XOR QA: Cross-lingual Open-Retrieval Question Answering across many languages

Akari Asai[1], Jungo Kasai[1], Jonathan H. Clark[2], Kenton Lee[2], Eunsol Choi[3], Hannaneh Hajishirzi[14]

[1]University of Washington, [2]Google Research, [3]The University of Texas at Austin, [4]Allen Institute for AI

Links ▶ Questions? @AkariAsai

Website & leaderboard

Repository

Summary

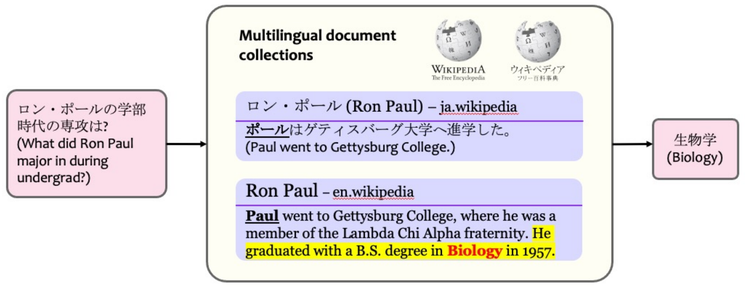
- A large-scale multilingual open-retrieval QA, **XOR-TyDi QA**
  - **40 k** newly annotated data across **7 languages**
  - Fine-grained **paragraph & span level annotations** & **human translated questions**
- A new task, **XOR QA** involving **cross-lingual IR and QA**.
- Experimental results show large spaces for improvements.

Background

Multilingual Open-retrieval QA is challenging:

- **Information scarcity**: limited web documents available in low resource languages (e.g., Telugu).
- **Information asymmetry**: bias towards their own cult ure or general interests

→ To build a QA system that works well across ma ny languages, **need to retrieve and read across m ultiple languages.**



Comparison with other datasets

Dataset	Asked by native speakers	Open-retrieval over millions of documents	Typological Diversity	Cross-lingual	train data
TyDi QA [1]	✓		✓		✓
MLQA [2]			✓	✓	
XQuAD [3]			✓		
MKQA [4]			✓		
MLQA-R [5]				✓	
QA@CLEF 2002-2008 [6]	✓				
XOR-TyDi QA	✓	✓	✓	✓	✓

Dataset Collections of XOR-TyDi

- For scalability and quality, **decompose tasks into subtasks**
- Introduce a new annotation framework, **collaborative model-in -the-loop QA**

Question Selection → **Question Translation** (Professional Translators)

Collect **"unanswerable" questions** in TyDi QA as we may find answers in an other language.

To find missing answers in English

Answer Retrieval (crowd-workers @AMT)

Collaborative model-in-the-loop QA

- A **model** helps **human** to quickly find relevant paragraphs
- **Humans** find paragraphs the model may overlook

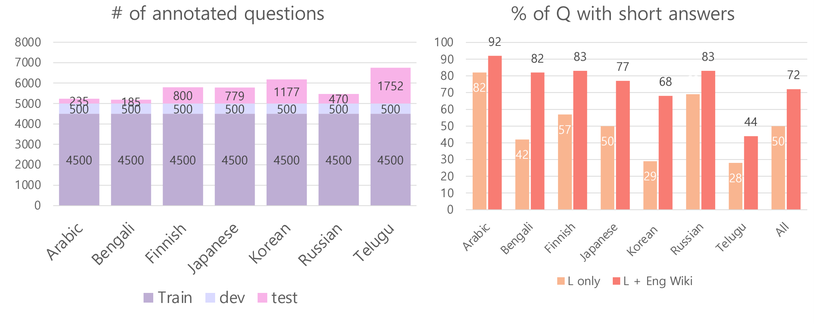
Answer Translation (Professional Translators)

Validate and translate answers back to **the target languages**

→ Cross-lingual data (Q_l, P_{eng}, A_l)

Dataset Statistics of XOR-TyDi

- **Train & evaluation data for 7 languages**
- Increase the answer coverage up to **40%**



XOR QA: 3 increasingly complex and useful tasks

XOR-Retrieve
- Input: Query in l and English Wikipedia
- Output: top k documents
- Metric: Top-k tokens recall ($k = 2000, 5000$)

XOR-EngSpan
- Input: Query in l and English Wikipedia
- Output: an answer in English, A_{eng}
- Metric: Token F1, EM

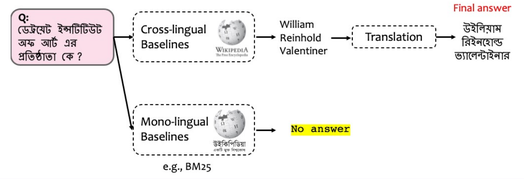
XOR-Full
- Input: Query in l, English Wikipedia and the Wikipedia in l
- Output: an answer in l
- Metric: Token F1, EM, BLEU

Cross-lingual baselines (for: XOR-Retrieve, XOR-English Span)

- **Translate-test:** translate query into English, run English DPR
- **Multilingual:** use multilingual BERT & XLM-R for encoders

Full baselines (for: XOR-full)

- **Full**: Cross-lingual Baselines + Monolingual baseline
- **English only**
- **Monolingual-only**



Results

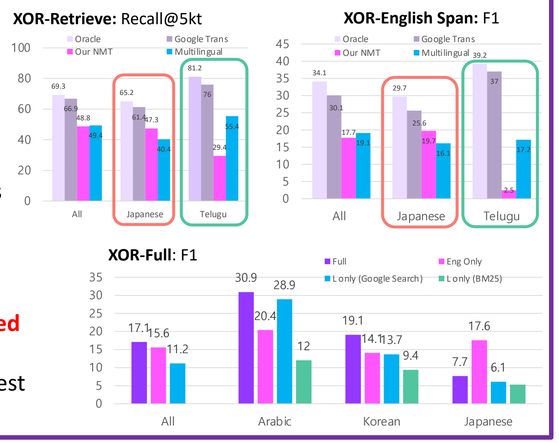
XOR-Retrieve

XOR-EngSpan

- **High resources L:** Tran slate-test (w/ human, GMT) > Multilingual
- **Low resources L:** Trans late-test suffers

XOR-Full

- **All models show limited performance**
- Full model performs best



References

[1]Clark et al., 2020. "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typoligucally Diverse Languages". *TACL*
2] Lewis et al., 2020. "MLQA: Evaluating Cross-lingual Extractive Question Answering". In *ACL*.
[3] Artetxe et al., 2020. "On the cross-lingual transferability of monolingual representations". In *ACL*.
[4] Longpre et al., 2020. "MKQA: A linguistically diverse benchmark for multilingual open domain question answering". *arXiv preprint.*
[5] Roy et al., 2020 "LAReQA: Language-agnostic answer retrieval from a multilingual pool". In *EMNLP*.
[6] Former et al., 2019. "Evaluating Multilingual Question Answering System at CLEF". In *LREC*.