Ethan Nguyen, Vu Pham, Mohan Rangarajan

Literary Review

XCS224U

Natural Language Understanding

Abstract

This document is a review of relevant existing literature for building different components of a Question Answering System

Table of Contents

[Literature Review 2](#_Toc36836904)

[Introduction 2](#_Toc36836905)

[Q & A System 2](#_Toc36836906)

[Attention is All You Need 2](#_Toc36836907)

[TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents 3](#_Toc36836908)

[SpanBERT: Improving Pre-training by Representing and Predicting Spans 3](#_Toc36836909)

[Personalizing Dialogue Agents: I have a dog; do you have pets too? 4](#_Toc36836910)

[Q & A with Knowledge Graph Integration 5](#_Toc36836911)

[Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks 5](#_Toc36836912)

[HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention 5](#_Toc36836913)

[Semantic Parsing for Single-Relation Question Answering 6](#_Toc36836914)

[Simple Question Answering by Attentive Convolutional Neural Network 6](#_Toc36836915)

[Knowledge Graph Embedding Based Question Answering 7](#_Toc36836916)

# Literature Review

## Introduction

Our team (Naturally Speaking) is interested in exploring the idea of building a domain specific question answering system (Q & A) that utilizes a knowledge graph to store and retrieve information specific to that domain. Our team comprising (Ethan Nguyen, Mohan Rangarajan, Vu Pham) researched a set of papers around two areas

1. Q & A – focusing on effective representation of the question posed and utilizing reading comprehension on retrieved paragraphs to determine the associated response
2. Integration between Q & A and the knowledge graph to efficiently retrieve responses from the knowledge graph

In each category, we have outlined the papers chosen, the motivation behind selecting the paper and what we anticipate leveraging in our final project from the paper.



Figure 1 : Initial idea for the project

## Q & A System

In this category, we focused on papers that help us capture the intent of the question by utilizing contextual embedding and attention mechanisms. In addition, we wanted to also use the transformer decoder to improve the reading comprehension and thereby extract the appropriate response to the question.

### Attention is All You Need

#### Motivation

This paper has the seminal work on utilizing attention mechanism and not rely on the sequential nature of the recurrence thereby allowing parallelization. If our Q & A system responses required us to train a large domain specific corpus, then employing this approach would reduce the training time required.

#### Problem/Task Definition

Sequential nature of the RNNs and GRUs prevent parallelization of computation required to accelerate the training process in preparation for an NLP task like language modeling and machine translation. The parallelization aspect is especially important when there are a lot of examples and memory limitations that constrain batching across examples.

#### Paper Summary

The paper proposes relying entirely on “attention” mechanisms. By using multi-head attention in conjunction with positional encoding, the paper takes advantage of order of the sequence in the input, dependencies within the input and dependencies between inputs/outputs. Multi-head attention also enables the model to attend to information from different representation subspaces at different positions. A position wise feed-forward network is used in conjunction with the attention sub-layer. The model comprises an encoder stack and a decoder stack with each stack having a self-attention and a feed-forward network. The decoder stack has an additional attention layer that allows the decoder to attend all positions in the input sequence.

#### Future Work

The paper looks at performance of translation and generalization as applied to English constituency parsing. Other NLP tasks such as Question Answering using Q & A datasets or Dialogue systems have not been evaluated. Image, audio and video modalities have not been tried and may be useful especially when encountering a combination of modalities in the corpus. When long sequences are involved, the impact on computational performance can be evaluated. To do so, we can restrict self-attention to consider only a neighborhood of limited size in the input sequence centered around the respective output position.

#### References

Authors: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, 2017, [*Attention is All You Need*](https://arxiv.org/pdf/1706.03762.pdf)

### TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents

#### Motivation

When the context (i.e. in the paragraph or in the knowledge graph from which to retrieve answers) does not have the answers, the system must rely on providing approximate answers or respond back with a “negative” response (like “I am sorry I don’t have the answer etc.”). While this may be appropriate in some scenarios, we may be able to elicit more clarity on the user’s intent (with a response like “Let me clarify if this is your question”). This paper proposes a “generative” model for responding to user’s questions by transferring learning from a relevant domain context in conjunction with dialog history between the user and the Q & A system.

#### Problem/Task Definition

Recurrent neural network based conversational agents with enough capacity and access to large data sets still have a tendency to produce generic responses, inconsistent outputs. In addition, these agents have difficulty in factoring more than the last dialogue utterance (the actual words communicated) as part of the response.

#### Summary

This paper utilizes a masked, multi-attention decoder-only transformer (based on Generative Pretrained Transformer). In addition to the user’s question, the model uses persona profile of the user and learned positional embeddings pretrained on the BooksCorpus dataset. The input representation utilizes a combination of utterance embedding, dialog state embedding and positional embeddings. For each user ‘s utterance, the sequence of input tokens for the model is concatenated with all the sentences in the user’s persona profile and the history of the dialog’s 3 -5 previous utterances. A multi-task learning approach is used by optimizing for a combination of next utterance classification loss and Language Modeling loss. Responses were generated using a beam search with sampling.

#### Future Work

Optimal settings and hyper parameters still need to be identified for the current transformer model chosen. It will be interesting to test against other data sets than just the PERSONA-CHAT training set. Additional transformer models can be used for evaluation.

#### References

Authors: Thomas Wolf, Victor Sanh, Julien Chaumond & Clement Delangue, 2019, [TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents](https://arxiv.org/pdf/1901.08149.pdf)

### SpanBERT: Improving Pre-training by Representing and Predicting Spans

#### Motivation

NLP tasks like Domain specific Question Answering may not have enough data. In those scenarios, contextual embeddings obtained from models pre-trained on large corpora may help mitigate issues. The approach outlined in SpanBERT extends the original Bi-directional Encoder Representation for Transformers (BERT) and has better performance especially in selecting answer spans in Question Answering task.

#### Problem/Task Definition

Answers required for NLP tasks frequently may involve reasoning that requires drawing inferences between two or more spans of text. Unlike approaches that propose increasing the size of the data or model size, the pre-training method used in this paper focuses on the training task and objective that is designed to better (a) represent the question and (b) better predict spans of text that will represent answers.

#### Summary

The paper proposes two key differences from BERT for pre-training – masking scheme and training objective. Random contiguous spans are masked (rather than individual tokens) and a “span-boundary” objective” is used. In-lieu of predicting individual masked tokens, the model learns to predict the entire masked span from the observed tokens within its boundary. Also, SpanBERT does not utilize the Next Sentence Prediction (NSP) objective in BERT. Only single segments of text are sampled from the corpus. SpanBERT exceeds BERT F1 score on SQuAD 1.1. and SQuAD 2.0 by 3.3% and 5.4%.

#### Future Work

It will be interesting to observe the performance of this model when masking budget (currently set at 15% of the input tokens) is modified. Also, the span boundary objective can also be modified to utilize information in the preceding token (not just the beginning of the span boundary) for predicting the next token in the span.

#### References

Authors: Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy,2020, TACL, [*SpanBERT: Improving Pre-training by Representing and Predicting Spans*](https://arxiv.org/pdf/1907.10529.pdf)

### Personalizing Dialogue Agents: I have a dog; do you have pets too?

#### Motivation

#### One major issue for conversation systems is their propensity to select the response with greatest likelihood, a consensus response of the humans represented in the training data. Outputs are frequently vague or non-committal and can be wildly inconsistent lack human identity. The paper proposes an encoded persona in distributed embeddings that capture individual characteristics such as background information and speaking style

#### Problem/Task Definition

Chatbots for social dialogue tend to have several problems: their responses are too generic and irrelevant while lacking context and consistent persona. The authors propose using models condition on profile information and interlocutors’ information to make the dialogue more human like.

#### Summary

The paper introduces the PERSONACHAT dataset which consists of crowd-sourced dialogues where each participant plays the part of an assigned persona; while each persona has a word-distinct paraphrase. The authors test various ranking and generative models on the PERSONACHAT dataset, and show that models that have access to their own personas in addition to the state of the dialogue are scored as more consistent by annotators but not as more engaging. They also show that models trained on PERSONACHAT (with or without personas) are more engaging than models trained on dialogue from other resources (movies, Twitter). Hence, making the PERSONACHAT dataset is a useful training set for open ended conversation system.

#### Future Work

Predicting the profiles from a conversation moves chitchat tasks in the direction of goal-directed dialogue, which has metrics for success. We hope that the PERSONACHAT dataset will aid training agents that can ask questions about users’ profiles, remember the answers, and use them naturally in conversation.

#### References

Authors: Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, Jason Weston

#### (Sep 2018), [Personalizing Dialogue Agents: I have a dog, do you have pets too?](https://arxiv.org/pdf/1801.07243.pdf)

#### Comparison of Q & A System

The four papers evaluated address different subtasks. “Attention is all you need” focuses on effectively reducing the training time associated with a large corpus by using the attention mechanism and facilitating the parallelization of the computation required for training. The tasks used for evaluating the performance of the model are machine translation and constituency parsing. The inherent architecture involves both an encoder and a decoder.

The SpanBERT paper focuses on the pre-training for more effectively representing the question and extracting the corresponding answer from a retrieved paragraph. As part of the fine-tuning phase, a Q & A task specific model should be used to predict the appropriate span-start and span-end locations for the response. Essentially it uses only the encoder stack of the transformer albeit with different parameters.

The TransferTransfo paper essentially starts by pre-training a language model on the BooksCorpus dataset but effectively using only the decoder portion of the transformer essentially using Generative Pretrained Transformer-2. As part of the fine-tuning, the language model is modified to produce dialog (response) to user interaction.

The Personalizing Dialogue Agents aims to improve the dialogue agents by providing them with a configurable persistent persona, encoded by multiple sentences of textual description. This profile can be stored  
in a memory-augmented neural network and then used to produce more personal, specific, consistent and engaging responses than a persona-free model, thus alleviating some of the common issues in open ended conversation models

## Q & A with Knowledge Graph Integration

### Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks

#### Motivation

There are lots of research in question answer over knowledge graphs with many complex systems that claim state of the art performance. Simple, strong baselines need to be examined in a rigorous manner and compared to complex systems to quantify performance gain.

#### Problem/Task Definition

Answers Baseline for Q&A with knowledge graph has not been explored adequately and unclear how much Neural Networks techniques actually help. The authors seek to establish strong baseline to objectively quantify the contribution of various DL techniques to many steps of Q&A problem which include entity detection, entity linking, relation prediction and evidence combination.

#### Summary

On SIMPLEQUESTION dataset, the authors find simple LSTMs and GRUs with few common heuristics yield accuracies that comparable with state-of-the-art techniques. They also show non deep learning techniques such as CRF and Logistic Regression perform reasonably well on entity detection and relation prediction. They conclude that some state-of-the-art NN architectures only improve modestly at the cost of significant complexity and heavy technical debt.

#### Future Work

Benefits of neural networks could be obtained with more “complex” systems, therefore isolating their advantages in a controlled manner is desirable. For the task of simple QA over knowledge graphs, the authors suggest to start with simple strong baselines of simple neural network or classical technique then move to more complexed system after adequately examine the baseline.

#### References

#### Authors: Salman Mohammed, Peng Shi, Jimmy Lin (Submitted on June 2018), [Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Network](https://www.aclweb.org/anthology/N18-2047.pdf)s

### HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention

#### Motivation

To exam an inner working of a real-world application of task-oriented QA system with knowledge graph and its system architecture with various NLP components.

#### Problem/Task Definition

A knowledge-based system holds clear advantages in providing targeted responses to well-defined questions and thus is a convenient and reliable approach in implementing a question answering system in knowledge centric domains such as medical fields. However, a knowledge-based system can sometimes be too rigid in a conversational context. The paper’s authors propose a neural network model which provides a more flexible way for various situations where questions are not matched in knowledge-based system.

#### Summary

The authors propose an online question and answer system for medical application. The hybrid system consists of a knowledge graph and a text similarity model to find the most similar question from a large QA dataset using hierarchical BiLSTM attention architecture. The text-similarity model is found to outperform MaLSTM and BERT due to the benefit of its attention layer and its embedding on the specific domain.

#### Future Work

The paper only considered the single-turn question-and-answer mechanism. An important future direction is to add user profiles into the system and provide a more precise and tailored medical assistant to each specific user.

#### References

#### Authors: Qiming Bao, Lin Ni, Jiamou Liu (Submitted on 8 Feb 2020), [HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention](https://arxiv.org/pdf/2002.03140.pdf)

### Semantic Parsing for Single-Relation Question Answering

#### Motivation

TBD.

#### Problem/Task Definition

This research proposes a two-step process: 1. separate a question into a relation pattern and an entity mention, then 2. use a semantic similarity model based on a twin CNN to find the best match of the pattern/mention from the KG

This model first applies a word-hashing layer onto the tri-grams of the input words then feeds them thru a twin (Siamese) CNN-based semantic model with max-pooling to extract the most important local features. (2014).

#### Summary

A new semantic parsing framework based on twin (Siamese) CNN with maxpooling and sub-word hashing ( letter-trigram count vector ) from the input words is used to match relation patterns (predicate) and entity mentions (entity). The model runs on the PARALEX knowlege base.

Due to using the letter-trigram vectors, the model handles the out-of-vocab problem better and also outperforms the BoW representations. Experiments show this framework achieve higher F1 and Recall on the QA task than the previous work PARALEX.

#### Future Work

#### Beside the two suggestions from the authors, which are to replace WikiAnwerss data and ReVerb KB that are used in the experiments with a dedicated entity linking system to improve performance and reduce the number of candidate predicate/entity from the KB and to tackle the problem of multiple-relation questions, we may replace the twin CNN network with a Bi-LSTM encoder+decoder or transformers aiming for higher similarity rate and therefore better F1 rate.

#### References

Authors: Thomas Wolf, Victor Sanh, Julien Chaumond & Clement Delangue, 2019, [Semantic Parsing for Single-Relation Question Answering](https://www.aclweb.org/anthology/P14-2105/)

### Simple Question Answering by Attentive Convolutional Neural Network

#### Motivation

TBD.

#### Problem/Task Definition

Using character-level and word-level convolutional neural networks to improve the accuracy when matching single subject and predicate questions (simple questions) with facts from Freebase.

#### Summary

To improve the simple QA problem, the authors use an effective entity linker ( either passive where longest consecutive common subsequence is applied, or active where a BiLSTM\_CFR is applied ) to find the possible ( mention, pattern) pairs and , and then use a fact selection with one character-level CNN to find the entity that matches the mention, and one word-level CNN with attentive max-pooling to find if a predicate is a paraphrase of the pattern. The model runs on Freebase KG, and trains on SimpleQuestions question set.

Their model shows better performance on the simple QA system as well as the predicate classification.

#### Future Work

#### TBD

#### References

Authors: Thomas Wolf, Victor Sanh, Julien Chaumond & Clement Delangue, 2019, [Simple Question Answering by Attentive Convolutional Neural Network](https://www.aclweb.org/anthology/C16-1164/)

### Knowledge Graph Embedding Based Question Answering

#### Motivation

TBD.

#### Problem/Task Definition

This research suggests using word embedding, neural networks, and a special distant metric function to find the closest KG triple to the relation/mention vectors from the question.

This research proposes a two-step process: 1. separate a question into a relation pattern and an entity mention, then 2. use a semantic similarity model based on a twin CNN to find the best match of the pattern/mention from the KG

#### Summary

Using the embedding word representations for the KG, the research can handle questions that have entities and predicates different from the ones in the training data by finding the closest entity and relation candidates. The embeddings also preserve the sentence structure and relation information which helps better predictions of predicate and entities from the KG. The model uses an attention-based BiLSTM to calculate the representations of the predicate and head entity. Instead of searching all the head entities in the KG which consumes time, a head entity detection model is used to select successive tokens in the question as the name of the head entity. A special distance function is used to find the final triple as the answer for the question. The model runs on two Freebase subsets FB2M, FB5M and the SimpleQuestions question set

- Experiments show this model has better performance than all state-of-the-art method.

#### Future Work

#### Three main factors of this research that increase the QA performance is to use embeddings - Glove in this research - , BiLSTM, and a special distant function to measure the distance between the pattern/mention of the question and the relation/entity of the triples from the KG. Based on these, some future works may be :

#### - change from Glove embedding representation to others such as Bert

#### - replace the BiLSTM with transformers to better have similarity and word paraphrase,

#### - tune the distance function.

#### References

Authors: Thomas Wolf, Victor Sanh, Julien Chaumond & Clement Delangue, 2019, [Knowledge Graph Embedding Based Question Answering](https://dl.acm.org/doi/10.1145/3289600.3290956)

#### Comparison of Q & A with Knowledge Graph Integration

Single-relation( or simple ) question answering is still far from perfect because the predicate of a question ( also called relation pattern ) can be expressed in many different ways, the question 's subject ( also called entity mention ) can also be ambiguity by the context, typo, spaces .... These makes the problem to find the best match of the pair (pattern/mention) from the question to the triple (head, relation, tail ) in the knowledge graph difficult.

The three papers try to solve the same problem: how to find the best answer from the KG when the input question may have may different paraphrases as well as ambiguous subjects. While all three models gain higher performance than their corresponding baseline systems, two of the papers (1 and 3) take into account the case of out-of-vocab situation, and the other (2) does not when it searches at the word level only. This is understandable because its KG is Firebase that had 44 million topics. With this huge size, out-of-vocab seems diminished. Also, the third paper uses word embedding representations and a special distance function while the first two papers work on word similarity that are calculated thru the models.