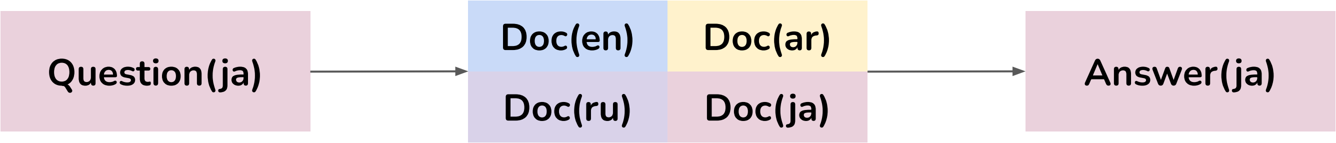
ZusammenQA:

Data Augmentation with Specialized Models for Cross-lingual Open-retrieval Question Answering System

Chia-Chien Hung[1], Tommaso Green[1], Robert Litschko[1], Tornike Tsereteli[1], Sotaro Takeshita[1], Marco Bombieri[2], Goran Glavaš[3], Simone Paolo Ponzetto[1]

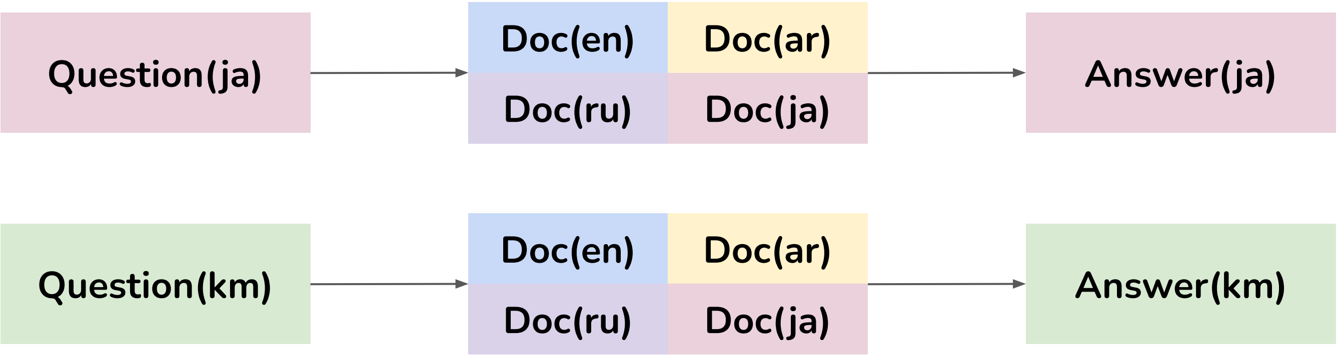
[1]Data and Web Science Group, University of Mannheim, Germany
[2]ALTAIR Robotics Lab, University of Verona, Italy
[3]CAIDAS, University of Würzburg, Germany

Cross-lingual Open-retrieval Question Answering (COQA)



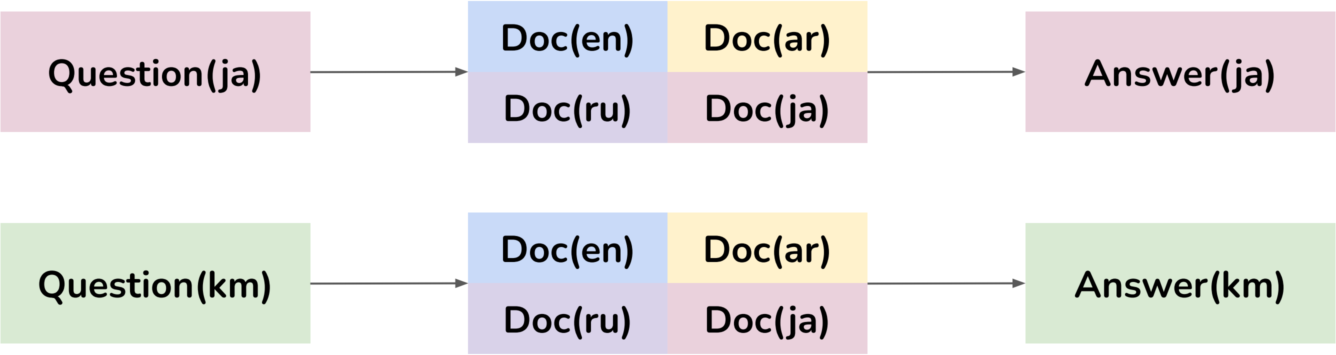
{
'id': '7931051574381133444',
'question': '『指輪物語』はいつ出版された ',
'answers': ['1954年から 1955年'],
'lang': 'ja',
'split': 'train',
'source': 'xor-tydi'
}

Cross-lingual Open-retrieval Question Answering (COQA)




```
{
    'id': '7931051574381133444',
    'question': '『指輪物語』はいつ出版された   ',
     'answers': ['1954年から 1955年'],
    'lang': 'ja',
    'split': 'train',
    'source': 'xor-tydi'
}
```

Cross-lingual Open-retrieval Question Answering (COQA)



{
'id': '7931051574381133444',
'question': '『指輪物語』はいつ出版された ',
'answers': ['1954年から1955年'],
'lang': 'ja',
'split': 'train',
'source': 'xor-tydi'
}

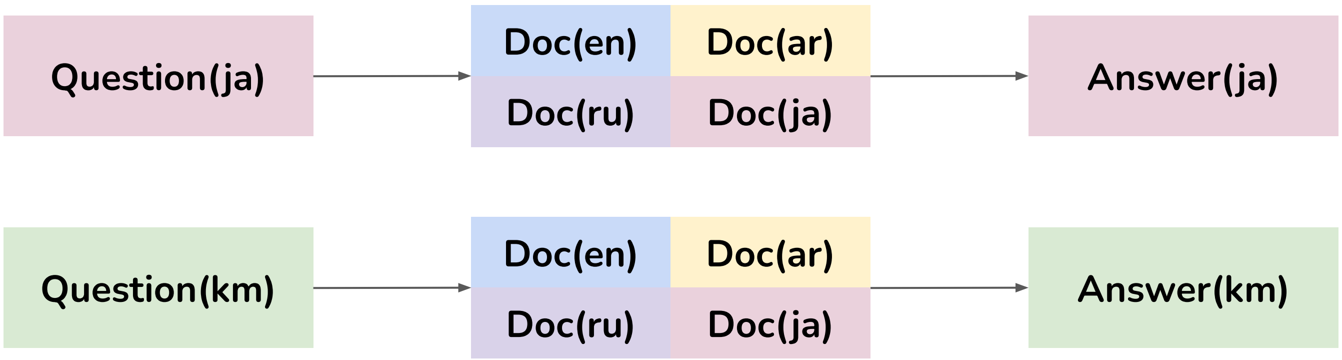
MIA-Shared Task: Constrained Track

- Train: Natural Questions *[Kwiatkowski et al., 2019]*, XOR-TyDi QA *[Asai et al., 2021]*
- Dev / Test: XOR-TyDi QA, MKQA *[Longpre et al., 2020]*, Surprise Languages
- Preprocessed Wikipedia Passages
- Unlabeled data

with training data	without training data
Arabic (ar), Bengali (bn), English (en), Finnish (fi), Japanese (ja), Korean (ko), Russian (ru), Telugu (te)	Spanish (es), Khmer (km), Malay (ms), Swedish (sv), Turkish (tr), Chinese (zh-cn), Tamil (ta)*, Tagalog (tl)*

Table 1: languages for the shared task datasets **surprise languages*

Cross-lingual Open-retrieval Question Answering (COQA)



{
'id': '7931051574381133444',
'question': '『指輪物語』はいつ出版された ',
'answers': ['1954年から1955年'],
'lang': 'ja',
'split': 'train',
'source': 'xor-tydi'
}

MIA-Shared Task: Constrained Track

- Train: Natural Questions *[Kwiatkowski et al., 2019]*, XOR-TyDi QA *[Asai et al., 2021]*
- Dev / Test: XOR-TyDi QA, MKQA *[Longpre et al., 2020]*, Surprise Languages
- Preprocessed Wikipedia Passages
- Unlabeled data

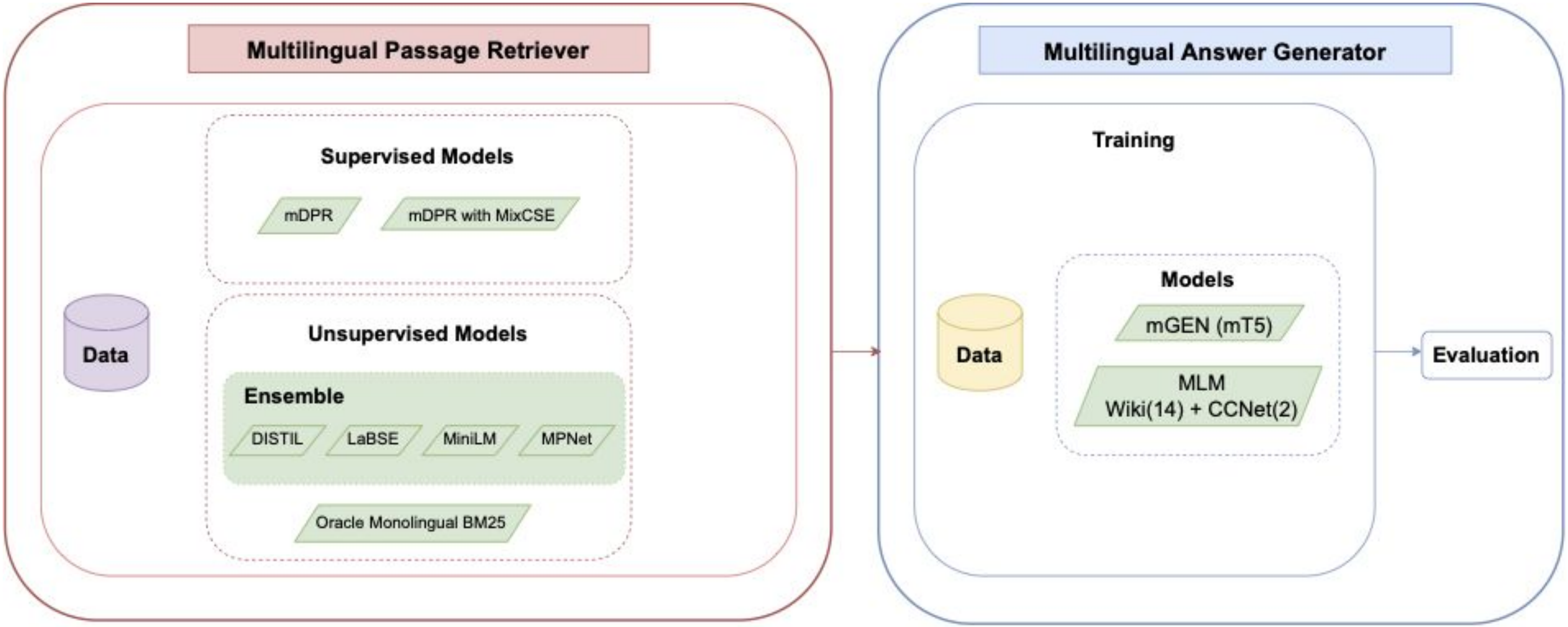
with training data	without training data
Arabic (ar), Bengali (bn), English (en), Finnish (fi), Japanese (ja), Korean (ko), Russian (ru), Telugu (te)	Spanish (es), Khmer (km), Malay (ms), Swedish (sv), Turkish (tr), Chinese (zh-cn), Tamil (ta)*, Tagalog (tl)*

Table 1: languages for the shared task datasets **surprise languages*

Will data augmentation help to improve the unseen languages performance?
Will language- and domain-specialization help to improve the COQA performance?

ZusammenQA: Data Augmentation with Specialized Models for COQA System

Data Augmentation → Passage Retrieval → Answer Generation



ZusammenQA: Data Augmentation with Specialized Models for COQA System

Data Augmentation → Passage Retrieval → Answer Generation

QA generation

The Lego Group began manufacturing the interlocking toy bricks in 1949. Movies, games, competitions, and six Legoland amusement [...]

▷ **QA-generator (T5)** ▷

Question: Who funded LEGO?
Answer: Ole Krik Christiansen

Further, we translated QA pairs and run heuristics rule-based filtering.

Data Augmentation ⇨ **Passage Retrieval** ⇨ Answer Generation

- **Supervised Models (mDPR variants)**
  - mDPR *[Asai et al., 2021]* with MixCSE loss *[Zhang et al., 2022]*
    - Contrastive learning can help to alleviate the *anisotropy problem*
    - As training goes by, the influence of the negatives fades
    - Mixed negatives can help in keeping a strong gradient

$$\mathcal{L}_{\text{mdpr}} = -\log \frac{\langle \mathbf{e}_{q_i}, \mathbf{e}_{p_i^+} \rangle}{\langle \mathbf{e}_{q_i}, \mathbf{e}_{p_i^+} \rangle + \sum_{j=1}^{n} \langle \mathbf{e}_{q_i}, \mathbf{e}_{p_{i,j}^-} \rangle}$$

$$\tilde{\mathbf{e}}_i = \frac{\lambda\, \mathbf{e}_{p_i^+} + (1-\lambda)\, \mathbf{e}_{p_{i,j}^-}}{\| \lambda\, \mathbf{e}_{p_i^+} + (1-\lambda)\, \mathbf{e}_{p_{i,j}^-} \|_2}$$

  - mDPR trained with augmented data

ZusammenQA: Data Augmentation with Specialized Models for COQA System

Data Augmentation ⇨ Passage Retrieval ⇨ Answer Generation

- **Unsupervised Models**
  - Dense Retrieval
    - Rank ensembling with sentence encoders
      - DISTIL *[Reimers and Gurevych, 2020]*
      - LaBSE *[Feng et al. 2022]*
      - MiniLM *[Wang et al., 2020]*
      - MPNet *[Song et al., 2020]*
  - Term-based Retrieval
    - Monolingual oracle BM25
      - We create **monolingual** indexes using BM25 representations
      - Language identification of the question
      - Querying of the indexes using the answer *(oracle)*

ZusammenQA: Data Augmentation with Specialized Models for COQA System

Data Augmentation → Passage Retrieval → Answer Generation

Sequence to Sequence

Question

レゴグループを設立したのは誰ですか？[1]

Retrieved Passages

Generator (mT5)

Answer

オレ・カーク・クリスチャンセン[2]

[1] Who funded the LEGO group? [2] Ole Kirk Christiansen

ZusammenQA: Data Augmentation with Specialized Models for COQA System

Data Augmentation → Passage Retrieval → Answer Generation

Model Specialization

Input Document

レゴグループは1949年に連動おもちゃのレンガの製造を開始しました。<X>、ゲーム、 競技 会 、および6つのレゴランド遊園地がこのブランドで <Y> されました。2015 年 7 月 現在、[...]

→ **Generator (mT5)** →

Masked Tokens Prediction

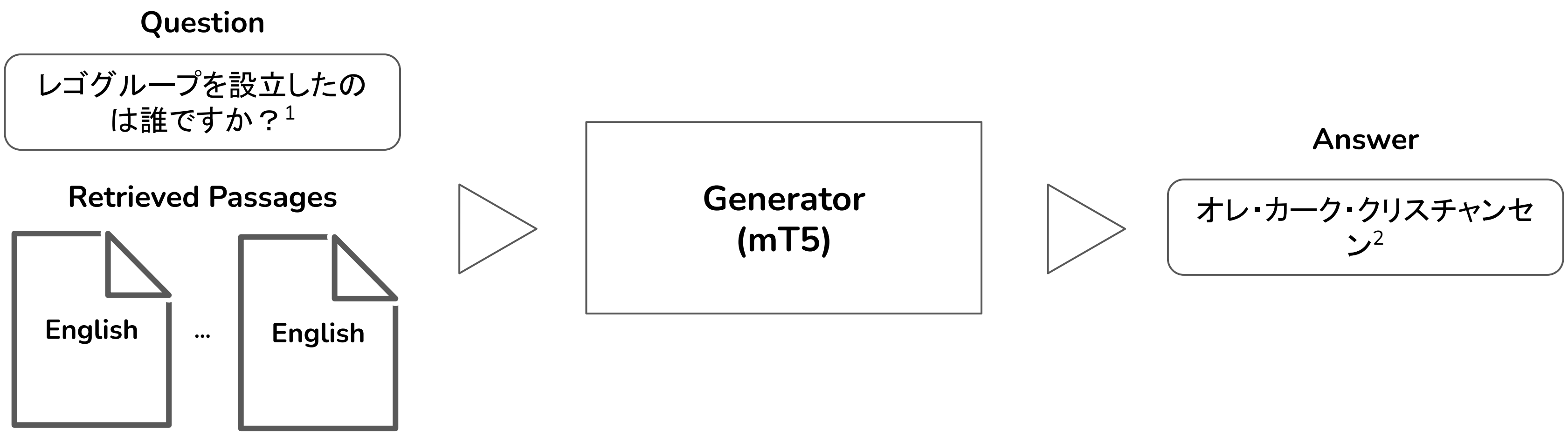
<X> 映画 <Y> 開発

Additional masked language modeling on documents for domain / language specialization.

ZusammenQA: Data Augmentation with Specialized Models for COQA System

Data Augmentation → Passage Retrieval → Answer Generation

Sequence to Sequence with Augmented Data - AUG-QA



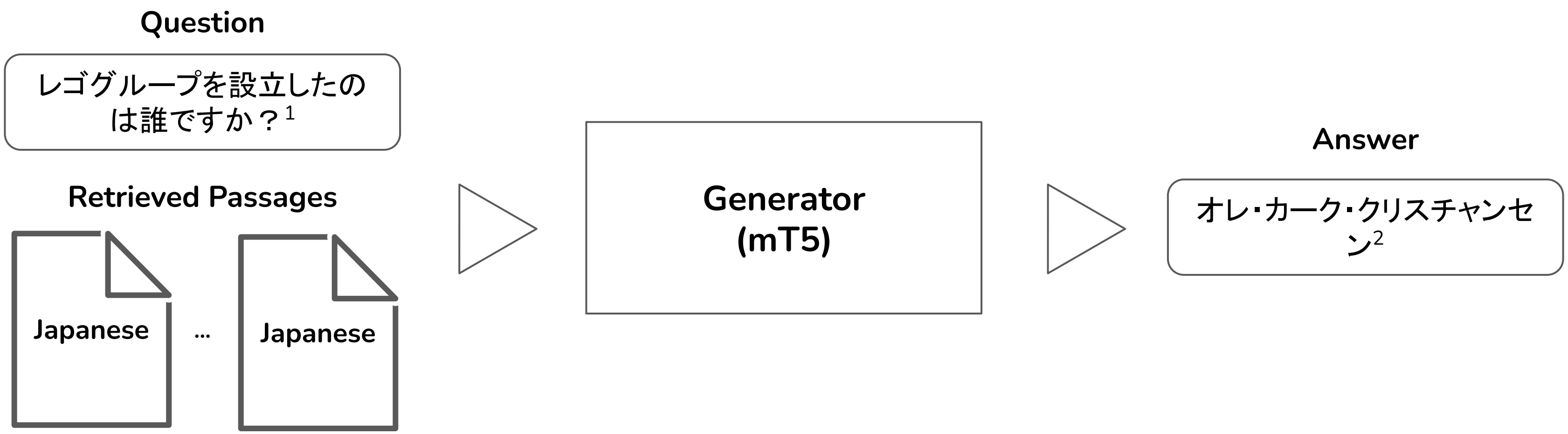
Translate only augmented question and answer

[1] Who funded the LEGO group? [2] Ole Kirk Christiansen

ZusammenQA: Data Augmentation with Specialized Models for COQA System

Data Augmentation → Passage Retrieval → Answer Generation

Two types of Augmented Data - AUG-QAP



Translate only augmented <u>question</u>, <u>answer</u> and <u>passages</u>

[1] Who funded the LEGO group? [2] Ole Kirk Christiansen

Evaluation: Overall Performance

Models	XOR-TyDi QA ar	bn	fi	ja	ko	ru	te	Avg.
mDPR + mGEN (baseline 1)	49.66	33.99	**39.54**	39.72	**25.59**	**40.98**	36.16	37.949
Unsupervised Retrieval								
OracleBM25 + MLM-14	0.34	0.49	0.52	2.56	0.19	0.57	5.16	1.404
EnsembleRank + MLM-14	0.34	0.49	1.33	2.56	0.38	6.27	16.21	3.161
Supervised Retrieval								
mDPR(AUG) *with* MixCSE + MLM-14	20.94	7.18	15.27	23.16	10.25	19.23	10.53	15.223
mDPR(AUG) + MLM-14	24.99	15.19	20.33	22.31	10.68	18.82	11.97	17.754
mDPR + MLM-14	**51.66**	31.96	38.68	40.89	25.35	39.87	**37.26**	**37.951**
mDPR + MLM-14(XORQA & AUG-QA)	49.41	32.90	37.95	**40.97**	24.22	39.29	35.76	37.213
mDPR + MLM-14(XORQA & AUG-QAP)	48.79	33.73	38.33	39.87	25.26	39.11	37.94	37.577
mDPR + MLM-16	49.92	31.16	37.20	39.92	24.63	38.78	34.30	36.558
mDPR + MLM-16(XORQA & AUG-QA)	49.45	31.59	38.33	40.44	23.83	38.67	35.92	36.889
mDPR + MLM-16(XORQA & AUG-QAP)	48.21	**34.20**	38.78	40.76	24.81	39.49	34.37	37.231

Table 2: Evaluation results on XOR-TyDi QA test data with F1 and macro-average F1 scores

Models	Avg.
mDPR + mGEN (baseline1)	**27.55**
Unsupervised Retrieval	
OracleBM25 + MLM-14	2.75
EnsembleRank + MLM-wiki14	7.94
Supervised Retrieval	
mDPR(AUG) *with* MixCSE + MLM-14	11.91
mDPR(AUG) + MLM-14	14.27
mDPR + MLM-14	27.00
mDPR + MLM-14(XORQA & AUG-QA)	26.56
mDPR + MLM-14(XORQA & AUG-QAP)	26.83
mDPR + MLM-16	26.00
mDPR + MLM-16(XORQA & AUG-QA)	26.47
mDPR + MLM-16(XORQA & AUG-QAP)	26.57

Table 4: Results of macro-average F1 for all QA datasets

Models	MKQA ar	en	es	fi	ja	km	ko	ms	ru	sv	tr	zh-cn	Surprise ta	tl	Avg.
mDPR + mGEN (baseline1)	**9.52**	**36.34**	**27.23**	**22.70**	**15.89**	6.00	**7.68**	**25.11**	**14.60**	**26.69**	**21.66**	13.78	0.00	12.78	**17.141**
Unsupervised Retrieval															
OracleBM25 + MLM-14	2.80	10.81	3.70	3.29	5.89	1.53	1.51	5.49	1.85	7.42	2.94	1.81	0.00	8.23	4.090
EnsembleRank + MLM-14	6.43	31.66	20.02	17.38	10.68	**6.24**	4.38	21.03	6.27	21.09	17.13	7.22	0.00	8.39	12.709
Supervised Retrieval															
mDPR(AUG) *with* MixCSE + MLM-14	4.71	28.06	12.78	8.22	7.92	5.44	2.74	12.90	4.65	13.86	8.38	3.99	0.00	6.72	8.599
mDPR(AUG) + MLM-14	5.64	29.23	17.27	15.51	7.81	5.83	3.38	16.57	6.80	17.21	13.10	4.53	0.00	8.09	10.785
mDPR + MLM-14	8.73	35.32	25.54	20.42	14.27	6.06	6.78	24.10	12.01	25.97	20.27	13.95	0.00	11.14	16.040
mDPR + MLM-14(XORQA & AUG-QA)	8.46	35.12	24.74	19.50	14.38	5.62	7.22	23.24	11.46	24.49	19.67	**15.79**	**0.86**	12.18	15.909
mDPR + MLM-14(XORQA & AUG-QAP)	8.48	34.73	25.46	20.09	14.61	5.00	7.42	24.16	12.04	25.61	19.62	15.60	0.00	12.41	16.089
mDPR + MLM-16	8.15	34.14	24.85	19.38	13.73	5.93	6.51	22.21	11.46	24.91	18.82	13.62	0.00	12.59	15.451
mDPR + MLM-16(XORQA & AUG-QA)	8.21	34.06	25.65	20.14	14.22	5.80	6.70	24.40	11.82	25.71	19.92	15.42	0.40	12.36	16.057
mDPR + MLM-16(XORQA & AUG-QAP)	8.08	33.89	24.94	20.50	14.11	5.15	7.15	22.95	12.95	24.93	19.68	15.27	0.14	**13.07**	15.915

Table 3: Evaluation results on MKQA test dataset and two surprise languages with F1 and macro-average F1 scores

Conclusions

- Data augmentation with language- and domain-specialized additional training helps to improve resource-lean languages
- Unsupervised vs Supervised retrieval models
- Batch-size for Dense Passage Retrieval methods