한영·영한 번역 병렬 말뭉치 품질 예측 모델 구축 및 적용

고원희, 최진혁, 최규동

트위그팜

{wonhee.go, jinhyuk.choi, ken.choi}@twigfarm.net

Wonhee Go, Jinhyuk Choi, Kyudong Choi

Twigfarm

요약문

본 연구를 통해 한영·영한 번역 병렬 말뭉치의 품질을 평가하는 모델을 제안하고, 이를 이용해 시중의 한영·영한 번역 병렬 말뭉치에 대해 품질 평가를 실시한다. 연구의 의의는 다음과 같다. 첫 번째로, DA(Direct Assessment) 점수로 평가하는 한영·영한 QE(Quality Estimation) 연구를 실시하며 관련 데이터를 최초 구축했다. 두 번째로 번역 병렬 말뭉치 품질 평가 모델 TwiQE를 구축했다. 세 번째로 복수의 전문가가 수립한 품질 기준을 제시하며 번역 병렬 말뭉치 품질 개선의 기틀을 마련했다. 마지막으로, 구축된 모델을 활용하여 AI HUB[1]에 공개된 한영 번역 병렬 말뭉치 2종, 영한 번역 병렬 말뭉치 2종[2]에 대한 품질 평가를 실시했다. TwiQE를 통해 도출한 품질 평가 점수를 바탕으로 개선이 필요한 말뭉치를 두 수준으로 나누어 파악했고, 각 수준에 미치지 못하는 말뭉치의 수량과 비율을 파악하여 품질 개선의 기틀을 마련했다.

TwiQE의 가장 큰 장점은 컴퓨팅 자원의 사용을 최소화한다는 점이다. 기학습된 언어 모델(Pre-trained Language Model, PLM)과 코사인 유사도 활용으로 컴퓨팅 자원의 사용을 최소화했다. 또한 원문-번역문 코사인 유사도와 원문-역번역문 코사인 유사도를 함께 사용하여 한국어와 같은 저자원(low-resource) 언어에서의 성능을 높였다.

키워드: 번역문 품질 평가(QE, Quality Estimaion), 역번역(Round-trip translation), 코사인 유사도(cosine similarity), DA 점수(Direct Assessment score)

[1] https://www.aihub.or.kr/

[2] 2022년 공개된 “기술과학 분야 한-영 번역 병렬 말뭉치 데이터”, “전문분야 영-한·중-한 번역 말뭉치 (식품)”, “일상생활 및 구어체 한-영 번역 병렬 말뭉치 데이터”의 3종 말뭉치에서 한-영 방향의 병렬 말뭉치 2종과 영-한 방향의 병렬 말뭉치 2종을 취해 품질 평가를 진행했다.

1. 서론

인공지능에 대한 뜨거운 관심으로 학습 데이터가 점차 확보되고 있는 가운데 데이터의 품질에 대한 논의가 점차 수면 위로 떠오르고 있다. 데이터의 품질을 개선하기 위해 선행되어야 할 것은 품질의 측정이다. 품질이 수치로써 명시되면 한정된 자원과 시간으로 최대한의 품질 개선 효과를 낼 수 있기 때문이다.

본 연구에서는 말뭉치 중에서도 번역 병렬 말뭉치에 주목했다. 그 어느 때보다 번역이 중요한 비즈니스로 대두되고 있기 때문이다. 글로벌 OTT 사업의 등장으로 나라 간 문화 콘텐츠 교류가 활발해진 가운데, 한국 콘텐츠가 급부상해 세계 시장에서도 1위를 차지하는 현상이 나타났다. 이에 한국어가 포함된 번역 업무가 전에 없이 중요해졌다.

이러한 배경을 바탕으로 본 연구에서는 번역 병렬 말뭉치 품질 평가 모델을 제안하고, 시중의 번역 병렬 말뭉치에 대한 품질 평가를 실시했다. 품질 평가 모델을 학습시키기 위해 먼저 번역문 품질 예측(Quality Estimation, 이하 QE) 데이터가 필요했다. 번역문에 대한 사람의 직접적 평가 점수(Direct Assessment, 이하 DA)가 붙은 한국어 QE 데이터는 현재 공개된 것이 없기 때문에 자체적으로 구축했다. 이후 번역문을 다시 원문 방향으로 기계번역하여 역번역문(round-trip translation, 이하 RTT)을 생성하고, 기학습된 언어 모델(Pre-trained Language Model, 이하 PLM)을 통해 각각의 문장에 대한 표현 벡터인 임베딩(embedding) 행렬을 생성했다. 도출된 임베딩을 바탕으로 내적(dot product) 연산을 통해 원문-번역문 간 코사인 유사도와 원문-역번역문 간 코사인 유사도를 구했다. 앞서 구한 두 가지 유사도 결과와 어절 수[3]를 자질(feature)로 이용하여 회귀분석모델과 부스팅 앙상블(ensemble) 모델 등, 다양한 종류의 머신러닝 모델을 예측기 후보로서 비교하고, 최종 모델을 선정했다. 이때 예측기 학습에 사용된 각 자질들은 절제 연구(Ablation study)를 통해 유효한 자질임을 확인했다. 이후 그리드서치(grid search)를 통해 가장 좋은 성능을 보이는 모델을 선정하였다. 이렇게 선정된 최종 모델은 AI HUB의 한영 번역 병렬 말뭉치 2종, 영한 번역 병렬 말뭉치 2종에 대한 품질 평가에 사용되었다.

[3] 어절 수에 관련한 자질로는 세 가지를 선정했다. '원문 어절 수'와 '역번역문 어절 수', 그리고 '원문 어절 수와 역번역문 어절 수의 차이'가 있다.

2. 선행 연구

2.1. 번역문 품질 예측 (QE, Quality Estimation)

번역문 품질 예측(Quality Estimation, 이하 QE)은 2012년 WMT(Workshop on Machine Translation)에 처음 등장한 이래로 현재 시점까지 꾸준히 학계의 관심을 받고 있다. QE는 어떠한 참조 문장 없이 원문과 번역문만을 가지고 품질 지표를 제공하는 것을 목표로 한다. 본 연구 역시 추가적인 참조문 없이, 제시된 원문-번역문 쌍을 이용해 말뭉치 품질 평가 모델을 설계하는 것이 그 목적이다.

QE의 하위 태스크는 역사에 따라서 구체적인 내용이 조금씩 변화했으나, 가장 최근의 연구 흐름[1]에서는 세 가지로 제시된다. 첫 번째는 문장 단위로 DA(Direct Assessment) 점수를 예측하는 것이고 두 번째는 단어 단위로 HTER(Human-mediated Translation Edit Rate) 점수를 예측하는 것이다. 마지막은 심각한 오류(CE, Critical Errors)를 번역문이 품고 있는지를 문장 단위로 예측하는 것이다. 본 연구에서는 문장 단위의 직접적 평가인 DA 점수 예측에 초점을 맞춰 연구를 진행했다.

대표적인 QE 모델 구조는 [2]에서 제안한 predictor-estimator 구조이다. 이는 모델을 두 단계에 나눠 설계하는 방식이다. 첫 번째 단계인 predictor는 번역 병렬 말뭉치를 이용해 학습시키며, 두 번째 단계인 estimator는 QE 데이터로 학습시킨다. 이후 많은 QE 모델이 predictor-estimator 구조를 따라 설계되었다.

Unbabel[3] 역시 기존에 제안된 predictor-estimator의 구조를 따르나, predictor 부분을 기학습된 언어 모델 종류 중 하나인 mBERT(multilingual BERT)[4] 모델, 또는 XLM(Cross-lingua Language Model)[5] 모델을 이용하여 수행했다.

TransQuest[6]는 코사인 유사도를 이용해 품질평가를 시도했다. 기학습된 언어 모델을 이용하며, 두 가지 모델을 앙상블 했다는 특징이 있다. 이때 임의의 여러 가중치 쌍을 적용한 가중 평균 앙상블(weighted average ensemble)을 통해 최종 품질 예측 점수를 얻었다. 2020년 당시 DA 점수 예측 태스크에서 SOTA(state-of-the-art)를 달성했다.

코사인 유사도를 활용하는 방식은 명백한 장점이 있다. 한국어처럼 QE 데이터가 많이 확보되지 않은 저자원(low-resource) 언어에서도 좋은 성능의 모델을 얻어낼 수 있다는 점이다[6]. 이에 본 연구에서는 코사인 유사도를 활용하되, 두 종류의 코사인 유사도를 활용했다. 첫 번째 코사인 유사도는 '원문과 번역문 간의 코사인 유사도(이하 원문-번역문 코사인 유사도)'이다. 두 번째 코사인 유사도는, '원문과 역번역문(RTT)과 간에 얻어낸 코사인 유사도(이하 원문-역번역문 코사인

유사도)' 이다. 여기서 역번역문이란 번역 병렬 말뭉치에서 번역문을 다시 원문의 언어로 기계번역을 하여 얻어낸 문장을 뜻한다. 두 종류의 코사인 유사도와 원문의 어절 수, 역번역문의 어절 수, 원문과 역번역문의 어절 수 차이 등을 자질로 이용한 앙상블 모델을 설계해 제일 성능이 높게 나오는 최적의 모델을 얻어냈다. 이는 TransQuest[6]에서 임의의 가중치 쌍을 부여하여 두 개 모델을 앙상블한 것과 차이가 있다.

기존의 QE 연구는 번역문을 완벽한 정답 문장(gold sentence)으로 가정하고 연구를 진행한다. 하지만 본 연구는 번역 병렬 말뭉치 자체에 대한 의문을 제기하고 품질을 평가하기 때문에, 기존의 연구에서 하는 방식인 BLEU 점수를 베이스라인으로 사용하는 것이 불가능하다. 따라서 코사인 유사도를 활용하는 TransQuest[6]의 다국어 모델을 베이스라인으로 사용하되, 제로샷 전이학습(zero-shot transfer learning)의 결과를 비교하여 TwiQE의 성능을 증명했다.

2.2. 한국어가 포함된 QE

영어-한국어 QE를 처음 시도한 [7]은 영한 QE 데이터 구축 방안을 제시하고, 심층학습 기반의 predictor-estimator 구조의 모델에 영어-한국어 언어쌍을 적용한 모델을 선보였다. 하지만 품질 예측이 HTER 점수로 이뤄지기 때문에 DA 점수를 예측하지는 않는다. [8]은 한국어-영어 QE에 대해 연구했으며, 데이터 증강을 통해 의사 데이터(pseudo-data)를 생성한 후 학습을 진행하는 방법을 제안했다. 하지만 이 역시 HTER 점수 기반으로 연구가 진행되었다.

DA 점수는 문장에 담긴 의미와 화용론적 정보, 문법적 오류 등을 고려해 종합적이고 직관적으로 품질을 판단하나, HTER은 정해진 참조 문장을 기준으로 기계 번역문이 참조 문장으로부터 얼마나 다른지를 측정한다. 이렇듯 번역의 품질에 대한 관점이 명백히 다르기 때문에 [9]에 따르면 두 점수의 상관관계 또한 낮게 나타난다. HTER 기반의 한영·영한 QE 연구만 수행된 가운데, DA 점수 기반의 한영·영한 QE 연구가 절실한 시점이다. 이에 본 연구에서 한영·영한 DA 점수 예측 모델 TwiQE를 제시하는 바이다.

3. 번역문 품질 평가 모델의 제안: TwiQE

본 연구에서 구축한 모델인 TwiQE는 크게 세 가지의 특징이 있다. 첫 번째 특징은 원문-번역문 코사인 유사도와 원문-역번역문 코사인 유사도를 동시에 사용해 모델의 성능을 높였다는 점이다. 두 번째 특징은 기학습된 언어 모델을 이용하기 때문에 비교적 적은 자원으로 고품질의 문장 임베딩을 얻어낼 수 있다는 점이다. 세 번째는 사람이 직접 평가하는 DA(Direct Assessment) 점수를 품질 점수로 예측한다는 점이다.

TwiQE는 원문-번역문 간 코사인 유사도와 원문-역번역문 문장 간의 코사인 유사도를 동시에 사용한다. [10]에 따르면 역번역문을 이용한 품질 예측은 관련된 데이터셋이 많이 존재하지 않는 저자원 언어에서 더욱 효과적이다. 그러나 원문-역번역문 문장 쌍의 유사도로만 품질 예측을 할 경우에는 기계번역기의 성능에 따라 역번역문의 품질이 결정된다는 단점이 있다. 즉, 기계번역의 오류율이 그대로 역번역문에 반영되는 것이다. 이를 보완하기 위해 원문-번역문 문장 쌍의 유사도 점수도 함께 참고하여 최종 품질 점수를 예측했다. 실제로 항목 4.3.2에서 절제 연구(ablation study)를 수행한 결과, 원문- 번역문 간 코사인 유사도와 원문-역번역문 간 코사인 유사도 둘 다 예측 성능에 유의미한 영향을 끼친다는 사실을 알 수 있었다.

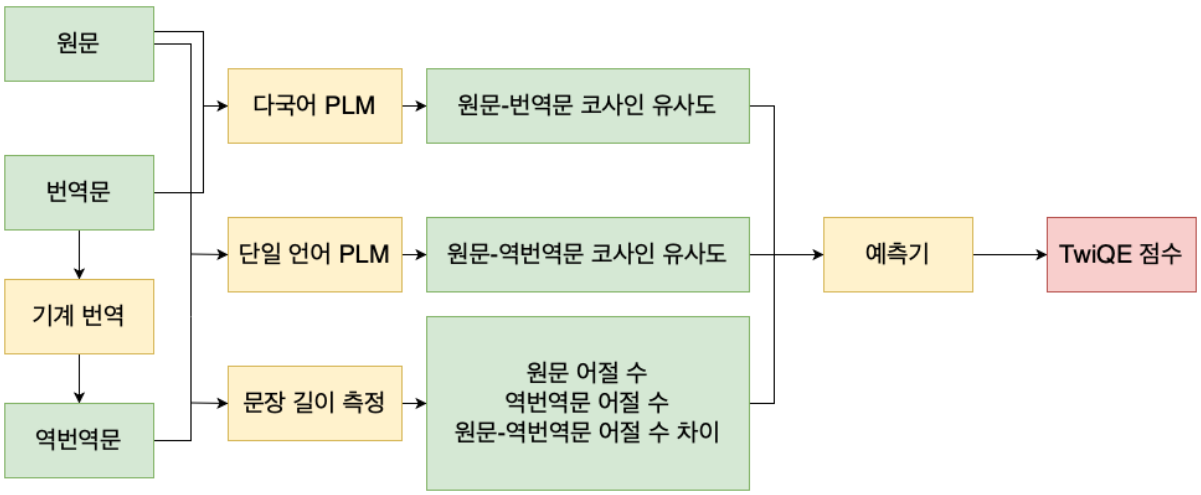

그림 1 TwiQE 모델 구조도.

원문-번역문 코사인 유사도와 원문-역번역문 코사인 유사도, 원문 어절 수, 역번역문 어절 수, 원문-역번역문 어절 수 차이는 모두 자질(feature)로써 예측기 훈련에 사용되었다.

실험 과정은 다음의 세 단계로 정리할 수 있다.

첫 번째는 데이터 구축 단계이다. 기존의 QE 데이터셋에는 한국어가 포함되어 있지 않기 때문에 언어 전문가에게 의뢰해 총 2,600 문장의 번역문에 대해 DA 점수를 평가하고 모델 학습과 검증을 진행했다.

두 번째는 TwiQE 모델 구축 단계이다. 모델의 구축 과정은 다시 두 단계로 정리할 수 있다. 첫 번째는 PLM(Pre-trained Language Model, 이하 PLM)의 검증 과정이다. TwiQE는 PLM의 문장 임베딩을 사용하는 모델이다. 따라서 PLM의 성능에 의존도가 높으므로 사전에 성능에 대한 판단은 필수이다. 두 번째는 예측기의 학습 과정이다. 이 과정은 다시 네 단계로 나뉜다. (i) 번역 병렬 말뭉치에서 기계번역으로 역번역문을 추가했다. (ii) PLM을 이용해, 원문, 번역문, 역번역문 각각에 대한 문장 임베딩을 얻어냈다. (iii) 앞서 얻어낸 임베딩들을 사용하여 원문-번역문 쌍의 코사인

유사도와 원문-역번역문 코사인 유사도를 얻어냈다. (iv) 두 개의 코사인 유사도에 어절 수에 대한 정보가 담긴 추가적인 자질을 포함하여 DA 점수를 예측하도록 모델을 훈련했다.

마지막 단계는 TwiQE 모델의 검증 과정이다. 성능 검증을 위해서 비교 모델로 TransQuest[6]을 사용했으며 성능은 상관 계수를 이용해 측정했다. 구축한 QE 데이터 중, 학습에 사용되지 않도록 분리한 테스트 데이터에 대해 단문 단위로 점수를 예측했다. 이후 모델이 예측한 DA 점수와 사람이 평가한 DA 점수의 상관 계수를 확인하여 모델의 성능을 확인했다. 또한 예측기 훈련에 사용된 자질(feature)들이 유효함을 보이기 위해 절제 연구(ablation study)를 수행했다. 이 연구를 통해 각 자질들을 하나씩 제거해보며, 개별 자질이 모델의 성능에 얼마나 영향을 미치는지 확인하였다.

4. 실험

4.1. 한국어-영어/ 영어-한국어 QE 데이터 구축

기존 DA 점수 예측 QE 데이터셋에는 한국어가 포함되어 있지 않기 때문에 새로이 구축할 필요가 있었다. QE 데이터 구축 과정은 다음과 같다. 먼저 AI HUB에 게시된 한영·영한 번역 병렬 말뭉치를 확보했다. 그 후 다국어 PLM[4]으로 원문-번역문의 코사인 유사도를 구하고, 이를 기반으로 점수대 별 동일한 수를 샘플링했다. 이후 전문가에게 의뢰해 총 2,600쌍의 원문-번역문 문장 쌍에 대해 사람이 평가한 DA 점수를 얻어냈다.

WMT21[1]에서 제시한 DA 점수 체계는 FLORES[11] 가이드라인에 일부 따르며, 100점 척도의 점수 체계를 가진다. 점수 체계는 다음과 같다.

- 0-10: 완전히 틀린 번역
- 11-29: 몇 개의 단어는 일치하지만 원문으로부터의 전체적인 의미는 다른 번역
- 30-50: 중대한 실수가 있는 번역
- 51-69: 이해 가능하고 전체적인 의미 전달은 되지만 철자나 문법 오류가 있는 번역
- 70-90: 원문의 의미를 거의 보존하고 있는 번역
- 91-100: 완벽한 번역

하지만 이 점수 체계의 몇몇 항목은 모호하게 표현되어 있어 주관적으로 해석될 위험이 크다. 또한 하나의 항목에 해당하는 점수의 범위가 넓어 어떤 점수대로 평가해야 적절할지 헷갈리는 상황이 발생한다. 이에 기존의 연구에 사용된 기준보다 더 촘촘한 새로운 척도를 제시하는 바이다. 이는 사람이 실제로 번역 병렬 말뭉치를 검수할 때 자주 보이는 오류 유형과 그 정도를 분별해서

[4] SBERT[16]를 사용하여 원문-번역문 코사인 유사도를 구했다.

귀납적으로 작성한 규칙으로, 처음 평가하는 평가자들에게도 방향을 제시할 수 있을 정도의 구체성을 지녔다.

귀납적으로 정립한 DA 점수 평가의 기준은 다음과 같다.

- 0-9: 문장이 아예 잘못 매치됨
- 10-19: 일부 단어가 일치하나 문맥이 다름
- 20-29: 구 단위로 의미가 누락되거나 불필요하게 추가됨
- 30-39: 용어의 주요한 오역, 누락, 추가로 인해 문맥 파악에 어려움이 있음
- 40-49: 용어의 주요한 오역, 누락, 추가가 있으나, 문맥 파악에는 어려움이 없음
- 50-59: 일부 단어가 대명사 등으로 대치됨
- 60-69: 고유 명사가 틀리거나 시제, 단어의 용법 등 문법적 오류가 있음
- 70-79: 문맥이 일치하고 누락이나 오역이 없으나 뉘앙스가 약간 다름
- 80-89: 번역 투 등, 원어민이 보기에 어색한 표현이 있을 수 있으나 문장 전체적인 의미가 누락이나 추가 없이 호응함
- 90-100: 어색한 표현이 없고 누락이나 중복, 잉여 표현이 없으며, 관용어나 속담 등이 원어민의 문화에 알맞은 표현으로 번역되었음

여기서 동일 항목의 범위 내의 세부 평가 기준은 유창성으로 한다. 즉, 같은 점수의 범위 내에서도 유창성이 떨어지고 번역 투가 강한 경우에 낮은 쪽으로 점수를 부여한다. 평가의 단위는 앞뒤 문맥이 없음을 가정한 단문 단위로 한다. 자세한 예시는 부록 1에 기술했다.

4.2. 모델 구축: TwiQE

4.2.1. 사전학습 언어모델(PLM, Pre-trained Language Model) 검증 및 선정

TwiQE 모델은 단일 언어 PLM과 다국어 PLM을 미세 조정(fine-tunning) 과정 없이 그대로 사용하므로 PLM의 성능에 의존한다. 따라서 어떤 PLM을 사용할지 결정하기 위해 검증하는 과정을 거쳤다.

단일 언어 PLM 검증은 두 가지 방식으로 이루어졌다. 첫 번째로 의미적인 유사도에 대한 성능을 평가하기 위해서 원문-역번역문 문장 쌍을 이용해 코사인 유사도를 구하고 QE 데이터의 DA 점수와의 상관관계를 분석했다. 두 번째로 해당 언어에 대한 일반적인 이해도, 즉 문법이나 통사적 구조에 대한 성능을 측정하기 위해 프로빙 태스크(probing task)[12, 13]를 이용했다. 프로빙 태스크의 여러 하위 태스크 중 다국어 동시 적용 가능성, 가용한 자원 등을 고려하여 SentLen, WC, TreeDepth, Top- Const, Tense, Negation의 여섯 가지 하위 태스크를 선정했다. 태스크에 적용한 모델의 성능 지표는 기본적으로 정확도(accuracy)를 이용했으나, 라벨별 데이터가 불균형한

TreeDepth와 TopConst의 경우 균형 정확도(balanced accuracy)[5]로 평가했다. 이에 덧붙여 다국어 PLM은 단일 언어 PLM의 두 가지 검증 항목에서 '다국어성(multilinguality)'을 추가 검증했다. 결과적으로 한국어 PLM으로는 KoSimCSE[14, 15]의 KoSimCSE-bert-multitask[6], 영어 PLM으로는 SimCSE[14]의 sup-simcse- roberta-base[7], 다국어 PLM으로는 SBERT[16]의 paraphrase-multilingual-mpnet-base-v2[8]를 선정했다. PLM 검증 및 선정에 대한 자세한 설명과 프로빙 태스크 데이터 구축 과정은 부록 2에 첨부하였다.

4.2.2. 예측기 선정 및 훈련

TwiQE에서는 다섯 가지의 요소를 자질로 사용하여 예측기 훈련을 진행했다.다. 즉, 원문-번역문 코사인 유사도, 원문-역번역문 코사인 유사도, 원문 어절 수, 번역문 어절 수, 원문-역번역문 간 어절 수 차이를 이용하여 예측기를 훈련했다. 역번역문의 경우 구글 번역 API(Cloud Translation API)[9][17, 18]를 이용했는데, 기계번역 중 결과 문장의 일부가 잘려서 나오는 경우가 있어 어절 수에 대한 정보를 추가했다.

모델은 선형 회귀 모델[10]과 그래디언트 부스팅[19] 앙상블 계열인 XGBoost(Extreme Gradient Boosting)[11] 모델[20]을 이용했다. 각 모델에 대해 그리드서치를 수행했고, 상관 계수를 기준으로 최적의 결괏값을 보이는 모델을 선정했다. 모델 간 비교의 결과는 표 1, 표 2와 같다. 모델의 성능은 사람이 측정한 DA 점수와 모델이 예측한 DA 점수와의 상관관계로 측정했다. 1에 가까울수록 강한 상관관계를 나타내고, 이는 모델의 성능이 좋음을 뜻한다.

상관관계 측정에는 피어슨 상관 계수와 스피어만 상관 계수, 켄달 타우 상관 계수를 모두 사용[12]했다. 각 상관 계수에 대한 p-value는 모두 1e-100 이하로 유의한 결괏값임을 보였다. 말뭉치의 방향과 관계없이 모두 XGBoost에서 피어슨 상관 계수가 동일하게 높았다. 비록 타 상관

---

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

[6] https://huggingface.co/BM-K/KoSimCSE-bert-multitask

[7] https://huggingface.co/princeton-nlp/sup-simcse-roberta-base

[8] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[9] 구글 번역 API는 2022년 09월 요청했음.
API 버전은 common-protos-1_3_1이며, 번역 API의 버전은 v2이다.

[10] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model

[11] https://xgboost.readthedocs.io/en/stable/

[12] [21]에 따르면 DA 점수 예측에 피어슨 상관 계수가 적합하다. 한편, 예측 대상인 DA 점수는 각 점수 사이의 등간 간격이 담보되지 않기 때문에 등간 척도보다는 서열 척도의 성격을 가진다. 따라서 서열 척도의 상관관계를 측정하는 데에 유리한 스피어만 상관 계수도 함께 제시했다. 또한 켄달 타우 상관 계수도 역시 서열 척도에 유리하며, 단조성까지 함께 고려하기 때문에 함께 제시했다.

계수에서 선형회귀 모델보다 상대적으로 낮은 값이 기록되었으나, 각 변수에 대한 설명력이 좋다는 뚜렷한 장점이 있으므로 XGBoost를 최종 TwiQE 모델의 예측기로 선정했다.

표 1 한국어-영어 예측기 후보 모델의 성능을 상관 계수를 통해 비교한 결과

	피어슨 상관 계수	스피어만 상관 계수	켄달 타우 상관 계수
선형회귀	0.774879	0.773738	0.580343
선형회귀-릿지	0.775882	**0.774285**	0.581228
선형회귀-라쏘	0.776260	0.774226	**0.581280**
XGBoost	**0.787127**	0.760178	0.568419

표 2 영어-한국어 예측기 후보 모델의 성능을 상관 계수를 통해 비교한 결과

	피어슨 상관 계수	스피어만 상관 계수	켄달 타우 상관 계수
선형회귀	0.888336	**0.849370**	**0.665742**
선형회귀-릿지	0.888221	0.849298	0.665655
선형회귀-라쏘	0.888033	0.848921	0.665212
XGBoost	**0.901194**	0.828153	0.644776

4.3. 모델 성능 평가

4.3.1. 베이스라인과의 성능 비교

코사인 유사도를 활용하는 TransQuest[6]를 베이스라인으로 활용해 TwiQE의 성능을 검증했다. TransQuest는 한국어 QE 데이터에 대한 미세 조정(fine-tunning)이 이루어지지 않은 다국어 PLM 기반 모델이므로, 제로샷 전이학습(zero-shot transfer learning)을 통해 성능을 비교했다. TwiQE 모델로는 5.1.1.의 과정을 통해 선정한 XGBoost 모델을 사용했다.

더욱 엄밀한 성능 비교를 위해 말뭉치의 언어 방향과 어투(문어체/ 구어체)에 따라 말뭉치를 나누어서 비교를 진행했다. 말뭉치별 비교 결과는 표 3과 같다. 상관 계수에 대해 유의수준은 모두 1e-20 이하로 낮게 나와, 유의미한 값임을 알 수 있었다.

베이스라인과의 비교 역시 사람의 DA 점수와 모델의 DA 점수와의 상관관계를 분석해 성능을 측정했다. 표 3을 참고하면 알 수 있듯이, 모든 언어 방향에서 TwiQE가 베이스라인보다 상관 계수가 더 높게 기록되며, 뛰어난 성능을 보였다.

표 3 말뭉치의 언어 방향과 성격에 따른 베이스라인과 TwiQE의 성능 비교 결과

			피어슨 상관 계수		스피어만 상관 계수		켄달 타우 상관 계수	
			baseline	**TwiQE**	baseline	**TwiQE**	baseline	**TwiQE**
한영	한-영 전체		0.644339	**0.787127**	0.665529	**0.760178**	0.484815	**0.568419**
	기술과학 분야 한-영		0.807929	**0.913217**	0.836671	**0.895681**	0.643002	**0.723066**
	일상생활 및 구어체 한-영	전체	0.578067	**0.712778**	0.590187	**0.678116**	0.421776	**0.489268**
		격식	0.735160	**0.853934**	0.675764	**0.744837**	0.493061	**0.553638**
		비격식	0.502368	**0.638278**	0.524477	**0.636709**	0.373502	**0.463598**
영한	영-한 전체		0.759642	**0.901194**	0.747476	**0.828153**	0.547468	**0.644776**
	전문분야 (식품) 영-한		0.748277	**0.925736**	0.677462	**0.765425**	0.493406	**0.596748**
	일상생활 및 구어체 영-한		0.774238	**0.882130**	0.791232	**0.858551**	0.597247	**0.660349**

4.3.2. 변수 검증

절제 연구(Ablation study)

각 변수의 기여에 대해 검증하기 위해 절제 연구(ablation study)를 실시하여 각각의 변수가 모델의 최종 성능에 얼마나 영향을 미치는지 알아보았다.

기존 모델을 상기해보자면 다음의 다섯 개의 자질이 있다. 원문-번역문 코사인 유사도(f0), 원문-역번역문 코사인 유사도(f1), 원문 어절 수(f2), 역번역문 어절 수(f3), 원문 어절 수와 역번역문 어절 수의 차이(f4)로, 이들 자질 중 일부를 제거해보며 모델의 성능을 측정했다. 성능 측정 결과는 사람이 평가한 DA 점수와 모델이 평가한 DA 점수의 상관 계수를 기준으로 비교했으며, 언어 방향에 따라 나눠서 비교했다. 결과는 표 4와 같다. p-value는 모두 1e-50 이하로 매우 낮게 나오며 결괏값이 유의함을 보였다.

표 4의 결과를 통해 본 연구에서 선정한 자질들이 모두 모델에 일정 부분 기여하고 있음을 알 수 있다. 특히 f1(원문-역번역문 코사인 유사도)을 삭제했을 때 f0(원문-번역문 코사인 유사도)을 삭제한 것보다 성능 하락이 심한 것을 보아, f1의 기여도가 더 높음을 알 수 있다. 어절 수와 관련된 자질(f2, f3, f4) 또한, 이들 모두가 한꺼번에 삭제되었을 때 모델의 성능 하락이 심각하게 나타난 것을 보아 모델 예측에 기여하고 있음을 알 수 있다.

한편, 언어 방향과 상관없이 다섯 개의 자질이 모두 유효했기 때문에 한국어-영어 모델과 영어-한국어 모델 둘 다 다섯 가지 자질을 모두 사용해 TwiQE를 구축했다.

표 4 한국어-영어, 영어-한국어 모델의 학습 자질 차이에 따른 성능 비교.

f0: 원문-번역문 코사인 유사도/ f1: 원문-역번역문 코사인 유사도/ f2: 원문의 어절 수/ f3: 번역문의 어절 수/ f4: 원문과 역번역문의 어절 수 차이

		피어슨 상관 계수	스피어만 상관 계수	켄달 타우 상관 계수
한영	**기본**	**0.787127**	**0.760178**	**0.568419**
	f0삭제	0.767900	0.731132	0.540452
	f1삭제	0.741600	0.731394	0.538543
	f2, f3 삭제	0.779488	0.755101	0.565469
	f2, f3, f4삭제	0.720912	0.687912	0.501484
	f0 only	0.688853	0.655107	0.471362
	f1 only	0.695139	0.663167	0.480948
영한	**기본**	**0.901194**	**0.828153**	**0.644776**
	f0삭제	0.880998	0.804579	0.622396

	f1삭제	0.875481	0.809065	0.620515
	f2, f3 삭제	0.875910	0.795425	0.603214
	f2, f3, f4삭제	0.835291	0.752705	0.553015
	f0 only	0.819115	0.736183	0.537239
	f1 only	0.794028	0.716751	0.515928

변수 중요도 파악

XGBoost 모델은 결정나무(Decision Tree) 기반의 모델이므로 훈련이 완료된 모델을 분석해 변수 중요도를 파악할 수 있다. 변수 중요도를 파악해본 결과 그림 2와 같은 결과가 나왔다. 언어 방향과 상관없이 두 가지 종류의 코사인 유사도 모두 품질 점수를 예측하는 데에 중요한 기여를 하고 있음을 확인할 수 있다.

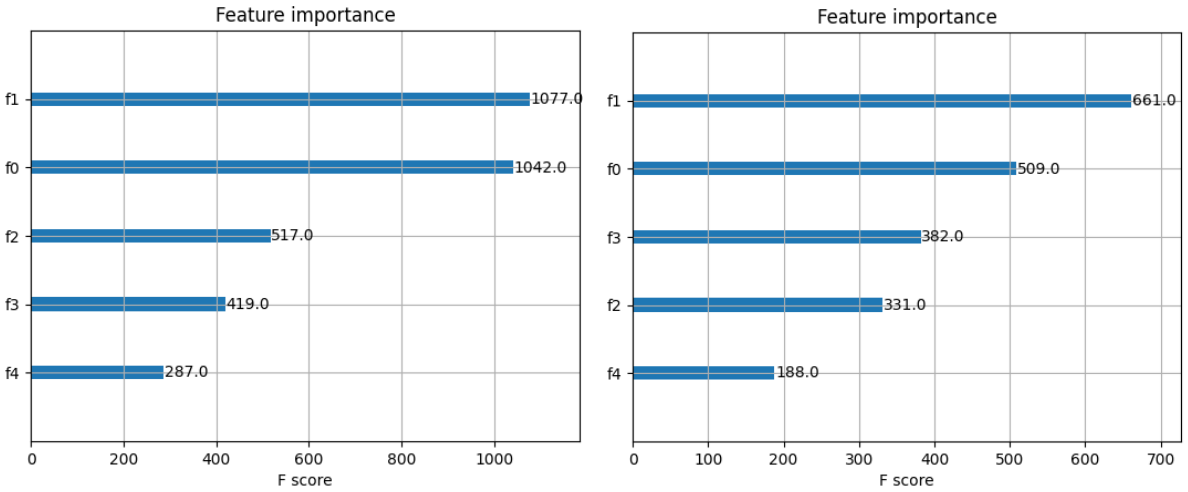

그림 2 (왼) 한-영에 대한 변수 중요도 그래프 (오) 영-한에 대한 변수 중요도 그래프.

f0: 원문-번역문 코사인 유사도/ f1: 원문-역번역문 코사인 유사도/ f2: 원문의 어절 수/ f3: 번역문의 어절 수/ f4: 원문과 역번역문의 어절 수 차이

5. 모델 적용

5.1. TwiQE을 이용한 번역 병렬 말뭉치 품질 점수 산정

앞선 과정을 통해 구축된 TwiQE를 이용해 실제 AI Hub에서 공개한 데이터의 품질을 평가했다. “전문분야 영-한 번역 말뭉치 (식품)”, “기술과학 분야 한-영 번역 병렬 말뭉치 데이터”, “일상생활 및 구어체 한-영 번역 병렬 말뭉치 데이터”를 품질 평가 대상으로 선정했다. 각 말뭉치에 대한 탐색 결과는 부록 3에 담았다.

“일상생활 및 구어체 한-영 번역 병렬 말뭉치 데이터”에는 “한-영”과 “영-한”의 두 가지 언어 방향이 포함되어 있다. 또한 한-영 방향의 번역 병렬 말뭉치 내에서도 “일상생활" 분야의 문장들과 그 이외의 분야의 문장들은 어절 수 평균이나 어투, 발화 상황 등의 면에서 판이하기 때문에 이를 또 “비격식”과 “격식”으로 구분했다. 따라서 “일상생활 및 구어체 한-영 (비격식)", “일상생활 및 구어체 한-영 (격식)", “기술과학 분야 한-영", “일상생활 및 구어체 영-한", “전문분야 (식품) 영-한"과 같이 다섯 가지 종류로 구분하여 분석을 실시했다.

각 말뭉치의 품질을 TwiQE를 통해 DA 점수를 예측한 결과 대부분의 번역문에 대해 높은 DA 점수를 예측하였고, 따라서 번역의 품질도 대부분 훌륭함을 알 수 있었다. 또한 점수대에 따른 번역문의 개수를 그래프로 그려봤을 때 대부분 80~90점 대의 점수를 가진 번역문이 가장 많았다(그림 3-7). 대체로 구어체 말뭉치가 문어체 말뭉치보다 높은 점수대가 많았다. 한편 “기술과학 분야 한-영” 말뭉치는 그림 5에 보이듯이, 다른 말뭉치와 비교했을 때 품질 점수가 대체로 낮은 점수대 쪽으로 치우쳐져 있었다.

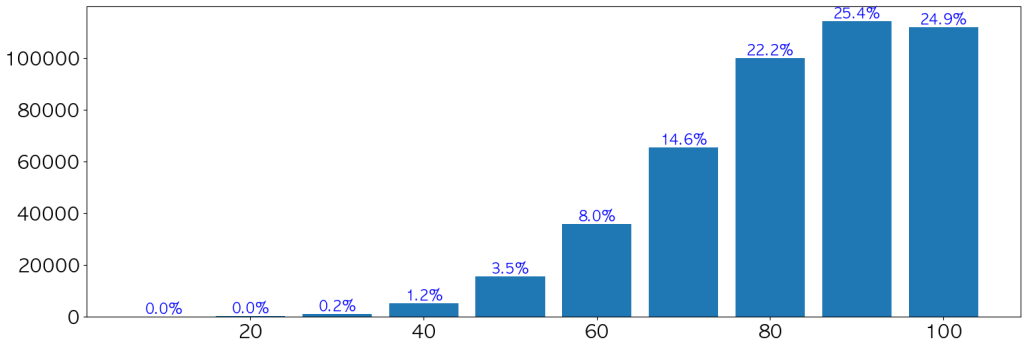

그림 3 일상생활 및 구어체 한-영 (비격식) 점수 별 빈도 막대 그래프

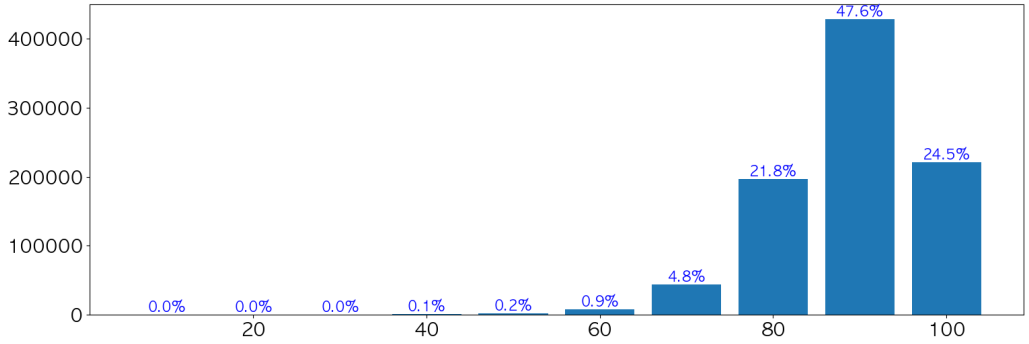

그림 4 일상생활 및 구어체 한-영 (격식) 점수 별 빈도 막대 그래프

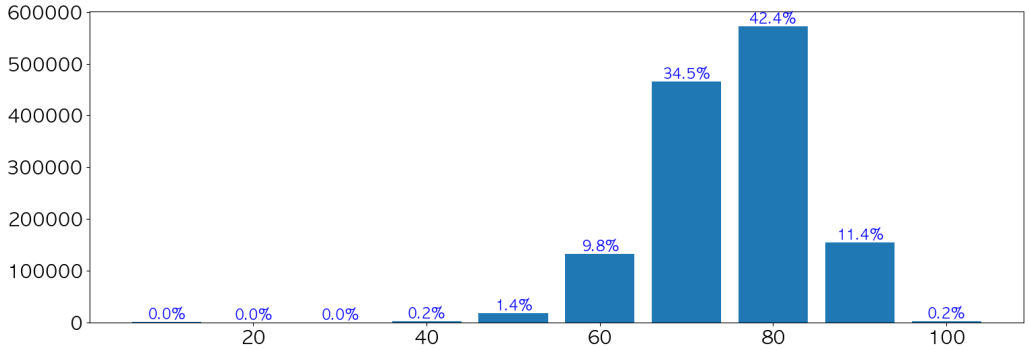

그림 5 기술과학 분야 한-영 점수 별 빈도 막대 그래프

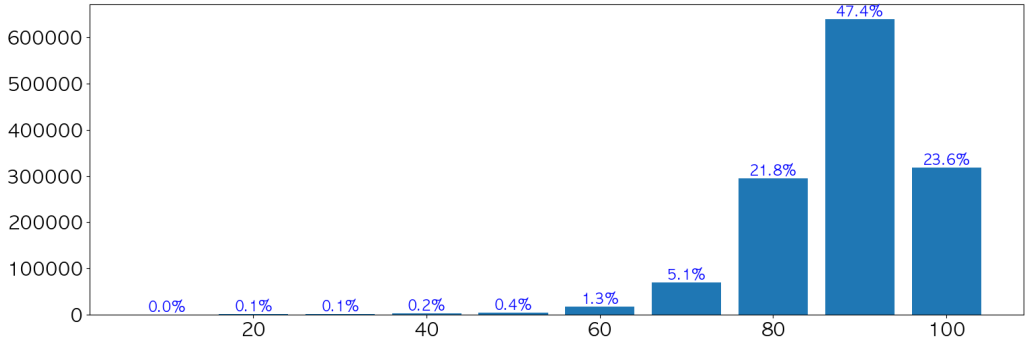

그림 6 일상생활 및 구어체 영-한 점수 별 빈도 막대 그래프

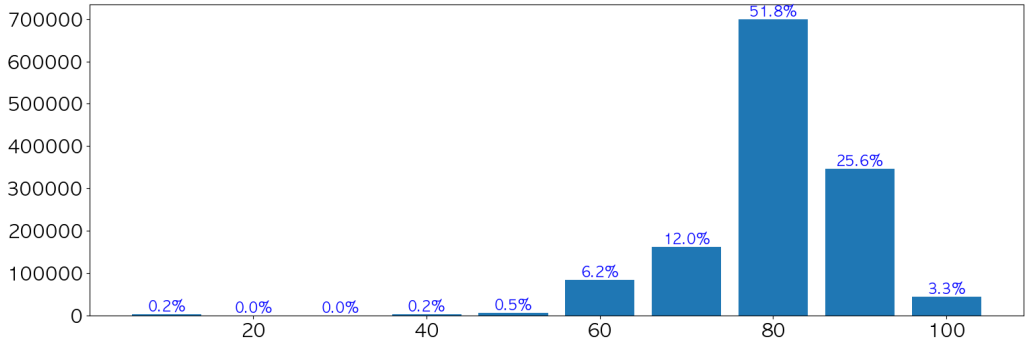

그림 7 전문분야 (식품) 영-한 점수 별 빈도 막대 그래프

5.2. 말뭉치 품질 평가

5.1에서 도출한 TwiQE 점수를 바탕으로 말뭉치 품질 평가를 위한 기준을 마련하였다. 그 기준은 두 가지의 임곗값(threshold)으로 제시된다. 첫 번째 임곗값은 '원문과 번역문이 판이하여 새롭게 번역 작업이 필요하다고 판단되는 점수(이하 1차 임곗값)'로, 1차 임곗값 이하의 점수를 가지는 번역문은 완전히 새로운 번역 작업이 필요한 정도의 좋지 않은 품질을 의미하도록 설정했다. 두 번째 임곗값은 '문장의 일부 단어 등 지엽적인 부분에서 수정이 필요하다고 판단되는 점수(이하 2차 임곗값)'로, 2차 임곗값 이상의 TwiQE 점수를 획득한 번역문은 문법이나 단어의 오류가 없는 번역문으로 분류했다.

번역 병렬 말뭉치를 언어 방향과 어투에 따라 다섯 종으로 구분하여 각각에 대해 복수의 전문가가 임곗값을 수립했고, 이를 평균 내어 각 번역 병렬 말뭉치에서 개선이 필요한 데이터의 수량과 비율을 구했다. 결과는 표 5와 같다.

1차 임곗값 이하는 모든 말뭉치에서 꽤 낮은 비율로 나타났다. "일상생활 및 구어체 한-영(비격식)"을 제외하고는 모두 1% 미만의 낮은 값을 보였다. 비율로 따지면 작은 수치지만, 숫자로 따지면 많은 수의 오류 데이터를 TwiQE를 이용하여 감지할 수 있다는 사실은 무척 고무적이다.

한편, 1차 임곗값 초과 2차 임곗값 이하 구간에 해당하는 오류 데이터는 말뭉치에 따라 그 비율이 판이하게 달랐다. 특히 "기술과학 분야 한-영"에서는 그 비율이 타 말뭉치에 비해 현저하게 높게 나타났다. 그림 5의 분포를 살펴보아도, 타 말뭉치와 비교해 "기술과학 분야 한-영" 말뭉치는 낮은 점수대에 품질 점수가 쏠려있는 경향을 보인다.

위의 현상에는 두 가지의 해석이 있을 수 있다. 첫 번째는 "기술과학 분야 한-영"이 타 말뭉치에 비해 낮은 품질의 말뭉치라는 해석이다. 두 번째는 "기술과학 분야 한-영" 데이터에 대해서는 전문가가 더 까다롭게 2차 임곗값을 설정했다는 해석이다.

표 5 TwiQE를 이용한 병렬 말뭉치 품질 평가 결과

	1차 임곗값 이하 (very bad)	1차 임곗값 초과 ~ 2차 임곗값 이하 (bad)	2차 임곗값 초과 (good)
일상생활 및 구어체 한-영 (격식)	2,326 쌍 (0.26 %)	94,954 쌍 (10.55 %)	802,938 쌍 (89.19 %)
일상생활 및 구어체 한-영 (비격식)	6,235 쌍 (1.39 %)	91,697 쌍 (20.39 %)	351,850 쌍 (78.23 %)
기술과학 분야 한-영	3,068 쌍 (0.22 %)	920,619 쌍 (68.19 %)	426,475 쌍 (31.59 %)
일상생활 및 구어체 영-한	2,720 쌍 (0.2 %)	64,719 쌍 (4.8 %)	1,282,906 쌍 (95.01 %)

전문분야 (식품) 영-한	6,032 쌍 (0.45 %)	45,164 쌍 (3.35 %)	1,298,804 쌍 (96.21 %)

6. 결론

6.1. 연구의 의의 및 한계점

TwiQE 모델의 구축과 적용을 통해, 해당 번역 병렬 말뭉치에서 개선이 필요한 데이터를 구분할 뿐 아니라, 개선을 위해 요구하는 수정 작업의 수준도 제시하여 번역 병렬 말뭉치에 대한 검수 작업의 편의성을 높였다. 이를 통해 기존에 공개된 번역 병렬 말뭉치의 품질을 향상하고, 차후 구축되는 번역 병렬 말뭉치에 대해서도 선제적인 평가가 가능할 것으로 기대한다.

또한, 한영·영한 QE(Quality Estimation)에 대해 단문 단위 DA 점수 예측 연구를 실시했다. 이 과정에서 한영·영한 DA 점수 QE 데이터를 최초 구축하였으며, DA의 평가 항목을 구체화·세분화하여 DA 평가자들에게 구체적인 지침을 제시할 수 있는 기준을 마련하였다.

고무적인 결과물을 산출했음에도 불구하고, TwiQE 모델은 내재적인 한계를 가졌다. 최종 모델 학습을 위한 주요한 자질로 코사인 유사도가 사용되는데, 이는 PLM에 대한 의존을 높이는 원인이 된다. 또한 TwiQE는 모델의 경제성을 위해 스칼라값으로 도출된 코사인 유사도를 조합하여 최종적인 평가 점수를 산출하는 방식을 택하고 있는데, 문장 임베딩에서 바로 점수를 도출하는 방식보다 정보 소실이 우려되는 상황이다. 마지막으로 역번역문 생성을 위해 기계 번역을 이용했는데, 이를 위한 자원의 투입이 필요하다는 실무적 한계점이 존재한다. 대규모의 데이터를 검수한다면 그에 비례하여 기계 번역 사용을 위한 비용도 증가할 것이다.

6.2. 후속 연구 제안

병렬 말뭉치 평가 결과에 대한 분석

앞서 5.2 항목에서 TwiQE 모델을 적용한 결과, “기술과학 분야 한-영”에서 전체의 68.19%를 차지하는 920,619 문장의 번역문이 1차 임곗값과 2차 임곗값 사이의 점수로 평가받는 것으로 나타났다. 이러한 결과가 2차 임곗값이 잘못 설정되어서 비롯된 것인지, 아니면 실제 데이터의 오류를 잡아낸 것인지 검증할 필요가 있다. 이를 위해 두 가지 연구를 시도해볼 수 있다. 첫 번째는 해당 920,619 문장의 번역문 중 일부를 무작위 비복원 추출하여 실제로 해당 구간(1차 임곗값 초과, 2차

임곗값 이하)에 합당한 품질인지 확인하는 것이다. 즉, 추출한 문장이 단어 등 지엽적인 수정이 필요한 부분이 있는지 아닌지를 이진 분류로 확인한다. 번역의 오류가 존재하는 데이터의 비율을 산출하여 모델이 오류 데이터를 제대로 골라낸 것이 맞는지 확인할 수 있다. 두 번째는 "기술과학 분야 한-영"과 유사한 도메인을 가지는 한-영 번역 병렬 말뭉치를 대상으로 TwiQE 모델을 적용하여 동일한 임곗값으로 말뭉치를 평가하는 것이다. 만약 기존과 동일한 점수 분포를 보인다면 2차 임곗값의 오류를 의심할 수 있지만, 서로 다른 점수 분포가 나타난다면 데이터의 품질에 문제가 있는 것으로 추측할 수 있다.

모델 발전의 제안

TwiQE는 두 가지 종류의 코사인 유사도뿐만 아니라 어절 수라는 형태적 속성을 자질로 추가하여 성능을 높였다. 이외에도 다양한 문장의 문법적, 형태적 속성을 자질로 추가하여 실험해볼 수 있을 것이다.

한편, TwiQE는 6.1 항목에서 지적했듯이 하나의 숫자로 도출된 코사인 유사도를 사용하기 때문에 문장 임베딩보다는 정보를 많이 소실할 위험이 있다. 이에 원문과 번역문, 역번역문의 임베딩 자체를 이용하여 DA 점수를 예측하는 신경망을 구성하여 성능 비교를 해볼 수 있으리라 기대한다.

감사의 글

-이 논문은 2022년도 과학기술정보통신부의 재원으로 한국지능정보사회진흥원이 추진하는 AI데이터 품질 개선 오픈랩의 지원을 받아 작성된 논문입니다.

-본 연구에 사용된 (데이터명)데이터셋에 대한 정보는 'AI허브(AI-Hub,www.aihub.or.kr)'에서 확인할 수 있습니다.

참고 문헌

[1] Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, André F. T. Martins., "Findings of the WMT 2021 Shared Task on Quality Estimation," In *Proceedings of the Sixth Conference on Machine Translation*, pp. 684-725, 2021.

[2] Hyun Kim, Hun-Young Jung, Hongseok Kwon, JongHyeok Lee, Seung-Hoon Na., “Predictor-estimator: Neural quality estimation based on target word prediction for machine translation,” In *ACM Trans. Asian Low-Resour. Lang. Inf. Process (in press)*, 2017.

[3] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, André F. T. Martins., “Unbabel’s Participation in the WMT19 Translation Quality Estimation Shared Task,” In *Proceedings of the Fourth Conference on Machine Translation*, vol. 3, pp. 78–84, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019.

[5] Guillaume Lample and Alexis Conneau., “Cross-lingual Language Model Pretraining,” In *Advances in Neural Information Processing Systems*, 2019.

[6] Tharindu Ranasinghe, Constantin Orasan and Ruslan Mitkov., “TransQuest at WMT2020: Sentence-Level Direct Assessment,” In *Proceedings of the Fifth Conference on Machine Translation*, pp. 1049-1055, 2020.

[7] 김현, 신재훈, 이원기, 조승우, 이종혁., “심층학습 기반의 Predictor-Estimator 모델을 이용한 영어-한국어 기계번역 품질 예측,” *정보과학회논문지*, 45(6), pp. 545-553, 2018.

[8] 어수경, 박찬준, 서재형, 문현석, 임희석., “단어 수준 한국어-영어 기계번역 품질 예측,” *한국정보과학회언어공학연구회 2021년도 제33회 한글 및 한국어 정보처리 학술대회*, pp. 9-15, 2021.

[9] Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, André F. T. Martins., “MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset,” In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4963-4974, 2022.

[10] Jihyung Moon, Hyunchang Cho and Eunjeong L. Park., “Revisiting Round-trip Translation for Quality Estimation,” In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 91–104, 2020.

[11] Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, Marc'Aurelio Ranzato., "The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English," In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 6098-6111, 2019.

[12] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, Marco Baroni., "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties," In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 2126-2136, 2018.

[13] 안애림, 고병일, 이다니엘, 한경은, 신명철, 남지순., "한국어 문장 임베딩의 언어적 속성 입증 평가," *한국정보과학회언어공학연구회 2021년도 제33회 한글 및 한국어 정보처리 학술대회*, pp. 161-166, 2021.

[14] Tianyu Gao, Xingcheng Yao and Danqi Chen., "SimCSE: Simple Contrastive Learning of Sentence Embeddings," In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

[15] Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, Hyungjoon Soh., "KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding," In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 422–430, 2020.

[16] Nils Reimers and Iryna Gurevych., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," In *Published at EMNLP 2019*, 2019.

[17] google api (2016): Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[18] google api (2022): Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. arXiv preprint arXiv:2205.03983.

[19] Jerome H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, pp. 1189–1232, 2001.

[20] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," *In Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining,* pp. 785-794, ACM, 2016.

[21] Nils Reimers, Philip Beyer and Iryna Gurevych., "Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity," In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 87–96, 2016.

[22] Philipp Dufter and Hinrich Schütze., "Identifying Elements Essential for BERT's Multilinguality," In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437, 2020.

[23] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, Melvin Johnson., "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization," In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

부록 1. DA 점수 평가의 기준과, 그 기준에 따른 구체적 예시

표 6 DA 점수 평가의 기준과, 그 기준에 따른 구체적 예시

DA 점수	설명	예시		
		한국어 원문	영어 번역문	참조 문장 (gold sentence)
0-9	문장이 아예 잘못 매치됨	구체적이어야 합니다.	Our guests always enjoy taking them home as a souvenir.	You must be specific.
10-19	일부 단어가 일치하나 문맥이 다름	완료되었으므로 지금 시청할 수 있습니다.	Thousands of Korean dramas await you!	It's done, so you can watch it now.
20-29	구 단위로 의미가 누락되거나 불필요하게 추가됨	이러한 방법을 통해 우리 학교를 깨끗이 하여 쾌적한 환경에서 생활할 수 있을 것입니다.	In these ways, we can live in a more pleasurable environment.	By keeping our school clean in this way, we can have more pleasurable school life.
30-39	용어의 주요한 오역, 누락, 추가로 인해 문맥 파악에 어려움이 있음	아이아스는 '장님'이 돼 소 떼와 양 떼를 마구잡이로 죽인다.	Ias becomes a "jang" and kills herds and sheep recklessly.	AIas becomes ‘blind’ and kills herds and sheep recklessly.
40-49	용어의 주요한 오역, 누락, 추가가 있으나 문맥 파악에는 어려움이 없음	그런 걸 할 수 있다는 말을 저희한테 안 하시고!	I can't believe you nev er told us you could produce something like that!	I can't believe you never told us you could do something like that!
50-59	일부 단어가 대명사 등으로 대치됨 (단문 단위의 품질 평가이기 때문에 문맥이 없다고 가정)	"진한 홍차로 주문하겠습니다."	"I'd like a strong one, please."	"I'd like strong black tea, please."

60-69	고유 명사가 틀리거나 시제, 단어의 용법 등 문법적 오류가 있음	향후 계속된 예후 관찰이 필요하다고 생각됩니다.	It was thought that it was necessary for constant watch out.	I believe that consistent follow-up is necessary. (시제가 틀리고 that절 주술 관계 왜곡)
70-79	문맥이 일치하고 누락이나 오역이 없으나 뉘앙스가 약간 다름	안녕하세요, 저는 이제 반 50살이 됐습니다.	Hello, I'm 25 years old this year.	Hello, I turned half-fifty this year. (‘이제' 막 되었다는 뉘앙스와 유머러스한 뉘앙스 상실)
80-89	번역투 등, 원어민이 보기에 어색한 표현이 있을 수 있으나 문장 전체적인 의미가 누락이나 추가 없이 호응함	네가 기대할 만하다고 해서 지금 엄청나게 기대하고 있어.	I'm can't wait now because you said it's worth looking forward to.	I’m looking forward to it because you said it’s worth the wait.
90-100	어색한 표현이 없고 누락이나 중복, 잉여 표현이 없으며, 관용어나 속담 등이 원어민의 문화에 알맞은 표현으로 번역되었음	자꾸 되풀이해서 말하지 말아 주세요.	Don't harp on the same string.	Don't harp on the same string. (왼쪽과 동일)

부록 2. PLM 검증 및 선정의 구체적 과정과 결과

한국어 PLM 검증 및 선정

검증을 진행할 한국어 PLM으로는 “KoSimCSE-bert-multitask”[14, 15]와 “KoSimCSE-roberta-multitask”[14, 15]를 선정했다. 이는 2019년에 제안된 SimCSE[14]의 한국어 버전이다. 표 7과 표 8의 결과를 바탕으로 한국어 모델로 “KoSimCSE-bert-multitask”를 선정했다. p-value는 모두 1e-50 이하 수준으로, 결괏값이 유효함을 보였다.

표 7 한국어 PLM의 유사 문장에 대한 성능을 측정하기 위한 원문-역번역문 코사인 유사도와 DA 점수와의 상관관계

	피어슨 상관 계수	스피어만 상관 계수	켄달 타우 상관 계수
(1)	**0.645156**	**0.680376**	**0.515402**
(2)	0.640873	0.676079	0.514319

(1): KoSimCSE-bert-multitask (2): KoSimCSE-roberta-multitask

표 8 한국어 PLM의 언어적 속성에 대한 성능을 측정하기 위한 프로빙 태스크

	SentLen	WC	TopConst	TreeDepth	Negation	Tense	Average
(1)	**62.30**	**40.33**	**34.93**	**36.80**	**93.54**	**95.00**	**60.48**
(2)	54.98	38.46	31.35	34.36	92.50	93.76	57.57

(1): KoSimCSE-bert-multitask (2): KoSimCSE-roberta-multitask

영어 PLM 검증 및 선정

영어 모델 중에서는 SimCSE[14]의 “sup-simcse-bert-base-uncased”와 “sup-simcse-roberta-base”, 그리고 SBERT[16]의 “all-MiniLM-L12-v2”를 선택했다.

표 9와 표 10의 결과에 따라, DA 점수와의 높은 상관관계를 보여주면서 프로빙 태스크에서 두 번째로 높은 점수를 기록한 “sup-simcse-roberta-base”를 선정했다. p-value는 모두 1e-30 이하 수준으로, 결괏값이 유효함을 보였다.

표 9 영어 PLM의 유사 문장에 대한 성능을 측정하기 위한 원문-역번역문 코사인 유사도와 DA 점수와의 상관관계

	피어슨 상관 계수	스피어만 상관 계수	켄달 타우 상관 계수
(1)	0.517203	0.543472	0.405463

(2)	**0.577669**	**0.605038**	**0.450498**
(3)	0.487008	0.497564	0.371495

(1): sup-simcse-bert-base-uncased (2): sup-simcse-roberta-base (3): all-MiniLM-L12-v2

표 10 영어 PLM의 언어적 속성에 대한 성능을 측정하기 위한 프로빙 태스크

	SentLen	WC	TopConst	TreeDepth	Negation	Tense	Average
(1)	**67.19**	**56.63**	**61.05**	**20.49**	**96.19**	**94.91**	**66.08**
(2)	61.65	52.42	56.39	19.49	93.97	94.33	63.04
(3)	52.01	54.48	43.42	17.16	89.57	87.29	57.32

(1): sup-simcse-bert-base-uncased (2): sup-simcse-roberta-base (3): all-MiniLM-L12-v2

다국어 PLM 검증 및 선정

다국어 PLM은 원문-번역문 문장 쌍 코사인 유사도를 구할 때 사용된다. 다국어 PLM은 단일 언어 PLM의 두 가지 검증 항목에서 '다국어성(multilinguality)'을 추가 검증했다. 다국어성은 교차언어 문장 검색(cross-lingual sentence retrieval)과 교차 언어 프로빙 태스크(cross-lingual probing task)를 통해 검증했다. 검증 대상이 된 모델은 SBERT[16]의 "distiluse-base-multilingual-cased-v2(이하 distiluse)" 모델과 "paraphrase-multilingual-mpnet-base-v2(이하 mpnet)" 모델이다.

- 다국어 PLM의 교차 언어 문장 검색(Cross-lingual sentence retrieval)

교차언어 문장 검색은 다국어성(Multilinguality)의 정도를 평가하는 유명한 방법[22]으로, 원문과 번역문 간의 코사인 유사도를 구하여 다국어성의 정도를 가늠하는 방식이다. 이 방법은 이상적인 원문 번역문 쌍을 전제하고 있다. 이에 전문가가 평가한 번역 병렬 말뭉치 중 DA 점수 95점 이상을 달성한 한국어-영어 110쌍, 영어-한국어 112쌍을 선정하여 이상적인 병렬 데이터를 생성하고, 원문과 번역문 간의 코사인 유사도를 구했다. 그 후 이를 평균 내어 다국어성을 평가하였다. 표 11은 그 결과이며, 1에 가까울수록 좋은 점수이다.

표 11 교차언어 문장 검색 방식을 통한 다국어 모델 평가

	피어슨 상관 계수	스피어만 상관 계수	켄달 타우 상관 계수
(1)	0.629354	0.656439	0.487733
(2)	**0.648913**	**0.674689**	**0.503452**

(1): distiluse-base-multilingual-cased-v2 (2): paraphrase-multilingual-mpnet-base-v2

- 다국어 PLM의 교차 언어 프로빙 태스크(Cross-lingual probing task)

다국어성을 평가하기 위해 통상적으로 [23]에서와 같이 개체명 분석이나 감성 분석과 같은 다운스트림 태스크를 제로샷 전이 학습으로 수행한다. 본 연구에서는 문장 임베딩을 직접 평가하되 비교적 간단한 자원으로도 다국어성을 평가하기 위해, 프로빙 태스크를 제로샷 전이 학습으로 수행했다. 이에 대한 결과는 아래 표 12, 표 13과 같다.

두 개의 모델 간에 성능이 좋은 태스크가 갈리므로, [13]에서 수행했던 방식과 동일하게 평균을 내어 최종 모델을 선택한다. 한영과 영한 데이터에서 모두 mpnet 모델이 우수한 성능을 보였다. 이는 앞선 다국어성 검증 결과와도 동일한 결과로, mpnet이 다국어에 강한 모델임을 알 수 있다.

표 12 다국어 PLM의 언어적 속성에 대한 성능을 측정하기 위한 한국어-영어 프로빙 태스크

	SentLen	TreeDepth	Negation	Tense	Average
(1)	33.57	27.82	**94.95**	**93.02**	62.34
(2)	**42.04**	**28.07**	92.5	92.15	**63.69**

(1): distiluse-base-multilingual-cased-v2 (2): paraphrase-multilingual-mpnet-base-v2

표 13 다국어 PLM의 언어적 속성에 대한 성능을 측정하기 위한 영어-한국어 프로빙 태스크

	SentLen	TreeDepth	Negation	Tense	Average
(1)	37.18	21.48	**89.08**	**94.1**	60.46
(2)	**45.85**	**21.7**	86.43	92.15	**61.53**

(1): distiluse-base-multilingual-cased-v2 (2): paraphrase-multilingual-mpnet-base-v2

- 다국어 PLM의 각 단일 언어에 대한 프로빙 태스크(Monolingual probing task)

개별 언어에 대한 다국어 PLM의 형식적, 문법적, 의미적인 측면에서의 성능을 측정하기 위해 각 단일 언어에 대한 프로빙 태스크를 수행했다. 그 결과 distiluse 모델이 WC를 제외한 모든 태스크에서 좋은 성능을 보였다.

표 14 다국어 PLM의 언어적 속성에 대한 성능을 측정하기 위한 한국어 프로빙 태스크

	SentLen	WC	TopConst	TreeDepth	Negation	Tense	Average
(1)	**74.76**	34.07	**31.08**	**38.33**	**93.54**	**95.34**	**61.19**

(2)	63.91	**35.87**	29.70	36.28	91.90	93.40	58.51

(1): distiluse-base-multilingual-cased-v2 (2): paraphrase-multilingual-mpnet-base-v2

표 15 다국어 PLM의 언어적 속성에 대한 성능을 측정하기 위한 영어 프로빙 태스크

	SentLen	WC	TopConst	TreeDepth	Negation	Tense	Average
(1)	**78.11**	45.83	**58.78**	**21.37**	**96.52**	**93.86**	**65.75**
(2)	65.62	**55.86**	53.90	21.08	95.34	93.47	64.21

(1): distiluse-base-multilingual-cased-v2 (2): paraphrase-multilingual-mpnet-base-v2

앞선 결과들을 바탕으로, 다국어성의 성능이 더 우수한 mpnet을 다국어 모델로 선택하여 연구를 진행했다.

프로빙 태스크 데이터셋 구축 방식

프로빙 태스크를 수행하기 위해, 몇 가지 규칙에 의해 프로빙 태스크 데이터를 준비했다. 이후 각 PLM에 분류기를 붙이고 준비한 프로빙 태스크 데이터를 학습시켜 성능을 평가하는 방식으로 PLM 검증을 진행했다. [12]에 제안된 방식을 따라 WC를 제외한 나머지 태스크에서는 분류기로 MLP를 사용했고, WC에서만 로지스틱 회귀 모델을 이용했다. 분류기가 MLP인 경우 validation 기준 정확도가 처음으로 하락하는 시점 이전의 학습 결과를 비교했으며, TreeDepth와 TopConst 데이터는 라벨별 데이터가 불균형하기 때문에 균형 정확도 점수를 사용하였다.

프로빙 태스크 데이터셋은 AI Hub에 공개된 “일상생활 및 구어체 한-영 번역 병렬 말뭉치 데이터”, “기술과학 분야 한-영 번역 병렬 말뭉치 데이터”, “전문분야 영-한 번역 말뭉치 (식품)”를 이용하여 구축했다.

- SentLen (sentence length)

SentLen은 문장 길이 분류 태스크로, 문장의 길이 정보가 임베딩에 얼마나 정확하게 저장되는지 평가하는 태스크이다. 어절 수 길이에 따라 7가지로 분류했으며, 그 범위를 넘어서는 데이터는 활용하지 않았다. 어절 범위는 1~4 어절, 5~8 어절, 9~12어절, 13~16어절, 17~20어절, 21~25어절, 26~28 어절로 나눴고, 수량은 라벨 별 20,000개로 통일하였다.

- WC (word content)

WC는 중빈도 어휘에 따른 분류 태스크로, 중빈도 어휘를 라벨로 하여 특정 문장이 주어졌을 때 어떤 중빈도 어휘를 가지는지에 따라 분류하는 태스크이다. 데이터를 구축하기 위해 각 문장을 어절 단위로 잘라 빈도수를 측정한 뒤, 고빈도 어휘에서 저빈도 어휘 순으로 정렬하였다. 그리고 6,000번째 어휘부터 6,999번째 어휘까지 1,000개의 어휘를 선정하였고, 그것을 바탕으로 다른 어휘와 중복되지 않는 한 개의 어휘만을 라벨로 가지는 문장들을 선별하였다. 라벨별로 데이터의 수를 측정하여 너무 적은 데이터를 가지는 라벨은 제거하였고, 결과적으로 한국어에서 996개의 라벨, 영어에서 915개의 라벨이 남았다. 최종 데이터에는 라벨별로 100개의 데이터를 저장하였다.

- TreeDepth (tree depth)

TreeDepth는 통사 구조의 최대 깊이 추론 태스크로, 문장의 의존 구조를 분석하여 루트 노드로부터 가장 긴 깊이를 파악하는 태스크이다. 한국어는 5에서 12까지 8개의 라벨, 영어는 4에서 9까지 6개의 라벨로 이루어졌다. 교차 언어 프로빙 태스크를 진행할 때는 각각 5에서 9까지 5개의 라벨만 사용했다.

- TopConst (top constituents sequence)

TopConst는 최상위 구성 성분의 하위 노드 예측 태스크로, 문장의 의존 구조에서 루트 노드의 하위 노드가 어떤 성분인지 파악하는 태스크이다. 각 라벨은 하위 노드의 빈출 순서에 따른 19개의 분류와 그 외 분류를 포함하는 'OTHER'까지 20개로 이루어졌다. 'OTHER'의 수량은 전체 데이터 수량의 5% 이하로 제한하였다.

- Tense

Tense는 문장의 시제를 분류하는 태스크이다. 라벨은 과거와 비과거로 나누었다. 데이터셋 구축을 위해 한국어에서는 "-었-", "-았-", "-였-"과 같은 시제 선어말어미, 영어에서는 과거 동사의 등장을 확인하여 과거 시제를 분류하였다. 전체 데이터에서 과거와 비과거를 나누고, 각 라벨의 수량을 동일하게 추출하여 학습할 데이터셋을 구축하였다.

- Negation

Negation은 긍정문과 부정문을 분류하는 태스크이다. 시제 분류와 마찬가지로 형태소 분석기를 이용하여 한국어에서는 "안", "못"과 같은 일반 부사, "않", "말", "못하"와 같은 보조 용언, "아니", "없", "모르", "안되" 등의 동사나 형용사의 등장을 확인하였고, 영어에서는 "not"이나 "Nothing" 등의 등장 여부로 부정문 여부를 파악하였다. 전체 데이터에서 긍정문과 부정문을 분류하였고, 각 라벨의 수량을 동일하게 추출하여 학습할 데이터를 구축하였다.

부록 3. 번역 병렬 말뭉치 탐색 결과

표 16 번역 병렬 말뭉치 파악

	데이터 수량 (문장쌍 수)	원문 언어 코드	번역문 언어 코드	원문 최대 어절 수	원문 최소 어절 수	번역문 최대 어절 수	번역문 최소 어절 수
(1)	1,350,000	ko	en	78	1	98	1
(2)	1,350,345	en	ko	75	2	55	1
(3)	1,350,162	ko	en	158	1	331	1
(4)	1,350,000	en	ko	142	1	111	1

(1): 일상생활 및 구어체 한-영 번역 병렬 말뭉치 데이터, (2): 일상생활 및 구어체 영-한 번역 병렬 말뭉치 데이터, (3): 기술과학 분야 한-영 번역 병렬 말뭉치 데이터, (4): 전문분야 영-한 번역 말뭉치 (식품)

표 17 번역 병렬 말뭉치 분야 별 수량 파악

	분야	데이터 수량 (문장 쌍 수)
일상생활 및 구어체 한-영	일상생활	449,782
	해외영업	630,197
	해외고객과의채팅	270,021
일상생활 및 구어체 영-한	해외고객과의채팅	270,138
	일상생활	449,952
	해외영업	630,255
기술과학 분야 한-영	정치	179,883
	기술과학	360,206
	세계	449,786
	기후	90,120
	경제	270,167
	글로벌동향정보	404,778

전문분야 (식품) 영-한	위해식품정보	540,110
	법제도정보	90,174
	연구평가정보	314,938

표 18 번역 병렬 말뭉치 별 구체적 예시

	원문	번역문
일상생활 및 구어체 한-영	내년도에는 한국의 자연을 배경으로 한 달력을 구상하고 있습니다.	Next year, we are planning a calendar set in Korea's nature.
	>들어가세요, 들어가세요.	>Get in, Get in.
	아, 답은 분명하죠.	Oh, it's a very obvious question.
일상생활 및 구어체 영-한	We have two types of stainless steel that match your requirements.	귀하의 요구에 맞는 두 종류의 스테인리스가 있습니다.
	Anyway, what's special about it?	어쨌든 특별한 점은 뭐죠?
	Thank you, and we hope you consider our offer.	감사합니다, 저희 제안을 고려해 주시기 바랍니다.
기술과학 분야 한-영	모형 4는 외향성과 심리적 계약 위반의 관계가 부정적일 것이라는 가설 3을 검정하고자 모형 4를 실증 분석하였다.	Model 4 was demonstrated to test hypothesis 3 that the relationship between extroversion and psychological contract violations would be negative.
	전술한 개인활동관리 방법 및 인맥정보 기반의 개인활동관리 방법은 기록 매체에 저장되는 컴퓨터에 의해 실행되는 컴퓨터 프로그램 또는 애플리케이션의 형태로도 구현될 수 있다.	The above-described personal activity management method and personal activity management method based on personal information may be implemented in the form of a computer program or application executed by a computer stored in a recording medium.

	독일 KSpG 제2장 제4조는 이산화탄소 수송에 관한 규정을 두고 있다.	Article 4 of Chapter 2 of the German KSpG provides for the transport of carbon dioxide.
전문분야 (식품) 영-한	The taxonomic identification and comparisons were performed at the genus level.	분류학적 식별 및 비교는 속 수준에서 수행되었다.
	Therefore, multiple transcriptional regulators appear to be involved in adaptation to acid stress.	따라서 여러 전사 조절자가 산성 스트레스에 대한 적응에 관여하는 것으로 보인다.
	Blotch mines—the typical damage symptom caused by T.absoluta—were observed on the leaf surfaces in all the survey locations.	토마토 나방에 의한 전형적인 피해 증상인 블로흐 마인은 모든 조사 위치의 잎 표면에서 관찰되었다.