서는 상관관계와 p값이 모두 높게 나왔던 것을 생각하면 한 사람의 번역능력을 자동평가하는 데에 최소 1인당 12개의 번역이 필요한 것으로 보인다. 본 연구에서도 개개 텍스트 평가에서는 자동평가-인간평가의 상관관계가 매우 낮거나 마이너스의 상관관계를 보이기도 했다. 그리고 믿을만한 전문번역사의 정답 번역이 있어야 한다. 정답이 되는 참조번역 없이는 자동평가 자체가 불가능하다. 또 참조번역, 후보번역 모두 기계가 인식할 수 있도록 언어단위 별로 배치, 정렬해야 하는 번거로움이 있다.

향후 이러한 어려움을 극복할 수 있다면 정량평가에 자동평가 결과를 활용해 볼 수 있을 것으로 기대해 볼 수 있다.

참고문헌

김유섭 (2017)「기계 번역에서의 평가」『한국외대 통번역연구소 제18회 국제학술대회 <디지털 시대의 통번역> 발표집』, 337-357.

김진아, 성초림, 이상원, 장현주 & 이향 (2002)「번역품질평가에 관한 소고」.『외국문학연구』11: 85-123.

박혜주 (2007)『문학번역 평가 시스템 연구』한국문학번역원.

이영훈 (2010)「프랑스 명작소설 한국어 번역 연구를 위한 번역평가 시스템 개발」.『통역과 번역』12(2): 149-179.

조성원 (2007)「번역비평의 가능성을 묻는다: 번역평가 기준으로서의 “충실성”과 “가독성”에 대하여 －영미문학연구회 번역평가사업에 대한 소고－」『안과 밖』23(0): 96-121.

한국외대 번역평가인증 연구팀 (2016)「번역인증제도 (실무편)」『한국외대 통번역연구소 학술대회 <언어, 통번역의 평가 및 인증> 발표집』, 23-33.

Banerjee, Satanjeev & Lavie, Alon (2005) ‘METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments’, in *Proceedings of the ACL 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) (to be published).

Denkowski, Michael & Lavie, Alon (2011) ‘Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems’, in *Proceedings of the 6th Workshop on Statistical Machine Translation*, 85-91.

Han, Aaron L. Wong, Derek F. & Chao, Lidia S. (2012) ‘LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors’, in *Proceedings of COLING 2012: Posters*, 441 – 450.

House, Juliane (1997) *Translation Quality Assessment: A Model Revisited*. Gunter Narr: Tübingen.

Jakobsen, Arnt Lykke (2011) 'Tracking Translators' Keystrokes and Eye Movements with Translog', *Methods and Strategies of Process research: Integrative Approaches in Translation Studies.* John Benjamins: Amsterdam, 37-55.

Kunilovskaya, Maria (2015) 'How far do we agree on the quality of translation?' *English Studies at NBU*, 1(1): 18-31.

Lai, Tzu-Yun (2011) 'Reliability and Validity of a Scale-based Assessment for Translation Tests', *Meta* 56(3): 713-722.

Lavie, Alon, Sagae, Kenji & Jayaraman, Shyamsundar (2004) 'The Significance of Recall in Automatic Metrics for MT Evaluation.' https://www.cs.cmu.edu/~alavie/papers/Recall-AMTA-04.pdf

PACTE (2008) 'Results of the Validation of the PACTE Translation Competence Model. Decision Making', in *AILA World Congress 2008. Multilingualism, Challenges & Opportunities*, 1-5.

Papineni, Kishore, Roukos, Salim, Ward, Todd & Zhu, Wei-Jing (2002) 'BLEU: a Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311-318.

Sasaki, Yutaka (2007) 'The truth of the F-measure' https://www.cs.odu.edu/~mukka/cs795sum10dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

Waddington, Christopher (2001a) 'Should translations be assessed holistically or through error analysis? Hermesi, *Journal of Linguistics.* 26: 15-37

Waddington, Christopher (2001b) 'Different Methods of Evaluating Student Translations: The Question of Validity', *Meta* 46(2): 311-325.

Williams, Malcolm (2001) 'The Application of Argumentation Theory to Translation Quality Assessment', *Meta* 46(2): 326-344.

[부록] BLEU와 METEOR 계산법

BLEU 계산식

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} Wn \ logPn\right)$$

$$\text{Brevity Penalty (BP)} = \begin{cases} 1, & \text{if } c > r \\ e^{\left(1-\frac{r}{c}\right)}, & \text{if } c \leq r \end{cases}$$

W (weight) - 가중치

n - 그램 수

P (Precision) - 정확도

c (candidate translation) - 후보번역

r (reference translation) - 참조번역

METEOR 계산식

$$\text{METEOR} = \text{Fmean} \times (1 - \text{Penalty})$$

$$Fmean = \frac{10\,\text{PR}}{\text{R}+9\text{P}}$$

$$\text{Penalty} = 0.5 \times \left(\frac{\#chunks}{\#unigram\_matched}\right)^3$$

R (recall) - 재현도

[Abstract]

Application of Automatic Evaluation to Human Translation

Kim, Bo-young, Kim, Yeon-joo, Seo, Seung-hee, Song, Shin-ae, Lee, Jin-hyun, Jeon, Kyoung-ah, Choi, Ji-soo, Hong Seung-bin, Chung, Hye-yeon
(Hankuk University of Foreign Studies, Hallym University),
Heo, Tak-sung (Hallym University)

This paper deals with two questions. The first is whether BLEU and METEOR, which were developed to evaluate machine translation, can also be used for the evaluation of human translations. The second is, how can these systems be adapted to evaluate human translations in a more variable way? These questions can be subdivided into the following questions: (1) What is the more variable evaluation system, BLEU or METEOR? (2) Is the present precision-recall ratio appropriate? (3) Between the grades and ranks of automatic evaluation, which correlates better with human evaluation? Five translator trainees majored respectively in Arabic, German, English, Japanese, and Spanish (a total of 25 students), translated four texts into Korean (a total of 100 texts). The translations were evaluated by two professional translators in each language and their evaluation results were compared with the outcome of the automatic evaluation. The results showed that the METEOR, recall and ranks correlated with the human ratings better than the BLEU, precision and scores. This and other findings from this experiment suggest that with the minimum number of ca. 12 translations, METEOR can be used at least when determining the order of student performance.

▸ Key Words: automatic evaluation, translation quality, validity, reliability, BLEU
▸ 주제어: 자동평가, 번역품질, 타당도, 신뢰도, BLEU

김보영
한국외국어대학교 통번역대학원 한아과 박사과정
byk0419@gmail.com
관심분야: 통번역사 역할, 통번역 교육

김연주
한국외국어대학교 통번역대학원 한아과 박사과정
delphi0817@hanmail.net
관심분야: 통번역, 통번역교육, 아랍어

서승희
한국외국어대학교 통번역대학원 한서과 박사과정
vero81@naver.com
관심분야: 외교통번역, 제도번역

송신애
한국외국어대학교 통번역대학원 한일과 박사과정
shine22.song@gmail.com
관심분야: 통번역학, 통번역교육

이진현
한국외국어대학교 통번역대학원 한일과 박사과정
0219jin@gmail.com
관심분야: 웹툰 번역, 현지화, 통역평가

전경아
한국외국어대학교 통번역대학원 한서과 박사과정
linguista@naver.com
관심분야: 언어학, 문학번역

정혜연
한국외국어대학교 통번역대학원 한독과 정교수
johanna2000@naver.com
관심분야: 통번역과정, 인지심리학, 기계번역

최지수
한국외국어대학교 통번역대학원 한독과 박사과정
jsuuch@gmail.com
관심분야: 인지언어, 담화분석, 기계번역

허탁성
한림대학교 융합소프트웨어학과 석사과정
gjxkrtjd221@gmail.com
관심분야: 기계번역, 자연어 처리, 딥러닝

홍승빈
한국외국어대학교 통번역대학원 한영과 박사과정
sebihong@gmail.com
관심분야: 언어전환, 통번역평가, 통번역교육

논문투고일: 2020년 2월 10일
심사완료일: 2020년 3월 1일
게재확정일: 2020년 3월 11일