CONCRETE: Improving Cross-lingual Fact-checking with Cross-lingual Retrieval

Kung-Hsiang Huang, ChengXiang Zhai, Heng Ji

COLING 2022

발제자: 김한성

23-06-02

Abstract

영어 Fact-checking 데이터 셋이 다수 이후 다른 언어는 데이터 희소성 존재.

Cross-lingual 기술은 상대적 희소성이 높은 언어에서 좋은 효과를 기대할 수 있는 기술

본 저자는 Cross-lingual Fact-checking을 개선하기 위한 Cross-lingual Retrieval 모델을 제안

ORQA의 ICT Pretraining을 착안한 X-ICT를 제안

이는 결국 X-Fact zero-shot task에서 2.23% 향상한 SOTA를 달성

Claim	*Muslimische Gebete sind Pflichtprogramm an katholischer Schule.* Muslim prayers are compulsory in Catholic schools.
Label	Mostly-False (*Grösstenteils Falsch*)
Claimant	Freie Welt
Language	German
Source	de.correctiv.org
Claim Date	March 16, 2018
Review Date	March 23, 2018

X-Fact 데이터 구조

Background X-fact task

Split	# claims	# languages
Train	19079	13
Development	2535	12
In-domain	3826	12
Out-of-domain	2368	4
Zero-shot	3381	12

X-FACT - Evaluating Generalization

- Three evaluation sets for measuring generalization of fact-checking systems
- In-Domain Test
  - Language and source both in training
- Out-of-Domain Test
  - Language in training, but source not in training
- Zero-Shot Test
  - Neither language nor source in training.

Experiments and Baselines

- Experiments performed with mBERT
- Models and Baselines:
  - **Claim-Only**: Determine rating only using the claim statement.
  - **Claim+Metadata**: Additional metadata such as the claimant along with the claim statement
  - **Evidence-based**:
    - Extract evidences using Google Search on the claim statement.
    - Aggregate evidence using Attention-based model.

Introduction

- Fact-checking의 필요성 언급
- Fact-Checking을 위해선 신뢰성 높은 Corpus 가 필요
- Fact Checker는 항상 low-resource language datasets을 만들어야 함.

문제 정의
leverage high-resource languages with **zero-shot cross-lingual transfer**

train on leverage language
and just test on low language

Introduction

Cross-lingual setup research

Claim to Claim : 주장의 언어와 다른 언어로 된 주장과 유사도 기준 매핑

하지만 이러한 접근 또한 fact-checked된 다른 언어의 주장과 라벨이 존재해야함

X-Fact에서도 google search Engine을 활용하여 evidence를 제공했음

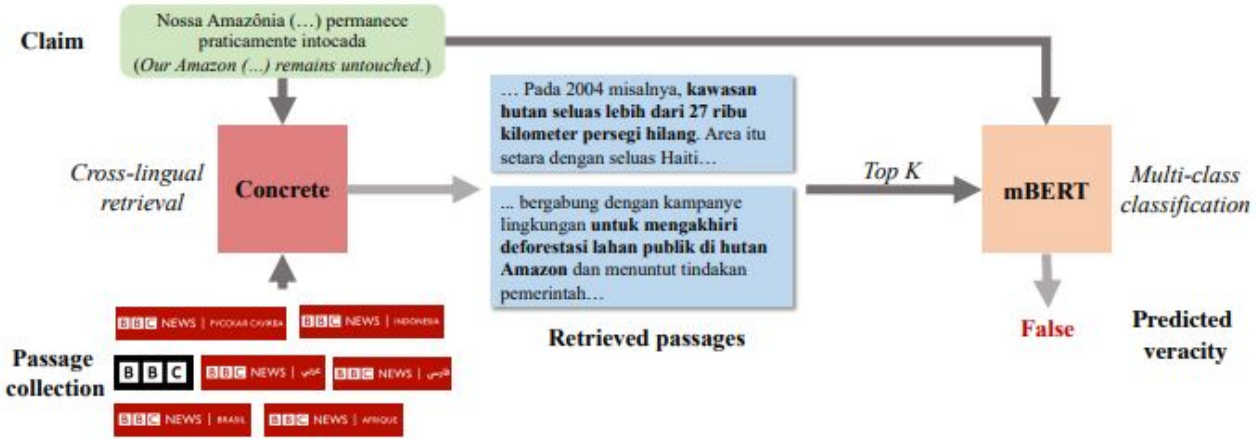
하지만 이 또한 수집된 문장의 신뢰성은 고려하지 않았음

Introduction

Cross-lingual setup research

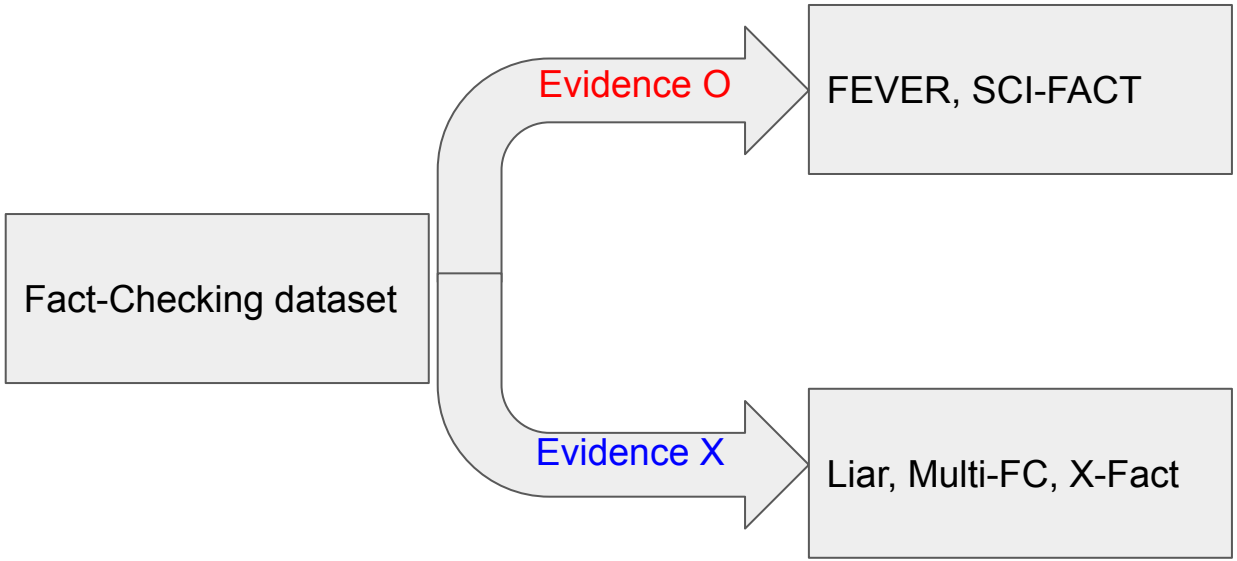
Claim-oriented Cross-Lingual Retriever를 위한 세팅의 필요성을 느꼈고

이러한 프레임 워크인 CONCRETE(**C**laim-**o**rie**n**ted **C**oss-lingual **Ret**rie**v**er)를 제안.



Related Work

Fact-checking dataset



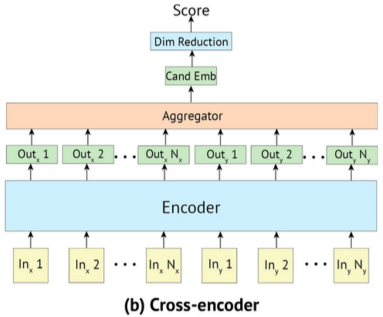
Related Work

Cross-lingual Retrieval

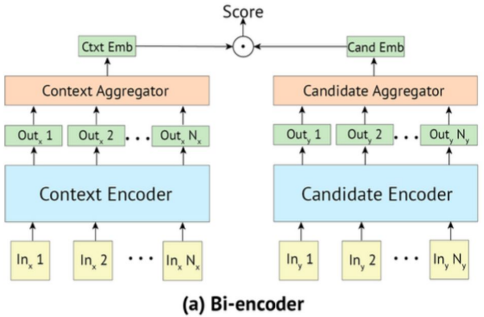
Large Multi-lingual model이 생기면서 **cross-encoder형태의 아키텍처**가 제안

이후 **bi-encoder 구조 시간 복잡도 또한 줄이고 효율도 챙긴 mDPR 제안**

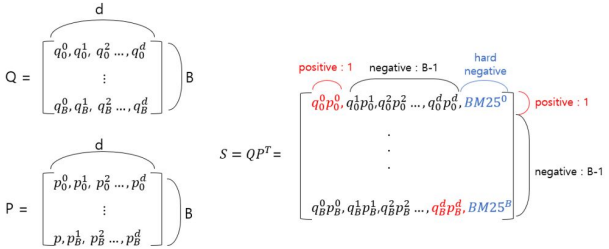
하지만 **domain discrepancy**로 claim evidence pair로 재학습한 CONCRETE등장



LLM Fine tune



DPR



DPR in-batch negative

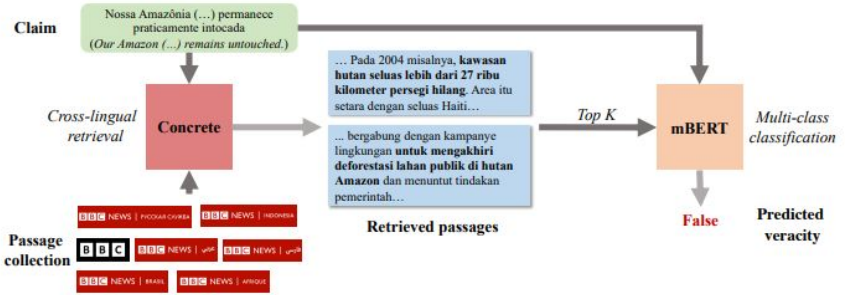
Task Definitions

Cross-lingual setup research

1. 크롤링 : 49,000개의 기사 추출 (7개 언어로 구성,BBC, 16년에서 22년)
   passage를 100토큰 가량으로 분리
   347,557개의 passage 제작

2. X-Fact 문제를 풀기 위해 2-stage 프레임 워크 고안
   - **Retrieval : Claim에 적절한 근거 Passage 찾기**
   - **Reader : Claim과 Passage를 기준으로 veracity를 확인**

Proposed Method

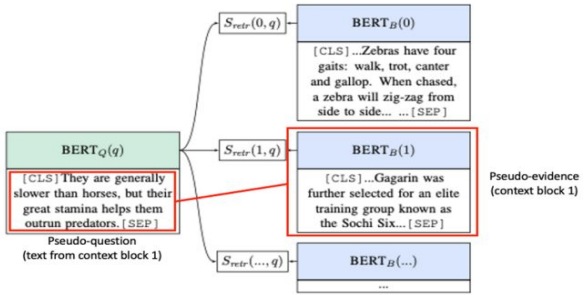


CONCRETE(Retrival)

- mDPR에서 학습한 multilingual IR datasets에서는 **'주장'**이 아닌 **query**가 주어진다.
  - \+ 데이터 셋은 현재는 접근이 불가능하다.

Modification

1. OrQA방법을 착안한 pseudo claim 생성



4. OrQA ICT training
   a. 문장 임베딩의 성능을 높이기 위한 방법론
   b. sudo-question을 만들어 Retrieval학습

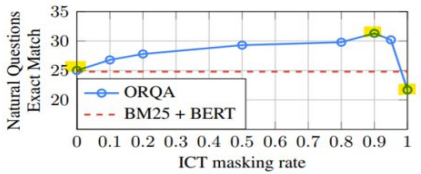


Figure 3: **Analysis**: Performance on our open version of the Natural Questions dev set with various masking rates for the ICT pre-training. Too much masking prevents the model from learning to exploit exact n-gram overlap. Too little masking makes language understanding unnecessary.

Proposed Method

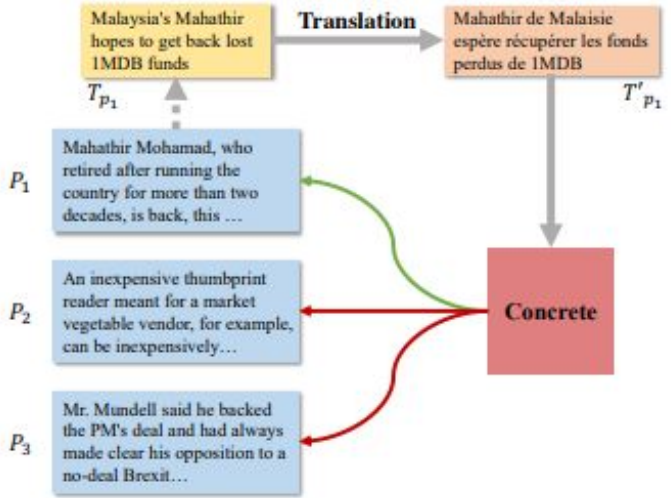
X-ICT : Cross-lingual Inverse Cloze Task

1. query랜덤 마스킹 문제점
   a. 랜덤 쿼리가 Claim 성향인지 알 수가 없다
   b. domain mismatch 확률이 높다. (문서 내 주제와 어긋나는 문장)
2. Claim의 성향에 제일 가까운 Document의 **Title**을 pseudo claim이라 정의
   -> domain mismatch가 적다.
3. mBART to translate title (동일 확률을 위해 언어별 1/7)

$$L_{X\_ICT} = -log \sum_{p_i \in P} \frac{exp(sim(T^{'}_{p_i}, p_i))}{\sum_{p_j \in BATCH} exp(sim(T^{'}_{p_i}, p_j))}$$

$$c : claim$$
$$p_i : i_{th}\ document$$
$$E_c(*) : Claim\ Encoder$$
$$E_P(*) : Passage\ Encoder$$

$$sim(c, p_i) = E_c(c)^T E_p(p_i)$$

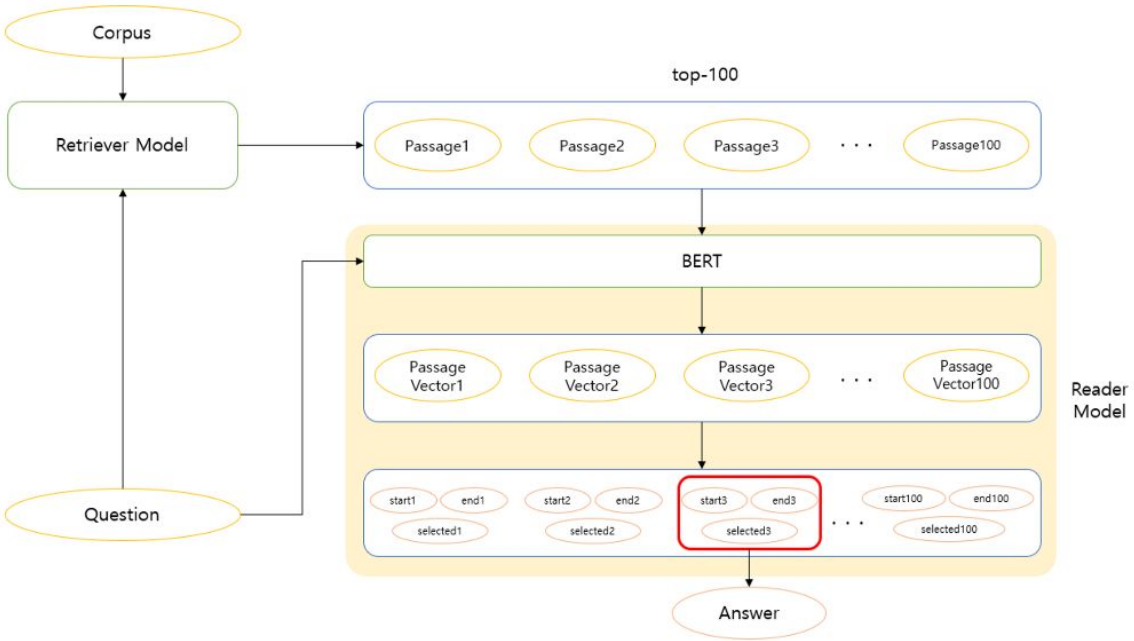


Proposed Method

Multilingual Reader

$$h_T = mBERT(T)[CLS]$$
$$h_{p_i} = mBERT(p_i)[CLS]$$

$$L = \frac{1}{N}\sum_{i=1}^{N} y_i \, log\hat{y}_i$$



PLM : bert-base-multilingual-cased

Baseline

- MT + DPR : translation input으로 모두 영어 변환 후 DPR
- BM25 : No train just test
- mDPR : 사전 학습된 Multilingual DPR
- Google Search

Implementation Details

X-ICT	AdamW, lr : 2e-5, 30 epoch. max_length : 256
READER	mBERT fine tune lr :5e-5, classifier lr : 1e-3 max_length :512

Results

train lan != test lan

train lan == test lan

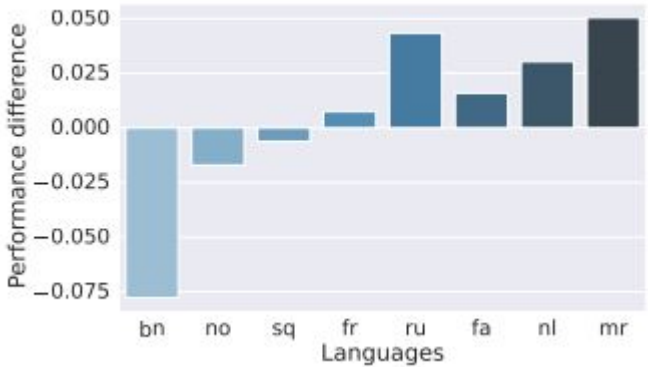
	Reader	Retrieval Method	Zero-shot F1 (%)	In-domain F1 (%)
Prior (Gupta and Srikumar, 2021)	Majority	None	7.6	6.9
	mBERT	None	16.7	39.4
	mBERT	Google Search	16.0	41.9
Ours	mBERT	None	17.25	36.91
	mBERT	Google Search	16.02	**42.61**
	mBERT	MT+DPR	15.01	35.29
	mBERT	BM25	17.43	38.29
	mBERT	mDPR	17.60	36.79
	mBERT	CONCRETE	**19.83***	40.53

for using Google Search suggests that the reader may exploit biases or patterns presented in Google Search's results that are not transferrable across languages. To validate this hypothesis, we an-

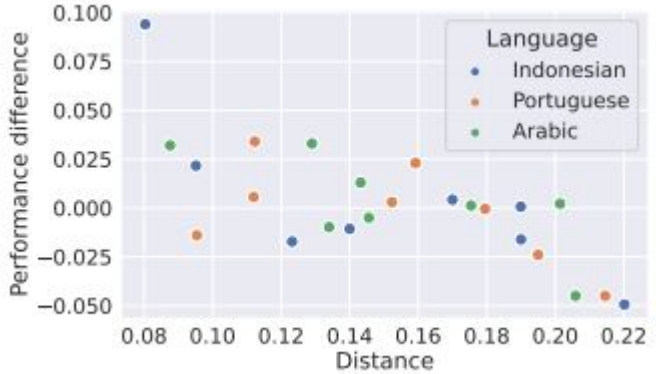
where Google Search results contain the string "SALAH" (WRONG), 50% of them are PARTLY TRUE and 45% of them are FALSE. Such patterns

Results

언어 거리간 음의 상관관계가 존재



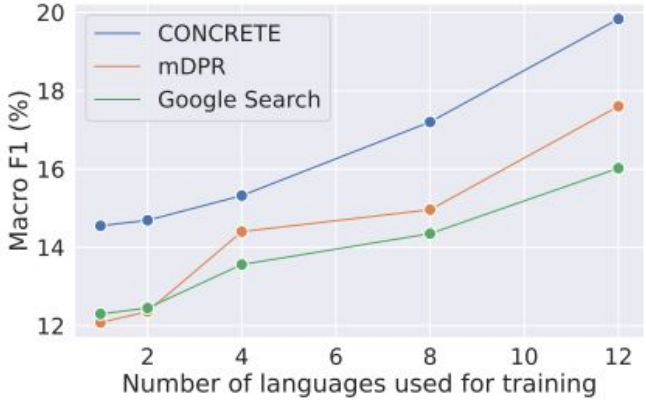
인도네시아어 문단이 문단 컬렉션에서 제거된 경우 각 언어의 결과
(언어유사도 기준으로 정렬)



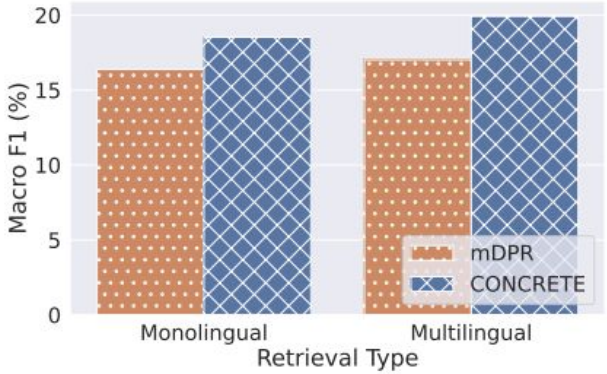
번역된 인도네시아어, 포르투갈어, 아랍어로 검색된 문단의 매크로 F1 점수의 성능 차이

Results

학습데이터 언어 수의 관계



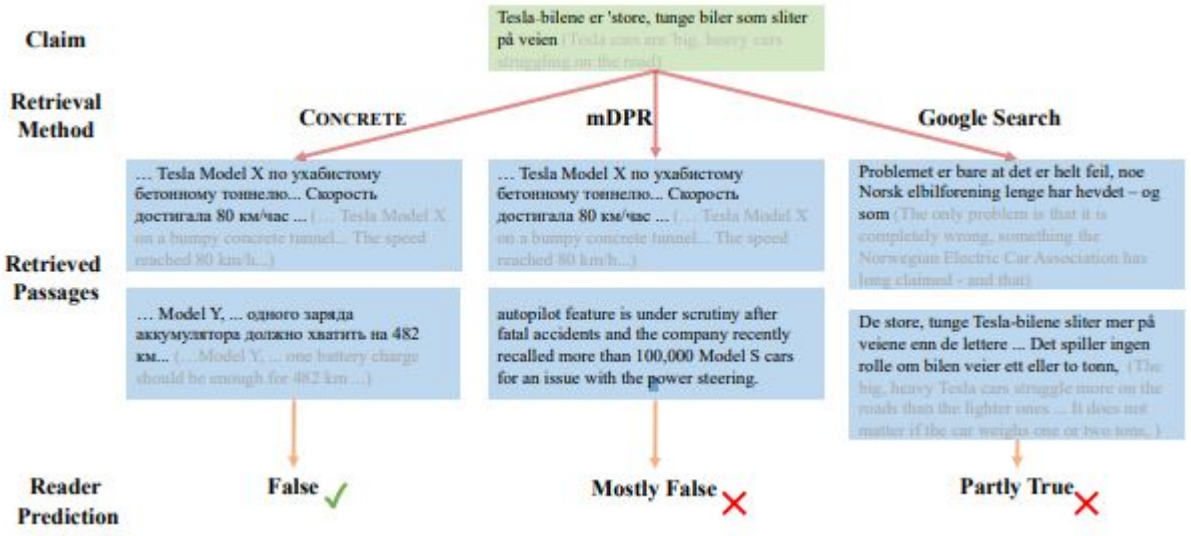
인도네시아어 문단이 문단 컬렉션에서 제거된 경우 각 언어의 결과
(언어유사도 기준으로 정렬)



번역된 인도네시아어, 포르투갈어, 아랍어로 검색된 문단의 매크로 F1 점수의 성능 차이

Impact of X-ICT

Domain mis match



Remaining Challenges

- Evidence cannot be retrieved : 실제 근거가 아예 없는 경우
- Under-specified context : 주어진 근거로 사실을 판단하기 불충분할때
- Require intent identification : claim의 의도가 불분명한 경우
- Reader failure
- Annotation error

Figure 7: Distribution of the remaining errors.

Conclusion and Future works

- CONCRETE는 새로운 Cross-lingual Retrieval을 제안 및 domain specific하게 문제를 적절히 정의 및 모델 제안
- Fact-checking 분야의 Language generalization을 확보한 모델이다.
- IR 분야에 Claim과 같은 쿼리가 없는 것을 극복해줄 좋은 방법이라 주장
- 또한 X-Fact zero-shot task에서는 SOTA를 달성

감사합니다.