파파고의 빨간펜 선생님 :
QE 모델 구축과 응용

김현중, 임승현
NAVER Cloud / Papago

NAVER DEVIEW 2023

CONTENTS

NAVER
DEVIEW
2023

1. QE(Quality Estimation) 기술의 중요성

1.1. 자동 번역 평가 모델의 필요성

NAVER
DEVIEW
2023

전문가 평가: 기계번역 모델간 품질을 가장 정확히 비교 및 평가하는 방법

- 시간 및 비용 측면에서 비쌈 (평가 오래걸림 + 요청 횟수에 비례하는 고정비용)
- 평가 데이터 구축 필요함
- 현 ML 산업은 매우 fast-pace, 서비스 모델 개선/업데이트 인터벌이 짧음

자동 번역 평가: 번역 모델 간 품질을 기계가 평가

- 전문가 평가 대비 적은 비용과 시간 투입
- BLEU를 일반적으로 많이 쓰지만 정밀한 평가 불가

1.1. 자동 번역 평가 모델의 필요성

NAVER
DEVIEW
2023

전문가 평가: 기계번역 모델간 품질을 가장 정확히 비교 및 평가하는 방법

- 시간 및 비용 측면에서 비쌈 (평가 오
  릿수에 비례하는 고정비용)
- 평가 데이터 구축 필요함
- 현 ML 산업은 매우 fast-pace, 서
  업데이트 인터벌이 줄어드는 추세 (?)

자동 번역 평가: 번역 모

- 전문가 평가 대비 적은 비용
- BLEU를 일반적으로 많이 쓰지만 정밀한 평가 불가

QE 모델

자동 번역 평가를 위해
사용되는 QE 기술!

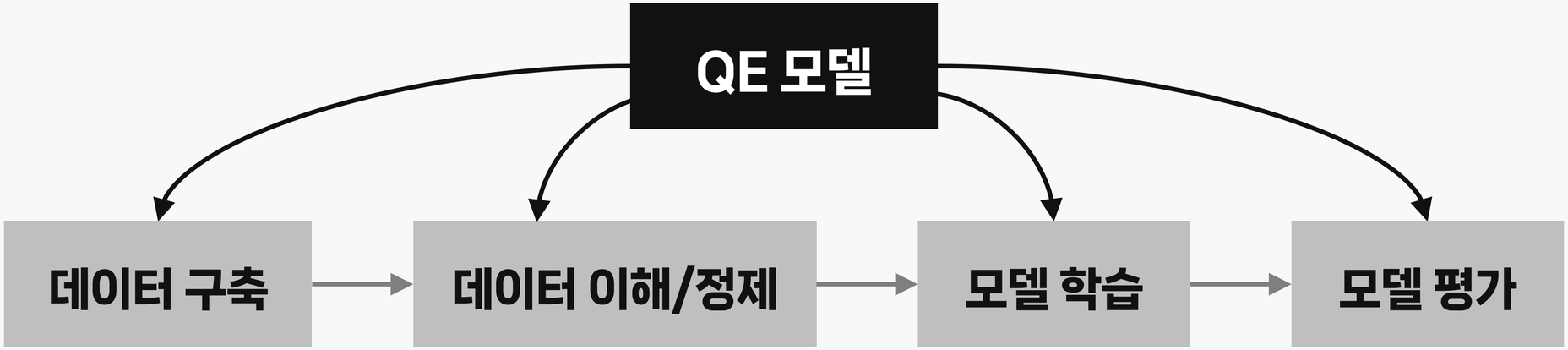
1.2. QE(Quality Estimation) 기술이란?

NAVER
DEVIEW
2023

원문-번역문 Pair의
번역 품질을 예측하는 기술

1.3. 다양한 QE 사용처

NAVER
DEVIEW
2023



단순 평가 용도를 넘어 모든 MT 프로세스에 활용되는 추세

1.3. 다양한 QE 사용처

NAVER
DEVIEW
2023

- 새로운 데이터 구축시 자동검수 기능
- 기 구축 데이터간 품질 비교 및 이해
- 품질 의심되는 데이터 정제 기능 (e.g. 인공 병렬 문장쌍 데이터)

1.3. 다양한 QE 사용처

NAVER
DEVIEW
2023

- 효율적인 학습에 이용 가능

(e.g. curriculum learning, scheduled training)

1.3. 다양한 QE 사용처

NAVER
DEVIEW
2023

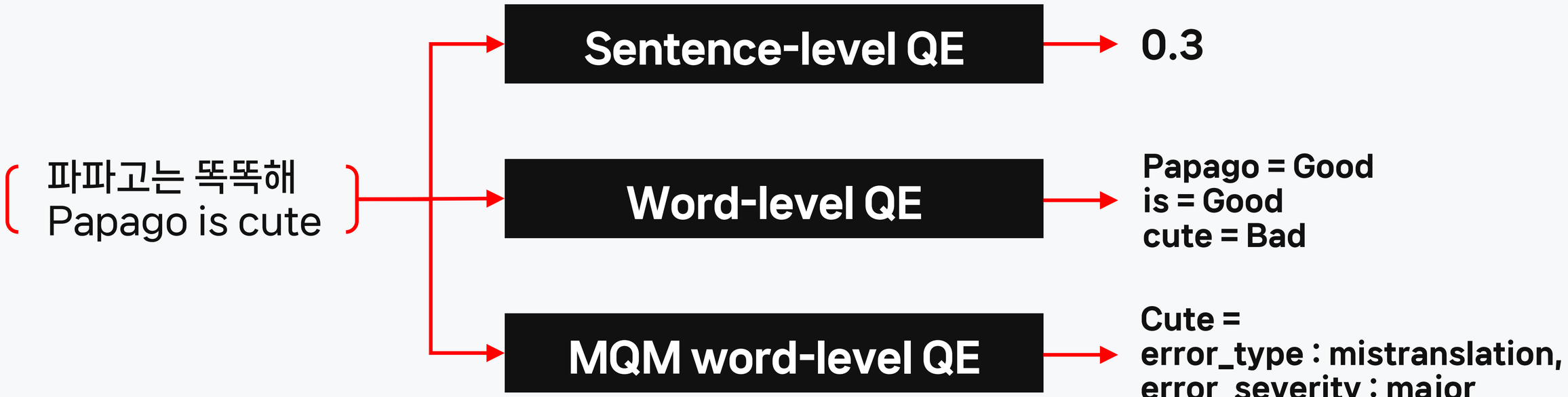
- 모델 자동 평가를 통해 빠른 의사결정 가능
- 전문가 평가와 QE 모델 목적에 맞게 병용

2. 파파고가 개발한 QE 모델

2.1. QE Task 소개

NAVER
DEVIEW
2023

- WMT 2020 ~ 2022 QE Shared Task 참여하며 자체 기술 고도화
- QE Task : 목적에 따라 QE 모델의 예측값 형태도 다양함



2.2 파파고 QE 모델 방향성

NAVER
DEVIEW
2023

사람이 만든 고품질 QE 데이터는 상당히 제한됨

- 다양한 도메인/언어쌍에서 고품질 데이터를 만드는 것의 어려움

파파고는 고품질 데이터가 제한된 상황에서의 QE 학습법에 집중

1. 인공 학습데이터 생성을 통해 데이터 양 / 언어쌍을 보강
2. 고품질 학습 데이터 부족한 환경에 특화된 학습 방법론
3. QE 특화 pretraining 방법론

2.3 파파고 QE 모델 기술

NAVER
DEVIEW
2023

1. 인공 학습데이터 생성을 통해 데이터 양 / 언어쌍을 보강

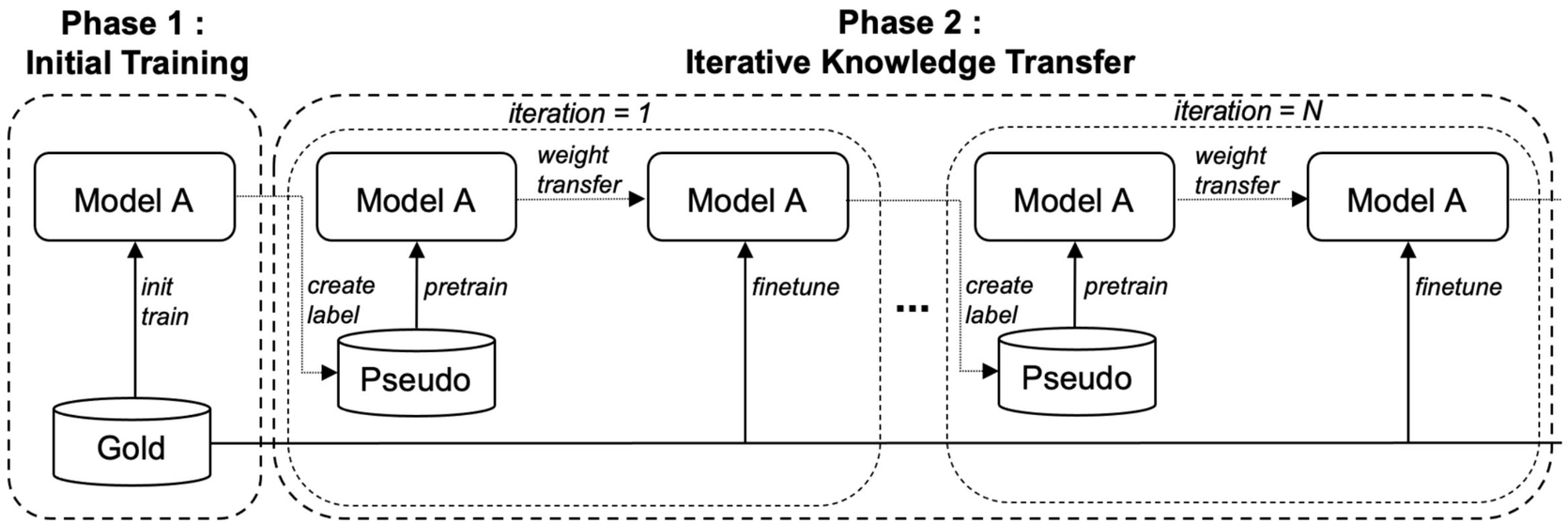
학습 데이터	En-De	En-Ru	Zh-En	En-Mr	Km-En	Ps-En	En-Ja	En-Cs	Multi
Gold	0.499	0.516	0.252	0.555	0.633	0.617	0.319	0.570	0.573
Gold + 인공데이터	0.576	0.584	0.287	0.682	0.639	0.627	0.385	0.594	0.611

- 공개된 QE 모델을 사용하여 (원문, 번역문) 문장쌍에 대한 인공 레이블을 생성
- 인공데이터를 학습에 추가 사용시, 모든 언어쌍에서 성능 향상을 관찰

2.3 파파고 QE 모델 기술

NAVER
DEVIEW
2023

2. 고품질 학습 데이터가 부족한 환경에 특화된 학습 방법론



학습 단계별 모델 성능

Phase 1 = 0.690 Phase2 (iteration=1)=0.719 Phase2 (iteration=2)=0.724

2.3 파파고 QE 모델 기술

NAVER
DEVIEW
2023

3. QE 특화 pretraining 방법론

- Downstream task 와 비슷한 Pretrain task 으로 언어 모델을 추가학습하여 성능을 높힘
  - 예시1: BLEURT pretraining for Metrics task
  - 예시2: MASS pretraining for MT task
- LM Pretraining -> (QE pretraining) -> QE task
- Monolingual/bilingual data 를 이용한 QE pretraining 방법론 연구 진행중

2.4 파파고 QE 모델 평가

NAVER
DEVIEW
2023

QE모델 품질 예측 성공 예시

원문	번역문	유형	QE 점수
덴마크 축구 협회는 1899년에 설립되었다.	The Korean Football Association was founded in 2002.	명사/숫자 오류	0.23
덴마크 축구 협회는 1899년에 설립되었다.	The Danish Football Football Association was founded in 1899 1899.	hallucination	0.43
그녀의 눈동자는 빛나는 별처럼 빛났다.	Her eyes shone like a shining star.	좋은 번역	0.73
그녀의 눈동자는 빛나는 별처럼 빛났다.	His eyes shines like a shining star.	성별/시제 오류	0.36
우리는 언제나 학생들을 환영합니다.	We always welcome students.	좋은 번역	1.00
우리는 언제나 학생들을 환영합니다.	We always don't welcome students.	반대 의미	0.35

2.4 파파고 QE 모델 평가

NAVER
DEVIEW
2023

QE모델 품질 예측 실패 예시

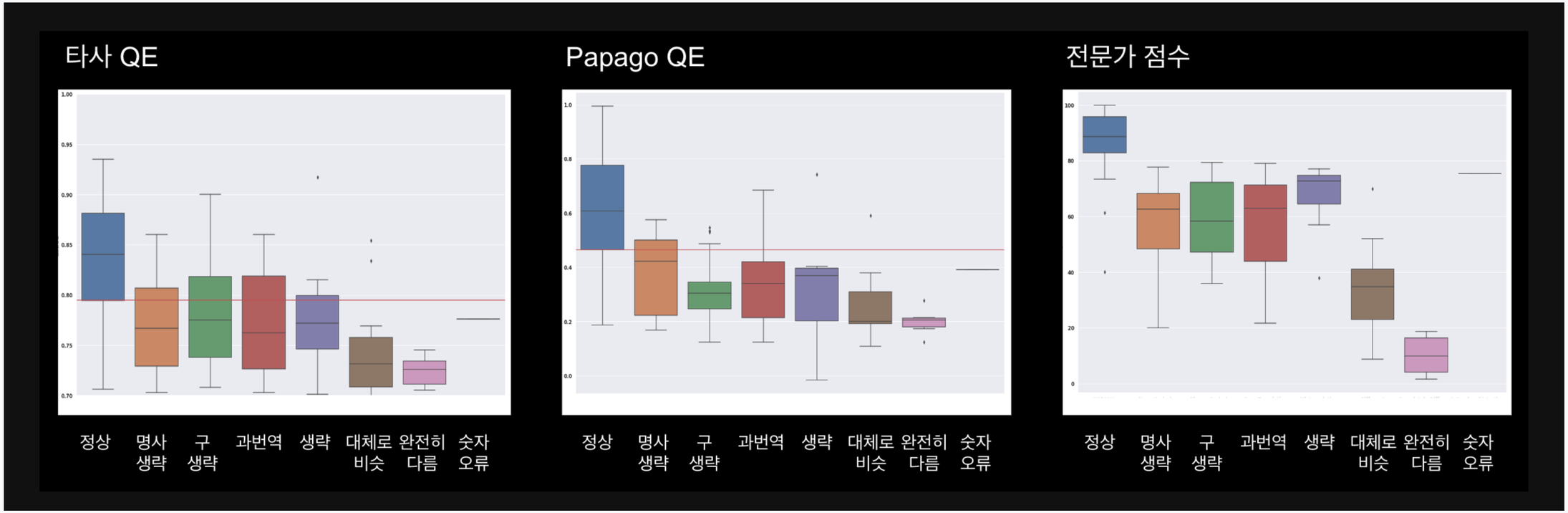
원문	번역문	유형	QE 점수
행간의 의미를 파악해야한다	You should read between the lines	좋은 번역	0.38
줄 사이를 읽어봐야한다	You should read between the lines	직역 오류	0.95
나 어제 밤 샀어	I bought chestnuts yesterday	좋은 번역	0.37
나 어제 밤 샀어	I bought last night	번역 오류	0.83
미국에서는 12개월~15개월 사이에 4회차 접종이 권장된다.	In the United States a fourth dose is recommended between 12 and 15 months of age.	좋은 번역	0.51
미국에서는 12개월~15개월 사이에 4회차 접종이 권장된다.	In the United States, four rounds of vaccination are recommended between 12 and 15 months.	번역 오류	0.93

2.4 파파고 QE 모델 평가

NAVER
DEVIEW
2023

QE 모델의 장/단점을 확인하며 지속적으로 모델 품질 개선 중

타사 대비 Papago QE 의 품질이 높음을 확인, 그러나 개선 여지가 남음

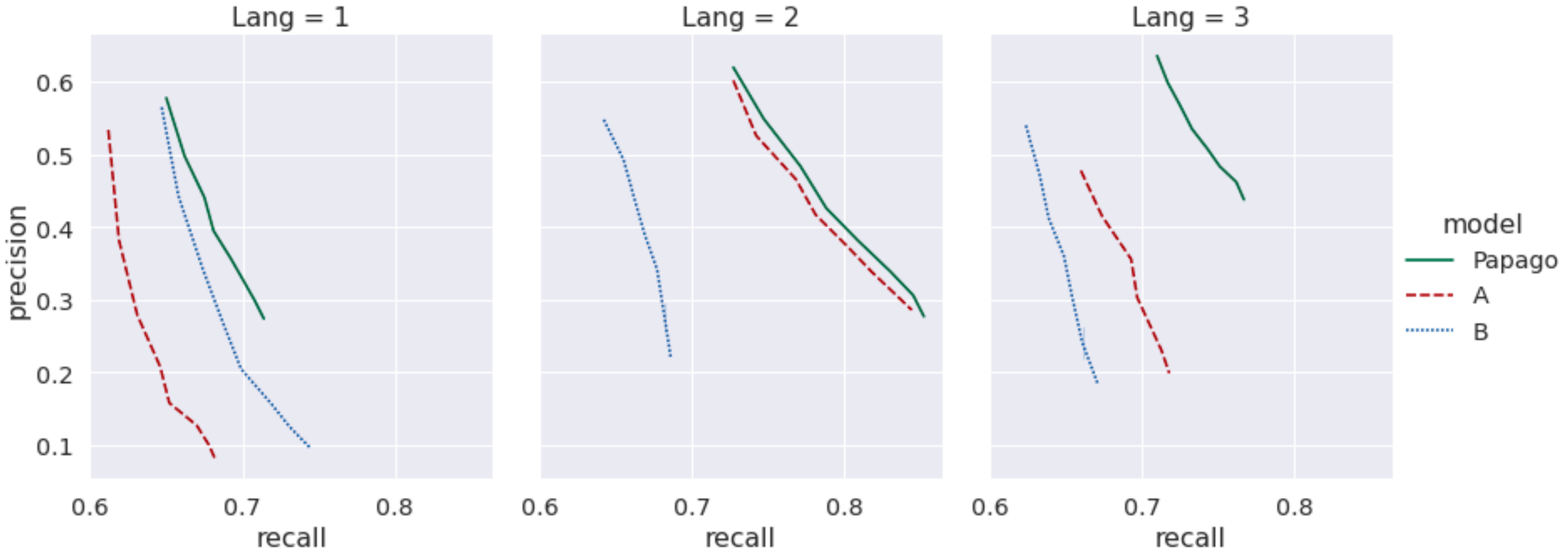


2.4 파파고 QE 모델 평가

NAVER
DEVIEW
2023

더 좋은 번역문을 구분하는 능력 (QE(src, hypo+) > QE(src, hypo-))

타사 대비 Papago QE 는 여러 언어쌍에서 구분 능력이 더 좋음

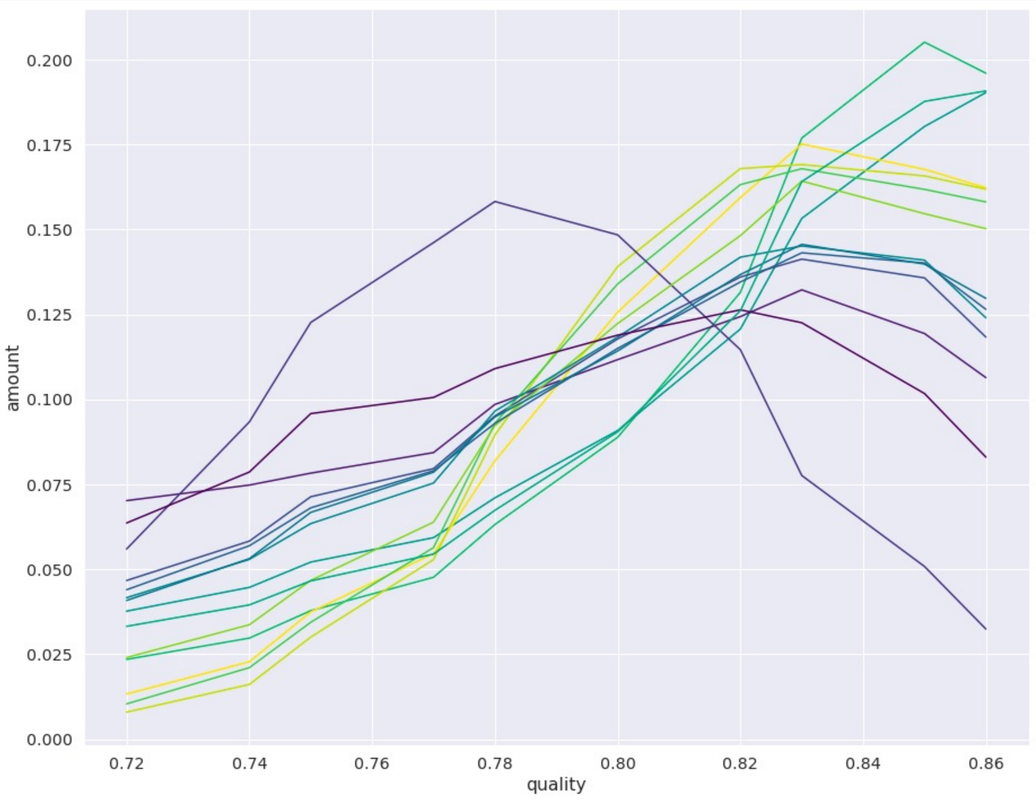


NAVER
DEVIEW
2023

3. QE 모델 활용 예시

3.1 학습데이터 품질 관리

NAVER
DEVIEW
2023



문장/코퍼스 단위로 중저품질 관리

- 보유한 데이터간 품질 비교 및 이해

- 의심되는 데이터 정제

- QE 점수를 활용한 효율적 학습

3.2 Parallel Mining

NAVER
DEVIEW
2023

Monolingual corpus 를 NMT 에 활용하는 방법

- 모델 기반: BART, MASS pretrain → NMT fine-tune

- 데이터 증강 기반: Back-translation
  - NMT(한>영) 학습 시 영어 데이터를 활용하기 위해 (한*, 영) 데이터를 생성 (한*=NMT(영>한))
  - (한*, 영) alignment 가 정확하나 한*이 NMT 모델 품질에 영향을 받음
  - 사람이 작성한 문장을 이용하여 병렬데이터를 생성하고 싶음

- 데이터 증강 기반: Parallel Mining
  - 대량의 한국어, 영어 말뭉치가 존재할 때 한-영 alignment 를 탐색
  - 사람이 생성한 (한, 영) 데이터를 확보할 수 있음

3.2 Parallel Mining

NAVER
DEVIEW
2023

PM 에서는 dual-encoder 를 이용하여 각 문장을 벡터화

- 문장 인코더는 cross / dual encoder 두 종류로 나뉨

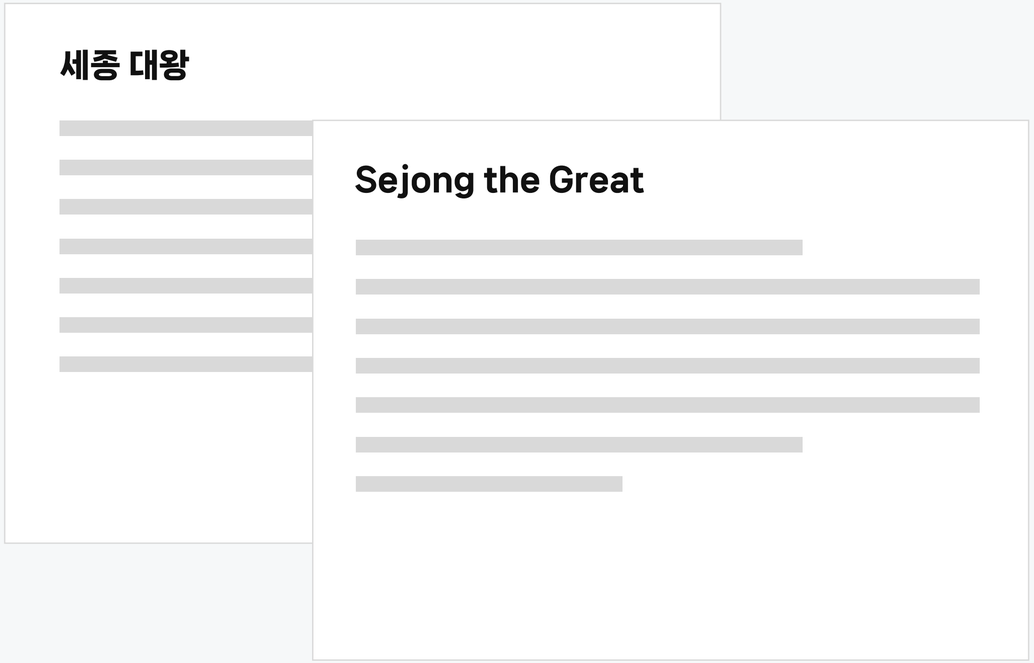
- Cross Encoder
  - M([s, t]) = E 처럼 두 문장을 concat 하여 모델에 입력.
  - Regressor(E) 를 이용하여 (s, t) 의 번역 품질/문장 유사도를 계산 (e.g.: PapagoQE, COMET)
  - Es, Et 를 만들 수 없기 때문에 s 와 유사한 t 를 탐색하는데 이용할 수 없음

- Dual Encoder
  - M(s) = Es, M(t) = Et 처럼 각 문장이 각각의 임베딩으로 표현
  - (s, t) 의 번역 품질/문장 유사도는 벡터 간 유사도(Cosine)를 이용하여 계산 (e.g.: LaBSE, LASER)

3.2 Parallel Mining

NAVER
DEVIEW
2023

PM은 대규모 텍스트에 존재하는 번역 문장쌍을 자동 추출



세종은 조선의 제4대 국왕이다.

Sejong the Great was the fourth king of the Joseon dynasty of Korea.

3.2 Parallel Mining

NAVER
DEVIEW
2023

[ko_1] ···
[ko_2] 세종은 조선의 제4대 국왕이다.
[ko_3] 세종대왕은 조선의 왕이다.
[ko_4] 세종의 업적에 대한 존경의 ···
[ko_5] ···

[vi_1] ...
[vi_2] Sejong là vị vua thứ tư của Joseon.
[vi_3] Joseon là quốc gia cổ đại của Hàn Quốc.
[vi_4] Vua Sejong là vua của Joseon.
[vi_4] ...

...
$sim(ko_2, vi_2) = 0.95$
$sim(ko_2, vi_3) = 0.31$
$sim(ko_3, vi_2) = 0.80$
...

[ko_2] 세종은 조선의 제4대 국왕이다.

[vi_2] Sejong là vị vua thứ tư của Joseon.
~~[vi_4] Vua Sejong là vua của Joseon.~~

유사 문서 탐색 → 문서 내 문장 분리 → 유사도 계산 → 정제 및 번역쌍 획득

다국어 문장 분리 기술

다국어 문장 인코더 기반
유사도 측정 기술

QE 모델 + 필터링 기반
데이터 정제 기술

3.2 Parallel Mining

NAVER
DEVIEW
2023

Dual Encoder 는 문장의 미세한 차이를 잘 표현하지 못함

- Cross encoder 는 모델이 두 문장 (s, t) 간의 차이를 직접 확인
- Dual encoder 를 이용하여 확보한 번역 문장쌍 후보에 cross encoder 기반 QE 모델 적용하여 PM 정확도를 높임

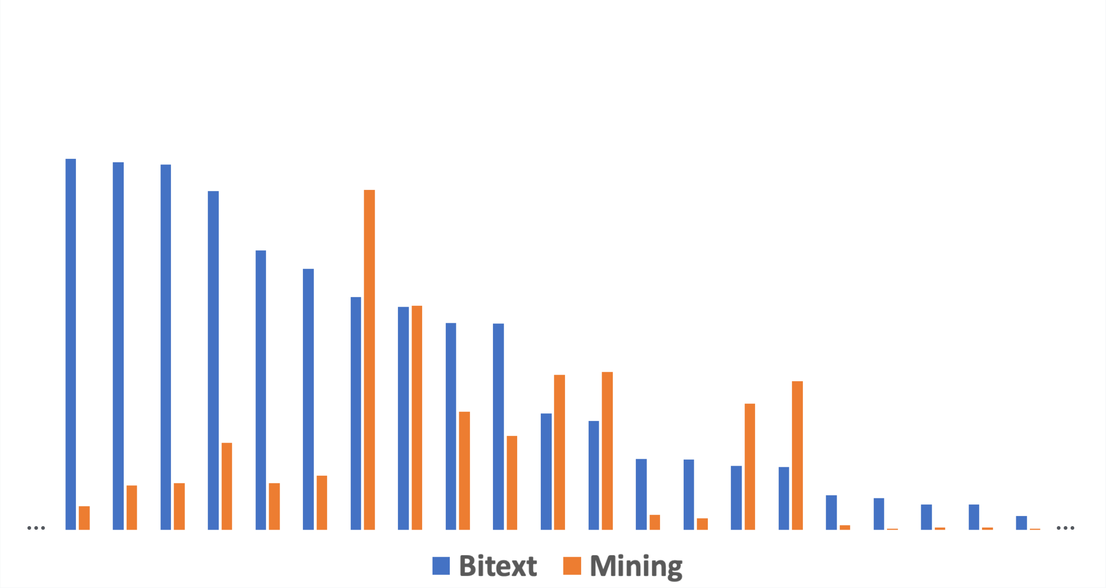
문장유사도	Ko	En
0.8816	1830년대 초반에 그는 여자 형제를 만나기 위해서 켄터키로부터 온 메리 오웬스를 만났다	In the early 1830s, he met Mary Owens from Kentucky.
0.8464	세종(世宗, 1397년 5월 7일 ~ 1450년 3월 30일)은 조선의 제4대 국왕(재위 : 1418년 9월 9일 ~ 1450년 3월 30일)이다	Sejong the Great (; 15 May 1397 – 8 April 1450) was the fourth king of the Joseon dynasty of Korea.
0.8313	9시간 동안 혼수상태에 빠진 링컨은 4월 15일 오전 7시 22분에 사망한다.	After remaining in a coma for eight hours, Lincoln died at 7:22 am on April 15.

3.2 Parallel Mining

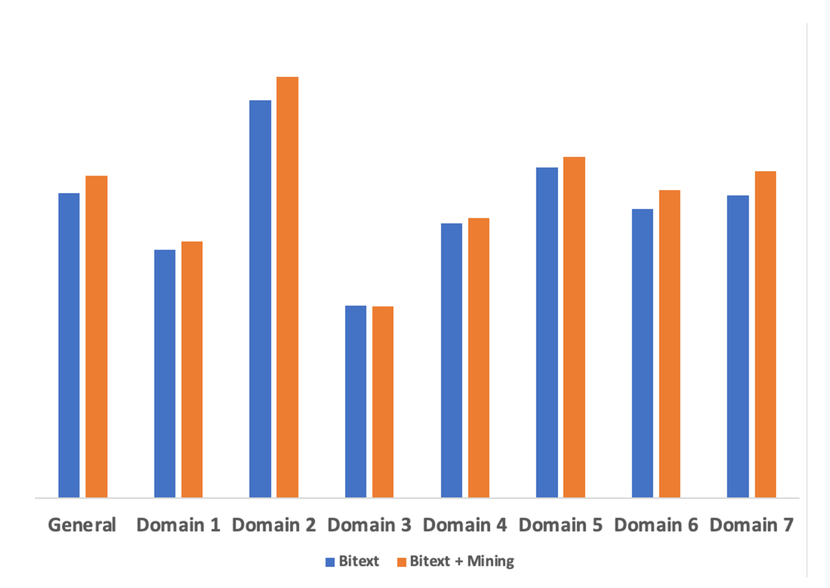
NAVER
DEVIEW
2023

마이닝을 통해 병렬문장쌍을 확보

- Low-resources 에서는 보유한 병렬데이터 보다도 많은 마이닝 결과를 획득하기도 함
- High-resources 에서도 보유하지 못한 도메인의 병렬데이터를 확보할 수 있음



PM 을 통하여 확보한 일부 언어별 문장쌍 양



High-resource 언어쌍의 NMT 품질 평가

3.3 Next Papago PM/QE

NAVER
DEVIEW
2023

문장생성 ML : 대량의 고품질 데이터 확보할수록 품질 향상

- LM 과 달리 NMT 에서는 병렬 문장쌍 형태의 데이터가 필요
- PM 은 적은 비용으로 병렬문장쌍을 확보함에 유용한 기술
- PM, QE 의 모델 품질 향상을 위한 방법을 지속적으로 탐색 중

PM 용 dual encoder 의 학습법 개선 노력중

- 일반적으로 PM 용 문장인코더는 병렬데이터를 이용하여 학습
- 병렬 데이터가 적은 언어쌍에서도 정확한 문장인코더를 학습하는 방법 개발 중

QE 용 cross encoder 의 학습법 개선 노력중

- 사람이 생성한 소규모의 (src, tgt, quality) 데이터 대신 다량의 병렬 문장쌍을 이용하는 QE 학습법 개발 중

Q & A

Thank You