번역학연구, 26(3), 395-423 (2025년 9월 30일)
https://doi.org/10.15749/jts.2025.26.3.013

온라인 회의 환경에서 실시간 음성 번역 성능 분석: FAR 모델 기반 EventCat 사례 연구

배문정(영남대학교)

1. 서론

최근 인공지능 기술의 비약적인 발전에 따라 기계 번역(Machine Translation, MT)의 성능이 크게 향상되었고, 이에 따라 인간 통역 없이도 자동 동시통역이 가능한 시대가 머지 않았다는 기대감이 확산되고 있다. 자동 통역 시스템은 네 가지 핵심 처리 과정을 순차적으로 거쳐 완성된 서비스를 제공한다. 먼저 음성 신호를 텍스트로 변환하는 음성 인식(ASR) 과정, 다음으로 전사된 텍스트를 기계 처리에 적합하도록 전처리하고 표준화하는 정제 과정, 이어서 정제된 원어 텍스트를 목표 언어로 변환하는 기계 번역(MT) 과정, 마지막으로 번역 결과를 음성으로 변환하는 음성 합성(TTS) 과정을 통해 최종 결과물을 생산한다(Sudoh et al., 2020). 현재로서는 기계 번역 결과를 음성으로 출력하기보다 텍스트로 제시하는 경우가 많아 음성 합성 단계가 생략되며, 따라서 음성을 다른 언어의 텍스트로 변환하는 음성번역이 자동 통역, AI 통역 또는 기계 통역과 동의어로 사용되는 경향이 있다(최문선, 2025). 특히 Zoom, Teams, Google Meet 등과 같은 온라인 회의 플랫폼에서 제공하는 실시간 자동 번역 기능은 자동 동시통역에 대한 기대를 대중에게 체감 가능한 형태로 제공하고 있으며, 일부 사용자들은 이를 통해

실제 회의나 세미나를 진행하기에 불편이 없을 것이라고 평가하기도 한다. 그러나 이러한 기대와 달리, 실시간 자동 번역의 품질에 대한 체계적이고 실증적인 검증 사례는 여전히 부족한 실정이다(최문선, 2025; Papi et al., 2024).

실시간 기계통역이 활용될 수 있는 맥락은 국제회의, 교육 콘텐츠, 방송 등 다양하지만, 대부분의 연구는 정제된 콘텐츠(예: 영화, 강의 영상, MOOC 자막 등)를 대상으로 번역 품질을 평가해왔다(Che et al., 2017; Hu et al., 2020). 그런데 현실적으로 많은 사용자가 경험하는 온라인 회의 플랫폼에서 실시간 번역이 제공되고 있지만 이러한 서비스의 번역 품질에 대한 평가 연구는 찾아보기 어려운 실정이다.

본 연구는 이러한 문제의식을 바탕으로, EventCat이라는 상용 실시간 회의 통역 플랫폼의 음성 번역 품질을 분석 대상으로 삼았다. EventCat[1]은 온라인 회의 플랫폼과 연동하여 실시간 자막 번역을 제공하는 시스템으로, 다양한 언어 방향에 대해 AI 기반의 음성 인식과 기계 번역 엔진을 결합하여 동작한다. 하지만 EventCat과 같은 실시간 번역 시스템이 실제 사용자 환경에서 얼마나 정확한 번역을 제공하는지는 충분히 분석되지 않았다. 이에 본 연구는 피더슨(Pedersen, 2017)이 제안한 FAR 모델을 변형 적용하여 EventCat의 번역 결과를 평가하고, 번역 오류의 유형과 원인을 고찰하고자 한다.

2. 이론적 배경 및 선행 연구

2.1 기계 번역의 진보와 실시간 자동 번역의 과제

기계 번역(Machine Translation, MT)은 초기의 통계 기반 시스템에서 현재의 신경망 기반 기계 번역(Neural Machine Translation, NMT)으로 진화하며 비약적인 발전을 이루었다(Stahlberg, 2020). 특히 Google Translate, DeepL,

1) https://www.eventcat.com/ko

Papago 등은 일상적인 텍스트 번역을 넘어서 자막 및 실시간 음성 기반 번역까지 적용 범위를 넓히고 있다. 이에 따라 자동 생성된 자막의 정확성, 사용성에 대한 연구들도 다수 이루어졌다.

예를 들어, 후 등(Hu et al., 2020)은 MOOC 콘텐츠에서 자동 번역 자막과 인간 번역 자막의 수용도를 비교한 연구를 통해, 자동 자막이 반드시 열등하지 않으며 일부 경우에는 오히려 인간 번역보다 더 높은 점수를 받는 경우도 있다고 밝혔다. 체 등(Che et al., 2017)도 MOOC에서 기계가 생성한 중국어 및 영어 자막은 원문 음성 인식에서 높은 정확도를 보이는 것으로 입증되었다고 하였다. 쉬얼(Schierl, 2023)은 자동 자막과 인간 자막의 수용도 비교 연구에서, 자동 번역 자막 품질도 수용이 가능하나 인간 번역 조건에서 사용자의 이해도가 높았고 전반적인 사용자 경험도 더 긍정적인 것으로 나타났다고 보고하였다. 그러나 이들 연구는 사전 녹화된 영상에 대한 전사본(transcription)을 후편집(post-editing)하여 철자나 문장 부호 오류, 음성 인식 오류 등을 바로 잡은 텍스트를 기계 번역한 결과물을 대상으로 했으며, 실시간 발화 기반의 번역 평가는 매우 드물다(Fantinuoli & Prandi, 2021; Papi et al., 2024).

판티누올리와 프란디(Fantinuoli & Prandi, 2021)가 그러한 연구 중 하나인데, TED 영상의 영어→독일어 자동 동시 음성 번역(Automatic Simultaneous Speech Translation, ASST) 품질과 인간 통역사의 수행을 수기로 비교 분석하여, 자동 번역은 정보 전달에는 상대적으로 우수했으나, 표현의 자연스러움에서는 인간 통역에 미치지 못한다고 지적하였다. 이 연구는 기존의 많은 평가 연구와 달리, 원문 발화를 사전에 전사·정제하지 않고, 실제 발화 음성을 ASST 시스템에 직접 입력하여 평가를 진행하여 기술의 실제 사용 환경에서의 성능을 보다 충실하게 반영한다는 점에서 의의가 있다.

2.2 FAR 모델

기계 번역 품질 평가는 전통적으로 참조 번역(reference translation)을 기준으로 기계 번역(MT) 또는 인간 번역 결과의 품질을 자동 혹은 수동으로 측정하는 다양한 방법이 제안되어 왔다. 예를 들어, BLEU(Papineni et al., 2002), TER(Snover et al., 2006), METEOR(Banerjee & Lavie, 2005)와 같은 전

통적 자동 지표는 주어진 번역 결과를 참조문과 비교하여 n-그램 일치율이나 편집 거리(edit distance)를 산출한다. 최근에는 COMET(Rei et al., 2020), BERTScore(Zhang et al., 2020) 등 대규모 사전학습 언어모델 기반의 의미적 유사성 평가 기법도 널리 사용되고 있다. 이러한 범용 품질 평가 기법들은 모든 형태의 번역에 적용 가능하다는 장점이 있다.

이에 비해 FAR 모델(Pedersen, 2017)은 이러한 범용 번역 품질 평가와는 성격이 다른 자막 번역(subtitling) 특화 평가 모델이다. 첫째, FAR 모델은 정답 참조문(reference translation)을 기준으로 하지 않고, 평가자가 기능적 등가(Functional equivalence), 수용성(Acceptability), 가독성(Readability)이라는 세 가지 핵심 범주에 따라 번역 결과를 직접 판단한다. 둘째, 시간·공간 제약, 화면 시각 요소와의 조화, 자막 표시 속도 등 자막 특유의 물리적·인지적 제약 조건을 평가 범위에 포함한다는 점에서, 문서 번역이나 일반 음성 번역 평가와 구분된다. 셋째, 평가는 전적으로 전문가의 판단에 기반하며, 오류의 심각도에 따라 가중 감점하는 방식으로 점수를 산출한다. 이러한 특성 때문에 FAR 모델은 대규모 자동화 처리가 어렵지만, 자막이라는 매체의 품질을 현실적으로 반영할 수 있다는 장점이 있다. 또한 평가자의 주관이 많이 개입된다는 단점이 있지만 오류를 유형별로 분류하고 계량화함으로써 번역 품질을 객관적으로 분석할 수 있다는 장점이 있다(Du & Lu, 2024). 코글린(Koglin, 2020)은 영어를 브라질 포르투갈어로 번역한 영화 예고편 자막의 후편집본을 대상으로 FAR 모델, 번역가 평가, 시청자 수용성의 세 가지 관점에서 품질을 분석하였다. 그 결과 FAR 점수와 평가자 및 시청자의 인상이 상관관계가 있음을 확인했으며, 이는 FAR 모델이 실제 인간의 질적 평가와 일치함을 보여주는 타당성 근거가 될 수 있다.

연속적인 발화에 대해 후편집 없이 제공되는 실시간 동시 번역의 품질을 FAR 모델로 분석한 연구는 많지 않다. 그나마 확인되는 몇몇 사례 중 두와 루(Du & Lu, 2024)는 FAR 모델을 활용하여 중국 내 AI 기반 자막 번역 플랫폼(iflyrec)의 품질을 인간 번역과 비교하였다. 연구자는 중국어→영어와 영어→중국어 두 방향에서 국제 행사 홍보 영상과 TED 강연 자막을 분석한 결과, 전반적으로 인간 번역이 기능적 등가, 수용성, 가독성에서 모두 우수하였으며, 특히 기능적 등가에서 의미 왜곡과 맥락 단절이 기계 번역에서

빈번하게 나타났다. 기계 번역은 영어→중국어 방향에서 중국어→영어보다 상대적으로 나은 품질을 보였으나, 국제 행사 홍보 영상과 같이 고품질이 요구되는 상황에서는 후편집 없이는 단독 활용이 어렵다고 결론지었다. 또한 야오(Yao, 2022)의 연구 역시 FAR 모델을 적용하여, AI 기반 번역 및 자막 생성 기능을 제공하는 NetEaseSight 플랫폼이 영어 TED 강연 20편을 중국어로 자동 번역한 자막을 분석하였다. 평가 결과, 문법 오류나 단어 누락은 드물었으나 의미 전달 오류와 부적절한 자막 분절·길이 문제가 빈번했다. 연구자는 번역 엔진의 희귀어·고유명사 처리, 어순 및 맥락 유지, 음성 인식 정확도 향상, 그리고 후편집 결합을 주요 개선 방향으로 제시하였다.

이 두 연구는 모두 FAR 모델을 활용해 실제 영상 자료를 기반으로 자막 번역 품질을 체계적으로 수치화했다는 공통점이 있다. 국내에서는 아직까지 실시간 발화를 대상으로 자동 번역 품질을 분석한 연구가 거의 이루어지지 않은 것으로 파악된다. 이에 본 연구는 온라인 회의 환경에서 AI 통역사를 Zoom 회의에 초대하여 실시간 번역 기능을 활용하는 상황을 상정하였다. 실제 공식 행사에서 이루어진 연설 영상을 Zoom 회의에서 재생하여 발화를 입력하였고 이 발화는 자동 번역 엔진인 EventCat을 통해 즉시 번역되었으며, 이를 바탕으로 번역 품질을 평가하였다. Zoom이나 Teams와 같은 온라인 회의 플랫폼에서도 자체적으로 자동 번역 서비스를 제공하고 있지만 유료 플랜에서만 사용이 가능하기 때문에 본 연구에서는 한 달 동안 무료 체험이 가능한 EventCat을 활용하였으며 따라서 연구 결과는 EventCat의 자동 번역 출력물에 국한된다는 점을 밝혀둔다.

3. 연구 방법

3.1 연구 대상

본 연구는 총 4편의 영상을 분석 대상으로 하였다. 영어 원어 영상 2편(미국 연방준비제도 의장, 싱가포르 총리 연설)과 한국어 원어 영상 2편(한국은행 총재, 금융감독원장 연설)으로 구성되었으며, EventCat을 활용해 각

각 한국어와 영어 자막을 생성했다. 모든 영상은 경제·금융 관련 공식 행사 연설로, 발표자가 준비된 원고를 바탕으로 발표했다. 영상의 길이는 모두 약 10분 내외로[2], 발화 속도는 각각 평균 분당 약 161, 156, 153, 166단어이다.

표 1
분석 대상 영상 목록

영상	연사	원어	연설 제목
1	미국 연방준비제도 의장	영어	Remarks on Fed's framework review
2	싱가포르 총리	영어	The impact of US tariffs on Singapore and the global economy
3	한국은행 총재	한국어	2025년 신년사
4	한국 금융감독원장	한국어	부산금융중심지 지정 15주년 기념 심포지엄 축사

3.2 자동 번역 생성 절차

본 연구에서 사용한 자동 번역은 EventCat의 온라인 회의 실시간 음성 번역 기능을 활용하여 생성하였다. 우선 Zoom에서 회의를 개설한 후, EventCat 홈페이지의 우측 상단 메뉴에서 ONLINE MEETING 하위 항목인 Invite AI Interpreter를 선택하였다. 이어서 Translation Model에서 Swift(Fast) 또는 Wise(Accurate) 중 후자를 선택하고, 다음 페이지에서 Meeting URL, 원어(Speaking language), 목표어(Target language) 등의 회의 정보를 입력하였다.

다음 단계에서 필요 시 Glossary 기능을 선택하여 용어집을 등록할 수 있으며, 이를 생략할 수도 있다. 이후 Invite AI Interpreter를 클릭하면 Zoom 회의에 Interpreter Bot이 초대된다. 회의가 시작되면 Zoom 채팅창에 "I'm interpreting this meeting in real time. Please click the link below to see the live captions and transcript."라는 안내 메시지가 표시되며, 제공된 링크를 클릭하

2) 한국은행 총재의 2025년 신년사는 약 22분 길이의 영상이지만 10분 분량만 분석 대상으로 하였다.

면 새로운 페이지에서 실시간 자동 번역을 확인할 수 있다. 번역문 표시 형식은 두 가지 중 한 가지를 선택할 수 있다. 아래 그림 1과 같이 하나는 원어와 도착어가 함께 표시되는 형식(형식 1)이고 다른 하나는 도착어만 표시되는 형식(형식 2)이다. 연속적인 발화에 대해 번역문이 지속적으로 표시되는데 형식 1을 선택하여 번역문을 확인해 보면 원어 발화를 자체적으로 분절하여 텍스트로 표시한 후 그 내용을 목표어로 번역하는 순서로 진행된다는 것을 알 수 있다.

그림 1
자막 표시 형식

3.3 분석 방법

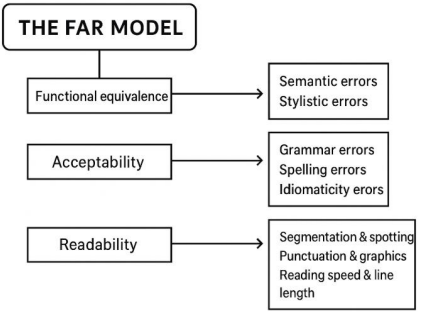
3.3.1 FAR 모델

본 연구에서는 피더슨(Pedersen, 2017)이 제안한 언어간 자막 품질 평가 모델인 FAR 모델을 활용하여 각 영상의 번역 품질을 평가하였다. FAR 모델은 자막 품질을 기능적 등가(Functional equivalence), 수용성(Acceptability), 가독성(Readability)의 세 측면에서 분석하는 평가 도구이다. 기능적 등가는 발화자의 의미가 목표어 자막에 얼마나 충실하게 반영되었는지를 나타내며, 의미 오류(Semantic errors)와 문체 오류(Stylistic errors)를 기준으로 판단한다. 수용성은 도착어 자막이 해당 언어의 규범과 관습에 얼마나 부합하는지를

평가하며, 문법 오류(Grammar errors), 철자 또는 맞춤법 오류(Spelling errors), 관용 표현 오류(Idiomaticity errors)를 포함한다. 여기서 '관용 표현'은 단순한 관용구 사용을 넘어 언어의 자연스러운 구사를 의미한다. 다시 말해, 해당 언어 원어민이 듣기에 자연스럽고 어색하지 않은 표현을 뜻한다. 가독성은 독자가 자막을 읽는 데 필요한 인지적 노력을 측정하는 것으로, 자막의 분절과 시점 설정(Segmentation and spotting), 문장 부호와 그래픽의 적절성(Punctuation and graphics), 읽기 속도와 행 길이(Reading speed and line length) 등을 종합적으로 고려하여 평가한다.

그림 2

FAR Model



이 모델은 감점 방식을 사용한다. 피더슨(Pedersen, 2017)은 의미 오류가 자막 품질에 가장 큰 영향을 미친다고 보고, 의미 오류의 경우 심각한 오류(Serious)는 2점, 보통 수준의 오류(Standard)는 1점, 경미한 오류(Minor)는 0.5점을 감점하도록 하였다. 나머지 오류 유형의 경우 심각도에 따라 1점, 0.5점, 0.25점을 감점한다. 기능적 등가(F), 수용성(A), 가독성(R) 각 영역별 감점 점수를 합산하여 총점을 구한 후 자막 개수(N)로 나누어 최종 점수를 산출한다(FAR = (N − F − A − R) / N). 따라서 산출된 숫자가 작을수록 번역의 품질이 높은 것이다. 심각도의 판단 기준도 부록 1과 같이 상세하게 제시하였다.

표 2
오류 유형에 따른 감점 점수

오류 유형	심각도	감점 점수
의미 오류	경미	0.5
	보통	1
	심각	2
기타 오류	경미	0.25
	보통	0.5
	심각	1

그런데 FAR 모델은 화면 하단 자막 품질을 평가하는 도구로 개발되었고, 본 연구의 분석 대상은 별도 페이지에 제공되는 실시간 번역이라는 차이가 있다. 그렇다 하더라도 FAR 모델을 이러한 실시간 번역 분석에 적용하는 데 큰 무리는 없었다. 다만 '분절 및 시점 설정(Segmentation and spotting)'에서 '시점 설정(spotting)'은 영상 대사에 맞춰 자막의 시작과 종료 시점을 정하는 작업인데, 본 연구의 실시간 번역에서는 발화자의 입 모양과 자막을 동기화할 필요가 없으므로 주로 분절(segmentation)을 기준으로 평가했다. 또한 본 연구에서는 한 번에 제시되는 번역 구간의 개수가 FAR 모델의 자막 개수에 해당한다. 그림 1을 예로 들자면, '여러 회원님이 싱가포르의 특정 산업에 대한 관세의 영향에 대해 문의해 주셨습니다', '상황을 신중하게 평가하고 있습니다. 그러나 우리가 더 걱정하는 것은 이들 기업이 직면한 직접적인 영향이 아니라 글로벌 무역 시스템과 세계 경제에 미치는 더 광범위한 영향입니다' 등이 각각 하나의 구간이 되는 것이다. 후자의 경우 두 개 문장으로 구성되어 있지만 한 번에 제시되었기 때문에 하나의 구간으로 본다.

3.3.2 채점 기준

표 2와 부록 1을 기본적인 채점 기준으로 하되, 본 연구에서 각 영상을 분석하면서 각 범주에 대해 다음과 같은 사항들을 추가로 고려하였다.

3.3.2.1 기능적 등가

FAR 모델에서 가장 감점 배점이 큰 항목은 '기능적 등가' 범주의 '의미 오류'이다. 본 연구에서 영상들을 분석하면서 의미 오류를 다시 '분절 오류로 인한 오역', '전사 오류로 인한 오역', '번역 엔진의 오역'[3]의 등으로 세분화할 수 있었다. 영상별로 어떤 유형의 의미 오류가 많이 나타나는지 분석하고 그 원인도 살펴보았다.

'문체 오류'는 도착어 표현이 맥락상 어색하거나 직역투 또는 어역이 맞지 않는 경우로 판단했다. 예를 들어, 영상 1에 나오는 'Several of our global peers have adopted similar approaches'가 '여러 글로벌 동종 업계에서도 유사한 접근 방식을 채택하고 있습니다'라고 번역되었는데 여기서 'peers'는 다른 국가의 중앙은행들을 의미하는 것이어서 '동종 업계'라는 표현은 적절하지 않아 0.5점을 감점했고, 연설 서두에 나오는 'Thanks for being here'가 '와 주셔서 감사합니다'로 번역되었는데 '참석해 주셔서 감사합니다'라고 하는 것이 공식적인 자리에서 하는 인사말로 더 적절할 것으로 판단되나 전자로 번역해도 의미는 통하기 때문에 0.25점을 감점했다.

3.3.2.2 수용성

'수용성' 범주에서 '문법'과 '철자 또는 맞춤법' 오류는 비교적 간단하게 파악할 수 있고 수도 적었다.

이 범주에서는 '관용 표현' 오류가 가장 많았고 판단에 있어 주관이 가장 많이 개입되는 항목이기도 했다. 예를 들어, 비유적인 표현에서 비유를 살려서 번역하지 못하고 직역을 했더라도 그 부분이 해당 번역 구간에서 차지하는 비중이 미미할 경우 0.25점을 감점했고, 해당 오류가 번역 구간 전체의 의미에 영향을 줄 경우 1점을 감점했다.

3.3.2.3 가독성

앞서 언급했듯이 '분절 및 시점 설정' 오류 중에서는 '분절'만 본 연구에 해당되며 심각도에 따라 0.25점, 0.5점, 1점으로 차등 감점하였다. 참고로, 분

3) 여기에서 '번역 엔진의 오역'은 분절 오류나 전사 오류가 없었는데 번역이 잘못된 경우를 말한다.

절 오류는 보통 두 개의 구간에 걸쳐 나타나는데 이런 경우, 한 구간에만 오류로 반영하였다. 아래 <예시 1>을 예로 들면, 분절 오류가 3개 구간에 모두 영향을 주기는 하지만 분절 오류 자체는 2개이기 때문에 2개 구간에 대해서면 오류로 처리했다.

‘문장 부호 및 그래픽’에서도 ‘문장 부호’만 본 연구에 해당되는데, 문장 부호가 잘못 찍히거나 대소문자가 잘못 나온 경우 감점하였고, 이 역시 의미에 영향을 주는 정도에 따라 차등 감점하였다.

‘읽기 속도 및 행 길이’ 항목에 대해서는 한 구간이 3줄을 넘어가면 번역을 눈으로 따라가기가 어려워진다고 판단하여 3, 4줄은 0.25점, 5줄은 0.5점, 그 이상은 1점 감점하기로 했다.

4. 분석 결과

각 영상에 대해 EventCat이 생성한 자동 번역의 품질을 FAR 모델을 활용해 분석하였다. 산출된 점수는 표 3에 제시되어 있다.

표 3
영상별 점수

영상		1	2	3	4
기능적 등가(F)	의미 오류	74	21.5	43	78
	문체 오류	6	3.25	3	0.25
수용성(A)	문법 오류	0	0	1.5	2
	철자 오류	0	0	0	0
	관용 표현 오류	2.25	0	4.75	1.5
가독성(R)	분절 및 시점 설정	22.5	3	6	5.75
	문장 부호 및 그래픽	9.75	0	1.25	0.25
	읽기 속도 및 행 길이	1	1	5	5.75
총 감점		115.5	28.75	64.5	93.5
자막 개수(N)		110	97	76	65
총점		1.05	0.30	0.85	1.44

다른 오류 유형은 비교적 간단하게 파악할 수 있고 오류의 수도 상대적으로 적었기 때문에 가장 배점이 크고 품질에 많은 영향을 주는 '기능적 등가' 범주의 '의미 오류'를 위주로 영상별로 상세하게 분석하고 다른 오류들에 대해서는 필요 시 간단히 부연 설명하기로 한다.

4.1 영상 1

영상 1은 미국 연방준비제도 의장의 연설로 영어에서 한국어로 번역된 것으로, 전체 110개 번역 구간 중 51개 구간에서 의미 오류가 발견되었다. 상기 3.3.2.1에서 설명한 대로 의미 오류를 다시 '분절 오류로 인한 오역', '전사 오류로 인한 오역', '번역 엔진의 오역'의 세 가지 유형으로 세분화 하였다. 분절 및 전사 오류로 인한 오역과 번역 엔진 오역이 각각 단독으로 나타난 구간의 수는 21, 8, 15개 발견되었고 전사 오류와 번역 엔진의 오역이 함께 나타난 경우가 2건, 분절 오류와 번역 엔진의 오역이 함께 나타난 경우가 4건 있었으며, 분절 오류, 전사 오류, 번역 엔진의 오역이 모두 중복되어 나타난 경우도 1건 있었다. 예를 들어, 'And in a welcome and historically unusual result, as this room knows, this disinflation has come without the sharp increase in unemployment that has often accompanied a campaign of rate hikes to reduce inflation.'이라는 문장이 아래와 같이 전사 및 분절되어 번역되었다.

<예시 1>
And in a welcome,
그리고 환영의 뜻을 전합니다,

and historically unusual result. As this room knows, this disinflation was has come without the sharp increase in unemployment that has often accompanied a CAMPA.
역사적으로도 이례적인 결과입니다. 이 방에서도 알다시피, 이번 디플레이션은 경기 침체기에 흔히 동반되는 급격한 실업률 증가 없이 발생했습니다.

Of rate hikes to Reduce Inflation,
인플레이션을 줄이기 위한 금리 인상,

위의 예에서 두 번째 구간의 경우, 분절과 전사 오류 그리고 번역 엔진의 오역이 모두 포함되어 있다. 우선 이 구간은 직전 구간의 ‘welcome’과 이어져야 하는데 분리가 되었고, 이 구간의 마지막 부분도 다음 구간과 이어져야 하는데 분리가 된 문제가 있다. 전사 오류로는 ‘result, as this room knows’를 연결된 하나의 문장으로 전사해야 하나 분리된 문장으로 처리한 경우와, ‘campaign’을 ‘CAMPA’로 잘못 전사한 후 이를 '경기 침체기'로 임의 번역한 경우가 있다. 또한 ‘As this room knows’를 ‘이 방에서도 알다시피’라고 한 것과 ‘disinflation’을 ‘디플레이션’이라고 한 것도 번역 엔진이 잘못한 번역이다.[4] 의미 오류의 유형 중 분절 오류가 26건으로 가장 많았는데 이것은 해당 연사가 한 문장을 길게 구사하는 경향 때문이기도 하다. <예시 1>의 문장은 34개 단어로 이루어져 있고 심지어 52개 단어로 구성된 문장도 있다. 이를 통해 AI 번역 엔진은 실시간 번역에서 아직 인간 통역사처럼 의미 단위로 문장을 분절하지 못하는 것을 알 수 있다. 인간 통역사가 <예시 1>의 문장을 처리했다면 다음과 같이 세 부분으로 분절했을 것이다.

(1) And in a welcome and historically unusual result,
(2) as this room knows,
(3) this disinflation has come without the sharp increase in unemployment that has often accompanied a campaign of rate hikes to reduce inflation.

이처럼 본 영상의 번역에서 나타난 분절 오류는 모두 인간 통역사라면 범하지 않을 오류였다. 인간 통역사는 의미 단위로 분절하는 반면, 자동 번역에서는 명확한 기준 없이 임의로 분절이 이루어지는 것이 원인으로 보인

4) ‘As this room knows’는 ‘이 방에 모인 사람들이 알다시피’라는 의미로 ‘여러분들이 아시다시피’ 정도로 번역할 수 있다. 또한 ‘디스인플레이션’은 물가 상승률은 둔화되지만 물가가 여전히 상승하는 상태를 의미하고, ‘디플레이션’은 물가가 하락하는 상태를 말하기 때문에 차이가 있다.

다. 반면 전사 오류와 번역 엔진 오역의 경우 인간 통역사의 실수와 유사한 사례가 상당수 있었다. 예를 들어, ‘at an appropriately low level’이 ‘at inappropriately low level’로 잘못 전사된 경우가 있었는데, 이는 인간 통역사도 할 수 있는 실수이다. 번역 엔진 오역의 예로, ‘~ fulfill our congressional mandate’를 ‘의회의 의무를 이행’이라고 번역한 것이 있는데 옳은 번역은 ‘의회가 부여한 의무를 이행’이며 이 부분 역시 인간 통역사라도 배경지식 유무에 따라 전자로 번역할 가능성이 없지 않아 보인다. 한편 전사 오류의 경우, 숫자를 잘못 전사하여 오역이 발생한 사례가 상당수 있었다. 인간 통역사도 숫자 통역에서 실수할 수 있지만 양상이 달랐다. 예를 들어, ‘1.6’에 소숫점을 붙이지 않아 ‘16’으로 번역되거나 ‘a two percent’를 ‘A2’로 전사하여 완전히 다른 의미로 번역되었다. 이러한 숫자 오류는 인간 통역사와 차별화되는 특징이다. 다만 본 연구는 음성 자동 번역의 품질 검증을 목적으로 하므로 인간 통역사와의 비교 분석은 이 정도로 갈음하기로 한다.

4.2 영상 2

영상 2는 싱가포르 총리의 연설로 영상 1과 마찬가지로 영어에서 한국어로의 번역이 이루어졌다.

본 영상은 총점이 0.30으로 영상 1에 비해 현저히 낮았고 ‘의미 오류’에서도 영상 1의 74보다 훨씬 낮은 19점이 감점되어 번역의 정확도가 높다고 추정할 수 있다. 총 97개 번역 구간 중 17개 구간에서 의미 오류가 발견되었는데, 그 중 분절과 전사 오류로 인한 오역이 각각 3건과 4건이었고 분절과 전사에 문제가 없는 순수한 번역 엔진 오역이 10건이었다. 분절 오류로 인한 오역이 영상 1에 비해 현저히 적었는데, 이는 연사가 구사하는 문장 길이에 영향을 받은 것으로 보인다. 영상 2는 최대 문장 길이가 27단어인 반면, 영상 1은 27단어를 넘는 문장이 23개로, 영상 2의 짧은 문장 구조로 인해 분절할 필요성이 줄어들면서 자연스럽게 오류도 적었던 것으로 보인다. 번역 엔진 오역의 예로, 원문에서 'tariff'를 'rate'로 표현했을 때 이를 한국어로 '요금'으로 번역하여 관세가 아닌 다른 의미로 해석되는 경우가 있었다. 하지만 대부분 이러한 단편적 오류였으며 구간 전체의 의미를 크게 왜곡하는 경우는 거의 없었다.

영상 1의 연사가 미국 억양을 구사하는 것과 달리 본 영상 연사는 싱가포르 억양을 구사함에도 번역 정확도가 더 높았다는 점을 고려할 때 억양보다 문장의 길이와 복잡성이 자동 번역 품질에 더 큰 영향을 미치는 것으로 판단된다.

4.3 영상 3

영상 3은 한국은행 총재의 연설로 한국어에서 영어로 번역되었다. 총 76개의 번역 구간 중 30개에서 의미 오류가 발견되었다. 의미 오류 유형을 세분화한 결과, 영한 번역과 달리 한영 번역에서는 추가 오류 유형을 확인할 수 있었다. 바로 한국어 원문에 주어가 없거나 불분명해서 발생하는 오류(이하 “주어 누락”)였다. 따라서 크게 분절 및 전사 오류, 주어 누락으로 인한 오역과 번역 엔진 오역의 네 가지 유형으로 구분할 수 있었다. 분절과 전사 오류, 주어 누락으로 인한 오역이 각각 6개, 12개, 3개, 그리고 순수한 번역 엔진 오류가 14개 발견되었다.[5)]

예를 들어, <예시 2>의 경우 분절과 전사 오류를 모두 가지고 있었다.

<예시 2>
원문
단적으로 한국과 미국의 매출액 상위 15대 기업을 10년 전과 비교해 보면 미국은 7대 기업이 신규로 진입한 반면에 우리는 2개 기업만이 바뀌었는데 그 2개 기업을 살펴보면 그중 신산업을 통해서 성장했다고 볼 수 있는 기업은 한개에 불과해 사실상 신규 기업이 진입이 거의 없었다고 말할 수 있습니다.

전사(분절)
단적으로 한국과 미국의 매출액 상위 15대 기업을 10년 전과 비교해 보면 미국은 7대 기업이 신규로 진입한 반면에 우리는 2개 기업만이 바뀌었는데, 그 2개 기업을 살펴보면 그중 신산업을 통해서 성장했다고 볼 수 있는 기업

5) 한 구간에 여러 유형의 오류가 중복되는 경우가 있기 때문에 오역의 개수는 오류 구간의 개수보다 많다.

개에 불과해 사실상 신규 기업의 진입이 거의 없었다고 말할 수 있습니다.

번역

Just to give you an idea, if you compare the top 15 companies by revenue in the U.S. and Korea from 10 years ago, the U.S. has seven new entrants, while we only have two, and if you look at those two, they're the ones that have grown through new industries.

With only a few, it's fair to say that there has been virtually no new entrants.

우선 '기업' 뒷부분이 분절되어 거기까지만 전사된 내용으로 번역을 하다보니 중요한 내용인 '한개에 불과하다'는 내용이 번역될 수 없었고 '한개' 마저도 '개'로 잘못 전사하여 'with only a few'로 오역이 되었다. 이처럼 한영 번역에서는 한국어의 후위 정보를 영어에서 전위로 표현해야 하는데, 분절 오류로 인해 중요한 서술어가 다음 행으로 넘어가면서 해당 구간에서 누락되어 임의 번역되는 경우가 많았다.

전사의 경우 단순히 단어 하나를 잘못 전사하는 단순한 종류의 오류도 있었지만, 연사가 잠시 말을 더듬거나 실수로 반복할 경우, 더듬는 부분의 발음이 명확하지 않아 임의로 전사하거나 의도치 않게 실수로 반복하는 말을 그대로 전사하여 번역 오류로 이어지는 경우가 있었다. 아래 <예시 3>이 그러한 사례이다. 이 예시의 한국어 전사에서 굵은 글씨로 된 '우리 경제 신생 산업이나 기업이 부족한 것은'은 '우리 경제에 신성장 산업이나 기업이 부족한 것은'을 잘못 전사한 것이다. 연사가 '신성장' 부분에서 잠시 멈칫하면서 '생'과 '성'의 중간 정도로 발음했기 때문이다. 그리고 '이 파괴 과정에서의'와 '사회적 새로'는 불필요하게 들어가서 문장 전체의 의미를 흐리게 하는 원인이 되었는데 이것이 그대로 전사되면서 영어 문장의 의미도 불분명해졌다.

<예시 3>

우리 경제 신생 산업이나 기업이 부족한 것은 창조적 파괴 과정에 수

> 반되는 **이 파괴 과정에서의** 퇴출이 **사회적 새로** 일으키는 사회적 갈등을 관리하려고 하기보다는 안정을 추구한다는 이유로 이를 피해 왔기 때문이 아닌지 돌아볼 필요가 있습니다.
> We need to ask ourselves if the lack of economic start-ups and enterprises is not because we have avoided the process of creative destruction that accompanies it in favor of stabilization rather than trying to manage emerging social conflicts.

또한 주어나 주체의 누락으로 영어 번역에서는 주어 등 대명사가 임의로 설정되는 경우가 있었다. 아래 <예시 4>의 한국어 원문에는 감사드리는 주체가 드러나 있지 않아 'We'라고 임의로 주어를 잡았는데 이 연설의 연사 개인이 감사를 표하는 것이므로 'I'가 적절하다. 또한 '가족분들'은 '직원들의 가족분들'을 의미하나 한국어 원문에 이것이 명시적으로 드러나지 않아 'our families'라고 잘못 번역되었다.

> <예시 4>
> 또한 한국은행 직원들이 업무에 집중할 수 있도록 도와주신 가족분들께도 진심으로 감사드립니다.
> We would also like to extend our sincere thanks to our families for helping us focus on our work.

추가로, '수용성' 범주의 '문법 오류'에서는 한국어에 단복수가 명시되지 않아 발생한 오류가 있었다. 한국어에서 '들'과 같은 복수 표현을 생략하는 경우가 많아, 인간 통역사들도 정확한 상황을 모를 때 'a/an'을 활용해 단수로 표현할지 's'를 붙여 복수로 표현할지 고민하게 되므로 이는 인간 통역사의 실수와 유사한 양상이다.

4.4 영상 4

영상 4는 금융감독원장의 연설로 영상 3과 마찬가지로 한국어에서 영어로의 번역이 이루어졌다. 65개의 번역 구간 중 55개에서 오류가 발견되었고 총점도 1.44로 가장 높아 4개 영상 중 정확도가 가장 떨어졌다. 의미 오류로 인한 감점도 78점으로 가장 높았다.

의미 오류의 가장 큰 원인은 전사 오류였다. 55개 오류 구간 중 49개에서 전사가 잘못되어 오역이 발생했다. 분절 오류도 많이 있었지만 전사 오류가 워낙 심각하기 때문에 의미 분절을 따져보는 것이 의미가 없을 정도이다. 예를 들어, <예시 5>의 원문은 다음과 같이 전사 및 번역되었다.

<예시 5>
원문
저는 이제 지난 15년 동안 우리 부산이 이루어낸 성과와 관련돼서 저희가 정부 당국에서 보고 있는 시각과 그런 어떤 그런 현상 분석을 설명을 드리고 어 거기에 앞으로 우리가 무엇을 해야 될지에 대한 어 저희 의견과 같이 오늘 인제 진행될 포럼의 내용들을 어떻게 저희가 바라보고 있는지 말씀드리겠습니다.

전사(분절)
아 근데 지난 15년 동안 회사의 성과와 관련해서 저희가 어 통계에서 보고 있는 사실과 어 설명을 드리고 어?

거기에 이제 앞으로 우리가 무엇을 해야 될지에 대한 어 저희 의견과 같이 그 어?

포럼의 내용들을 바라보고 있습니다.

번역
Uh, but what we're seeing in the stats and uh, what we're seeing in the company's performance over the last 15 years, uh, and I'll give you an explanation, uh?

There you have it, like our opinion on what we should be doing going forward, uh?

I'm looking at the content of the forum.

우선 연사가 말한 그대로 전사가 되지 않아서 전사 자체에 문제가 있기도 하고, '어'라는 필러(filler)가 나올 때 그것을 기점으로 문장을 분절해 버

려서 심각한 오역이 발생하기도 했다. 또한 여기에서 '같이'는 '함께'라는 뜻인데 '~처럼'의 뜻으로 영어로 번역되는 등 순수한 번역 엔진 오역도 있었다. 다음 예시는 의미 없는 말이 그대로 전사되고 번역되어 오역이 발생한 경우이다.

<예시 6>
다만 최근 환경 변화와 여러 가지 우리가 중심지 정책을 한번 점검을 해보기 위한...
However, with the recent changes in the environment and the many ways in which we've changed our center of gravity policy, we wanted to take a look at...

제반 여건에 대해 점검해볼 필요가 있습니다
We need to check all the right boxes

이 예시에서는 전사 자체는 비교적 잘 되었는데 '여러 가지'라는 말이 의미 없는 필러처럼 들어갔다는 것을 식별하지 못해 'the many ways'라고 그대로 번역 하면서 번역문 자체도 이해하기가 어려워졌다.

한국어가 모국어인 필자가 듣기에는 영상 3과 비교했을 때 영상 4의 연사의 발화가 크게 듣기 어렵지는 않았기 때문에 전사에서 이 정도로 많은 오류가 발생하는 원인을 설명하기는 사실 어렵다. 다만 연사가 원고를 읽으면서 연설을 함에도 불구하고 '어', '이제', '지금', '어떤 그런'과 같이 의미 없는 필러를 많이 사용하고 스스로 말을 정정하는 경우가 많았던 것이 주요 원인으로 보인다. 인간 통역사라면 이런 것들을 자연스럽게 걸러냈을 것인데 기계 번역에서는 모든 음성이 전사되기도 하고 무의미한 부분을 구별하지 못하기 때문에 발생하는 문제로 판단된다. 또한 연사의 발음이 부정확해지는 부분이 있고 문장 말미에서 목소리가 작아지거나 얼버무리는 특성도 전사 정확도를 떨어뜨린 요인으로 보인다. 또한 영상 1, 2에 대한 전사 및 번역을 보면 분절이 잘못되어 오역이 발생하는 경우는 많았지만 전사 자체가 심각하게 잘못된 경우는 많지 않았다는 점을 고려할 때, 아직 한국어의 다양한 억양이나 어투에 대해서는 EventCat이 충분히 훈련되지 않았을 가능성도 고려할 수 있겠다.

5. 논의

본 연구에서는 피더슨(Pedersen, 2017)이 제안한 FAR 모델을 변형·적용하여 EventCat이 생성한 실시간 음성 번역의 품질을 분석하였다. 분석 결과, 연설 영상의 언어 방향(영→한, 한→영), 연사의 발화 특성(문장 길이, 발화 명료도, 억양), 발화 구조(주어 표출 여부, 후위 서술 여부) 등이 번역 품질에 유의미한 영향을 미치는 것으로 나타났다. 또한 오류 유형별로 분절 오류, 전사 오류, 번역 엔진 오역, 그리고 한영 번역의 경우 주어 누락도 오역의 주요 원인으로 확인되었다.

5.1 언어 방향에 따른 오류 양상의 차이

영→한 번역(영상 1, 2)과 한→영 번역(영상 3, 4)은 오류 양상과 비중에서 뚜렷한 차이를 보였다. 영→한 번역에서는 분절 오류가 품질 저하의 주된 요인이었으며, 한→영 번역에서는 전사 오류와 주어 누락으로 인한 오류가 두드러졌다. 특히 한국어 원문에서 주어가 명시되지 않은 경우, 번역 엔진이 문맥을 잘못 해석하여 부적절한 대명사를 선택하거나 의미를 왜곡하는 경향이 나타났다. 이는 한국어의 주어 생략 특성이 기계 번역 품질에 부정적인 영향을 미칠 수 있음을 보여준다.

5.2. 문장 길이와 발화 구조의 영향

연사의 문장 길이와 구조는 번역 품질에 직접적인 영향을 미쳤다. 영상 1과 2를 비교하면, 문장 길이가 전반적으로 짧고 구조가 단순한 영상 2에서는 분절 오류가 현저히 적었다. 반면 장문이 빈번히 등장하는 영상 1에서는 AI가 의미 단위로 적절히 분절하지 못하고 임의로 구간을 나누어 오역을 유발하였다. 이는 인간 통역사가 의미 단위로 분절하는 전략과 대조적이며, 실시간 자동 번역이 여전히 문장 내부 구조를 효율적으로 처리하지 못함을 시사한다.

5.3 전사 오류의 특성과 원인

한→영 번역(영상 3, 4)에서는 전사 오류가 품질 저하의 주요 요인이었다. 특히 영상 4에서는 65개 구간 중 49개에서 전사 오류가 발견되어 품질에 치명적인 영향을 미쳤다. 발음 부정확, 발화 말미의 음성 약화, 불필요한 필러 사용, 자기 수정 등이 전사 오류의 주된 원인으로 추정되며, 무의미한 발화를 AI가 걸러내지 못하고 그대로 번역문에 반영하는 사례도 다수 확인되었다. 인간 통역사는 불필요한 발화를 걸러내는 능력을 갖추고 있으나, 기계 번역은 원문 음성을 기계적으로 전사하여 번역하기 때문에 불필요한 발화도 모두 번역되는 경향이 있었다.

5.4 번역 엔진 오역의 유형과 특성

전사나 분절 오류가 없더라도 번역 엔진 오역은 지속적으로 나타났다. 영→한 번역에서는 전문 용어를 문맥과 다르게 해석하거나(예: ‘tariff’를 ‘요금’으로 번역), 한→영 번역에서는 관용 표현이나 어법 선택에서 어색한 결과가 생성되기도 했다. 일부 오역은 인간 통역사도 청취 오류나 배경지식 부족으로 범할 수 있는 유형이었으나, 기계 번역에서는 반복적으로 나타나는 특정 오류 패턴이 있었다. 특히 본 분석에서 사용된 EventCat은 구간 단위로 번역을 처리하며, 분절된 구간 간 의미 연결이 거의 이루어지지 않았다. 즉, 특정 구간에서 필요한 참조 정보가 이전 발화에 있더라도 이를 반영하지 않고 해당 구간의 텍스트만을 기반으로 번역이 이루어졌다. 이러한 특성은 긴 문장이나 맥락 의존도가 높은 발화에서 번역 정확도를 크게 떨어뜨릴 수 있다. 또한 한국어 원문 특유의 주어 생략, 단복수 표지 부재, 후위 서술어 구조는 한영 번역 과정에서 주체의 임의 설정, 수일치 오류, 정보 누락을 빈번하게 유발하였다. 이는 언어 구조의 차이와 자동 번역 엔진의 학습 데이터 편중으로 인해 발생하는 문제로 볼 수 있다.

5.5 FAR 모델의 적용 가능성과 한계

FAR 모델은 원래 화면 하단 자막 품질 평가를 위해 개발되었으나, 본 연구에서는 실시간 번역 평가에 변형·적용하였다. 자동 번역 결과를 구체적 기준으로 평가할 수 있는 유용한 도구임에는 분명하지만 FAR 모델에는 다음과 같은 한계가 있었다.

- **평가자의 주관성**: 감점 점수 부여가 평가자의 판단에 크게 의존하며, 특히 '기능적 등가' 범주의 '문체 오류'와 '수용성' 범주의 '관용 표현 오류' 간 경계가 모호하였다. 다만, 이 경우에는 경계가 모호하더라도 한 개 항목에만 반영되어 점수가 한 번만 차감되기 때문에 총점에 영향을 주지는 않았다. 또한 심각도에 대한 판단 역시 주관이 개입될 수밖에 없다는 한계가 있다.

- **항목 중복 가능성**: '기능적 등가'의 '의미 오류'가 분절 오류로 인해 발생할 경우 그것을 '의미 오류'에도 반영하고 '가독성' 범주의 '분절 및 시점 설정 오류'에도 반영되어 중복 감점되는 문제가 있었다.

- **절대 기준 부재**: 몇 점 이상이면 문제가 있고 몇 점 이하면 양호한지에 대한 명확한 절대 기준이 없어, 번역문 간 상대 비교에는 유용하나 절대적 품질 판단에는 제약이 있다. 다만 전체 번역 구간 중 오역 구간 비율을 파악하면 실전에 활용할 수 있는 품질인지 여부를 판단할 수 있다. 또한 본 연구 대상인 4개 영상의 경우, 품질이 가장 우수했던 영상 2의 점수를 기준점으로 삼아 다른 번역 결과물의 품질을 판단하는 방법도 고려할 수 있다.

6. 결론

본 연구에서는 피더슨(Pedersen, 2017)이 개발한 FAR 모델을 활용하여 EventCat의 실시간 번역 성능을 검증하고, 발생한 오류의 유형과 발생 원인을 상세히 분석하였다. 분석 대상은 영→한과 한→영 방향의 총 네 편의 연설 영상이었으며, 오류 유형을 분절 오류, 전사 오류, 번역 엔진 오역, 그리고 한영 번역의 경우 주어 누락 오류도 포함하여 세분화하였다.

연구 결과, 언어 방향과 연사의 발화 특성이 번역 품질에 영향을 미치는

것으로 나타났다. 영→한 번역에서는 장문의 복잡한 문장에서 분절 오류가 빈번하게 발생하였으며, 한→영 번역에서는 전사 오류와 주어 누락으로 인한 오류가 품질 저하의 주된 요인이었다. 또한 발음 부정확, 불필요한 필러 사용, 자기 수정과 같은 발화 특성이 전사 및 번역 오류를 유발하는 주요 요인으로 확인되었다. EventCat의 번역 엔진은 분절된 구간을 독립적으로 번역하는 경향이 있어 장문이나 문맥 의존도가 높은 발화에서 의미 단절이 발생하기 쉽다는 점도 발견되었다. FAR 모델은 실시간 번역 평가에 적용 가능성을 보였으나, 범주 간 중복 가능성, 평가자의 주관 개입, 절대적 품질 기준 부재와 같은 구조적 한계가 확인되었다.

본 연구의 의의는 다음과 같다.

첫째, 대부분의 기존 연구에서 주로 후편집된 번역물이나 준비된 자막을 대상으로 품질을 분석한 것과 달리, 본 연구는 실시간 번역 환경에서의 자동 번역 품질을 체계적으로 평가하였다는 점에서 차별성을 가진다.

둘째, FAR 모델을 실시간 번역 분석에 맞게 변형·적용하고, 기존 범주를 세분화(예: 의미 오류를 분절 오류·전사 오류·번역 엔진 오역으로 분류, 한영 번역의 경우 주어 누락 오류 추가)함으로써 기존 평가 틀의 확장 가능성을 제시하였다.

셋째, 언어 방향, 발화 특성, 문장 구조와 같은 발화 변인이 번역 품질에 미치는 구체적 영향을 실증적으로 확인함으로써, 향후 실시간 번역 엔진 개발과 통역 교육 현장에 활용 가능한 실무적 시사점을 제공하였다.

넷째, 두와 루(Du & Lu, 2024)와 야오(Yao, 2022)와 같은 선행연구에서도 드러났듯이 본 연구에서도 후편집 없이 제공되는 원문 그대로의 기계 번역은 실제 활용 맥락에서 여전히 품질적 한계가 있음을 확인하였다. 이는 특히 공식 회의나 국제 행사 등 고품질 번역이 요구되는 상황에서, 후편집 또는 인간 통역사의 개입이 여전히 필요함을 시사한다.

다만, 본 연구는 분석 대상 영상 수가 제한적이어서 결과를 일반화하기 어렵고, EventCat의 내부 구조와 학습 데이터가 공개되지 않아 오류 원인을 완전히 규명하기 어려웠다는 점에서 한계를 가진다. 또한 오류 심각도 평가에 주관이 개입될 수 있는데 연구자 1인의 판단에 의존했다는 점도 한계로 작용한다. 향후 연구에서는 복수의 평가자를 참여시켜 평가자 간 신뢰도를

검증한다면 연구 결과의 객관성과 신뢰성을 더욱 높일 수 있을 것이다. 나아가, 다양한 주제·발화 특성을 가진 발화를 대상으로, 다른 실시간 번역 시스템과의 비교를 통해 번역 품질에 영향을 미치는 변인을 보다 정밀하게 규명한다면 흥미로운 연구가 될 것으로 기대한다.

참고문헌

최문선. (2025). AI 시대의 통역 서비스 재편: 새로운 통역 유형론에 대한 시론(試論). 번역학연구, 26(2), 169-197.

Banerjee, S., & Lavie, A. (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* (pp. 65-72). Association for Computational Linguistics.

Che, X., Luo, S., Yang, H., & Meinel, C. (2017). Automatic lecture subtitle gen eration and how it helps. *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT),* 34-38. https://doi.org/10.1109/IC ALT.2017.11

Fantinuoli, C., & Prandi, B. (2021). Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)* (pp. 245–254). Association for Computational Linguistics.

Du, J., & Lu, J. (2024). A comparative study on the translation quality between human and machine-generated subtitles. In *Proceedings of the 2024 6th International Conference on Natural Language Processing (ICNLP)* (pp. 62–66). Xi'an, China.

Hu, K., O'Brien, S., & Kenny, D. (2020). A reception study of machine translated subtitles for MOOCs. *Perspectives, 28*(4), 521-538.

Koglin, A., Silveira, J. G. P. d., Matos, M. A. d., Silva, V. T. C., & Moura, W. H. C. (2022). Quality of post-edited interlingual subtitling: Far model, translator's assessment and audience reception. *Cadernos De Tradução, 42*(01), 1-26.

Papi, S., Polák, P., Macháček, D., & Bojar, O. (2024). How "real" is your real-time simultaneous speech-to-speech translation system? arXiv preprint, arXiv:2412.18495(cs).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics.

Pedersen, J. (2017). The FAR model: Assessing quality in interlingual subtitling. *The Journal of Specialised Translation*, *28*, 210–229.

Rei, R., Stewart, C., Farinha, A. A., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics.

Schierl, F. (2023). Reception of machine-translated and human-translated subtitles: A case study. In *Proceedings of Machine Translation Summit XIX: Vol. 2. Users track* (pp. 42–53). Asia-Pacific Association for Machine Translation.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp. 223-231). Association for Machine Translation in the Americas.

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research, 69,* 343–418.

Sudoh, K., Kano, T., Novitasari, S., Yanagita, T., Sakti, S., & Nakamura, S. (2020). Simultaneous speech-to-speech translation system with neural

incremental ASR, MT, and TTS. ArXiv, abs/2011.04845.

Yao, G. (2022). Evaluation of machine translation in English-Chinese automatic subtitling of TED talks. *Modern Languages, Literatures, and Linguistics, 1*(1), 12-22.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. arXiv:1904.09675.

<부록 1>

FAR 모델 평가 기준

Category	Error Type		Severity Score(penalty point)	Description
Functional Equi	Semantic Error	Minor	0.5	Lexical error, low impact on meaning or plot
		Standard	1	Meaning is distorted but still partly interpretable
		Serious	2	Subtitle misleads or severely disrupts comprehension
	Stylistic Error	Minor	0.25	Mild stylistic inconsistency, e.g., register mismatch
		Standard	0.5	Register or tone clearly inappropriate for context
		Serious	1	Stylistic clash severely disrupts tone or immersion
Acceptability	Grammar	Minor	0.25	Annoying but not harmful grammar issue (e.g., whom/who)
		Standard	0.5	Moderate grammar error, affects flow or clarity
		Serious	1	Grammatical error makes subtitle hard to read
	Spelling	Minor	0.25	Typo or minor spelling issue
		Standard	0.5	Spelling alters meaning slightly
		Serious	1	Unintelligible due to severe misspelling
	Idiomaticity	Minor	0.25	Unnatural phrasing due to source interference
		Standard	0.5	Noticeably awkward or unidiomatic expression
		Serious	1	Translationese so strong it hinders understanding
Readability	Segmentation/Spotting	Minor	0.25	Unnatural line breaks or minor segmentation fault
		Standard	0.5	Subtitle breaks syntactic/semantic unit across lines
		Serious	1	Out of sync by more than one utterance
	Punctuation/Graphics	Minor	0.25	Improper punctuation but meaning preserved
		Standard	0.5	Incorrect formatting affecting readability
		Serious	1	Punctuation creates confusion or visual distraction
	Reading Speed/Line Length	Minor	0.25	Slightly fast but still readable subtitle speed
		Standard	0.5	Reading speed forces viewer to choose between reading and visuals
		Serious	1	Speed too fast; viewers likely miss content or give up reading

Evaluating real-time voice translation in online meetings: A FAR Model analysis of EventCat performance

Munjung Bae (m.bae7811@gmail.com)
School of Global Interpretation and Translation, Yeungnam University

Abstract

This study investigates the quality of real-time voice translation without post-editing in an online meeting context, focusing on the AI interpretation engine, EventCat. While recent advancements in neural machine translation have led to widespread expectations of near-perfect automated interpretation, few empirical studies have examined its performance in real-time spoken scenarios. Unlike most prior research that primarily analyzed post-edited content, this study evaluated translation quality in a live-like setting by inviting the AI interpreter to a Zoom meeting and streaming actual speeches from formal events as input. Translation quality was evaluated through Pedersen's (2017) FAR model, a framework that measures functional equivalence, acceptability, and readability. Results indicate that unedited machine translation (MT) output still does not meet the quality standards required in high-stakes situations, particularly in maintaining semantic coherence and contextual fidelity. This research contributes to the growing body of literature on real-time MT evaluation by extending the scope to spontaneous speech in virtual conferencing environments and providing empirical evidence on the limitations and potential of AI interpretation tools in professional settings.

Keywords: Real-time voice translation; AI interpretation; FAR Model; online meetings; machine translation quality assessment
키워드: 실시간 음성 번역, AI 통역, FAR 모델, 온라인 회의, 기계 번역 품질 평가

배문정(https://orcid.org/0000-0002-8847-4176)
영남대학교 글로벌통번역학부 영어통번역전공 조교수
m.bae7811@gmail.com

논문 투고일: 2025년 8월 15일
1차 심사 완료일: 2025년 9월 1일
2차 심사 완료일: 2025년 9월 7일
게재 확정일: 2025년 9월 15일