Reading Comprehension System using Natural Language Processing

Authors

Chinnu Pittapally
cpittapa@hawk.iit.edu

Mayuri Kadam
mkadam@hawk.iit.edu

Shorabh Dhandharia
shorabhd@hawk.iit.edu

Abstract

This project attempts to build a reading comprehension system that answers the questions based on text paragraphs related to different topics. This project utilizes machine learning and NLP techniques to classify different questions by their category and provide the best possible answer for them. We focus on providing short answers for all the fact-based questions related to each context paragraph.

We have used 2 different approaches to get the possible answer sentences. In first approach, we check the similarity between questions and paragraph sentence using n-gram model and apply rule-based selection for collecting possible answer sentences for each question category. In the second approach, we use Word Mover's Distance (WMD) to check similarity between question and paragraph sentences. We select relevant sentences with highest similarity and apply different machine learning models on both approaches to get to a single answer from the possible answers for each question.

We applied logistic Regression, Multilayer Perceptron and Keras Sequential model to predict one answer from the relevant answers chosen for each question. Questions like ‘who’, ‘where’ ‘which’ and ‘when’ were answered with good f1 score comparatively, in both approaches. The ‘what’ category answers were predicted better with the WMD similarity approach and rarely picked the correct answer from possible answer sentences using n-gram.

Introduction

Reading comprehension is the ability to read text, process it, and understand its meaning and answer the questions based on the text. This requires understanding of natural language and semantic/syntactic knowledge. There has recently been a strong increase in the research of question answering, which identifies and extracts answers from text paragraph.

At a high level, the goal here is to provide a system that can answer questions posed by humans in natural language for a given text. This includes reading the paragraph text, which may be a collection of sentences forming a story, or a set of facts forming a knowledge base. The system is expected to read the question and output an answer which may be a single word or a natural language sentence. We are focusing on predicting the right sentence in the paragraph which has the correct answer.

Since information retrieval has acquired a lot of importance with the continuous flow of data, reading comprehension systems plays a significant role in extracting the correct information by querying the retrieved data in less time with accurate information. This system could be used for aiding in faster summarization of text by answering the questions based on context. This could also be used as a grading system to generate the correct answers for text paragraphs and compare them with students' answers.

Background/Related Work

1. SQuAD: 100,000+ Questions for Machine Comprehension of Text

This paper analyzes the dataset to understand the types of reasoning required to answer the questions, leaning heavily on dependency and constituency trees. It implements a logistic regression model with lexicalized and dependency tree path features. We are trying to implement some of the experiments in this paper for our baseline model.

2. Exploiting Syntactic and Shallow Semantic Kernels for Question/Answer Classification

In this paper, answers for questions of reading comprehension are provided using question classification and answer classification. Apart from Parse Trees, BOW, POS and NERs, a tree structure - Predicate Argument Structures (PAS) and new kernel functions using SVMs to extract semantic information for better results I defined. We are referring this paper for questions and answers classification.

3. Question Classification using Head Words and their Hypernym

This paper talks about extracting the question head word and to use WordNet semantic features to the hypernyms introduction within six depths, which implicitly employs WordNet hypernyms. They proposed using Word shape in each question for question classification. They use Support Vector Machines and Maximum entropy models which can integrate features from many heterogeneous information sources for classification. Some of the approaches using head word and its features might be helpful for us to answer the what questions.

4. Parsing and Question Classification for Question Answering

This paper deals with augmenting the Penn treebank with more parse trees that can be used for the classification of questions in a better way. Having more question parses help classifying the questions better which in return improves the accuracies for the answer prediction in the reading comprehension. As an outcome of this research the Penn Treebank was scaled up from 250 parses to 1153 parses. We had considered this paper for utilizing the parses tree results for identifying sentence structure.

5. Answering Reading Comprehension Using Memory Networks

This paper extends earlier approaches for utilizing memory networks by combining LSTM model for reference. It uses a hybrid two-pass approach of first effectively pruning the input statements using traditional NLP feature engineering, and then passing the pruned statements and question through a memory network to obtain the answer.

Approach

N-gram - Rule Based Approach

1. Read text paragraph and each question
2. Match question unigrams with every sentence of the paragraph
3. Collect sentences for which the n-gram match was >= threshold value
4. Build feature vectors for the combination of all possible answer sentences combined with their respective questions
5. Label the sentence as positive if the n-grams for that sentence matched to the true answer and rest sentences as negative
6. Feed these feature vectors and labels to different classification models

Word2Vec - WMD Similarity Approach

1. Read text paragraph and each question
2. Match question to each sentence in the paragraph using WMD similarity
3. Collect 4 sentences from the paragraph with highest WMD similarity score for each question
4. Build feature vector for the combination of all possible answer sentences combined with their respective questions
5. Label a sentence positive if the true answer lies in that sentence and other sentences as negative
6. Feed these feature vectors and labels to different classification models

The above approaches are used to find the possible sentences that could contain the answers and collect the resulting feature vectors that are processed with Logistic Regression, Multilayer Perceptron and Keras Sequential models. We set the parameters for these models with their best parameter setting obtained through parameter tuning on validation set. For implementing these models, we have used gensim, keras and scikit-learn.

Experiment

- **Datasets:**

Training data can be fetched from:
https://drive.google.com/file/d/0B6hJIF_NL2HsY2dyb0V1SVlCSjA/view?usp=sharing

Testing data can be fetched from:
https://drive.google.com/file/d/0B6hJIF_NL2HsOG1qZmNHbTZydWM/view?usp=sharing

Source: SQuAD: Stanford Question Answering Dataset https://rajpurkar.github.io/SQuAD-explorer/

Stanford Question Answering Dataset (SQuAD) is a new reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is

a segment of text, or *span*, from the corresponding reading passage. SQuAD contains over 100,000+ question-answer pairs for 500+ articles.

Total number of paragraphs in train set: 18896
Number of question answer pairs in train set: 87599

Question type: what	Count: 48705
Question type: where	Count: 3766
Question type: which	Count: 6458
Question type: who	Count: 9034
Question type: whom	Count: 230
Question type: why	Count: 1246
Question type: how	Count: 9509
Question type: when	Count: 6667

The dataset is in JSON format, where each topic consists of various paragraphs with their respective questions and true answer pairs.

- **Sample data:**
-

{**'context'**: "As at most other universities, Notre Dame's students run several news media outlets. The nine student-run outlets include three newspapers, both a radio and television station, and several magazines and journals. Begun as a one-page journal in September 1876, the Scholastic magazine is issued twice monthly and claims to be the oldest continuous collegiate publication in the United States. The other magazine, The Juggler, is released twice a year and focuses on student literature and artwork. The Dome yearbook is published annually. The newspapers have varying publication interests, with The Observer published daily and mainly reporting university and other news, and staffed by students from both Notre Dame and Saint Mary's College. Unlike Scholastic and The Dome, The Observer is an independent publication and does not have a faculty advisor or any editorial oversight from the University. In 1987, when some students believed that The Observer began to show a conservative bias, a liberal newspaper, Common Sense was published. Likewise, in 2003, when other students believed that the paper showed a liberal bias, the conservative paper Irish Rover went into production. Neither paper is published as often as The Observer; however, all three are distributed to all students. Finally, in Spring 2008 an undergraduate journal for political science research, Beyond Politics, made its debut.",
'qas': [{'answers': [{'answer_start': 248, 'text': 'September 1876'}],
'id': '5733bf84d058e614000b61be',
'question': 'When did the Scholastic Magazine of Notre dame begin publishing?'}]

For n- gram approach we performed the n-gram match for each question with each sentence and found the similarity of each sentence to the question. After finding the similarity of sentences we applied the rules for collecting the correct answer based on the question category and it's expected NER or POS tag in the answer. We also applied classification models to predict correct answers for these relevant sentences. In Word2Vec approach we tried cosine similarity and WMD similarity to find the relevant

sentences for each question and observed WMD provides better sentence match for a given question. We picked top 4 sentences and predicted the correct answer using different models. We evaluated our models on validation set that is 10% held out data from the training set. We ran cross-validation on our training set and checked the performance of the models on validation set. We collected the best parameter settings using grid search for parameter tuning. Utilizing the best combination return by grid search we fitted the models on our training set and predicted the correct answers for each question from the possible answers.

- **Baseline method**

We considered the n-gram rule based model as our baseline and proceeded with higher models using different approaches for relevant sentence selection and correct answer prediction. Our baseline model gives .5091 F1 on test set.

- **Results of experiments**

Table 1: Evaluation results for predicting the true answer from possible answer on test set

Model Approach	Accuracy on Training	Precision on Test	Recall on Test	F1 on Test
N-gram rule based - Logistic Regression	0.75411	0.7447	0.7447	0.5091
N-gram rule based - MLP Classifier	0.5626	0.4790	0.6374	0.5470
N-gram rule based - Keras Sequential	0.7216	0.59	0.55	0.57
Word2Vec WMD - Logistic Regression	0.5673	0.2919	0.5442	0.3800
Word2Vec WMD - MLP Classifier	0.5626	0.4790	0.6374	0.5470
Word2Vec WMD - Keras Sequential	0.7693	0.60	0.77	0.67(avg)

- **Hyper parameter tuning using grid search**

```
0.657 (+/-0.037) for {'C': 1, 'random_state': 2, 'penalty': 'l1', 'class_weight': None}
0.657 (+/-0.037) for {'C': 1, 'random_state': 42, 'penalty': 'l1', 'class_weight': None}
0.657 (+/-0.037) for {'C': 1, 'random_state': None, 'penalty': 'l1', 'class_weight': None}
0.663 (+/-0.026) for {'C': 1, 'random_state': 2, 'penalty': 'l2', 'class_weight': None}
0.663 (+/-0.026) for {'C': 1, 'random_state': 42, 'penalty': 'l2', 'class_weight': None}
0.663 (+/-0.026) for {'C': 1, 'random_state': None, 'penalty': 'l2', 'class_weight': None}
0.647 (+/-0.021) for {'C': 0.1, 'random_state': 2, 'penalty': 'l1', 'class_weight': 'balanced'}
0.647 (+/-0.021) for {'C': 0.1, 'random_state': 42, 'penalty': 'l1', 'class_weight': 'balanced'}
0.647 (+/-0.021) for {'C': 0.1, 'random_state': None, 'penalty': 'l1', 'class_weight': 'balanced'}
0.636 (+/-0.024) for {'C': 0.1, 'random_state': 2, 'penalty': 'l2', 'class_weight': 'balanced'}
0.636 (+/-0.024) for {'C': 0.1, 'random_state': 42, 'penalty': 'l2', 'class_weight': 'balanced'}
0.636 (+/-0.024) for {'C': 0.1, 'random_state': None, 'penalty': 'l2', 'class_weight': 'balanced'}
0.672 (+/-0.022) for {'C': 0.1, 'random_state': 2, 'penalty': 'l1', 'class_weight': None}
0.672 (+/-0.022) for {'C': 0.1, 'random_state': 42, 'penalty': 'l1', 'class_weight': None}
0.672 (+/-0.022) for {'C': 0.1, 'random_state': None, 'penalty': 'l1', 'class_weight': None}
0.669 (+/-0.027) for {'C': 0.1, 'random_state': 2, 'penalty': 'l2', 'class_weight': None}
0.669 (+/-0.027) for {'C': 0.1, 'random_state': 42, 'penalty': 'l2', 'class_weight': None}
0.669 (+/-0.027) for {'C': 0.1, 'random_state': None, 'penalty': 'l2', 'class_weight': None}
```

- **Comparison of predicted answers with true answers**

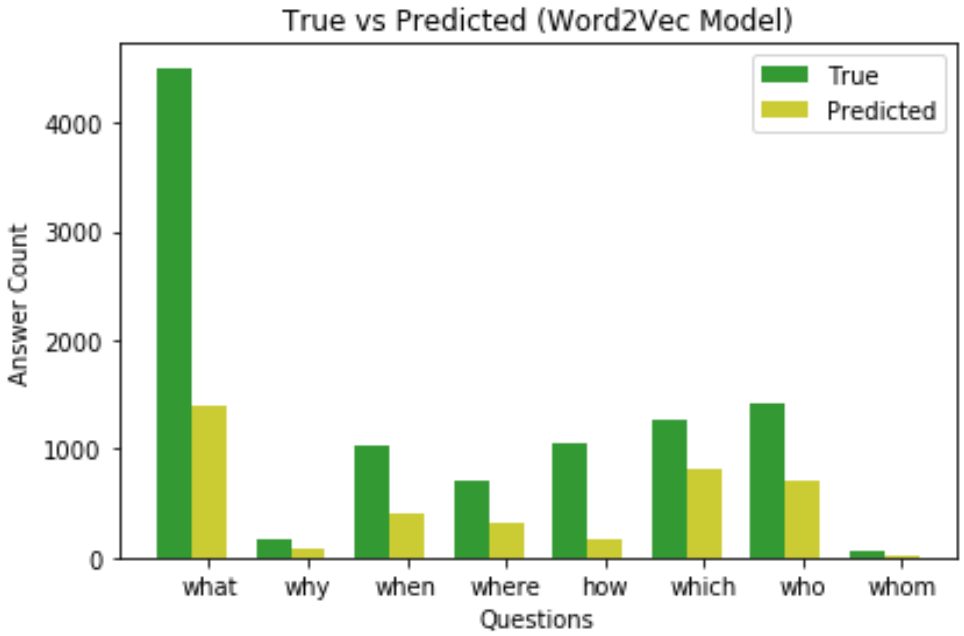

Figure 1: Comparison of true answers and predicted answer using WMD similarity

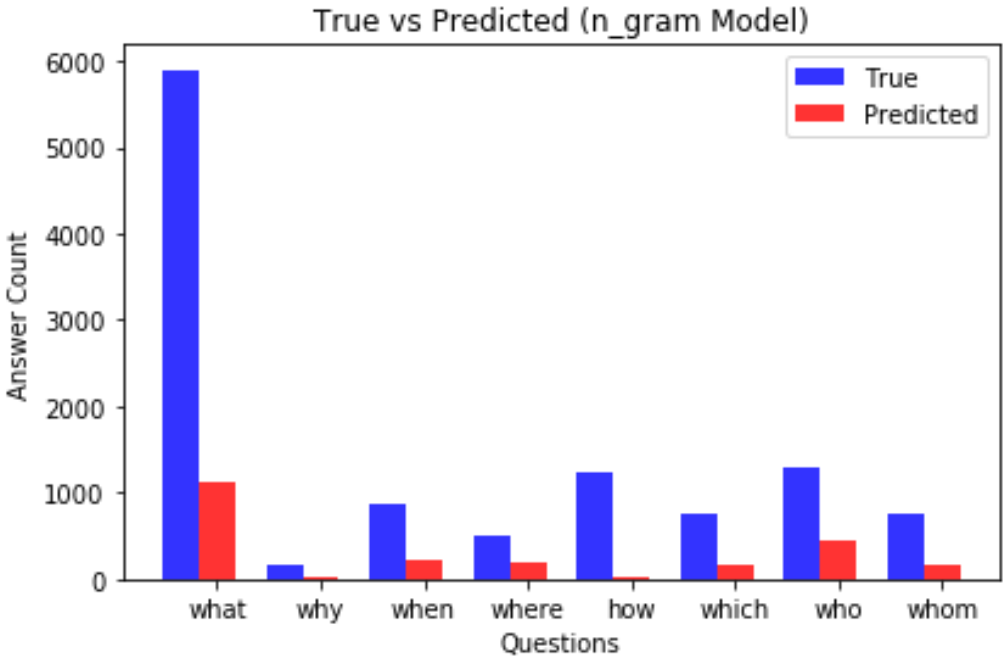

Figure 2: Comparison of true answers and predicted answer using N-gram similarity

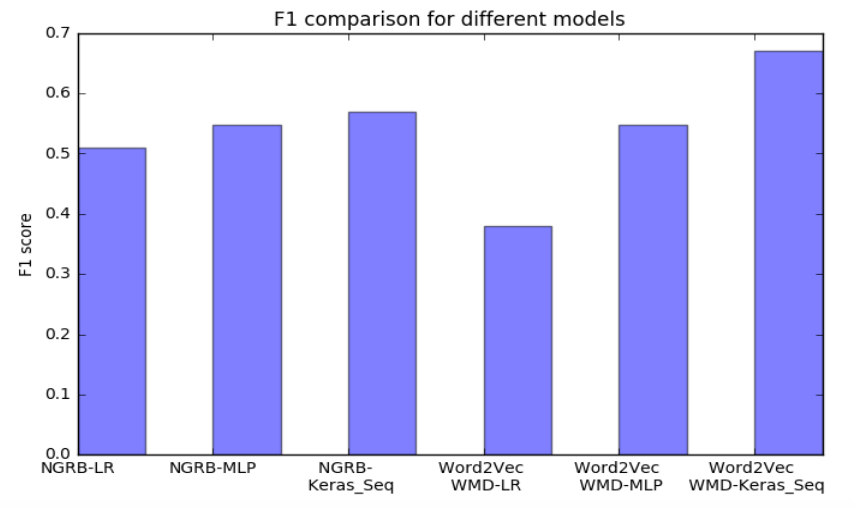

Figure 3: Comparison of different models

- **Sample results**

Q1: Who is the quarterback for the Panthers?

A1: The Panthers finished the regular season with a 15–1 record, and quarterback **Cam Newton** was named the NFL Most Valuable Player (MVP).
Predicted: True Probability: 0.562216605174

A2: They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of four teams that have made eight appearances in the Super Bowl.
Predicted: False

A3: The Broncos finished the regular season with a 12–4 record, and denied the New England Patriots a chance to defend their title from Super Bowl XLIX by defeating them 20–18 in the AFC Championship Game.
Predicted: False

A4: They defeated the Arizona Cardinals 49–15 in the NFC Championship Game and advanced to their second Super Bowl appearance since the franchise was founded in 1995.
Predicted: False

Q2: When were the finalists announced?
A1: The league announced on **October 16, 2012**, that the two finalists were Sun Life Stadium and Levi's Stadium.
Predicted: True Probability: 0.511527817711

A2: The San Francisco Bay Area last hosted in 1985 (Super Bowl XIX), held at Stanford Stadium in Stanford, California, won by the home team 49ers.
Predicted: False
A3: The South Florida/Miami area has previously hosted the event 10 times (tied for most with New Orleans), with the most recent one being Super Bowl XLIV in 2010.
Predicted: False

- **Model Errors:** Predicted true for all sentences which were selected as possible answers

Q3: Who was the game's top receiver?
A1: Sanders was his top receiver with six receptions for 83 yards.
Predicted: True
A2: Anderson was the game's leading rusher with 90 yards and a touchdown, along with four receptions for 10 yards.
Predicted: True
A3: Manning finished the game 13 of 23 for 141 yards with one interception and zero touchdowns.
Predicted: True

Conclusion

We observed the n-gram approach performed qualitatively better compared to all other models but it could rarely predict the 'what category questions. The second approach of using WMD similarity enabled us to find the possible answers for 'what' categories and quantified the identification of relevant answer sentence selection for other categories as well. Comparing the models in n-gram approach shows a high performance using the Keras Sequential model than Logistic Regression and MLP. In the WMD approach we found the prediction of MLP classifier gave us 0.5470 F1 for predicting correct answer being the maximum with this approach. Implementing the Word2Vec features for these models. Comparing predictions across the question categories we found that predictions for 'which', 'who' and 'when' categories were higher in the word2vec approach and 'where', 'who', 'when' predicted better using n-gram approach. The current approach can be extended by implementing coreference resolution between the entities and sentences in the paragraphs instead of just basing it on the sentence similarities for predicting the correct answers for what categories. Implementing deep learning and ensemble classification could be another approach to take this work forward in future.

References:

1. https://arxiv.org/pdf/1606.05250.pdf
2. http://www.aclweb.org/anthology/P/P07/P07-1.pdf#page=814
3. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.228.9777&rep=rep1&type=pdf#page=957
4. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.57.6834&rep=rep1&type=pdf
5. http://cs224d.stanford.edu/reports/KapashiDarshan.pdf