Question-Answering on SQUAD 2.0

CSE 676: DEEP LEARNING

PROJECT DESCRIPTION

Addressing the challenge of question answering (QA) in the realm of natural language processing (NLP), our focus lies on the Stanford Question Answering Dataset (SQuAD) Version 2.0. Our objective revolves around identifying answer spans within an extensive collection of Wikipedia articles in response to provided questions. QA systems serve as pivotal tools in information retrieval and comprehension, with wide-ranging applications such as virtual assistants, search engines, and educational aids. By harnessing machine comprehension techniques, our aim is to elevate the precision and efficiency of QA systems, thereby contributing to broader enhancements in NLP . Our methodology entails training end-to-end systems adept at handling the QA task. We leverage cutting-edge methodologies in machine comprehension and deep learning to extract pertinent answer spans from text passages. Users can get answers from massive amounts of text data, while these systems also form the foundation for chatbots and virtual assistants that can have natural conversations. By developing these techniques, you're not only improving information access but also contributing to advancements in NLP with applications like machine translation and sentiment analysis. The work on this problem is important because it contributes to developing AI systems that can understand and respond to human language more effectively, with potential benefits in various fields.

Background

In the landscape of natural language processing (NLP) and question answering, our project intersects with key works such as DrQA, BiDAF, and QANet. DrQA stands out for its holistic approach to open-domain question answering, integrating information retrieval from sources like Wikipedia with reading comprehension. BiDAF addresses the issue of premature summarization in attention-based models, proposing a hierarchical network that computes attention for every time step to prevent information loss. QANet, on the other hand, breaks away from recurrent neural networks (RNNs), relying solely on self-attention and convolutions to boost training speed and inference efficiency, inspired by the "Attention is all you need" paradigm shift.

In comparison to traditional approaches, our project adopts a multi-faceted strategy. While DrQA and BiDAF inform our understanding of leveraging external knowledge sources and optimizing attention mechanisms, respectively, our approach shares common ground with QANet in embracing self-attention but diverges in architecture and implementation details. We aim to harness the efficiency of QANet's self-attention mechanisms while possibly incorporating elements from DrQA and BiDAF to enhance comprehension and performance in question answering. Thus, our project amalgamates insights from various seminal works while charting its own course toward advancing the frontier of question answering systems.Our approach diverges from traditional methods by avoiding the use of seq2seq and encoder-decoder models for question answering.

Preprocessing

- Our approach provides a straightforward manual pipeline for QA systems.
- We rely on Spacy for tokenization, eliminating the need for additional libraries.
- Our pipeline covers all preprocessing tasks, from loading and parsing dataset files to filtering and numericalizing text.
- We filter large examples based on sequence lengths to optimize processing.
- Gathering text for vocabulary creation and mapping words to IDs are integrated into our pipeline.

Dataset

- We have used SQuAD2.0 dataset for implementing our models.

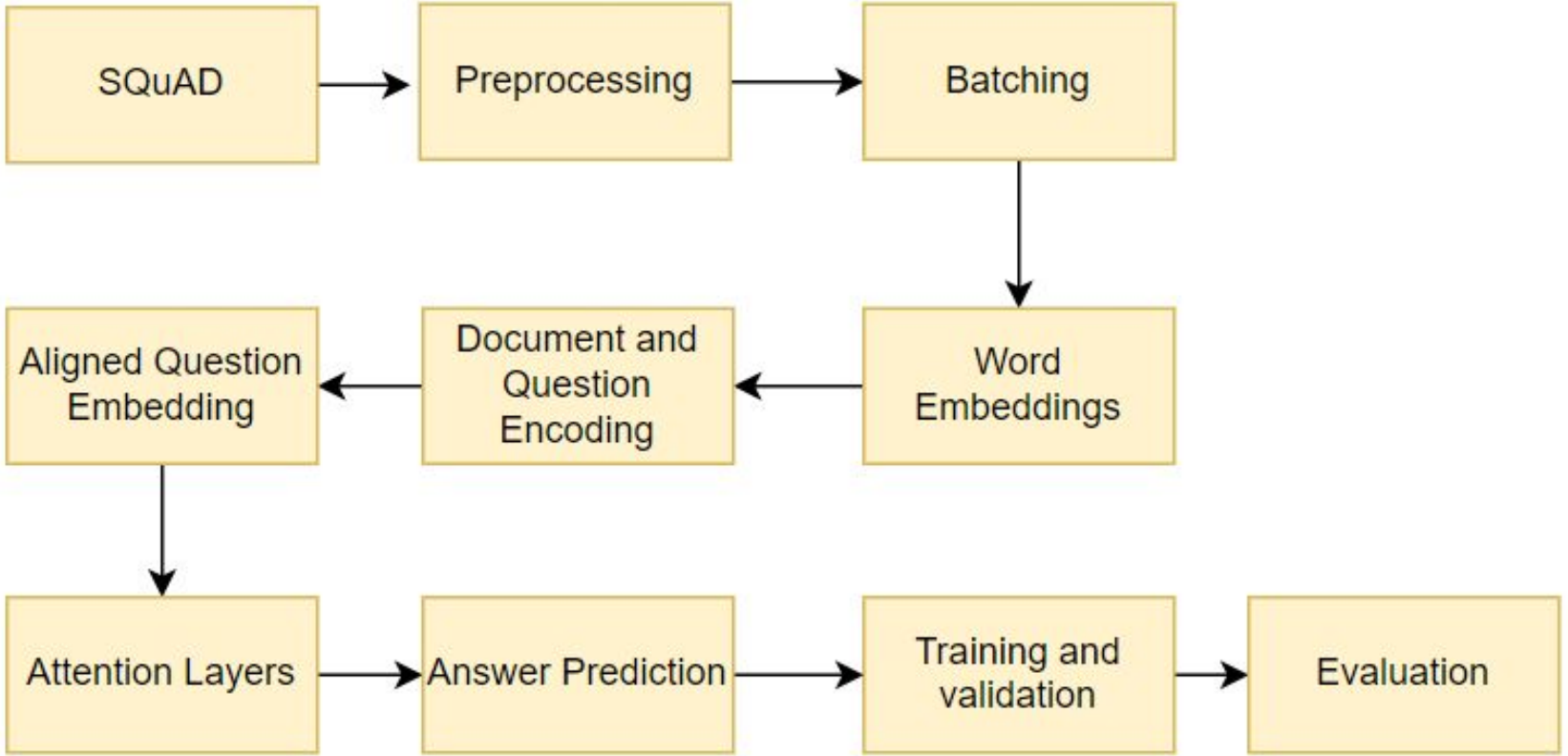
What is SQuAD?

- SQuAD is a dataset for reading comprehension. Crowdworkers ask questions about Wikipedia articles, and the answer to each question is a piece of text from the article. Sometimes, questions might not have an answer.

SQuAD Version 2 Enhanced Features Over Version 1:

- SQuAD2.0 merges 100,000 questions from SQuAD1.1 with 50,000 challenging questions created by crowdworkers to resemble answerable ones.
- SQuAD 2.0 acts as a standard for testing question answering models, tracking NLP advancements.

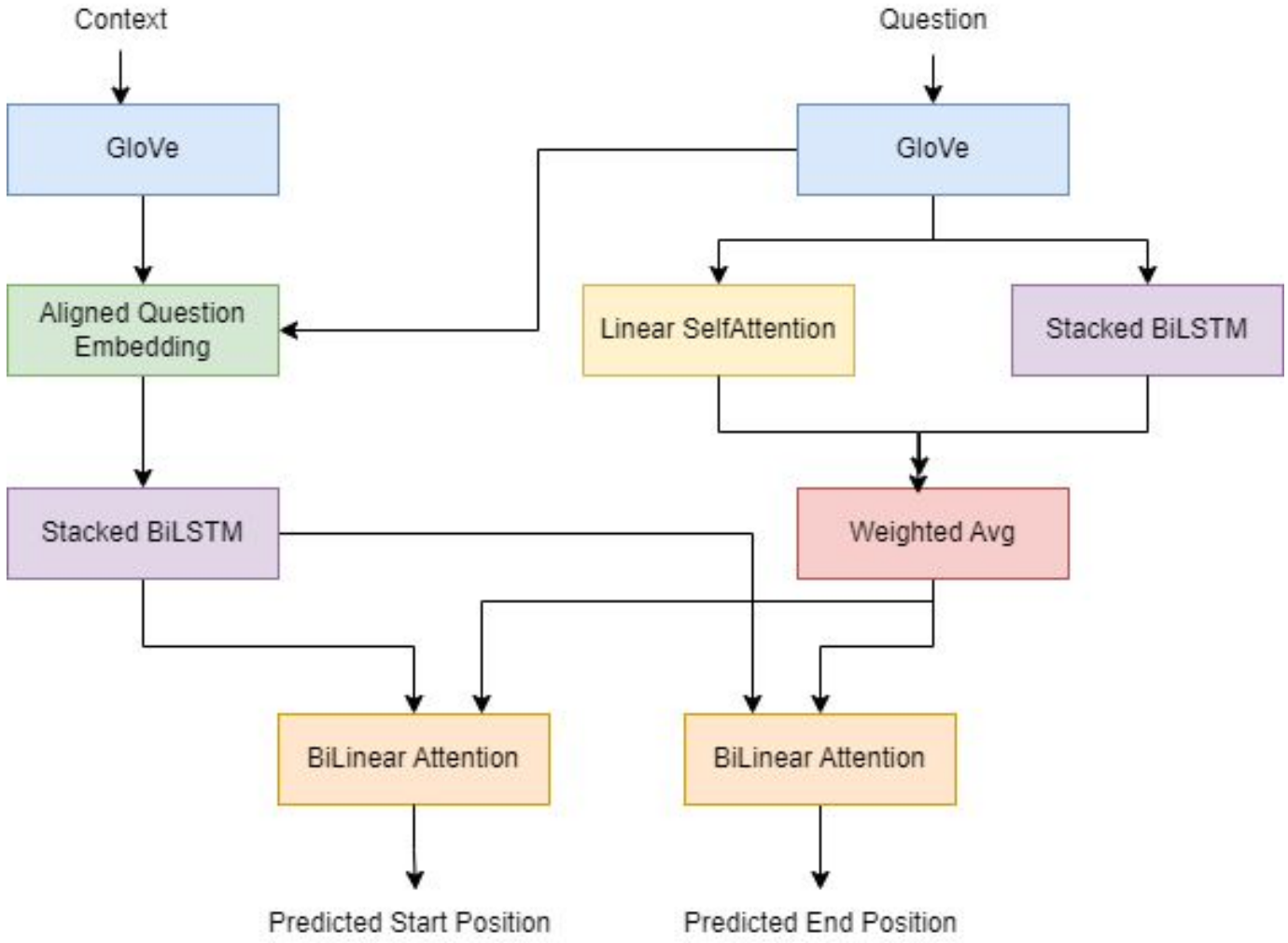
Model 1 - DRQA Overview



Model 1 - DRQA Architecture

.

- **GloVe:** Pre-trained word vectors that capture semantic meaning.
- **Stacked BiLSTM:** Captures meaning from both directions (past & future) in the context and question.
- **Aligned Question Embedding:** Refines the question's meaning to better align with the context using attention.
- **Linear Self-Attention & weighted avg (Weighted Average):** Assigns importance scores to question words based on their relevance to the context.
- **BiLinear Attention:** Calculates attention scores between each context word and every question word.
- **Predicted Start Position and End Position:** Identifies the most likely word where the answer begins and ends in the context.



Model 2- BiDAF

BiDAF, or Bidirectional Attention Flow for Machine Comprehension, tackles the challenge of premature summarization inherent in previous models employing attention mechanisms. Unlike earlier approaches that summarized input values and queries into fixed-size vectors, BiDAF introduces a novel approach. Instead of condensing the context paragraph prematurely, the attention is computed dynamically for each time step. This ensures that information loss is minimized, addressing a significant limitation of prior methods.

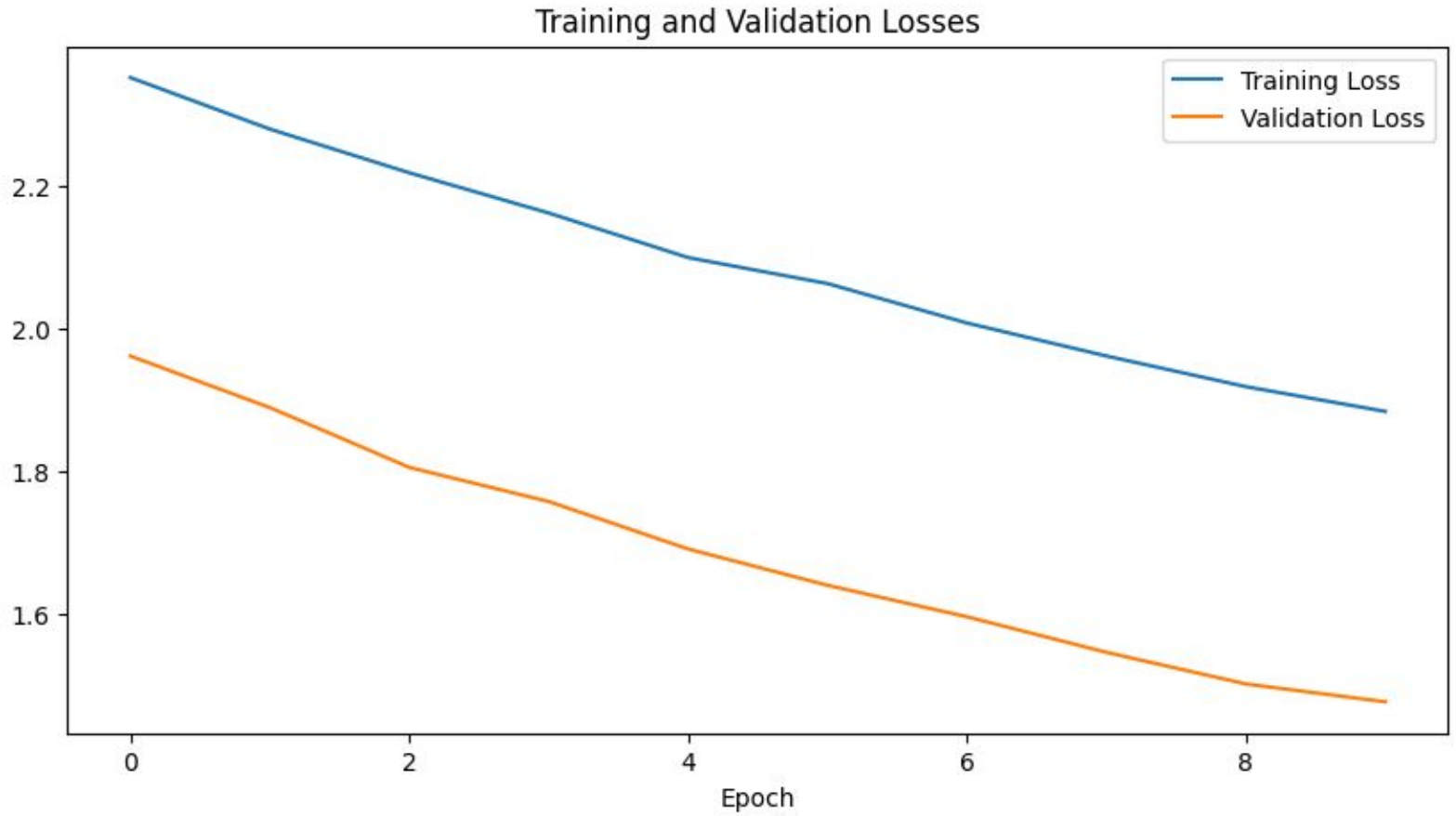
Additionally, the model integrates bidirectional attention, allowing information flow in both directions, which enhances comprehension. The proposed hierarchical, multi-stage network facilitates the flow of attended vectors and representations from previous layers to subsequent modeling layers, optimizing information retention and utilization throughout the network.

Model 3- QANet

- Prioritizes self-attention and convolutions for faster training and better performance.
- Dissects text, capturing details and context swiftly with self-attention and convolution techniques.
- These convolution techniques guarantee efficient calculation, better accuracy and faster insights.
- Simplifies complex text, facilitating smooth human-machine communication.
- Evolves constantly, integrating effective strategies like "Highway Networks" for improved performance.

Model 1 - DRQA Results

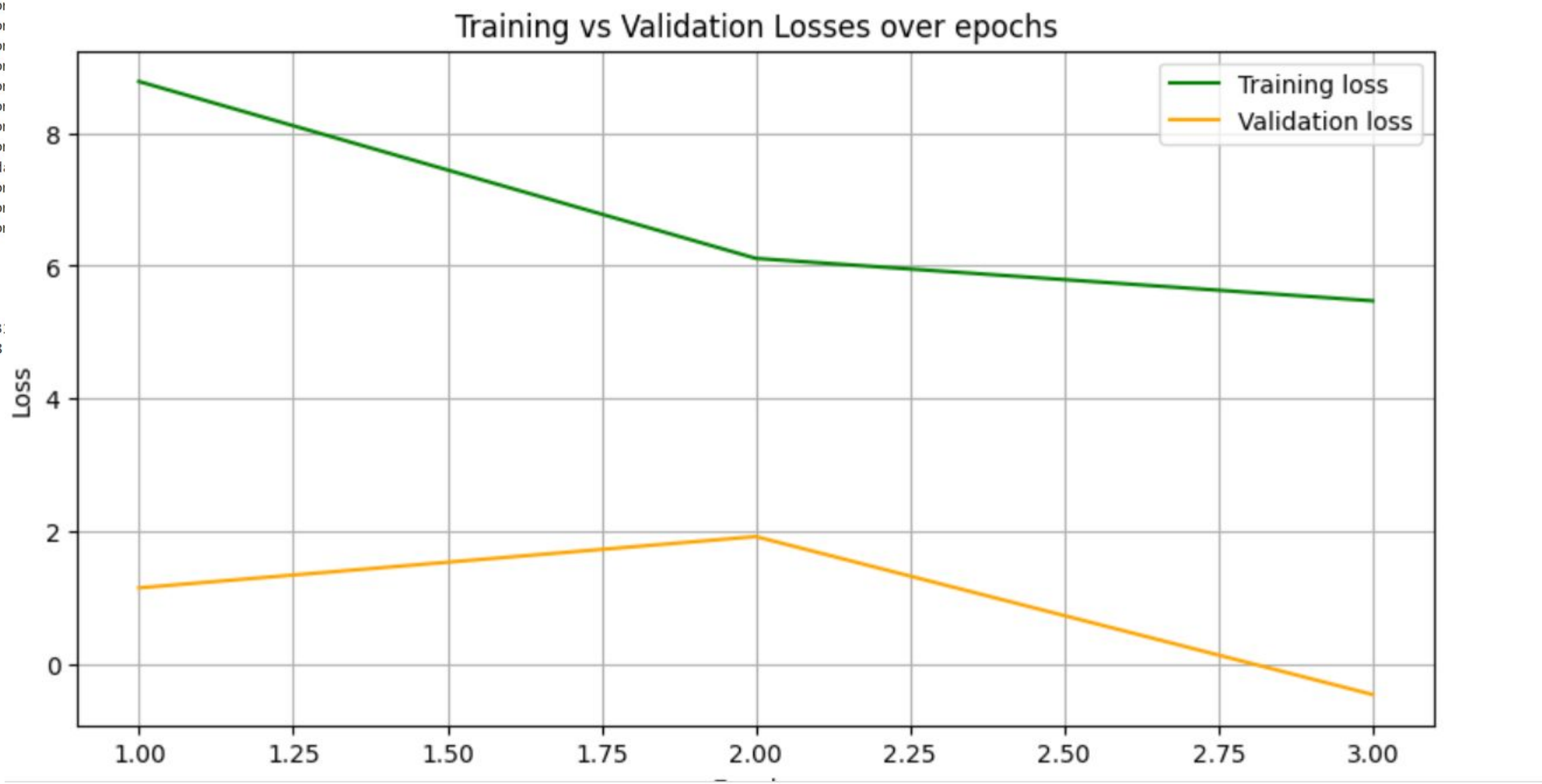
```
Loss on Training Set : 2.0621848696436964
Loss on Validation Set : 1.6397164655963803
Resultant EM Value : 36.58238343274687
Resultant F1 Score: 36.47555231009588
```



```
or Epoch  1
aining Phase


tch Starting from: 0
tch Starting from: 500
tch Starting from: 1000
tch Starting from: 1500
tch Starting from: 2000
tch Starting from: 2500
tch Starting from: 3000
tch Starting from: 3500
tch Starting from: 4000
tch Starting from: 4500
tch Starting from: 5000
lidation Phase


tch starting from : 0
tch starting from : 500
tch starting from : 1000
ss on Training data : 1.858750248920984
ss on Validation data : 2.9545606068703707
 Score obtained is  45.265602538747075
lue of EM(Excat Match score) obtained is  37.196159353154215
or Epoch  2
```

```
Loss on Training data : 1.6180625340514565
Loss on Validation data : 2.825645291327529
F1 Score obtained is  46.160237242102056
Value of EM(Excat Match score) obtained is  38.122631179988204

Loss on Training data : 1.4139476219119307
Loss on Validation data : 2.911216235061249
F1 Score obtained is  46.103087643847154
Value of EM(Excat Match score) obtained is  38.20685589151857

Loss on Training data : 1.2398962687263944
Loss on Validation data : 2.983005181066366
F1 Score obtained is  46.604556201513475
Value of EM(Excat Match score) obtained is  38.81327381453718
```

```
Epoch 1
Entered Training phase
/usr/local/lib/python3.10/dist-packages/spacy/pipeline/lemmatizer.py:211: UserWarning: [W108] The rule-based lemmatizer did not find POS annotation fo
  warnings.warn(Warnings.W108)
Starting batch from index: 0
Starting batch from index: 500
Starting batch from index: 1000
Starting batch fro
Starting batch fro
Starting batch fro
Starting batch fro
Starting batch fro
Starting batch fro
Starting batch fro
Starting batch fro
Entered into Valid
Starting batch fro
Starting batch fro
Starting batch fro
Training:
  Loss: 8.8131
Validation:
  Loss: 1.1943
  Exact Match: 5.3
  F1 Score: 9.8498
```



Summary

Our project enhances question answering using SQuAD 2.0. By leveraging advanced machine comprehension and deep learning, we extract precise answers from Wikipedia articles. This improves information retrieval and lays the foundation for virtual assistants and search engines. Our work advances NLP, enhancing AI's ability to understand and respond to human language effectively. We have got the best results for the model 2 Following are the metrics:

```
Loss on Training data : 1.2398962687263944
Loss on Validation data : 2.983005181066366
F1 Score obtained is  46.604556201513475
Value of EM(Excat Match score) obtained is  38.81327381453718
```

THANK YOU