AVADHAN: System for Open - Domain Telugu Question Answering

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

Priyanka Ravva*, Ashok Urlana, Manish Shrivastava
Language Technologies Research Center,
International Institute of Information Technologies - Hyderabad

November 18, 2023

Overview

Introduction and Motivation

- In the world around 6500 spoken languages are in existence, in those 1652 are from India.
- Telugu is the 4^{th} most spoken language in India.
- Exploration of low resource language
- Ease of accomplishing web based applications in native language
- Always the core aim of QA lies on the extraction of suitable answers only, not all the related documents to the query.

What are the challenges?

- Creation of pre-tagged dataset for question classification
- Building a base-line model for Telugu QA system
- Which classifier works better on Telugu?

Corpus Creation

- Manual dataset creation task involves more human intervention, mixed with various intuitions and analogies.
- Data creation performed based on the web crawling `https://upscgk.com/APPSC-gk`, `http://services.indg.in/online_quiz/index_te.php`
- We have created 1037 triplets [Question, Answer, Label] with the help of three annotators.
- The labeling of the data set was performed completely based on the answer type (Person, Location, Number, Organization, Time, Date, Percentage) related to the query.
- **Github :** `https://github.com/priyanka-ravva/Telugu-Question-Answering`

Examples

Answer Category	Query type	Outcome
పేరు (PERSON)	ఆనంద్ మఠ్ పుస్తకం రాసినది ఏవరు ? (Who wrote the book 'Anandamath'?)	బంకీమ్ చంద్ర చటోపాధ్యాయ (Bankim Chandra Chattopadhyay)
సంస్థ (ORGANIZATION)	ప్రపంచంలో అత్యంత లాభదాయక సంస్థ ఏది? (What is the world's most profitable company?)	అరంకో (Aramco)
నగరం (LOCATION)	ప్రపంచ ఆరోగ్య సంస్థ యొక్క ప్రధాన కార్యాలయం ఎక్కడ ఉంది? (Where is the headquarters of the World Health Organization?)	జెనీవా (Jeneva)
సంఖ్య (NUMBER)	మానవ శరీరంలో ఎన్ని ఎముకలు ఉంటాయి ? (How many bones are there in the human body?)	206
తేదీ (DATE)	ప్రతి సంవత్సరం ప్రపంచ పర్యావరణ దినోత్సవం ఎప్పుడు జరుపుకుంటారు? (when is the world environment day celebrated every year?)	జూన్ 5 (June 5)
సమయం (TIME)	చంద్రుడు భూమిని ఒకసారి చుట్టిరావడానికి ఎన్ని రోజులు పడుతుంది? (How many days does it take for the moon to encircle the earth?)	27 రోజులు (27 Days)
శాతం (PERCENTAGE)	సముద్రపు నీటి యొక్క సగటు లవణీయత ఎంత? (What is the average salinity of seawater?)	3.5 శాతం (3.5%)

Table 1: Categories of queries and answers

Model Description

AVADHAN has mainly three modules

- Information Retrieval
- Question Classification
- Answer Extraction

AVADHAN Architecture

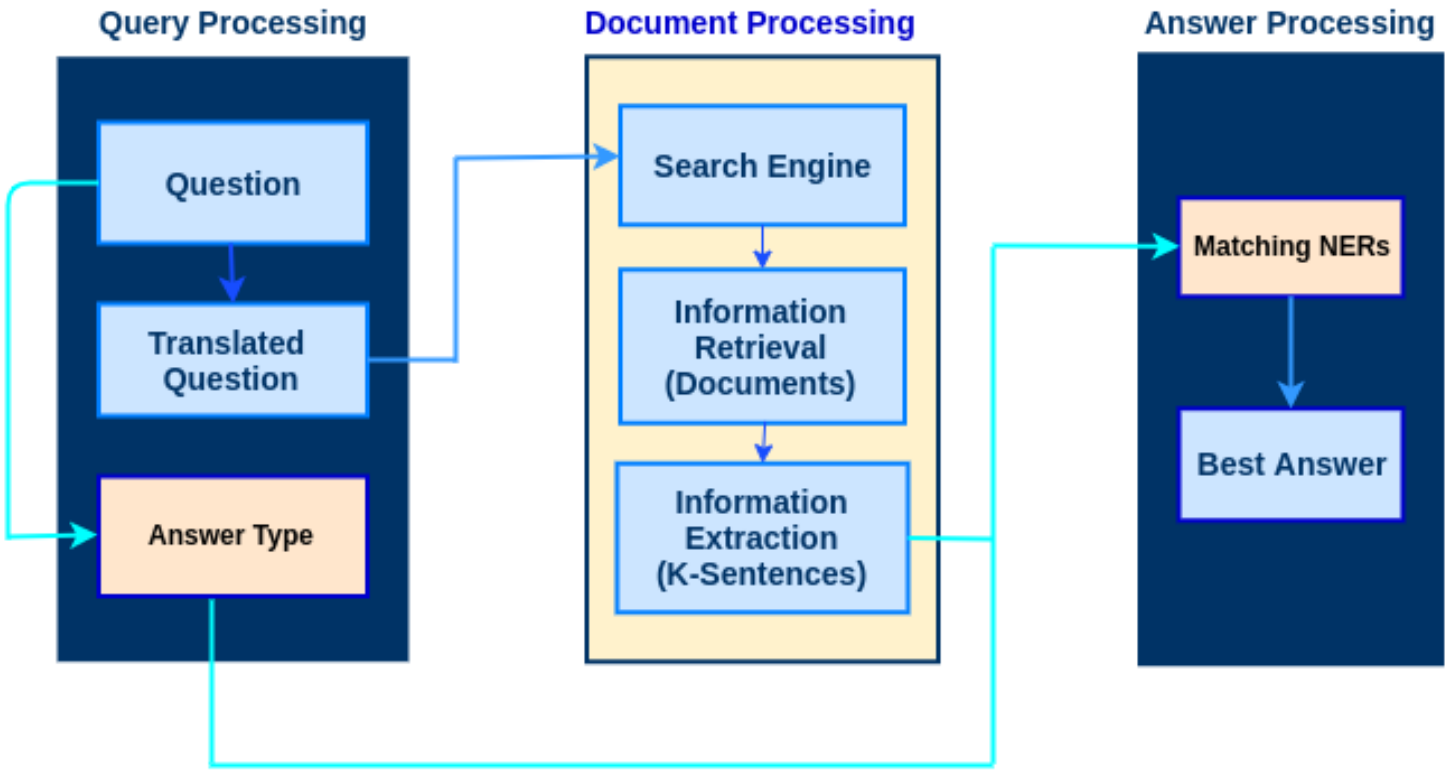

Figure 1: AVADHAN Architecture

Information Retrieval

- Web scraping technique is used to extract the unstructured data on the web into structured form using the "Bing" search engine.
- To avoid noise in data, we considered top 10 URLs with the most relevant information for the query.
- To find out the important sentences with respect to the query, we used the cosine similarity approach.

Question Classification

- Dataset was divided into train and test queries as 725 and 312 respectively.
- We have used TF-IDF for the input vector representation.
- Experiments were performed with different baseline neural network classifiers like LR, MLP and SVM for finding which classifier is better for Telugu question classification.
- Obtained accuracies are 71%, 72% and 73% for MLP, LR and SVM respectively.

Answer Extraction

- NERs applied on each of the top K-ranked sentences to extract the answer for the given query.
- If more than one answers occurs based on frequency of occurrences we have extracted the answer for the query.

Experiments and Results

- How many sentences(K) are essentially required to answer a query?
- Experiments were performed with varying number of sentences for fetching the correct answer.

Classifiers		K=10	K=20	K=30	**K=40**
SVM	EM	21.9	25.9	27.8	**31.6**
	PM	46.7	58.4	61.5	**68.5**
LR	EM	21.1	25.7	29.4	**31**
	PM	46.3	57.5	63.6	**66.6**
MLP	EM	21.5	24.5	29.4	**30**
	PM	47.3	57.4	62.8	**67**

Table 2: Overall performance of AVADHAN (in terms of %) by varying the number of sentences, **EM:** Exact match, **PM:** Partial match

Performance of AVADHAN

Category	Number of samples	Exact match (%)			Partial match (%)		
		C1	C2	C3	C1	C2	C3
Location	353	34.6	35.7	36.4	70	72	72.3
Person	317	18.9	20.8	20.5	68.8	66.4	68.5
Number	170	44.1	44.7	42.6	58.2	59.4	56.8
Date	125	28	32.8	37.2	80	76	87.6
Organization	29	27.6	17.2	20	37.9	31	26.7
Percentage	25	28	26.1	24	40	39.1	48
Time	18	22.2	5.6	18.8	55.6	61.1	75
Overall	**1037**	**30**	**31**	**31.6**	**67**	**66.6**	**68.5**

Table 3: Performance of AVADHAN for individual categories with K=40, **C1**- MLP, **C2** - LR, **C3**- SVM

Observations on Exact Match Cases:-

- Experiments were conducted with respect to MLP, SVM, LR (classifiers) to compare the accuracies.
- Particularly for TIME, PERSON, PERCENTAGE and ORGANIZATION categories the accuracy percentage is very low because of the uncertainty involved in the answer context.

Causes of low accuracy:

- Prefixes, suffixes and affiliations added to the NAME category
  Example:
  "ఆలిస్ ఇన్ వండర్ల్యాండ్" - పుస్తక రచయిత ఎవరు? ("Alice in Wonderland" - Who is the author of the book?)
  Predicted answer: లూయిస్ కారోల్ (Louis Carroll)
  Correct answer: లెవిస్ కారోల్ (Lewis Carroll)

Examples Continuation

- Possibility of occurrence of more than one possible answer for the same query

  Example:

  భారతదేశం యొక్క క్షిపణి మనిషి ఎవరు? (who is the missile man of India?)

  Possible Predicted answers:-

  1. అవూల్ పకీర్ జైనులాబ్దీన్ అబ్దుల్ కలాం (Avul Pakir Jainulabdeen AbLR.dul Kalam),
  2. ఎ. పి. జె. అబ్దుల్ కలాం (A. P. J. Abdul Kalam),
  3. డాక్టర్. అబ్దుల్ కలాం (Doctor. Abdul Kalam)

Observations on Partial Match Cases:-

In partial match, by fixing some threshold value, prediction of the final answer will be decided.

- The threshold was fixed as 0.7.
- In this approach also ORGANIZATION and PERCENTAGE answer categories produces low accuracy.

Conclusions

- This paper broadly explained the perplexities involved in the Telugu data set and also demonstrated various kind of query categories based on the resulting answer.
- SVM classifier performed better on AVADHAN as compared to MLP and LR.

What's ahead?

- Resolving of translation dependency.
- Reduction of time consumption with Google snippets data
- Extend to multilingual open-domain QA
- Improving the size of dataset to increase model performance

Type of query	Example
More than one answer possible query	2014 లో నోబెల్ శాంతి బహుమతి ఎవరికి వచ్చింది? (who got nobel peace prize in 2014?) Answers: కైలాష్ సత్యార్థి, మలాలా యూసఫ్‌జాయ్ (Kailash Satyarthi, Malala Yousafzai)
Miscellaneous type Query	మీరు సరిగ్గా ఏమీ చేయలేరా? (Can't you do anything right?) Answer: Rhetorical question
Time bounded Query	హాంగ్ కాంగ్ యొక్క ప్రస్తుత జనాభా ఎంత? (what is the current population of hong kong?) Answer: 73.9 లక్షలు (73.9 lakhs)

Table 4: Special types of Queries based on answers

Queries?