http://dx.doi.org/10.15749/jts.2022.23.1.003

번역학연구
23권 1호
(2022. 3. 31)

인간 번역평가에서 재현도(recall)의 중요성*

정 혜 연** · 최 지 수 · 허 탁 성 · 서 수 영***
(한국외대 · 한국외대 · 한림대 · 한림대)

1. 들어가며

정확도(precision)와 재현도(recall)는 원래 정보추출에 사용되는 개념으로(van Rijsbergen 1979), 기계의 성능을 평가하는 데 쓰인다. 예를 들어 불량품을 골라내는 기계가 있다고 했을 때 여기서 결정적인 정보는 제품이 불량품이냐 아니냐의 여부이다. 그리고 기계의 성능은 이 기계가 불량품이냐 아니냐의 정보를 얼마나 많이, 그리고 얼마나 정확하게 찾아내느냐에 달려 있다. 그리고 이것을 판가름하는 것이 정확도, 재현도, 이 두 가지 지표이다. 여기서 기계가 불량품을 있는 대로 모두 찾아낸다면 그 기계는 재현도가 높은 것이고, 내가 정의한 불량품에 더 잘 맞는 불량품을 잘 찾아낸다면 그 기계는 정확도가 높은 것이다.

* 이 논문은 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020S1A5A2A0104676412). / 이 연구는 (2021학년도) 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것임.
** 1저자, 교신저자
*** 본 논문 및 인간번역 자동평가 프로젝트에서 자문을 맡아주신 한림대학교 융합소프트웨어 학과 김유섭 교수님께 이 자리를 빌려 진심어린 감사말씀을 전한다.

정확도, 재현도는 번역평가 분야에도 활용된다. 이 두 지표는 어떤 번역이 얼마나 좋은 번역인지를 평가하는 자동평가 '기계 의 성능을 평가하는 역할을 하는 것이다.[1] 여기서 중요한 정보는 '좋은 번역 이냐, 아니냐 이고 자동평가의 성능을 판가름하는 것은 바로 이 정보를 얼마나 잘 찾아주느냐의 여부이다.

자동평가 시스템 중 널리 쓰이는 BLEU, METEOR에도 정확도, 재현도 지표가 사용된다. BLEU는 정확도만을 가지고, METEOR는 정확도와 재현도 모두를 사용해 좋은 번역을 골라내는데, 최근 연구에서는 정확도, 재현도가 기계번역 뿐 아니라 인간번역을 평가하는 데에도 사용될 수 있음을 보여주기도 하였다(김보영 외 2020; 정혜연, 박헌일, 우경조, 서수영 2021; Chung 2020).

문제는 두 지표를 인간번역 평가에 사용하려면 어떤 지표를 더 중요시해야 하는가이다. 정확도, 재현도 사용에 대해 기계번역 평가 분야에서도 이론이 있지만, 의견이 조금씩 다르고, 또 이론을 뒷받침하기 위해 주로 수식과 데이터만을 제시하고 있어, 직관적 해석이 빠져있다는 인상을 준다. 다시 말해 정확도와 재현도가 실제 번역평가에서 어떠한 의미를 갖는지 실무적인 관점에서 추가적 해석이 필요할 수 있다는 얘기다. 또 관련 선행연구는 주로 기계번역을 다루고 있는데, 본고에서는 인간번역을 주제로 하기 때문에, 두 지표를 인간번역에 사용하려면 기계번역과 인간번역의 차이도 별도로 고민할 필요가 있다. 바로 여기가 인간평가자가 기여할 부분이다. 인간번역의 특징을 잘 알고 다양한 번역을 평가해 본 번역평가자가 정확도와 재현도의 특징을 파악한 후, 본인의 평가 직관과 비교해보고 두 지표를 인간번역에 어떻게 사용할지를 고민하는 것이다.

정확도, 재현도는 인문학에서는 낯선 개념이지만, 향후 기계번역을 이해하는 데에 중요한 의미를 가질 수 있다. 다양한 기계번역기 성능 테스트에 지속적으로 사용되기 때문이다. 최근에도 가장 최신 언어모델인 BERT를 활용한 정확도, 재현도 평가모델이 사용되는 등(Zhang et al. 2020), 꾸준히 기계번역평가를 포함한 자연어처리 성능평가 분야에 활용되고 있다. 이렇듯 두 지표가 자동평

1) 일각에서는 자동평가의 필요성에 대해 의구심이 제기되기도 하는데 예로써 많은 학생이 수강하는 번역수업에서 자동평가가 유용하게 쓰일 수 있다. 한 번역수업을 학생 50명이 수강한다 할 때, 중간, 기말고사에 5문항만 출제한다고 해도 500개의 번역을 평가해야 한다. 교사 한 명이 이를 빠르고 일관되게 채점하는 건 어렵지만 (여럿이 나눠하는 경우도 일관성을 보장하기 어렵다), 기계의 도움을 받으면 가능하다.

가에 많이 사용되는 만큼 향후 자동평가를 활용하기 위해서는 본고의 주제를 꼭 한 번은 고민해보아야 한다.

본 논문은 기존 기계번역 평가에 사용된 정확도, 재현도의 사용 방법이 인간번역 평가에도 타당한지(인간평가에 사용하기 적합한지)를 이론적으로 검토하고, 실험을 통해 정확도, 재현도 각각의 중요성을 알아본다. 그리고 번역평가에 있어 정확도와 재현도의 비율을 어떻게 하는 것이 바람직한지 고민한다.

이는 BLEU나 METEOR와 같은 자동평가 시스템 사용에도[2] 중요한 문제이지만, 나아가 인간의 번역평가에도 시사하는 바가 있다.

2. 정확도(precision)와 재현도(recall)

정확도과 재현도의 개념을 이해하기 위해서는 먼저 네 가지 경우의 수를 설명해야 한다. 기계의 성능을 평가하려면 기계가 내가 필요한 정보를 얼마나 잘 찾는지를 판단해야 하는데, 이 때 아래의 네 가지 경우의 수가 발생하기 때문이다. 예를 들어 기계가 실제 내가 찾는(positive) 정보를 맞게 예측하는 경우, 이를 True Positive(TP)라고 한다(우측 하단). 그 이외에도 TN, FP, FN의 경우가 있는데, 이 네 가지 경우의 수의 수치에 따라 기계의 성능이 판가름 난다.

〈표 1〉 정확도와 재현도

		예측	
		Negative	Positive
실제	Negative	Negative를 Negative라 예측 True Negative (TN)	Negative를 Positive라 예측 False Positive (FP)
	Positive	Positive를 Negative라 예측 False Negative (FN)	Positive를 Positive라 예측 True Positive (TP)

2) 자동평가는 아직 개별 텍스트 차원에서 인간평가만큼 타당성이 높지 않다(Chung 2020). 한 사람이 여러 텍스트를 번역했을 때, 그 텍스트를 모두 평가하여 그 사람의 번역 수행능력을 평가하는 것만이 가능하다. 따라서 자격시험처럼 중요한 시험보다 수업에서 학생 수행평가에 활용하는 것이 더 안전하다. 따라서 본 논문의 논의도 학생 수행평가 상황을 상정하는 것으로 하겠다(교사번역-학생번역 비유도 그 이유).

$$\text{정확도} = \frac{TP}{FP + TP} \quad (1)$$

$$\text{재현도} = \frac{TP}{FN + TP} \quad (2)$$

기계의 성능이 좋은 경우, 기계는 실제 내가 찾는 정보를 맞게 예측한다(TP). 따라서 정확도, 재현도 모두가 이 경우의 수(TP)를 공통적으로 포함한다. 차이가 있다면 정확도의 경우는 실제로는 아닌데 맞다고 오판한 경우(FP)를 포함하고(식 (1)), 재현도의 경우는 실제로는 맞는데 아니라고 오판한 경우(FN)를 포함시킨다는 것이다(식 (2)).[3)]

이 개념은 다소 추상적이어서 이해를 돕기 위해 번역평가의 예를 들어보겠다. 여기서 내가 찾는 정보는 '좋은 번역 이다. 위의 네 가지 경우의 수도 번역평가의 예로 바꾸어 다시 설명하려 한다. 이 네 가지 경우의 수는 혼동하기 쉽다 보니 좋은 방법은 아니지만 기억하기 쉽게 아래와 같이 설명해보겠다.

True (실제로 좋은 번역을) Positive (좋은 번역으로 판단)
True (실제로 나쁜 번역을) Negative (나쁜 번역으로 판단)
False (잘못해서 나쁜 번역을) Positive (좋은 번역으로 오판)
False (잘못해서 좋은 번역을) Negative (나쁜 번역으로 오판)

기계는 좋은 번역을 실제 좋은 번역이라고 예측해야(TP) 성능이 좋은 것이다. 정확도, 재현도 모두 이 경우의 수를 분자로 갖는다. 차이는 정확도의 경우, 실제로는 나쁜 번역인데 좋은 번역이라고 오판하는 경우(FP)가 분모에 포함된다는 점이고, 재현도는 실제는 좋은 번역인데 나쁜 번역이라고 오판하는 경우(FN)가 분모에 포함된다는 점이다.

이를 교사, 학생번역의 예로 설명하면 이해가 쉽다. 교사번역은 정답이고, 학생번역은 오답이라고 단순화 해 보자. 교사번역은 좋은 번역이라고 했으니,

3) 실제 인간의 번역평가는 이와 같이 이분법적으로 단순화하기에는 무리가 있다. 그럼에도 이분법적 개념인 정확도, 재현도를 사용하려는 이유는 (번역양이 많은 경우에 한하지만) 그 결과가 쓸 만했기 때문이다. 실제 실험에서 인간평가와 자동평가의 상관관계가 0.8 이상이 나왔다(Chung 2020).

실제 좋은 번역으로 판단된 번역(TP)과 잘못해서 나쁜 번역으로 오판된 번역(FN)을 더한 값이고, 학생번역은 나쁜 번역이라 했으니 실제 나쁜 번역으로 판단된 번역(TN)과 잘못해서 좋은 번역으로 오판된 번역(FP)을 더한 값이다.

여기서 값이란 유니그램 수(본고에서 형태소 단위)를 말하는 것이다. 즉, 번역평가에서 재현도는 교사번역 전체 중에서 교사번역과 학생번역이 일치하는 형태소 수를 측정한 것이고, 정확도는 학생번역 전체 중에서 두 번역이 일치하는 형태소 수를 측정한 것이다. 위 설명으로 정확도, 재현도를 번역평가에 맞추어 다시 정의하면 아래와 같다.

$$\text{정확도} = \frac{\text{교사번역} \cap \text{학생번역}}{\text{학생번역}}$$

$$\text{재현도} = \frac{\text{교사번역} \cap \text{학생번역}}{\text{교사번역}}$$

3. 번역평가에서 정확도와 재현도 사용의 문제

문제는 정확도과 재현도, 두 지표가 마이너스 상관관계를 갖는다는 점이다. 즉, 한쪽이 높아지면, 다른 한쪽이 낮아지는 경향이 있다(권철민 2020: 159). 두 지표 모두 좋은 번역을 판단하는 데에 중요한 기준이지만, 두 지표가 동시에 높게 나오기는 어렵단 얘기이다(짧은 구간이지만 불가능하지는 않다고 한다. Buckland and Gey 1994). 그래서 보통 두 지표 중 하나만 쓰든지, 한 쪽에 가중치를 두든지 혹은 평균을 내어 쓴다(Sasaki 2007의 설명, Han 2018의 예).

어느 지표가 더 중요한지는 과제의 성격에 따라 달라질 수 있다. 과제마다 기계가 오판을 할 경우, 발생하는 비용이 다른데, 이 비용 중 어느 쪽이 더 큰지에 따라 지표의 중요성이 결정된다는 것이다. 관련 문헌에서 자주 언급되는 암 진단의 사례를 보면, 실제 암인데 암이 아니라고 오판(FN)하는 경우가 암이 아닌데, 암이라고 오판하는 경우(FP)보다 더 치명적이다. 후자에서는 다른 병원에서 다시 진단 받으면 그만이지만, 전자는 치료 시기를 놓쳐 목숨을 잃을 수도 있기 때문이다. 이 경우, 중요한 정보인 FN를 포함한 재현도가 더 중요한 지표

가 된다. 더 치명적인 실수인 FN 값(분모)이 낮아질수록 재현도 값이 커지니(식 (1) 참조) 재현도는 이 치명적 실수를 줄일 수 있는 지표이다.

반대로 스팸메일을 골라내는 경우에서는 스팸메일이 아닌데 스팸메일이라고 오판하는 경우(FP)가 반대의 경우보다 더 치명적이다. 후자의 경우는 스팸메일이 일반 메일함으로 왔으니 삭제해버리면 그만이지만, 전자의 경우는 중요한 비즈니스 메일인데 스팸메일함으로 가서 받지 못하는 경우가 생기기 때문이다. 따라서 여기는 FP를 포함한 정확도가 더 중요한 지표가 된다. 여기서도 더 치명적 실수인 FP(분모)가 낮아질수록 정확도 값이 커지니(식 (1) 참조) 정확도는 이 치명적 오판을 줄여주는 지표가 된다(권철민 2021: 157).

문제는 본고의 연구주제인 번역평가가 암 진단처럼 암이다, 암이 아니다 라고 정확히 구분 지을 수 있는 성격의 과제가 아니란 점이다. 번역의 좋고 나쁨은 많은 경우 정도의 문제이고 분명한 오류 0점부터 완벽한 정답 100점 사이의 스칼라(scala) 상에서나 표현될 수 있다. 예를 들어 BLEU에서 좋은 번역, 나쁜 번역을 구분하기 위해 교사번역 같은 정답을 기준으로 두는데, 이 정답이 진짜 정답이라는 보장이 없다. 평가대상인 학생번역도 충분히 좋은 번역일 수 있고, 오히려 더 좋은 번역일 수도 있다. 따라서 정답번역만을 기준으로 하는 재현도만 포함시키는 것은 위험할 수 있다. 나쁜 교사번역을 포함시키고 좋은 학생번역을 놓치는 위험을 피하기 위해서는 차라리 재현도와 정확도를 적절히 섞어서 평가하는 게 더 좋은 방법일 수 있다. 어느 한쪽이 전체 평가점수에 절대적 영향을 미치지 못하도록 하기 위함이다. 이렇게 되면 정확도나 재현도 개개 영향력은 줄지만, 둘 중 하나만 반영함으로써 야기되는 위험도 피할 수 있다.

4. 번역자동평가에서 정확도와 재현도

그렇다면 지금까지 기계번역 평가에서는 이 두 지표를 어떻게 사용해왔을까? 먼저 대표적 자동평가 시스템인 BLEU에서는 두 지표 중 정확도만을 사용해왔다. 파피네니 외(Papineni et al. 2002)는 그 이유에 대해 재현도가 전체 점수를 왜곡시킬 수 있기 때문이라고 하였다.[4] 즉, 재현도는 정답번역을 많이 포함할수록 높기 때문에 무조건 정답과 일치하는 번역이 수적으로 많은 번역에

더 높은 점수를 부여한다는 것이다. 실제 아래의 기계번역의 예를 보면, 후보1이 정답과 일치하는 단어 수가 더 많지만, 번역품질은 오히려 더 좋지 않다.

정답1: I always do. / 정답2: I invariably do. / 정답3: I perpetually do.
후보1: I always invariably perpetually do.
후보2: I always do. (Papineni et al. 2002)

그러나 인간번역은 다르다. 인간번역은 후보1처럼 인간의 직관에 어긋나는 번역을 하는 경우는 드물다. 따라서 재현도를 배제할 이유가 딱히 없어 보인다.

한편, 또 다른 자동평가 시스템인 METEOR는 파피네니 외(Papineni et al. 2002)를 비판하며 정확도, 재현도를 모두 사용했다. 약 1800개의 문장에 대해 인간평가와 정확도, 재현도 평가를 비교해 보고 실제 재현도가 인간평가와 더 높은 상관관계를 보임을 입증하였다(정확도: 0.573, 0.666 / 재현도: 0.954, 0.923). 그래서 라비에, 새게이, 재야러맨(Lavie, Sagae and Jayaraman 2004)은 METEOR를 개발하면서 재현도를 단순히 포함시키는 데에서 그치지 않고, 재현도에 9배라는 큰 가중치를 두었다(Banerjee and Lavie 2005).

그리고 이렇게 재현도에 9배의 가중치를 둔 평균을 Fmean이라고 하였다. Fmean은 정확도와 재현도의 조화평균(간단하게 비율의 평균)인데, 재현도에 9배의 가중치를 둔 것이다. 두 값을 같은 비중으로 평균을 낸 조화평균은 F1이라고 부른다(이상 <표2> 참조). 라비에, 새게이, 재야러맨(Lavie, Sagae and Jayaraman 2004)에서는 Fmean이 F1에 비해 인간평가와 다소 높은 상관관계를 보였으나, 우위가 뚜렷하지 않았다(F1: 0.881, 0.950 / Fmean: 0.954, 0.940).

〈표 2〉 정확도(P)와 재현도(R)의 비율

P와 R의 조화평균	R에 9의 가중치를 둔 조화평균
$F1 = \frac{2PR}{R+P}$	$Fmean = \frac{10PR}{R+9P}$

4) 그래서 파피네니 외(Papineni et al. 2002)는 재현도를 버리는 대신, 여러 개의 정답번역(reference translations)과 짧은 번역 벌점(brevity penalty)을 둠으로써 원문의 내용이 되도록 빠짐없이 반영되도록 하였다.

라비에, 새게이, 재야러맨(Lavie, Sagae and Jayaraman 2004)은 재현도의 중요성을 강조하면서 대략 세 가지 이유를 제시하고 있다. 첫째, 재현도 자체가 갖는 의미이다. 재현도는 정답번역을 기준으로 하기 때문에(분모에 정답번역) 정답의 내용이 얼마나 평가에 반영되는지를 보는 지표이다(Lavie, Sagae and Jayaraman 2004:3). 후보번역을 기준으로 하는 정확도와는 다르다는 의미이다. 둘째, BLEU에서는 정답번역을 기준으로 하는 재현도를 사용하는 대신에 여러 개의 정답번역을 사용하고 있지만, 후보번역과 정답번역은 어차피 개별 텍스트 차원에서만 비교되고, 그 중 가장 매치율이 높은 경우만을 수식에 사용하기 때문에 후보번역의 좋은 번역 모두가 좋은 번역으로 평가되는 것은 아니다. 즉, 재현도를 배제한 BLEU는 좋은 번역 커버율이 낮다는 것이다(동일 문헌 2-3, 7). 셋째, 라비에, 새게이, 재야러맨(Lavie, Sagae and Jayaraman 2004)에서는 이렇게 구체적으로 설명되지는 않았지만) 짧은 번역 벌점도 단순히 전체 텍스트의 길이를 기준으로 후보번역이 정답번역보다 더 짧은 경우에만 벌점을 주기 때문에 개별 문장에 좋은 번역의 내용이 얼마나 많이 반영되는지를 가늠하는 지표가 되기는 부족하다(동일 문헌 3).

라비에, 새게이, 재야러맨(Lavie, Sagae and Jayaraman 2004)이 제시한 이유 이외에도 앞서 설명한 네 가지 경우의 수로도 재현도의 중요성을 따져볼 수 있을 것 같다. 좋은 번역을 나쁜 번역이라고 오판하는 것(FN)과 나쁜 번역을 좋은 번역이라고 오판하는 것(FP) 중 어느 쪽이 더 치명적인 실수일까? 전자의 경우라면 암 진단 예시에서 보았듯 재현도가 더 중요한 지표이겠고, 후자의 경우라면 스팸메일 분류 예시에서 보았듯 정확도가 더 중요한 지표가 된다.

5. 인간 번역평가에서 중요한 건 재현도일까 정확도일까

정확도, 재현도의 중요성을 논의하기에 앞서 인간평가에 대해 먼저 논해보자. 인간평가자는 머릿속에 저마다 일종의 정답을 두고, 그 정답과 비교해 번역을 평가한다. BLEU와 같은 자동평가도 이러한 인간의 평가 원리를 닮아 있다. 차이점이 있다면 인간의 정답은 '좋은 번역 이라는 범위 내에서 다양하다는 것이다. 그러다 보니 누구나 동의하는 절대적 기준의 '좋은 번역은 찾기 어렵다.

상황에 따라(평가 목적, 평가자 등) 좋은 번역의 기준이 조금씩 다르고, 좋다-나쁘다의 판단도 '정도 의 문제가 될 수 있다. 상황에 따라 이 기준이 달라지면, 평가도 조금씩 달라질 수 있다.[5)]

이 좋은 번역의 기준은 상황을 보고 결정해야 할 필요가 있다. 예를 들어 번역사 자격시험 혹은 통번역대학원 졸업시험과 같은 상황에서는 좋은 번역의 판단이 매우 엄격해야 하겠고, 외국어 능력만을 측정하는 대학입학시험에는 오히려 더 너그러운 평가가 평가목적에 적합할 수 있다. 그래야 잠재성을 가진 학생을 모두 선발할 수 있기 때문이다. 전자의 경우는 나쁜 번역을 좋은 번역이라고 오판(FP)하는 경우가 더 심각한 오류가 될 수 있고, 후자의 경우는 좋은 번역을 나쁜 번역이라고 오판(FN)하는 경우가 더 치명적일 수 있다.

결국 중요한 것은 평가 목적이 무엇이고, 그 상황에서 평가자가 원하는 '좋은 번역이 무엇이냐'이다. 즉 번역평가의 기준을 어디에 두느냐에 따라 재현도-인간평가, 정확도-인간평가와의 상관관계는 달라질 수 있다고 생각한다. 인간평가 자체가 상황에 따라 달라질 수 있기 때문이다.

이상의 내용을 정리하면 다음과 같다. 높은 수준의 번역평가에서는 정확도가, 그렇지 않은 경우는 재현도가 더 중요한 평가 지표가 될 수 있다. 아래에서는 실제 평가를 통해 위 가설이 맞는지 검토해보고자 한다.

6. 실제 평가

6.1 연구질문

본 평가에서는 위 가설 검증을 위해 통번역대학원생과 학부생의 번역에 대한 평가를 비교한다. 그리고 본 가설 외에도 추가로 두 가지 질문에 대한 답을 구한다. 첫째, 본 평가에 수집된 코퍼스가 문학, 비문학 텍스트 각각 47.71%,

5) 그렇다고 번역평가가 자의적이라는 뜻은 아니다. '정답 과 평가기준은 다양할 수 있지만 경험 많은 번역사들이 생각하는 정답, 즉 좋은 번역의 범위와 중요한 평가기준은 어느 정도 정해져 있다. 이는 다양한 선행연구에서 드러난 높은 평가자 신뢰도를 보아도 알 수 있다(Kunilovskaya 2015; Lai 2011; Waddington 2001a; Waddington 2001b 등) 그 범위를 벗어나면 점수의 차이는 있겠지만 '나쁜 번역으로 분류된다.

52.29%로 균형을 이루고 있어, 문학, 비문학 텍스트 평가의 장르별 비교가 가능하다. 문학번역은 텍스트의 성격도 비문학과 크게 다르고 번역평가 역시 비문학 번역에 비해 좀 더 정성적인 성격을 띠는 경향이 있어(박혜주 2007)[6] 비교가 흥미로울 것으로 보았다. 둘째, 모든 경우에 대해 점수, 등수도 비교한다.

6.2 데이터

아래의 4개의 코퍼스가 사용되었다. 이 코퍼스는 본고의 실험을 위해 별도로 수집된 것이 아니어서 코퍼스의 성격이 통일되지 않았다. 가능한 한 많은 자료로 가설을 검증하기 위해 수집 가능한 텍스트를 모두 수집해 분석하였다.

〈표 3〉 실험 코퍼스

	언어	장르	원문 / 번역문 수	단어 수 (원문 평균)	번역
1	독일어	비문학	12 / 120	216.33	통역대학원생 (8인) 학부생 (2인)
2	독일어	문학	6 / 120	215.67	통역대학원생
3	독일어, 서어, 아랍어, 영어, 일본어	비문학	5 / 100 (언어 당 20)	203.13	통역대학원생
4	영어	문학	5 / 119	326.60	학부생
합계			26 / 459	240.43	

본고에서는 통번역대학원생 코퍼스인 1, 2, 3번을 상급번역으로 간주했다. 1번 코퍼스에도 학부생이 포함되기는 하였으나 그 수가 적고, 평균 이상의 번역 실력의 학부생을 선발해 번역을 시킨 관계로 이를 상급번역에 포함시켰다. 반면, 4번 코퍼스의 학부생은 보통 대학생으로 구성된 번역 학급으로 학생 편차가 커서 초급번역으로 분류했다.

6) 2번 코퍼스는 95% 정량, 5% 정성의 원칙을 따라 평가했고, 4번 코퍼스 60% 정확성, 40% 유창성의 기준에 따라 평가했다. 2번 문학번역 평가에서는 정량, 정성평가의 가중치를 달리해 보았지만 95:5 일 때 평가자 신뢰도가 가장 높아 이 비율을 선택했다.

6.3 방법

인간평가는 통번역대학원을 졸업하고 번역 강의 및 평가 경력이 있는 전문번역사가 맡았다(1번: 3인 / 2번: 4인 / 3번: 언어당 2인 / 4번: 2인). 한국외대 번역평가인증 연구팀(2016)에 따라 평가는 95% 오류 감점 방식, 5% 거시평가 방식으로 이루어졌고 (4번은 자체 평가방식: 정혜연, 박헌일, 우경조, 서수영 2021), 평가자의 평균을 자동평가와 비교하였다. 정확도, 재현도 계산에서 필요한 정답번역은 통번역대학원 출신의 경력 2~12년의 전문번역사가 맡았고 이들은 평가도 수행하였다.[7)]

자동평가는 한림대학교 융합소프트웨어 학과에서 수행했다. 모든 코퍼스에 대해 형태소 분석(1, 3번: kkma, 2, 4번: utagger) 후, 파이썬(python)으로 정확도, 재현도, 그리고 두 지표의 평균인 F1, Fmean을 각각 계산하였다.

전술했듯이 인간평가와 자동평가의 비교는 총 세 가지 측면에서 이루어졌다. 통번역대학원생과 학부생, 문학과 비문학, 점수와 등수 비교가 그것이다. 각 분야에서 인간평가의 평균과 자동평가의 간의 상관관계를 구해 자동평가가 얼마나 인간평가와 비슷한지를 알아보았다. 그리고 자동평가의 타당도 확보를 위해 인간평가자 간의 신뢰도(인간평가 간의 상관관계)도 별도로 계산하였다.

본고의 주제인 정확도, 재현도 이외에 두 지표의 평균인 F1, Fmean을 구해 이를 인간평가와 비교함으로써 정확도:재현도=1:9 비율이 적절한지도 보았다.

〈표 4〉 번역평가 시 비교대상

비교대상	
정확도-인간평가 vs 재현도-인간평가	F1-인간평가 vs Fmean-인간평가

6.4 실험 결과 및 분석

아래는 인간평가와 자동평가(정확도, 재현도, F1, Fmean)의 상관계수(의 평균)이다. 자동평가가 인간과 얼마나 비슷하게 평가했는가, 즉, 네 가지 자동평가

7) 정답번역을 제공하는 자가 전문번역사일 경우, 정답번역자가 누구냐가 인간평가-자동평가 상관관계에는 큰 영향을 미치지 않는다는 연구결과도 있다 (정혜연, 명혜정, 최혜림, 허탁성 2021).

의 지표가 인간번역을 평가하기에 얼마나 타당한가를 보여주고 있다.

〈표 5〉 코퍼스 1 – 독일어 비문학

정확도		재현도		F1		Fmean	
점수	등수	점수	등수	점수	등수	점수	등수
0.799	0.6	**0.800**	**0.77**	0.855	0.806	0.823	**0.83**

〈표 6〉 코퍼스 2 – 독일어 문학

정확도		재현도		F1		Fmean	
점수	등수	점수	등수	점수	등수	점수	등수
0.239	0.1	0.238	**0.211**	**0.266**	**0.206**	0.250	0.19

〈표 7〉 코퍼스 3 – 언어별 비문학

정확도 (독어)		재현도 (독어)		F1 (독어)		Fmean (독어)	
점수	등수	점수	등수	점수	등수	점수	등수
0.775	**0.7**	0.325	**0.7**	**0.683**	**0.8**	0.410	0.7
정확도 (서어)		재현도 (서어)		F1 (서어)		Fmean (서어)	
점수	등수	점수	등수	점수	등수	점수	등수
0.373	0.2	**0.475**	**0.6**	0.446	**0.3**	**0.472**	**0.3**
정확도 (아랍어)		재현도 (아랍어)		F1 (아랍어)		Fmean (아랍어)	
점수	등수	점수	등수	점수	등수	점수	등수
0.563	0.4	**0.580**	**0.5**	**0.723**	**0.9**	0.699	0.8
정확도 (영어)		재현도 (영어)		F1 (영어)		Fmean (영어)	
점수	등수	점수	등수	점수	등수	점수	등수
0.710	0.3	**0.853**	**0.7**	0.834	**0.9**	**0.854**	0.7
정확도 (일어)		재현도 (일어)		F1 (일어)		Fmean (일어)	
점수	등수	점수	등수	점수	등수	점수	등수
-0.317	-0.3	**-0.250**	**-0.1**	-0.308	-0.4	**-0.270**	**-0.1**

〈표 8〉 코퍼스 4 – 영어 문학

정확도		재현도		F1		Fmean	
점수	등수	점수	등수	점수	등수	점수	등수
0.138	0.1	**0.346**	**0.36**	0.323	0.335	**0.345**	**0.37**

<표5> ~ <표8>까지에서 코퍼스 1번과 3번(비문학)의 상관계수가 2번과 4번(문학)에 비해 전반적으로 높은 것을 관찰할 수 있는데, 이는 1, 3번의 경우

모든 텍스트를 같은 그룹의 학생이 번역했기 때문이다(예를 들어 10명의 학생이 12개의 텍스트를 번역해 120개의 번역문을 만들었다). 나머지의 상관관계가 비교적 낮은 것은 텍스트마다 번역자가 달랐기 때문이다(예를 들어 6개의 텍스트를 수십 명의 학생이 번역하였다). 그럼에도 정확도와 재현도, 그리고 F1과 Fmean의 비교는 개별 코퍼스 안에서 이루어지기 때문에 비교에 문제가 없다.

6.4.1 전체

총 16건 중 14건에서 재현도의 상관계수가 높았다(87.5%). 정확도와의 상관계수가 높은 건은 3건(1회는 재현도와 동점)뿐 이었는데 2번 독일어 문학 코퍼스의 점수, 3번 독일어 문학 코퍼스의 점수, 등수에서 였다.

재현도와 정확도의 비교에서는 재현도가 분명한 우위였으나 F1과 Fmean 비교는 그렇지 않았다. 상관계수가 높은 재현도에 매우 큰 가중치를 두었음에도 Fmean은 8건에서만(50%) 우위를 보였다.

이 결과만 놓고 보면 재현도가 인간번역평가에 더 중요한 지표라고 하더라도 재현도에 9배의 가중치를 주는 것은 오히려 평가의 타당성을 떨어뜨릴 수도 있다는 해석이 가능하다.

6.4.2 통번역대학원생 vs 학부생

통번역대학원생 평가에서는 15건 중 12건에서 재현도가 더 우세했다(80%). 학부생 평가에서도 모두 재현도가 더 좋은 결과를 보여주었다(100%). F1, Fmean의 경우, 전체 비율을 놓고 볼 때에는 뚜렷한 차이를 나타내지 않았다. 하지만 학부생 평가에서만큼은 Fmean이 뚜렷한 우위를 보였다. 통번역대학원생 평가에서는 15건 중 6건에서만 Fmean이 더 높은 상관관계를 보였다(40%). 주지했듯 학부생 평가에서는 2건 모두 Fmean이 더 좋은 결과를 내었다(100%).

가설과는 달리 번역자 수준과는 상관없이 재현도가 인간평가와 더 높은 유사성을 보여주었다. 인간번역평가에서 재현도가 더 중요한 지표라는 의미라고 해석할 수 있겠는데, 그 이유에 대해서는 더 깊은 고민이 필요해 보인다. F1, Fmean 비교에서는 점수, 등수를 합하여 볼 때, 둘이 거의 동률의 우위를 보여(9건, 8건) 재현도에 큰 가중치를 두는 것이 큰 의미가 없을 수 있다는 점을 다시 한 번 보여주었다. 다만, 케이스의 수는 적었지만 학부생 평가에서는 Fmean

이 모두 더 우세했다는 것은 학부생 번역평가에서 재현도가 더 중요한 지표일 가능성을 남겨두고 있다.

6.4.3 문학 vs 비문학

문학, 비문학 번역에서 모두 재현도가 더 우세했다. 비문학은 13건 중 11건에서(84.62%), 문학은 4건 중 3건에서(75%) 재현도가 더 높은 상관관계를 보여주었다. F1과 Fmean는 장르별 비교에서도 큰 차이를 보이지 않았다. 비문학은 13건 중 6건에서(46.15%), 문학은 4건 중 2건에서(50%) Fmean이 더 높은 상관관계를 보였다.

결과적으로 문학과 비문학 번역평가 간에 큰 차이가 없는 것으로 나타났다. 비문학 번역에 비해 문학번역의 상관관계가 낮은 것은 앞서 언급했듯이 문학번역 평가의 어려움에서 비롯되었다기 보다 텍스트마다 번역자가 달랐기 때문이라고 보는 것이 타당할 듯하다.

6.4.4 점수 vs 등수

점수, 등수에서도 재현도의 우위는 뚜렷했다. 점수에서는 8건 중 6건이(75%), 등수에서는 8건 모두에서 재현도의 상관관계가 더 높았다. F1과 Fmean은 점수, 등수에서도 거의 동률이었다. 점수에서는 8건 중 4건에서(50%), 등수에서는 9건(동점 있음) 중 4건에서(44.44%) Fmean의 상관관계가 더 높았다.

7. 논의 및 결론

본 실험결과는 우리에게 몇 가지 생각할 문제를 던져주었다. 첫째, 위 실험결과에서 보았듯 모든 비교분야에서 재현도가 우세한 비율이 압도적으로 높았다. 재현도의 우위가 뚜렷하다는 것은 실무 관점에서 무슨 의미일까? 라비에, 새게이, 재야러맨(Lavie, Sagae and Jayaraman 2004)에서 언급했듯이 후보번역보다는 정답번역을 기준으로 평가하는 게 안전하다는 의미이기도 하겠고, 다른 한편으로는 재현도의 분모인 False Negative(좋은 번역을 나쁜 번역으로 오판하는 경우)의 비율을 줄여야 한다는 뜻이기도 하겠다. 얼핏 후자의 설명은 번역평

가자의 직관과 배치되는 것처럼 느껴지기도 하지만, 초급번역 평가에서는 가능한 해석이다. 주지했듯 초급번역(대학입학시험의 외국어 평가)에서는 상급번역에 비해 평가기준을 좀 너그럽게 적용할 수 있다. 그래야 잠재성 있는 학생을 모두 선발할 수 있기 때문이다. 이 경우, 좋은 번역을 나쁜 번역으로 오판하는 FN을 줄여야 원하는 학생들을 모두 선발할 수 있다.

이번엔 번역평가를 앞서 소개한 스팸메일 골라내기와 암 진단과 비교해보자. 재현도가 인간평가와 더 비슷했다는 얘기는 번역평가가 스팸메일 골라내기보다는 암 진단의 성격과 더 유사하다는 의미가 된다. 이 해석은 직관적으로 크게 와 닿지 않지만, 인간평가의 성격을 곱씹어보면 일면 이해가 가기도 한다.

인간의 번역평가는 많은 경우, 오류감점방식으로 이루어진다. 그리고 실제 오류감점방식이 평가자 신뢰도가 가장 높은 것으로 나타나기도 했다 (Waddington 2001b). 오류감점이란 나쁜 번역을 나쁘다 판단하는 True Negative 성격을 갖는데, 이 오류감점 방식이 가장 타당하고 믿을만한 평가방식이라고 한다면 그 대척점에 선 오판인 False Negative(FN)가 가장 치명적 오류일 수도 있겠다. 따라서 FN의 비율을 줄였을 때 높아지는 재현도가 더 중요한 지표일 수 있다.

둘째, 거의 모든 조건에서 Fmean이 우세한 비율은 F1과 거의 동률로 나타났다. 재현도가 뚜렷하게 우세였지만, 그 재현도에 9배의 가중치를 둔 Fmean은 오히려 인간평가와의 상관관계가 낮았다. 심지어 재현도의 중요성을 주장했던 라비에, 새게이, 재야러맨(Lavie, Sagae and Jayaraman 2004:9)에서 조차 비슷한 결과를 볼 수 있다. 이는 실무적으로 무슨 의미를 가질까? 번역평가에서 재현도의 중요성은 압도적이지 않으며 따라서 재현도와 정확도를 둘 다 사용하려면 둘의 비율을 다양하게 조정해서 황금비를 찾아보아야 한다는 의미로 보인다.

셋째, 통번역대학원생과 학부생 비교도 다른 비교 카테고리와 큰 차이는 없었지만, 학부생 평가에서 예외적으로 Fmean 우위가 뚜렷했다. 나머지 케이스에서는 Fmean 우위가 40~50% 정도였는데 학부생 번역에서만 100% 우위를 보였다. 이건 학부생 번역에서 특히 재현도가 더 중요할 수도 있다는 의미로 해석된다. 다만, 본 실험은 학부생 코퍼스 크기가 작아 해석에 조심할 필요가 있다.

넷째, 재현도가 우세한 케이스가 압도적으로 많긴 했지만, 개별 케이스에서 정확도와 상관관계 차이가 미미했다. 차이가 작기는 Fmean과 F1도 마찬가지였

다. 다만, 정확도나 재현도나 두 번역의 일치도를 기반으로 하는 지표이고, 그 차이가 0~1사이의 상관관계로 나타나기 때문에 차이가 뚜렷해 보이지 않을 수 있다. 실제 기계번역을 대상으로 한 라비에, 새게이, 재야러맨(Lavie, Sagae and Jayaraman 2004)에서도 정확도, 재현도의 차이는 크지 않았다.

다섯째, 인간평가자 신뢰도가 낮은 코퍼스에서는 인간-자동평가 신뢰도도 낮았다. 이는 자동평가를 사용하려면 우선 인간평가를 신뢰할 수 있어야 한다는 의미로 해석할 수 있다. 이를 위해서는 어느 장르의 번역이든 정량평가, 그 중에서도 신뢰도가 높았던 오류감점 방식을 사용해 평가하는 것이 안전할 수 있다. 실제 문학성, 오락성(이상 정성평가 지표)을 중시해 평가했던 독일어 문학번역 평가에서도 평가자 신뢰도가 낮게 나와, 사후에 정성평가 : 정량평가(오류감점방식)의 비율을 각각 50:50, 70:30, 10:90, 5:95로 조정해 보았는데, 정량평가의 비율이 가장 높은 방식(5:95)에서 평가자 신뢰도가 가장 높게 나타났다.

본고에서 던진 질문은 본 실험으로 완전히 해결되지 않았다. BLEU나 METEOR를 인간번역평가에 맞게 사용을 하기 위해서는 더 다양한 번역평가 상황(텍스트 유형, 번역의 수준, 평가자 프로필 등)에서 실제 재현도가 항상 우위를 유지하는지 확인할 필요가 있다. 본고에서는 초급번역과 상급번역의 수준 차이가 뚜렷하지 않았고, 무엇보다 초급번역 코퍼스 규모가 작았다. 또 무엇보다 인간번역을 평가하는 데에 있어 정확도와 재현도의 적절한 비율이 무엇인지를 찾아볼 필요성이 있다 사료된다.

참고문헌

권철민 (2020)『파이썬 머신러닝 완벽 가이드』. 파주: 위키북스.

김보영, 김연주, 서승희, 송신애, 이진현, 전경아, 정혜연, 최지수, 허탁성, 홍승빈 (2020)「번역자동평가에서 풀리지 않은 과제」,『번역학연구』21(1): 9-29.

박혜주 (2007)『문학번역 평가 시스템 연구』, 서울: 한국문학번역원.

정혜연, 명혜정, 최혜림, 허탁성 (2021)「인간번역 자동평가에서 정답자와 평가자가 다르다면」,『독일언어문학』93: 75-95.

정혜연, 박헌일, 우경조, 서수영 (2021)「임베딩을 활용한 인간번역의 자동평가」, 『통번역학연구』 25(3): 141-162.

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

Banerjee, Satanjeev and Alon Lavie (2005) 'METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments , *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65-72.

Buckland, Michael and Fredric Gey (1994) 'The Relationship between Recall and Precision , *Journal of the American Society for Information Science* 45(1): 12-19.

Chung, Hye-Yeon (2020) 'Automatische Evaluation der Humanübersetzung: BLEU vs. METEOR , *Lebende Sprachen* 65(1): 181-205.

Han, Lifeng (2018) 'Machine Translation Evaluation Resources and Methods: A Survey , *IPRC-2018 (Ireland Postgraduate Research Conference).* Available at https://arxiv.org/pdf/1605.04515.pdf

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

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

Sasaki, Yutaka (2007) The Truth of the F-measure. Available at 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? *HERMES Journal of Language and Communication in Business* 26: 15-37.

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

van Rijsbergen, Cornelius (1979) *Information Retrieval,* London: Butterworth.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Weinberger and Yoav Artzi (2020) ‘BERTScore: Evaluating Text Generation with BERT , *Conference Paper at ICLR 2020*, 1-14.

[부록]

[정확도와 재현도 - 코퍼스 4의 예문]

[원문] Caden held the red ball by his side.
[정답] 케이든은 옆에 있던 빨간 공을 집어 들었다.
[후보1] 케이든이 옆에 있는 빨간 공을 쥐었다.
[후보2] 카덴은 빨간색 공을 옆에 두고 있었다.
[정확도] 후보1 = 0.750 후보2 = 0.643
[재현도] 후보1 = 0.615 후보2 = 0.571
[인간평가] 후보1 = 40.75 후보2 = 37.25 (50점 만점 - 텍스트 전체 기준)

[Abstract]

Significance of Recall in Automatic Metrics for HT Evaluation

Hye-yeon Chung*, Ji-soo Choi*, Tak-sung Heo** & Soo-young Seo**
(Hankuk University of Foreign Studies*, Hallym University**)

In the automatic evaluation of translations, precision and recall are two indices that show how precisely (precision) and how much (recall) the system is able to recognize the well-translated portion in a translation. It would be ideal if two indices could be equally weighted in the evaluation system, since both accuracy and completeness are important criteria in evaluation of human translations (HT). This is, however, not easy, as both indices are negatively correlated. Papineni et al. (2002), for example, opted for precision, while Lavie et al. (2005) used both indices, giving recall nine times more weight than precision. The aim of this work is to examine which of the two indices correlates better with evaluation of professional evaluators and how much weight should be given each to precision and to recall. For this purpose, 459 translated texts were rated with precision, recall, F1 (harmonic mean of precision and recall) and Fmean (nine times higher weight on recall) as well as by professional evaluators. The results show that recall correlates better with human evaluation than precision in almost all cases, but not Fmean than F1, which were equivalent in all but one case. They indicate that recall is indeed a more important metric, but the weight as high as nine on recall is not ideal for HT evaluation.

Keywords: automatic evaluation, translation quality, precision, recall, F1, Fmean
주제어: 자동평가, 번역품질, 정확도, 재현도, F1, Fmean

정혜연(1저자, 교신저자)
한국외국어대학교 통번역대학원 한독과 교수
johanna2000@naver.com
관심분야: 통번역과정, 인지심리학, 기계번역

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

허탁성(공동저자)
한림대학교 융합소프트웨어학과 석사
gjxkrtjd221@gmail.com
관심분야: 기계번역, 자연어처리, 딥러닝

서수영(공동저자)
한림대학교 융합소프트웨어학과 석사과정
syseo96@gmail.com
관심분야: 자연어처리

논문투고: 2022년 2월 5일
1차심사 완료: 2022년 2월 24일
2차심사 완료: 2022년 3월 17일
게재 확정: 2022년 3월 22일