Ethan Nguyen, Vu Pham, Mohan Rangarajan

Literature 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](#_Toc37009055)

[Introduction 2](#_Toc37009056)

[Motivation 2](#_Toc37009057)

[Problem/Task Definition 3](#_Toc37009058)

[Paper Summary 4](#_Toc37009059)

[Comparison 5](#_Toc37009060)

[Future Work 6](#_Toc37009061)

[References 7](#_Toc37009062)

[Appendix One 8](#_Toc37009063)

[Appendix Two 9](#_Toc37009064)

# 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 i.e. we wanted to 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
2. Integration between Q & A and the knowledge graph (KG) to precisely retrieve responses from the KG

In each category, we have outlined the papers chosen and the motivation behind selecting the papers.



Figure 1 : Initial idea for the project

## Motivation

This section outlines why we chose the different papers as part of our literary review.

Utilizing contextual embeddings is an important factor contributing to the success of many NLP tasks. Pre-training with large corpora, though important in the realm of deep learning based NLP approaches to achieve a good performance, can be time-consuming. The transformer concept initially published in the Attention is All You Need paper outlines a set of approaches to reduce the training time required thereby setting the stage for downstream NLP tasks like Question Answering. Our interest in exploring a domain specific conversational Q & A system led us to SpanBERT: Improving Pre-training by Representing and Predicting Spans paper which extends the original Bi-directional Encoder Representation for Transformers (BERT) paper and yields better performance in selecting answer spans in the Q & A task. User experience during a dialog can be enhanced when the system tries to elicit more clarity on user’s intent by generating a response based on past dialog history. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents proposes a “generative” model for responding to user’s questions. The model uses transferring learning from a relevant domain context in conjunction with dialog history between the user and the Q & A system. A major issue faced by conversation systems is that their outputs are vague, inconsistent and lack human identity. Personalizing Dialogue Agents: I have a dog, do you have pets too? proposes an encoded persona in distributed embeddings that capture individual characteristics such as background information and speaking style.

In general, single-relation or simple question answering is still far from perfect because the predicate of a question can be expressed in many different ways, and the question's subject (entity mention) can also be ambiguous due to lack of context, typos or space. These problems at best hinder the pair matching (pattern/mention) between the question and the expected triple (head, relation, tail) in the KG and may cause the system to return the wrong answer. These four papers [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](https://www.aclweb.org/anthology/P14-2105/), [Simple Question Answering by Attentive Convolutional Neural Network](https://www.aclweb.org/anthology/C16-1164/), and [Knowledge Graph Embedding Based Question Answering](https://dl.acm.org/doi/10.1145/3289600.3290956) present various approaches on how to find the best answer from KG even when the question may have different paraphrases or an ambiguous subject. Furthermore, the HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention paper explores the inner working of a real-world application of task-oriented QA system with KG and its system architecture with various NLP components.

More details about the motivation for selecting some of the individual papers is provided in Appendix One.

## 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. Attention is All You Need examines the parallelization aspect that is especially important when there are a lot of examples and memory limitations that constrain batching across examples.

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. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents attempts to address these issues by generating responses based on past dialogue history and persona profile of the speaker.

Chatbots for social dialogue tend to have similar problems raised by the TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents paper i.e. responses are too generic, irrelevant while lacking context and consistent persona. Authors of Personalizing Dialogue Agents: I have a dog, do you have pets too? propose using model’s condition on profile information and interlocutors’ information to make the dialogue more human like.

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 SpanBERT: Improving Pre-training by Representing and Predicting Spans focuses on the training task and selecting a training objective (span boundary objective) that is designed to better (a) represent the question and (b) better predict spans of text that will represent answers.

Baseline for Q&A with KG has not been explored adequately and it is unclear how much Neural Networks techniques actually help. Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks paper seeks to establish a strong baseline to objectively quantify the contribution of various deep learning techniques to many steps of the Q&A problem. Steps including entity detection, entity linking, relation prediction and evidence combination are evaluated by this paper.

A knowledge-based system holds clear advantages in providing targeted responses to well-defined questions. Such a system provides 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 be too rigid in a conversational context. HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention authors propose a neural network model which provides a more flexible way for various situations where questions are not matched in knowledge-based system.

A simple semantic parser tailored to single-relation questions, powered by advanced semantic similarity models to handle the paraphrase issue is used in Semantic Parsing for Single-Relation Question Answering to find best matching answer from the KG.

Another novel approach, proposed by the authors of Simple Question Answering by Attentive Convolutional Neural Network paper, involves 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.

Research in Knowledge Graph Embedding Based Question Answering proposes using word embedding, neural networks, and a special distance 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 from the KG.

## Paper Summary

Attention is All You Need proposes relying entirely on “attention” mechanisms. By using multi-head attention in conjunction with positional encoding, the paper takes advantage of the order of 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.

TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents 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 are generated using a beam search with sampling.

SpanBERT: Improving Pre-training by Representing and Predicting Spans 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%.

Personalizing Dialogue Agents: I have a dog, do you have pets too? 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. These tests show that models that have access to their own personas and dialog state 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 dialog from other resources (movies, Twitter). Hence PERSONACHAT dataset is a useful training set for open ended conversation system.

On SIMPLEQUESTIONS dataset, Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks finds that simple LSTMs and GRUs with few common heuristics yield accuracies that are comparable with state-of-the-art techniques. They also show non deep learning techniques such as Conditional Random Fields (CRF) and Logistic Regression perform reasonably well on entity detection and relation prediction. They conclude that some state-of-the-art NN architectures only provide modest improvements at the cost of significant complexity and heavy technical debt.

HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention authors propose an online question and answer system for medical application. The hybrid system consists of a KG and uses 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.

Semantic Parsing for Single-Relation Question Answering proposes a new semantic parsing framework based on twin (Siamese) CNN with max-pooling and sub-word hashing. A 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 knowledge base. The letter-trigram vectors enable the model to handle out-of-vocab problem better and also outperform the Bag of Words representations. Experiments show this framework achieves higher F1 and Recall on the QA task than the previous PARALEX work.

To improve the simple QA problem, authors of Simple Question Answering by Attentive Convolutional Neural Network 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. They then use fact selection with two CNNs - a character-level CNN to find the entity that matches the mention and a 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 dataset. Their model shows better performance on the simple-question QA system as well as the predicate classification.

The research proposed in Knowledge Graph Embedding Based Question Answering can handle questions that have entities and predicates different from the ones in the training data. The model uses embedding word representations and a special distance function to find the closest entity and relation candidates from the KG.

The embeddings also preserve the sentence structure and relation information which helps with 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. A special distance function is used to find the final triple as the answer for a question. The model runs on two Freebase subsets FB2M, FB5M and the SIMPLEQUESTIONS dataset. Experiments show this model has better performance than all state-of-the-art methods.

## Comparison

Attention is All You Need paper leverages the encoder/decoder framework in conjunction with attention mechanism and positional encoding to effectively reduce the training time associated with the corpora. SpanBERT: Improving Pre-training by Representing and Predicting Spans provides an approach for effectively identifying the question intent and an approach for locating the answer span by extending the BERT model. The TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents effectively uses the decoder portion of the GPT2 to generate responses dialog. Personalizing Dialogue Agents: I have a dog, do you have pets too? provides an approach for using a configurable, persistent persona to produce a specific, consistent, engaging and personalized responses during a conversation.

The four papers [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](https://www.aclweb.org/anthology/P14-2105/), [Simple Question Answering by Attentive Convolutional Neural Network](https://www.aclweb.org/anthology/C16-1164/), and [Knowledge Graph Embedding Based Question Answering](https://dl.acm.org/doi/10.1145/3289600.3290956)

aim at solving the same problem: how to find the best answer from the KG when the input question may have different paraphrases or ambiguous subjects. While all four models gain higher performance than their corresponding baseline systems, the [Semantic Parsing for Single-Relation Question Answering](https://www.aclweb.org/anthology/P14-2105/) and [Simple Question Answering by Attentive Convolutional Neural Network](https://www.aclweb.org/anthology/C16-1164/) papers take into account the case of out-of-vocab situation, while the other do not when operating at the word level only. This is understandable because the KG Firebase and Quora Duplicate Questions have huge corpus sizes. With these corpus sizes, out-of-vocab problem seems diminished. Also, the [Knowledge Graph Embedding Based Question Answering](https://dl.acm.org/doi/10.1145/3289600.3290956) paper leverages word embedding representations and a special distance function while the others work on word similarity that are calculated through the models.

One difference between [Knowledge Graph Embedding Based Question Answering](https://dl.acm.org/doi/10.1145/3289600.3290956) and HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention papers is how they handle questions whose responses are not in the KG. Knowledge Graph Embedding Based Question Answering solves the problem by comparing the question’s entity and predicates embedding vector to those in KG and selecting the best candidate with distance metric called Joint Distance Metric. The HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention paper also uses similarity model but with Siamese framework and Manhattan distance to compute the sentence level semantic similarity.

More details of the comparison are provided in Appendix Two

## Future Work

Attention is All You Need 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.

Optimal settings and hyper parameters still need to be identified for the current transformer model chosen in TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. 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.

Modifications to the masking budget (currently set at 15% of the input tokens) and the potential impact on performance of the model proposed in SpanBERT: Improving Pre-training by Representing and Predicting Spans is an interesting candidate for further research. Also, the span boundary objective proposed can be further researched to examine the impact of utilizing information in the preceding token (not just the beginning of the span boundary) for predicting the next token in the span.

Looking forward, the authors of Personalizing Dialogue Agents: I have a dog, do you have pets too paper foresee human generated profiles such as PERSONACHAT dataset can be used to predict users’ profiles and guide goal-directed dialogue, which has wide application in business areas. They hope that the dataset will aid training agents that can ask questions about users’ profiles, remember the answers, and tailor them naturally in conversation.

Classical techniques such as random forest or Bayesian approaches can be explored to select baselines for more simple question answering scenarios, in addition to the approaches suggested by the authors of Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks.

Since HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention 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 specific to each user.

#### Authors of Semantic Parsing for Single-Relation Question Answering propose replacing WikiAnswers data and ReVerb KG used in the experiments with a dedicated entity linking system. This will improve performance, reduce the number of candidate predicate/entity from the KG. They also propose addressing the problem of multiple-relation questions. We can also replace the twin CNN network with a Bi-LSTM encoder+decoder or transformers to aim for higher similarity rate and therefore better F1 rate.

The entity linker used in Simple Question Answering by Attentive Convolutional Neural Network can also utilize the character-level neural network (LSTM or CNN) instead of word-level neural network to increase the possible match from the KG. The word-level CNN in the fact selection can also use a transformer to take advantage of keeping the information of the word structure and order in the question.

#### Three main factors in the model proposed by Knowledge Graph Embedding Based Question Answering that increase the Q&A performance are to use Glove Embeddings, BiLSTM and a special distance function to measure the distance between the pattern/mention of the question and the relation/entity of the triples. We can also explore a contextual embedding like BERT (in-lieu of Glove), replace the BiLSTM with Transformers based model. These will help to better handle similarity and paraphrases. Another approach to explore is to tune the distance function.

## References

1. 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), arXiv:1706.03762v5 [cs.CL] 6 Dec 2017
2. 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, rXiv:2002.03140v1 [cs.CL] 8 Feb 2020
3. Xiao Huang, Jingyuan Zhang, Dingcheng Li, Ping Li Cognitive Computing Lab (CCL), Baidu Research, USA, Jan 2019, [Knowledge Graph Embedding Based Question Answering](https://dl.acm.org/doi/10.1145/3289600.3290956), [Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining](https://dl.acm.org/doi/proceedings/10.1145/3289600)January 2019 Pages 105–113
4. 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), arXiv:1801.07243v5 [cs.AI] 25 Sep 2018
5. Wen-tau Yih, Xiaodong He, Christopher Meek, Jun 2014, [Semantic Parsing for Single-Relation Question Answering](https://www.aclweb.org/anthology/P14-2105/), Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 643–648
6. Wenpeng Yin, Mo Yu, Bing Xiang, Bowen Zhou, Hinrich Schütze, Dec 2016, [Simple Question Answering by Attentive Convolutional Neural Network](https://www.aclweb.org/anthology/C16-1164/), Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1746–1756
7. 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), arXiv:1907.10529v3 [cs.CL] 18 Jan 2020
8. 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, Proceedings of NAACL-HLT 2018, pages 291–296
9. 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), arXiv:1901.08149v2 [cs.CL] 4 Feb 2019

## Appendix One

Attention is All You Need 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.

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, the user experience may be enhanced if we are able to elicit more clarity on the user’s intent (with a response like “Let me clarify if this is your question”). TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents 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.

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: Improving Pre-training by Representing and Predicting Spans extends the original Bi-directional Encoder Representation for Transformers (BERT) and has better performance especially in selecting answer spans in Question Answering task.

#### 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. Personalizing Dialogue Agents: I have a dog, do you have pets too? proposes an encoded persona in distributed embeddings that capture individual characteristics such as background information and speaking style

There are lots of research in question answer over KGs with many complex systems that claim state of the art performance. Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks helps provide a rigorous analysis and compares to complex systems to quantify performance gain.

## Appendix Two

### Q & A System

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Title | Attention is all you need | TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents | SpanBERT: Improving Pre-training by Representing and Predicting Spans | Personalizing Dialogue Agents: I have a dog, do you have pets too? |
| Architecture | 6 identical layers of Encoder each comprising a multi-head self-attention and a fully connected feedforward network with a residual connection around each and a normalization layer. Decoder has a similar architecture (but masked multi-head attention and feed forward layer) with an additional multi-head attention for the output of the encoder stack. | 12 layer **decoder only transformer** with masked 12 self-attention heads and positional embedding. Input sentences re pre-processed and tokenized using byte pair encoding vocabulary with 40K merges | A stack of 24 layers similar to a transformer encoder stack. Each layer has an attention layer and a feed-forward network with a 16 heads for the multi-head attention | Ranking model, ranking profile memory model, key-value profile model, seq2seq, generative model |
| Key Concepts | Attention, Positional Encoding | Combination of transfer learning based training scheme and fine-tuning using multi-task objective | Masked Language Model with Contiguous Span Masks and Span Boundary Objective | Persona stored in a memory-augmented neural network |
| Training Data Set | English German (4.5M Sentence Pairs, 37K tokens/English French (36M Sentences,25K tokens) and Berkeley Parser 17M sentene=ces/32K tokens and WSJ portion of Penn Tree Bank (40K sentences/16K tokens | Pretraining using BooksCorpus dataset and FineTuning with PERSONA-CHAT dataset using augmented input representation | BooksCorpus and English Wikipedia using cased Wordpiece tokens | PERSONA-CHAT dataset of 1155 possible personas, 162,064 utterances over 10,907 dialogs,15,602 utterances (1000 dialogs) of which are setaside for validation, and 15,024 utterances |
| Performance Measured on Data Set | English -to-German, English to French | PERSONA-CHAT test | SQuaAD 1.1, SQuAD 2.0, Five MRQA Tasks | PERSONA-CHAT test dataset (968 dialogs) |
| Performance Metric | BLEU score, Training Cost (in FLOPs) | Perplexity, Hits@1, F1 | Exact Match, F1 Score | Fluency, Engagingness, Consistency, Persona Detection |
| Performance Value | Base Model (E--> G, 27.3, E--> F , 38.1, 3.3 X 10^18) Big Model (E-->G, 28.4, E--F, 41.8,2.3X10^19) | 16.28, 80.7, 19.5 | 1.1 - (EM 88.8, F1 94.6) 2.0 - (EM -85.7, 88.7) MRQA - Avg (F1 -81.5)) | Fluency (4.31), Engagingness (4.25), Consistency (4.26), Persona Detection (.95) |

### Q & A System with Knowledge graph

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Title | HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi Directional Attention | Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks | Semantic Parsing for Single-Relation Question Answering | Simple Question Answering by Attentive Convolutional Neural Network | Knowledge Graph Embedding Based Questioning Answering |
| Architecture | Hierarchical BiLSTM with attention (The hyperparameters include batch\_size of 1024, the n\_epoch of 9, the n\_hidden of 100, the embedding\_dim of 300 and the max\_seq\_length of 10) | BiLSTM, BiGRU (hidden size of 300 and .3 dropout rate), CRF, CNN (300 output channel and .5 dropout rate), Logistic Regression | Convolutional NN | Entity Linker, Convolutional NN | Bi LSTM |
| Key Concepts | Attention, Siamese framework and Manhattan distance | Baseline, Entity Detection, Entity Linking, Relation Prediction | Letter-trigram vectors for word-hasing, CNN-base semantic model | Passive/Active Entity Linkers, character-level CNN for mentions, and word-level CNN with attentive max-poling for relations | word embeddings, Head Entity Detection Model, Joint Distance Metric |
| Training Data Set | Quora medical questions dataset | SIMPLEQUESTIONS dataset with of 75.9k/10.8k/21.7k training/validation/test questions with a (subject, predicate, object) triple from a Freebase | PARALEX, train(1.19M pairs of patterns and relations), validation(12K pairs) | SimpleQuestions: train(75,910), dev(10,845), FB2M, FB5M | SimpleQuestions: train(75,910), dev(10,845), FB2M, FB5M |
| Performance Measured on Data Set | SQuAD v1.1 and MNLI, medical questions-and-answer pairs (171 from ehealthforumQAs, 5679 from questionDoctorQAs and 23437 from webmdQAs) | SIMPLEQUESTIONS test dataset | test questions from PARALEX, answers from ReVerb | SimpleQuestions: test(21,687) FB2M,FB5M | SimpleQuestions: test(21,687) FB2M,FB5M |
| Performance Metric | Average Answer Accuracy Score | F1 Score, Answer accuracy | F1, Precision, Recall, MAP | Accuracy | Accuracy |
| Performance Value | 81.3% for ehealthforumQAs. 80.9% for questionDoctorQAs and 81.2% for webmdQAs | 91% F1 score for Entity Linking with BiLSTM. 82.8% F1 socre for Relation prediction with CNN. 74.9 % for answer accuracy with biLSTM | F1 0.57, P 0.58, R 0.57, MAP 0.28 | passive linker + attentive max-pooling CNN: 68.3(FB2M), 67.2(FB5M) active linker + attentive max-pooling CNN: 76.4 (FB2M), 75.9(FB2M) passive linker + traditional maxpooling CNN: 67.5(FB2M), 66.6(FB5M) active linker + traditional maxpooling CNN: 75.4(FB2M) 74.6(FB5M) | 0.754 (FB2M) 0.749(FB5M) |