Project Proposal
11-797: Question Answering
Team Name: The Flying Riddlers

Vasu Sharma, Srividya Pranavi Potharaju, Nitish Kulkarni, Gabriel Bayomi
Language Technologies Institute
Carnegie Mellon University

February 12, 2018

1 Introduction

In this proposal, we present the basic plan we intend to follow for our course project towards building a complete Question Answering system. We specify the team composition, team vision, data sources we plan to use, some challenges we expect to face, research questions we plan to answer, proposed plan, responsibilities for individual members and a time-line for the same.

2 Team Members

- Vasu Sharma
- Srividya Pranavi
- Nitish Kulkarni
- Gabriel Bayomi

3 Team Vision

To build a robust, state of the art end-to-end Question Answering system with novel improvements over the existing benchmarks, analyze its relative performance over multiple datasets (BioASQ and MS MARCO) and explain its shortcomings.

4 Datasets

We plan to initially target the MS Marco [4] and BioASQ [1] datasets for this course project. BioASQ and MS Marco offer a complimentary set of problems that we expect to face while working on them and hence we feel will allow us to learn to apply a diverse set of skills while building the system. BioASQ is also an ongoing challenge which aligns really well with the course and we expect to participate in the same to test our models in a competitive environment. MS Marco on the other hand is a much more comprehensive dataset on which we will be competing with models proposed by tech giants like Microsoft. The size of data also differs significantly between the 2 datasets which makes it an interesting comparative study to see how neural vs non neural models fare on the 2 datasets.

5 Expected Challenges

Before working with these datasets we foresee some challenges and roadblocks we expect to face while building Question Answering Systems for the same. Some of these challenges are ass follows:

1. BioASQ dataset uses a very specific Biomedical vocabulary which will render a lot of standard pre-trained embedding techniques useless. We would have to tune our embeddings to the specific type of biomedical vocabulary that BioASQ entails. This very different type of vocabulary also restricts the use of pre-trained models on other larger datasets as most other datasets will use a vocabulary which is vastly different from the vocabulary used here.

2. The small size of BioASQ dataset poses several challenges for using several of the state of the art Neural Attention models since they contain a very large number of parameters which will need a huge amount of data to train. Hence we need to smartly choose our approaches while respecting the amount of data we have to train those models.

3. The abstractive style of answers in both MS Marco and BioASQ pose an interesting language generation problem which is known to be a fairly hard problem in NLP. Most generative models suffer from the problem of not being able to effectively model sentence level syntax and semantics while also having problems in generating rare words from the vocabulary. These are problems we will have to think about while generating an abstractive answer generation framework.

4. The largely varying nature of the datasets prohibits the use of a standard model for all QA datasets. We will try to construct a more generalizable approach which can perform well on each of these datasets while simultaneously experimenting with more tailored approaches for each of the datasets.

6 Individual Responsibilities

In this section we detail the components each of us wants to work on and a basic outline of how we want to go about solving that specific problem.

6.1 Vasu

I am highly excited about trying out various state of the art neural approaches to perform a number of subtasks which question answering entails.
One research Question I wish to explore is to test if Neural modules can replace the existing standalone modules to create an end to end neural system which can outperform the existing modularized architecture?
Another research question I have is whether imposing sentence level linguistic constraints can improve generation quality by forcing the network to adhere to nuances of natural language?
In light of this the parts I want to work on are:

- Trying out **Neural Attentional models** for extractive answer generation from passages. Since the release of SquaD dataset [7], Neural attentional models have enjoyed tremendous success on the Machine comprehension task which entails extracting an answer span which can best answer a given question. I am particularly interested in exploring the use of R-Net [3], [9], [8] and [10] style architectures to perform answer span extraction.
- **Dual Co-attention** is something I wish to explore further. Most recent approaches perform an attention computation using the question as the reference and generate attention over the passage to detect which part of the passage is most relevant to the question. However, I feel a dual attention mechanism which makes this attention mechanism bi-directional and explores the attention over the question itself. This will allow more nuanced encoder representations for the answer decoder to work on.
- The next thing I want to work on is an **abstractive answer generation** framework. Most present approaches are extractive in nature leading to great BLEU [5] and ROUGE [2] scores on BioASQ dataset but attaining terrible human evaluation scores. On MS Marco datasets these approaches fail miserably as the ground truth answers are not spans from the passages themselves but paraphrases of the same.
  The way I intend to go about the generative mechanism is to start with a basic encoder-decoder seq2seq model which uses an LSTM based Language model to generate the answer word by word or character by character. However, we know that such approaches suffer from problems like lack of semantic and syntactic coherence across the sentence and also frequent repetition of words common across the corpus. I plan to remedy this by the use of explicit language constraints while also using better generation mechanisms like pointer networks and copy-generate mechanisms. The traditional log likelihood of generation sequence based loss function is also which can be improved to directly incorporate linguistic constraints.
  The small size of BioASQ is a potential problem for training such networks. What we plan to avoid this issue is to use transfer learning and data augmentation from richer data sources like PubMed.
- Another major thing that I want to experiment with is to have an **end to end neural QA pipeline** which doesn't fragment the pipeline into components. Fragmentation of the pipeline prevents the learning signals from one module to be effectively backpropogated to lower level modules and also don't allow the later modules to fix the errors made by the earlier modules. A fully end to end neural module can remedy both of these problems by allowing more compact integration between the various network modules and allowing them to learn collaboratively rather than individually. Such an end to end module should also be easier to train and will not have the problem of compounding errors from various standalone modules.
- **Document and sentence ranking** is also something that interests me and I wish to try out triplet ranking loss based siamese ranking networks and other novel architectures to see if we can improve the document and sentence ranking techniques being used in most present architectures.
- I also intend to try out **ensembling** neural and modularized non neural architectures and see if there respective strengths can be combined into a joint ensemble model which can distill their strengths into a single model for Question Answering.

6.2 Srividya Pranavi

I am interested in working at the intersection of Information Retrieval and Neural Text Generation models, leveraging both the techniques to build a Question Answering system aiming at improving human readability and accuracy of the correct answers. I want to understand various modules in QA pipeline and improve on the existing BioASQ architecture. I am also interested in exploring different neural model techniques for MS Marco dataset as the data set is large and it is better aligned to real world queries.

> *General Hypothesis*: A QA system that provides accurate, non-repetitive and human readable answers must be supported by a strong Information Retrieval system and abstractive methods.
>
> *Specific Hypothesis 1*: Ranking of an IR system both at the snippet level and the sentence level can be enhanced using better retrieval algorithms, Learning to Rank Models. LeToR models can be trained using previous year's data and the relevance document information provided.
>
> *Specific Hypothesis 2*: Diversification plays a significant role, as the answer that covers more intents, the better. This retrieval module should ideally take care of repetition.
>
> *Specific Hypothesis 3*: Human readability is a harder problem to tackle with. There are two possible methods I plan to try, paraphrasing and text-summarization. I intend to do both at the last step of the pipeline based on my general hypothesis.

In this regard, I would like to focus on:

- **Query Reformulation**: I would like to work on query reformulation by expanding the original query with potential query terms.This method primarily suits MS Marco dataset where the queries are from real users. Real world queries are messy, short and ambiguous. Hence the need of query expansion comes into play.I would like to implement the standard pseudo relevance feedback (Okapi BM25) based query expansion.
- **Enhanced Retrieval using BM25**: I want to expand the original relevant snippet set by using an enhanced retrieval algorithm like BM25 which takes into account both the term frequency and inverse document frequency normalizing over document length. The intuition is to not limit to the relevant snippets and documents given originally, rather expanding it using other information sources. Metrics: MAP, P@5
- **Diversification using xQUAD**: I plan to explore diversification methods that explicitly diversify the results in contrast to Maximum Marginal Relevance method which does an implicit diversification. The main idea is to rank documents in the order of maximum coverage and minimum redundancy or repetition.xQUAD is proven to give good results when combined with BM25 retrieval method, the primary reason to explore this combination. Metrics : NDCG, P-IA@5
- **Text Summarization**: I intend to explore different summarization techniques both extractive and abstractive (sequence to sequence learning, Generative Adversarial Network based approach) for answer generation, to improve sentence tiling for better human readability. Metrics: Rouge-N
- **Paraphrasing using Neural Models**: Paraphrasing is important in the context of human readability as most of the extracted text is not human readable. So, as a last step in the pipeline, I want to try generating paraphrases of the extracted answer which can help a human understand the clinical jargon in simple terms.There are different neural architectures proposed to generate paraphrases which I would explore during the course time.
- **Memory Networks**: I want to explore Dynamic Memory Networks, a recent neural architecture (also used in MS Marco baseline) which processes input sequences and questions, forms episodic memories, and generates relevant answers. I intend to try this model for MS Marco data set along with exploring its relevance in the context of BioASQ dataset before implementing this end to end model.

6.3 Nitish Kulkarni

My primary hypothesis is that there can be a generalized architecture of a well-performing Question Answering system would whose parameters can be tuned based on the specific data-set (size and nature) and the evaluation mechanism. I am also keen on exploring the interdependence of the different modules in the QA pipeline, and understand the behaviour of different combinations of specific modules.

With these goals in mind, I intend to work on the following hypotheses:

Hypothesis 1 : A well-performing QA system would comprise an ensemble of unsupervised and supervised document retrieval models. For smaller data-sets, the traditional IR techniques would yield better performance, while for large data-sets, LETOR and neural ranking algorithms would be better alternatives.

Hypothesis 2 : The retrieval performance as well as text summarization can be enhanced by incorporating unsupervised models trained on larger unlabelled data-sets. This could be implemented by learning distributed representations for domain-specific words or sentences to capture syntactic and semantic relationships, or by training generative models for assessing the quality of summarization.

Hypothesis 3 : The summarization of an answer can be improved by incorporating the evaluation criterion (or a proxy thereof) in the objective/loss function being optimized. A combination of extractive and abstractive techniques might yield a better over-all summarization performance.

To verify the stated hypotheses, I propose to implement the following techniques:

- **Query Processing**

  I wish to explore the effect of variants of query processing techniques such as stemming, POS tagging, normalization and query term expansion on the quality of the retrieved snippets. I also plan to experiment with different query restructuring and expansion methods using pseudo relevance feedback and term expansion.

- **Relevance Ranking**

  To begin with, I want to implementing benchmark IR models such as *Indri*. Subsequently, I am interested in implementing unsupervised neural extraction techniques using pre-trained word embeddings and soft matches of words instead of TF-IDF similarities. As an alternative to word similarity, I also wish to implement a sentence-similarity based model with pre-trained sentence embeddings, for sentence embeddings tend to contain more contextual information.

  In addition , I would also want to address the issue of multiple intents in the case of an ambiguous query using implicit diversification techniques like *PM2*.

- **Classification**

  For many of the answer types such as yes/no, factoid and list, I wish to see if the retrievals can be improved in a supervised setting using traditional methods Rank SVM and more recent methods such as LSTM/CNN based neural architectures.

- **Text Summarization**

  Since abstractive summarization methods typically tend to yield low ROUGE-N scores and extractive models have poor human readibility, I intend to work on developing a model that's a cross between the two, with a loss function to optimize for more meaningful and smoother transitions while retaining many n-grams from the extracted sentences. One way to implement this is by combining a language model for readable summary with policy-learning to maximize ROUGE using reinforcement learning [6]. At the same time, I am also interested in exploring more advanced extractive methods based on neural architectures and graphical models.

6.4 Gabriel Bayomi

I'm particularly interested on the end of the Question Answering pipeline: learning to rank, text summarization via abstraction or extraction and the exploration of new techniques to improve the final results given the information retrieved from the previous method. Moreover, I want to explore the current platform of BioASQ and understand the overall architecture in order to tackle small problems that could have a strong impact (following the pareto principle where 80% of the issues might come from 20% of the causes). These methods include:

- **Learning To Rank**: Given a particular pair of a query q and a sentence s, there are numerous ways of developing a function $f(q, s)$ where f represents a ranking (an attributed numerical value to represent relevance). From simple methods like BM25 to newer machine learning approaches, different formulations tend to present a strong effect on the final results of a ranking module. I intend to explore these differences, hopefully combine effective approaches, and optimize the current methodologies by providing a better rank estimator for sentences.

- **Extraction-based Methods for Summarization**: Most of the most successful method for text summarization come from extraction methods, where specific excerpts from the corpus are extracted and combined/reused to formulate a summary of the document collection. I intend to explore these methods by developing a module where the input are the top-k ranked sentences and, given a query, the output is an optimal answer (taking an ideal answer as training basis).

- **Abstraction-based Methods for Summarization**: Abstraction methods, usually based on neural networks, are still an open and difficult language problem, due to the many issues related to generation methods: guaranteeing a reasonable, grammatically and semantically correct text output is a non-trivial process. However, the capability of generating answers that are not necessarily dependent on extracted excerpts of text has the potential of better human readability as the process is closer to what an actual human would produce. I believe that the careful combination of abstraction and extraction methods could provide interesting results.

- **General Improvements**: Tiling + Search/IR: I believe that there still gaps to explore the full information given by queries (query expansion), improving the generalization of the IR methods. Moreover, I believe it's possible to upgrade sentence tiling techniques. For example, understanding how LSA could improve the current structure. I'm a firm believer of the Pareto Principle in relation to project development. Therefore, there are likely small changes on the current baseline environments of BioASQ and MS MARCO that could potentially have more impact than big systematic changes. I intend to pursue these tasks during the semester.

  General Hypothesis: A strong QA system would be able to take advantage of the whole available information adequately, ignoring meaningless content and summarizing informational data.

  Specific Hypothesis 1: Final results of an IR system both at the snippet level and the sentence level can be enhanced using an ensemble/fusion of the top-k possible answers.

  Specific Hypothesis 2: The Ranking of a QA system can ideally prioritize the most informational content and ignore meaningless data.

7 Tentative Timeline

Week	Date	Task
1	14 Feb	Explore code and dataset
2	21 Feb	Get an initial working end-to-end system
3	28 Feb	IR Module Enhancements
4	5 March	LETOR + Overall Improved Working system for 1^{st} submission
5	19 March	Improved Extractive Summary + Neural Retrieval Methods, tuned system for 2^{nd} submission
6	25 March	TBD

References

[1] A challenge on large-scale biomedical semantic indexing and question answering. *http://www.bioasq.org/* (2013).

[2] Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (Barcelona, Spain, July 2004), S. S. Marie-Francine Moens, Ed., Association for Computational Linguistics, pp. 74–81.

[3] Natural Language Computing Group, M. R. A. R-net: Machine reading comprehension with self matching networks. *ACL 2017* (2017).

[4] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. Ms marco: A human generated machine reading comprehension dataset. *CoRR abs/1611.09268* (2016).

[5] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. *ACL 2002* (2002).

[6] Paulus, R., Xiong, C., and Socher, R. A deep reinforced model for abstractive summarization. *CoRR abs/1705.04304* (2017).

[7] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100, 000+ questions for machine comprehension of text. *CoRR abs/1606.05250* (2016).

[8] Seo, M. J., Kembhavi, A., Farhadi, A., and Hajishirzi, H. Bidirectional attention flow for machine comprehension. *CoRR abs/1611.01603* (2016).

[9] Shen, Y., Huang, P., Gao, J., and Chen, W. Reasonet: Learning to stop reading in machine comprehension. *CoRR abs/1609.05284* (2016).

[10] Xiong, C., Zhong, V., and Socher, R. Dynamic coattention networks for question answering. *CoRR abs/1611.01604* (2016).