A COMPARARTIVE STUDY OF TRANSFORMER-BASED LANGUAGE MODELS' PERFORMANCE FOR ENGLISH & HINDI QUESTION ANSWER SYSTEMS

Final Dissertation Presentation

INDEX

BACKGROUND (QA SYSTEM)

The 4 questions for a Question Answering(QA) system

What

- ✓An information retrieval NLP task
- ✓Generates selective information specific to the question input.
- ✓Open-domain QA & Closed-domain QA systems

Why

- ✓For efficient information retrieval
- ✓For automated and intelligent query response mechanisms

Where

- ✓Online customer service systems
- ✓Chatbots
- ✓Virtual assistance systems
- ✓Automated query response mechanisms

How

- ✓Input feed to the model
- ✓Query or question processing
- ✓Document or passage retrieval
- ✓Computing similarity index and answer extraction

EVOLUTION OF QA SYSTEM

1. Manual QA systems (Till 2014)

- Built by humans, it cannot account for large datasets.
- Developing manual datasets is time consuming and excessive proof-reading required

2. RNNs, CNNs & LSTMs (2014-present)

- The sequential nature results parallelization limitations
- RNNs yet to resolve long-range dependencies

3. Self-attention Transformers (2017-present)

- Able to resolve the challenge of parallelization
- Able to resolve long-range dependencies
- Can handle transduction with attention instead of sequential framework

AIM & OBJECTIVES

Aim

To promote transformer-based approaches to build QA systems for low-resource languages.

- Compare RoBERTa & m-BERT for answer prediction of comprehension dataset ReClor in English & Hindi languages.
- Evaluate effectiveness of transformer model for an English-Hindi translation of the context-question pair for a comprehension dataset.
- Evaluate the effectiveness of transfer learning on the scale of fine-tunings required by the transformer model.

LITERATURE REVIEW

1st section:- About QA components of an MQA system

2nd section:- Existing applicability of transformer models in other domains

3rd section:- Advantages of transfer learning approach for language models and its accomplishments

4th section:- Different datasets used for QA tasks in monolingual, multilingual, and cross-lingual settings

PROBLEM STATEMENT

- Scarcity of training data for low-resource languages.

Challenge 1

- Most of the QA systems generate short or one-word answers and lack in predicting long sentences or subjective answers.

Challenge 2

- More quality work needed to address reasoning questions for complex text like reading comprehension.

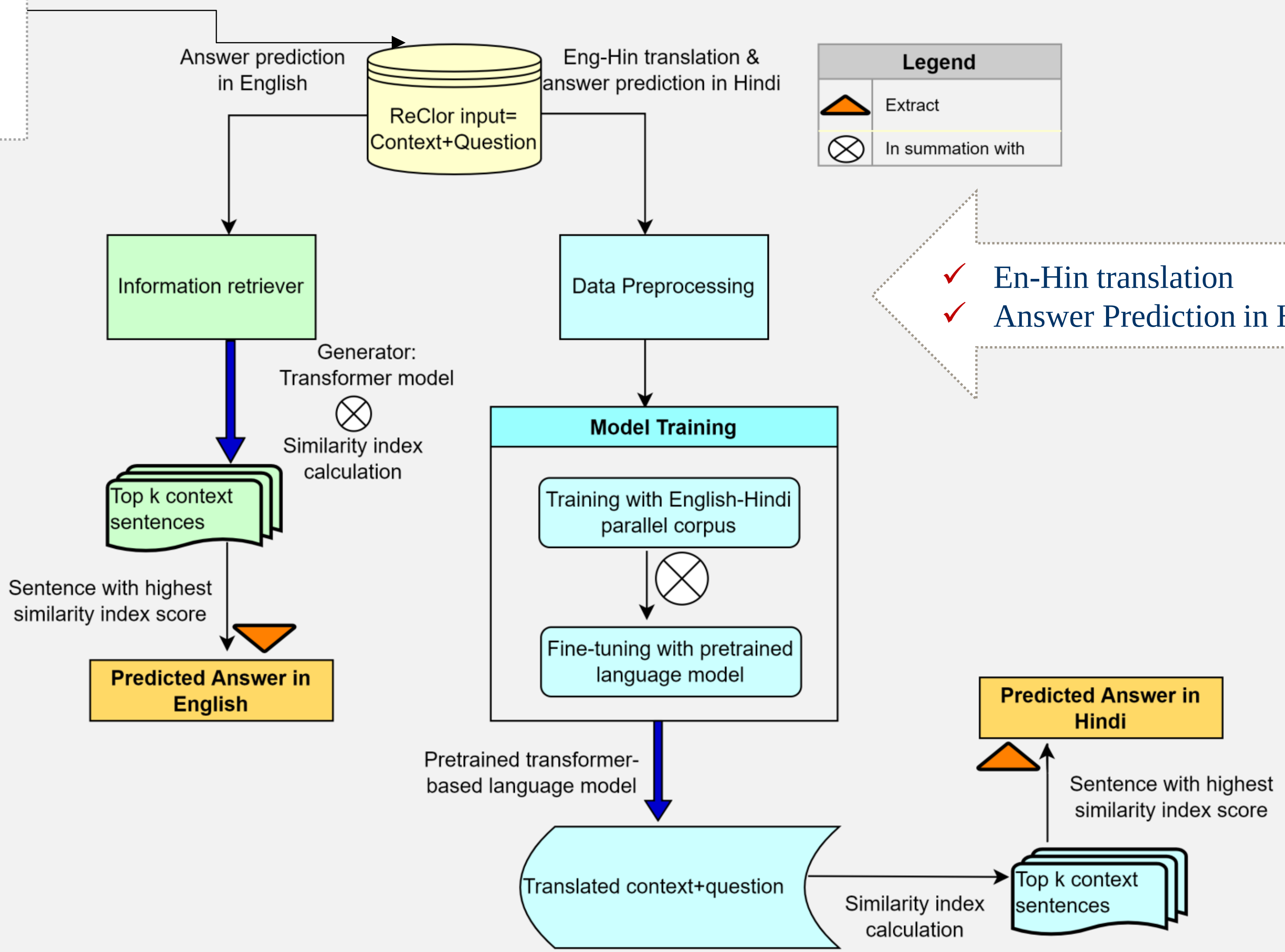
Challenge 3

METHODOLOGY
A TWO-PARTS APPROACH

6138 QA pairs with comprehension passages:- 4638 pairs (train data) with multiple answer options, 1000 pairs (test data), 500 pairs (validation data)

- ✓ Answer prediction in English

- ✓ En-Hin translation
- ✓ Answer Prediction in Hindi



ENGLISH→HINDI TRANSLATION-RESULTS

Metrics for English-Hindi translation	High	Average	Low	Inference/Example
Able to generated long sequence sentences	☑			English Question: Those whose view is described above hold inconsistent beliefs if they also believe that Hindi Question: जिनके विचार ऊपर वर्णित हैं, वे असंगत विश्वास रखते हैं यदि वे यह भी मानते हैं कि
Translated sentences’ proximity to ground truth		☑		Translation result from Google Translate: जिन लोगों के विचार ऊपर वर्णित है असंगत विश्वासों पकड़ अगर वे भी विश्वास है कि
# of parameter fine-tuning			☑	Most parameters unchanged from the pretrained model
Transliteration & Code-switching	☑			English word: ‘gray’ Transliterated Hindi version: ‘ग्रे’
Computational time			☑	Perhaps due to the transfer learning approach
POS tagging for superlatives*			☑	English superlative adjective: “and the smartest people” Hindi translation: “और स्मार्ट लोग” Ground Truth: “और सबसे स्मार्ट लोग”

* The metric result icon displayed in red font infers poor results

Created by: Suchitra A Gupta

ANSWER PREDICTION-RESULTS

Comparison Metrics	Results for Answer Prediction in English		Results for Answer Prediction in Hindi	
	RoBERTaBASE	m-BERT	xlmRoBERTa	m-BERT
Able to generated long sequence answers	↓	↓	↑	↓
Predicted answers proximity to ground truth	↑	↓	↑	↓
# of parameter fine-tuning	↑	↓	↑	↓
Transliteration	----	----	↑	↓
Code-switching	----	----	↑	↓
Computational time	↑	↑	↓	↑
Virtual Environment	GPU	TPU v3-8	GPU	TPU v3-8

↑ Good results ↑ Average Results ↓ Below Average

CONCLUSIONS & CONTRIBUTIONS

- Both $\text{RoBERTa}_{\text{BASE}}$ and m-BERT models are able to predict answers mostly in the form of short phrases or one-word answers.
- $\text{RoBERTa}_{\text{BASE}}$ is a better model to be used for answer prediction in English.
- xlmRoBERTa is an easier model to than m-BERT for answer prediction in Hindi.
- Transfer learning approach reduces computational time and can fetch better results with optimum fine-tuning.
- The transformer-based pretrained En-Hin parallel corpus contributed by the University of Helsinki supports transliteration but needs further pretraining to support POS for superlative adjective.
- m-BERT model demonstrates limited results for code-switching texts and phrases.
- m-BERT requires more fine-tuning than xlmRoBERTa.
- Both the models need further training to be able to generate subjective answers.

FUTURE WORKS AND RECOMMENDATIONS

RECOMMENDATIONS

- ✓ Transformer models like m-BERT that are trained using the Devanagari script need to be further trained with code-switching examples.
- ✓ Enhanced application of transformer models to be able to generate intelligent plausible distractors for comprehension passages.
- ✓ Answer label classification tasks for multiple answer options to be explored for ReClor dataset for Hindi and other languages

FUTURE WORKS

- ✓ RoBERTa and m-BERT need to be pretrained with more training examples to increase learning.
- ✓ Evaluate RoBERTa and m-BERT towards predicting ‘no answer available’ for a context-question pair.
- ✓ Further model training required to learn tokenization for long sequences to predict answers beyond one-word or short texts.
- ✓ Explore if there is any impact of the question-type on the answer predictability of the transformer model.

THANK YOU