Intelligent Question Answering Engine

TEAM MEMBERS

12IT28   KALYANAM RAJASHREE

12IT52   R GOKUL SHANTH

12IT63   S ASHISH BHARADWAJ

Introduction

- Need for QA : Huge volume of data – need for concise answers instead of just relevant documents

- Open Domain Vs. Closed Domain

- Issues to be tackled :

  - Question classification

  - Context and QA

  - Data sources for QA

  - Interactive QA

  - Advanced reasoning for QA

Related Work

Paper	Pros	Cons
Yih et. al	1. Using lexical Semantic resources 2. System is better regardless of algorithm 3. Outperforms syntactic analysis through dependency tree matching	1. Uncovered or inaccurate entity relations. 2. Lack of robust question analysis 3. The need of high-level semantic representation and inference
Mikolov et al.	1.Paragraph Vector learns fixed-length feature representations from variable-length pieces of texts 2. Overcome the weaknesses of bag-ofwords models 3. state-of-the-art results on several text classification and sentiment analysis tasks	1. Paragraph Vector can be expensive 2. Takes 30 minutes to compute the paragraph vectors test set, using a 16 core machine
Lei Yu et al.	1. Using Distributed representations to match questions with answers by considering their semantic encoding 2. Model easily applicable to a wide range of domains and languages 3. Model matches state of the art performance on the answer sentence selection task	1. Distributed representations are not very well equipped for dealing with cardinal numbers and proper nouns.
Callison-Burch et al.	1. Introduced novel features based on TED boosted F1 score by 10% compared with the use of more standard features 2. Modified design of the TED model is the state of the art in the task of ranking QA pairs	Without POS/NER/DEP features, F1 score drops 2. If CRF fails to find an answer, it is forced to tag an answer
Severyn et. al	1. Model is based on a convolutional neural network 2. A 3% absolute improvement in MAP and MRR2. 3. System requires no external parsers or resources.	1. Lower accuracy on long text pairs 2. strong overfitting on the training set will give poor generalization on the test data

Outcome of Literature Survey

► Current methods have not fully explored a collection of syntactic, semantic and string matching features.

► No coherent link between question and answer domains

►

Model	QASENT		WIKIQA	
	MAP	MRR	MAP	MRR
Word Cnt	0.5919	0.6662	0.4891	0.4924
Wgt Word Cnt	0.6095	0.6746	0.5099	0.5132
LCLR	**0.6954**	0.7617	0.5993	0.6086
PV	0.5213	0.6023	0.5110	0.5160
CNN	0.5590	0.6230	0.6190	0.6281
PV-Cnt	0.6762	0.7514	0.5976	0.6058
CNN-Cnt	0.6951	**0.7633**	**0.6520**	**0.6652**

Problem Statement

► The goal of the project is to build a robust and efficient Intelligent Question Answering Engine which is able to answer simple factoid questions by answer sentence extraction making use of a sentence retrieval system working under a document retrieval system.

- ► Question Processing and Classification
- ► Document Processing
- ► Document Retrieval
- ► Answer Sentence Retrieval
- ► Interaction Management and Driver
- ► Intelligent Answer Prediction

Question Classification

▶ Multiclass Logistic Regression Classifier

▶ Syntactic features – word, POS and chunks

▶ Semantic features – named entities and entity types

Question Type	Question example
Descriptive(DESC)	What are liver enzymes?
Human(HUM)	Name the scar-faced bounty hunter of The Old West?
Numeric(NUM)	How many Jews were executed in concentration camps during WWII?
Abbreviation(ABBR)	What does YMCA stand for?
Entity(ENTY)	What is considered the costliest disaster the insurance industry has ever faced?
Location(LOC)	What sprawling U. S. state boasts the most airports?

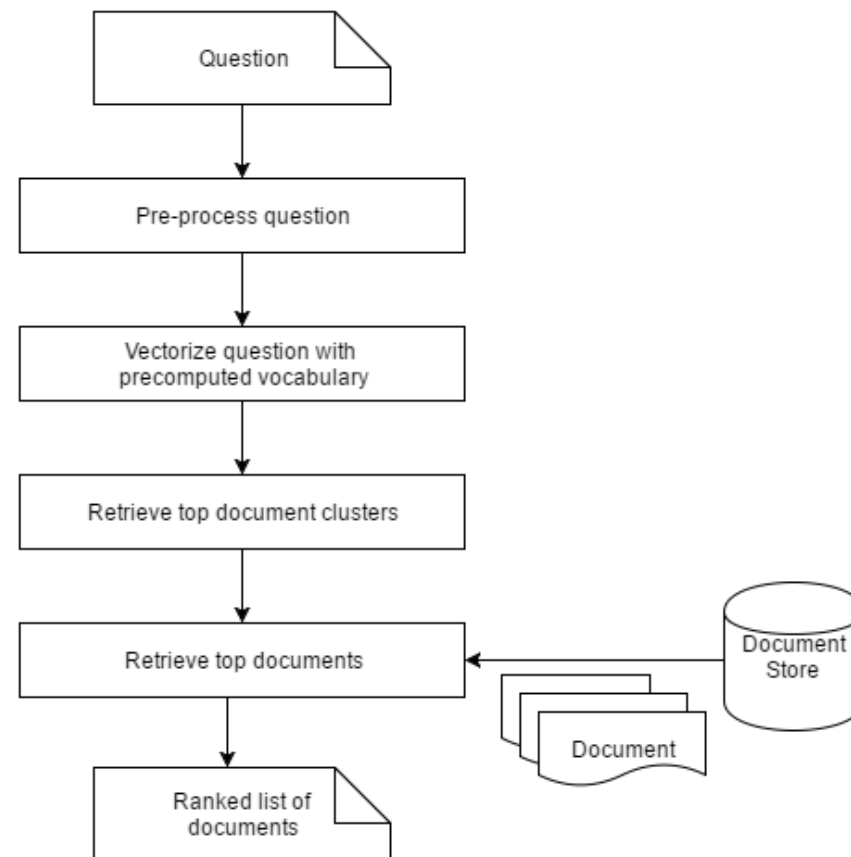# Document Clustering

▶ Tf-idf vectors of each document

▶ Kmeans clustering

▶ Silhouette score

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Plot of Silhouette Score vs. # Clusters
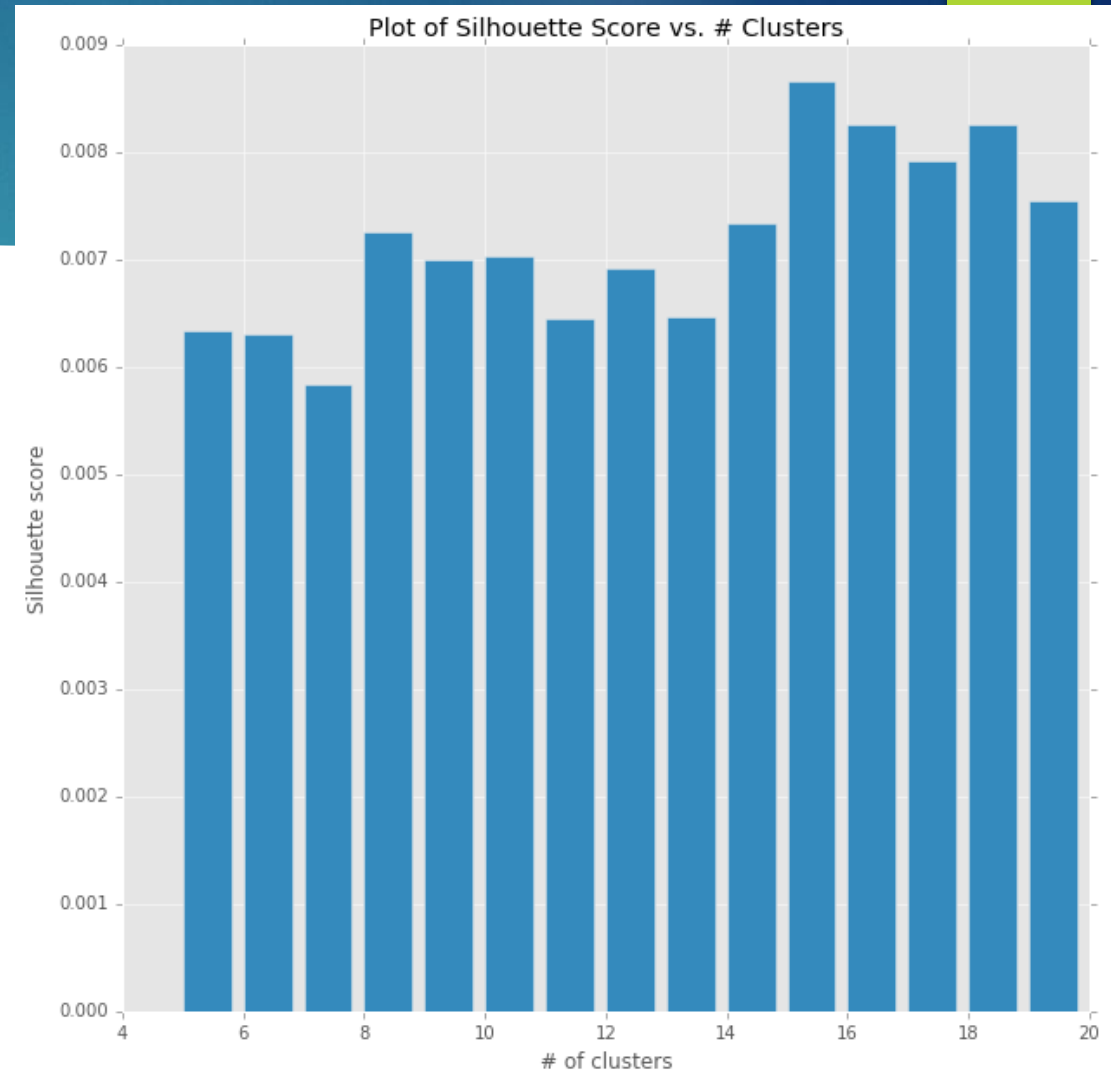
Document Retrieval

Answer Sentence Retrieval

- Logistic Regression Classifier
- Probabilistic binary classifier

	Train	Dev	Test	Total
# of ques.	2,118	296	633	3,047
# of sent.	20,360	2,733	6,165	29,258
# of ans.	1,040	140	293	1,473
Avg. len. of ques.	7.16	7.23	7.26	7.18
Avg. len. of sent.	25.29	24.59	24.95	25.15
# of ques. w/o ans.	1,245	170	390	1,805

Features
Unigram Count
Bigram Count
Trigram Count
Lemma Count
Weighted Word Count
Synonym Count
Antonym Count
Hypernym Count
Hyponym Count
Levenshtein Distance
Postive Similarity
Negative Similarity

|	
QuestionID	Q1
Question	how are glacier caves formed?
DocumentID	D1
DocumentTitle	Glacier cave
SentenceID	D1-0
Sentence	A glacier cave is a cave formed within the ice of a glacier.
Label	1
Type	train

Sample row from Dataset

| Unigram	2
Bigram	0
Trigram	0
Lemma Count	3
W-idf	1.832581464
Synonym	0
Antonym	0
Hypernym	0
Hyponym	0
Levenshtein distance	4
Positive Similarity	0.479220407
Negative Similarity	0.395776447
Question Type	DESC
Question	how are glacier caves formed
Sentence	a glacier cave is a cave formed within the ice of a glacier

Sample feature vector

# Pipeline
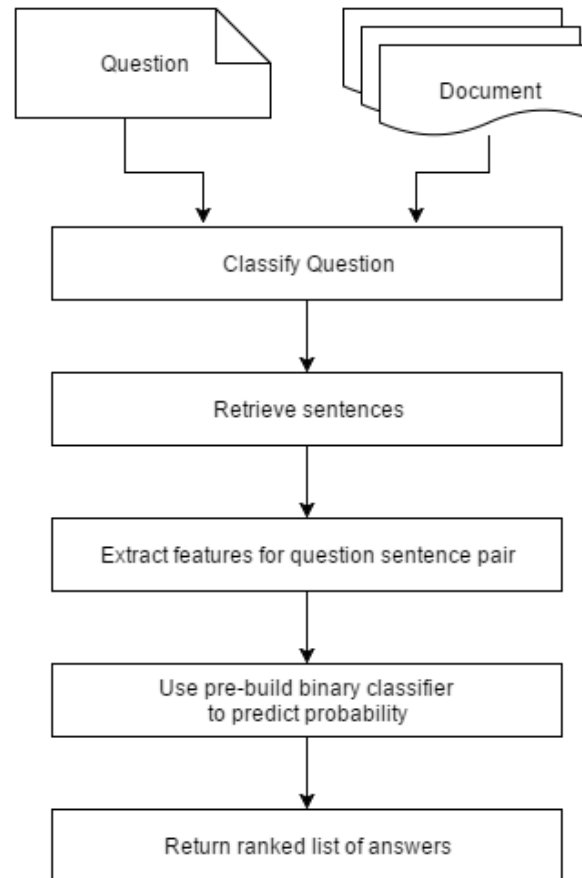
Intelligent Answer Prediction
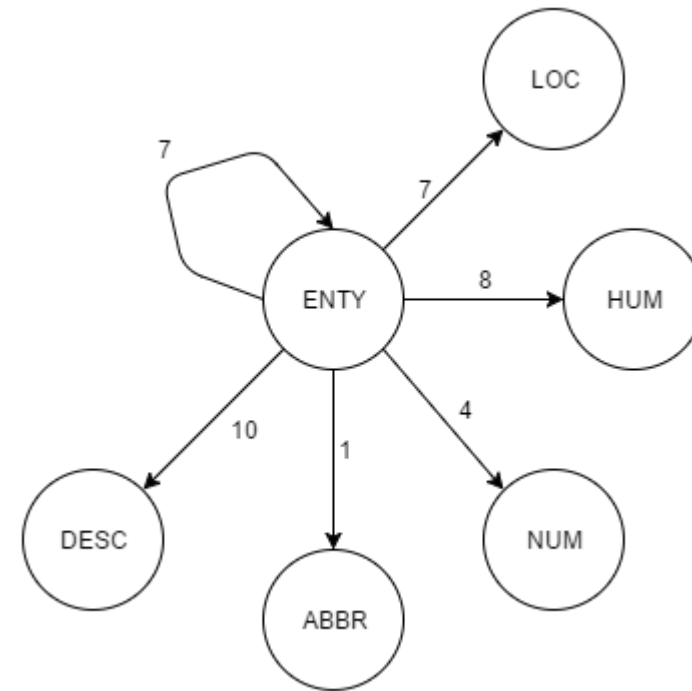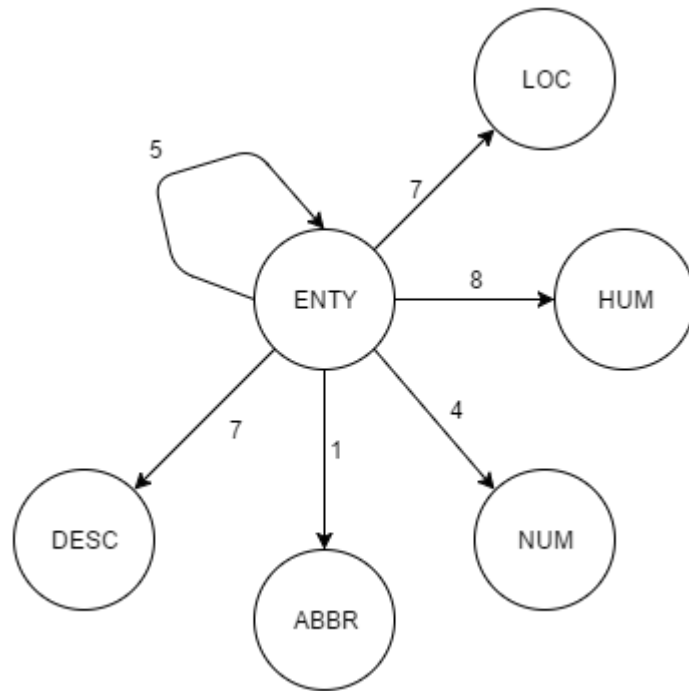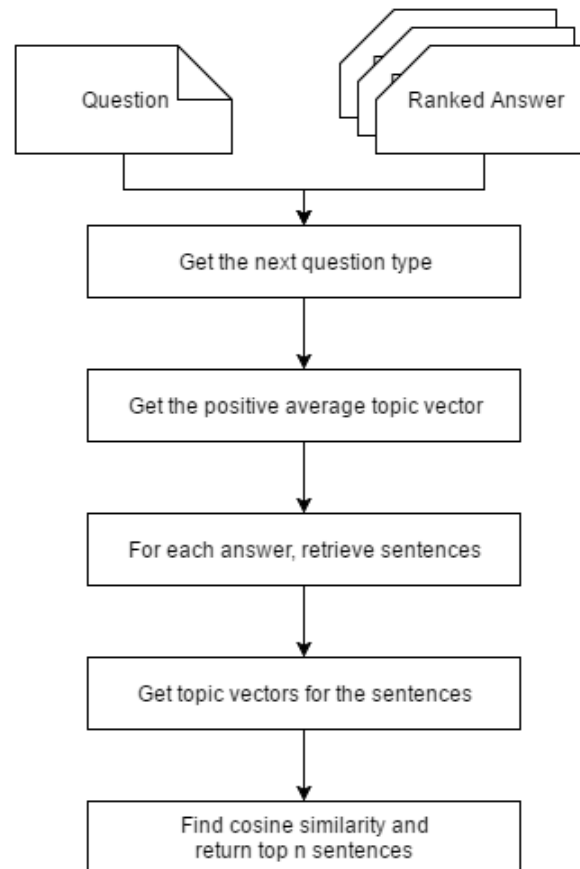
Results and Analysis

Performance Comparison

Model	WikiQA	
	MAP	MRR
Word Cnt	0.4891	0.4924
Wgt Word Cnt	0.5099	0.5132
LCLR	0.5993	0.6086
PV	0.5110	0.5160
CNN	0.6190	0.6281
PV-Cnt	0.5976	0.6058
CNN-Cnt	0.6520	0.6652
Proposed Method(Dev)	**0.6810**	**0.6858**
Proposed Method(Test)	**0.6789**	**0.6900**

$$Mean\ Reciprocal\ Rank = \frac{1}{|Q|} \sum_{q=1}^{Q} \frac{1}{Rank\ of\ answer}$$

$$Mean\ Average\ Precision = \frac{\sum_{q=1}^{Q} AvgPrecision(q)}{|Q|}$$
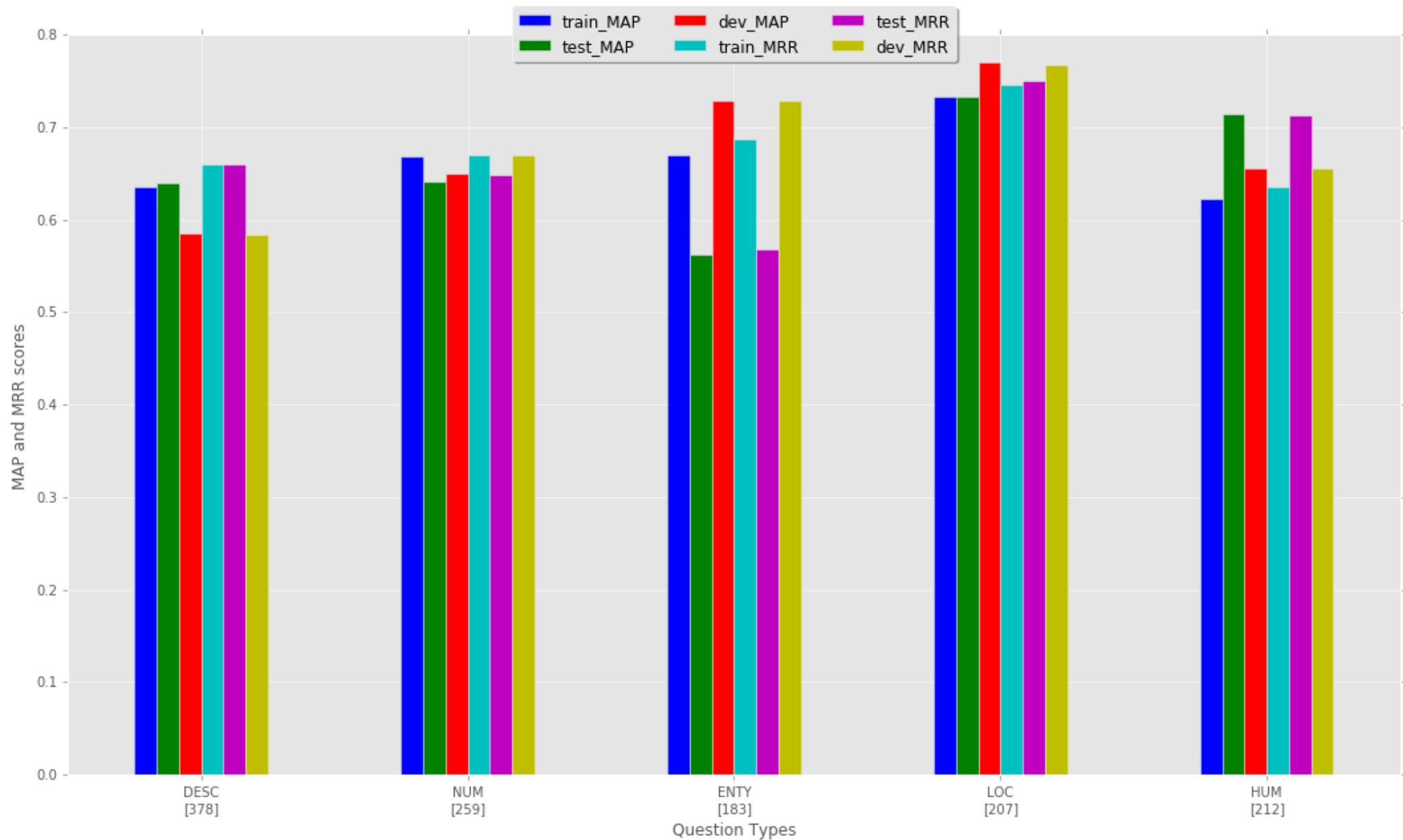
Question type wise performance analysis

Demo

Watson Jr.

What is the main story arc of Harry Potter?

ANSWER   REFRESH

ANSWER

The main story arc concerns Harry's quest to overcome the Dark wizard Lord Voldemort , whose aims are to become immortal, conquer the wizarding world , subjugate non-magical people, and destroy all those who stand in his way, especially Harry Potter.

- There are also many other themes in the series, such as prejudice and corruption.
- According to Rowling, the main theme is death.
- The series has also had some share of criticism, including concern for the increasingly dark tone.

1 2 3

Watson Jr.

What is the main story arc of Harry Potter?

ANSWER   REFRESH

ANSWER

A story arc is an extended or continuing storyline in episodic storytelling media such as television , comic books , comic strips , boardgames , video games , and in some cases, films .

- Webcomics are more likely to use story arcs than newspaper comics , as most web comics have readable archives online that a newcomer to the strip can read in order to understand what is going on.
- On a television program , for example, the story would unfold over many episodes.
- Many American comic book series are now written in four or six-issue arcs, within a continuing series.

Watson Jr.

What is the main story arc of Harry Potter?

ANSWER   REFRESH

ANSWER

Although story arcs have existed for decades, the term "story arc" was coined in 1988 in relation to the television series Wiseguy , and was quickly adapted for other uses.

- Webcomics are more likely to use story arcs than newspaper comics , as most web comics have readable archives online that a newcomer to the strip can read in order to understand what is going on.
- On a television program , for example, the story would unfold over many episodes.
- Many American comic book series are now written in four or six-issue arcs, within a continuing series.

Conclusion and Future Work

- Wide spectrum of features used – syntactic, semantic and string matching

- Topic vector and average topic vector introduced

- Question type wise analysis done

- Intelligent prediction of next question using weighted edge graphs

- Making use of advanced learning algorithms

- Connecting to the web to make the system dynamic

- DESC questions need to be more adequately handled