AVADHAN: System for Open-Domain Telugu Question Answering

Priyanka Ravva, Ashok Urlana, Manish Shrivastava
Language Technologies and Research Center, IIIT-Hyderabad,
CODS-COMAD 5-7 January 2020, Hyderabad, India

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.

Corpus Creation

- Manual dataset creation involves more human intervention, mixed with various intuitions and analogies.
- We have created 1037 QA pairs with the help of three annotators.
- The labeling of the dataset was performed completely based on the answer type (Person, Location, Number, Organization, Time, Date, Percentage) related to the query.
- We got the Fleiss kappa score of 0.85.
- **Github :** https://github.com/priyanka-ravva/Telugu- Question-Answering

Model Description

Our model named as 'AVADHAN', which has mainly three modules

- Information Retrieval
- Question Classification
- Answer Extraction

1. **Information Retrieval:** Using "bing" search engine we have extracted the most relevant information for the query, later using this information we have considered **top-K sentences** with the help of **Cosine Similarity method**.
2. **Question Classification:** To get the answer type for the query we need a Telugu question classifier, for that experiments were performed with different classifiers like LR, MLP and SVM.
3. **Answer Extraction:** NERs applied on top-K sentences to extract the predefined NERs classes, then match these classes with answer type to get the best answer for the given query.

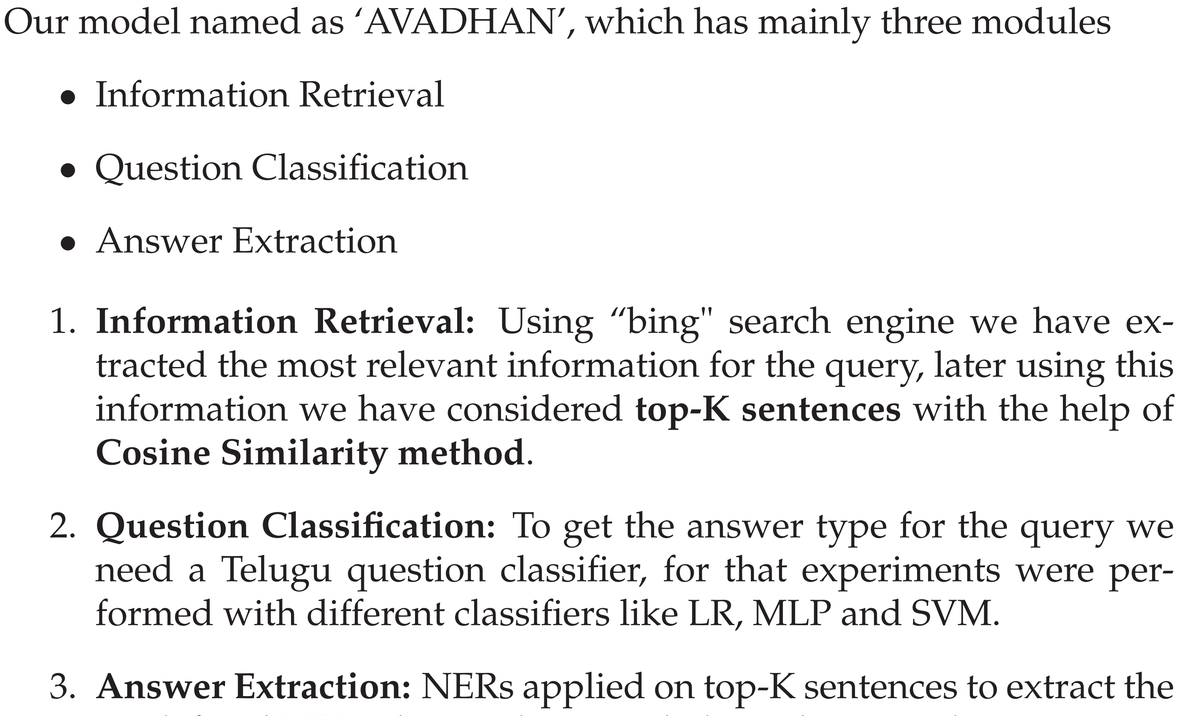

Figure 1: AVADHAN Architecture

Experiments and Results

- Experiments were performed with varying number of sentences (K) for fetching the correct answer.
- Gradually increasing K upto 40, we observed better predictions.

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 1: Performance of AVADHAN for individual categories with K=40

Causes for Low Accuracy:

- Prefixes, suffixes and affiliations added to the NAME category
  "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)
- Possibility of occurrence of more than one possible answer for the same query
  భారతదేశం యొక్క క్షిపణి మనిషి ఎవరు? (who is the missile man of India?)
  Possible Correct answers:-
  1. అవుల్ పకీర్ జైనులాబ్దీన్ అబ్దుల్ కలాం (Avul Pakir Jainulabdeen Abdul Kalam),
  2. ఎ. పి. జె. అబ్దుల్ కలాం (A. P. J. Abdul Kalam),
  3. డాక్టర్. అబ్దుల్ కలాం (Doctor. Abdul Kalam)

Observations on Exact and Partial Match Cases

- In Exact match, TIME, PERSON, ORGANIZATION and PERCENTAGE categories obtained low accuracy due to the uncertainty involved in the answer context.
- In Partial match method also ORGANIZATION and PERCENTAGE answer categories produces less accuracy.

Conclusions and Future work

- This paper broadly explained the perplexities involved in the Telugu data set and also demonstrated various kinds of query categories based on the resulting answer.
- Various experiments were performed for exact and partial match cases with different classifiers.
- With SVM classifier, AVADHAN produces better accuracies compared to MLP and LR.
- Possible to extend multilingual open- domain QA.
- To reduce the time consumption for retrieving the best answer, we can directly take the Google snippets data.

References

[1] Zhiping Zheng. Answerbus question answering system. In *Proceedings of the second international conference on Human Language Technology Research*, pages 399–404. Morgan Kaufmann Publishers Inc., 2002.

[2] Rami Reddy Nandi Reddy and Sivaji Bandyopadhyay. Dialogue based question answering system in telugu. In *Proceedings of the Workshop on Multilingual Question Answering*, pages 53–60. Association for Computational Linguistics, 2006.