UQA

Corpus for Urdu Question Answering

LREC-Coling 2024

Samee Arif
LUMS

Sualeha Farid
LUMS

Awais Athar
EMBL-EBI

Agha Ali Raza
LUMS

Introduction

- Systematic Inequalities in Language Technology Performance across the World's Languages (ACL, 2022):

"…which are both populous and under-served, along with also large but severely under-served languages like Kurdish, Urdu…"

- The shortage of resources in Urdu language hinders AI development for Urdu
- Limited AI access for Urdu speaking communities
- UQA: Corpus for Urdu Question Answering

Motivation

- Address the shortage of Urdu NLP resources
- Promote technological inclusivity
- Improve user interaction and information retrieval for Urdu

Methodology

- Based on Stanford Question Answering Dataset (SQuAD2.0)
- Evaluated translation models
- Introduced Enclose to Anchor, Translate, Seek (EATS) technique and translated SQuAD2.0 to Urdu
- Fine-tuned mBERT, XLM-R, mT5 on UQA achieving a **85.99 F1-Score** and a **74.56 Exact Match** score

Translator Evaluation

- Google Translator and Seamless M4T were evaluated for Urdu
- Two experiments were performed:
    1. Pilot evaluation on 100 sentences
    2. Final evaluation on 1,512 sentences
- Annotators were asked to pick between the translation produced by the two models

Translator Evaluation

Experiment 1

- Three annotators (computer science researches and native Urdu speakers)
- Krippendorff's Alpha was calculated to determine inter-rater reliability and was found to be 0.688

Seamless M4T	Google Translator	Both
51.67%	14.33%	34.0%

Translator Evaluation

Experiment 2

- Twelve annotators (undergraduate students, native Urdu speakers with English as medium of instruction)

Seamless M4T	Google Translator	Both
54.37%	37.43%	8.20%

Seamless M4T

- Seamless M4T tends to summarize or omit sentences when a paragraph of more than 1000 character length is translated
- Manual splitting of 3,307 paragraphs to prevent data loss

SQuAD2.0

- Data point:

Context: *The further decline of Byzantine state-of affairs paved the road to a third attack in 1185…*

Question: When did the Normans attack Dyrrachium?

Answer: 1185

Answer Index: 86

Is Answerable: True

Challenges of Translating SQuAD

- English:

Context: *The further decline of Byzantine state-of affairs paved the road to a third attack in 1185*

Answer Index: 86

- Translated:

Context: *بازنطینی ریاست کے مزید زوال نے 1185 میں تیسرے حملے کی راہ ہموار کی*

Answer Index: 31

Challenges of Translating SQuAD

- English:

Context: *The Raoulii were descended from an Italo-Norman named Raoul, the…*

Answer: *Raoulii were descended from an Italo-Norman named Raoul*

- Translated:

Context: *…جس راؤلی ایک اٹالو نارمن کی نسل سے تھے*

Translated Answer: راؤل نامی ایک اٹالو نارمن سے تعلق رکھتا ہے

Challenges of Translating SQuAD

- Linguistic differences between source and target language
- No one-to-one mapping between the words in source and translated language

Enclose to Anchor, Translate, Seek (EATS)

- Extracts the answer index in the translated text by using delimiters to tag the answer in the source text
- English:

Context: *Infrared radiation is used in industrial, ••scientific•• and medical applications.*

- Translated:

Context: *انفراریڈ تابکاری صنعتی ••سائنسی••اور طبی ایپلی کیشنز میں استعمال ہوتی ہے۔*

UQA Corpus

	Dev	Train
Answerable Questions	5,811	83,018
Unanswerable Questions	5,655	41,727

Fine-tuned Models

Model	F1 Score (%)	Exact Match (%)
mBERT	64.72	45.50
mT5-Small	67.24	52.37
mT5-Large	84.20	71.26
XLM-R	78.00	65.67
XLM-R-Large	84.42	72.24
XLM-R-XL	**85.99**	**74.56**

Conclusion and Future Work

- Established a pipeline to generate question answering dataset for low-resource languages
- Train a question generation model on UQA
- Generate domain specific datasets using the question generation model, removing the translation module from the pipeline

Thank You

Code and Data

- Code is publicly available at: github.com/sameearif/UQA
- Fine-tuned models and dataset is available at: huggingface.co/uqa