Project Report
Answer Sentence Selection (AS2)

Team Number: 12
Team Members:
- **Aditya Kondai**
- **Anantha Yadavalli**
- **Yashwanth Kottu**
- **Durga Koppisetti**

In this report, we will get to know about Answer Sentence Selection and the methodologies involved with it

Problem Statement :

The task of **Answer Sentence Selection (AS2)** can be formalised as follows:
given a question q and a set of answer sentence candidates C = {c1, c2, ..., cn },
assign a score si for each candidate ci such that the sentence receiving the highest score is the one that most likely contains the answer, i.e., the answer sentence. It is interesting to note that AS2 can be, in fact, modelled as a re-ranking task.

Although re-ranking is a structured output problem, most **state-of-the-art** approaches treat it as **point-wise classification**, i.e., classifying answer sentences as positive and all the others as negative.

The papers solve the AS2 problem using binary classification

Dataset

Dataset: Wiki-QA

Main Components/ Columns of Wiki-QA

Question - English Sentence representing the Question
Candidate - English Sentence representing the candidate for the question
Label - 0/1 denoting the whether this candidate is the answer to the question or not
● 1 denotes Answer
● 0 denotes not an Answer

Metrics :

MAP and MRR metrics are used in the answer-selection task to evaluate models and methods. These measures show the rating quality of candidate answers. The MRR measure

only considers the rank of the first relevant answer, but the MAP measure considers the order of all relevant answers (Manning et al., 2008). These two measures are shown below.

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk})$$

$$MRR(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} r_j$$

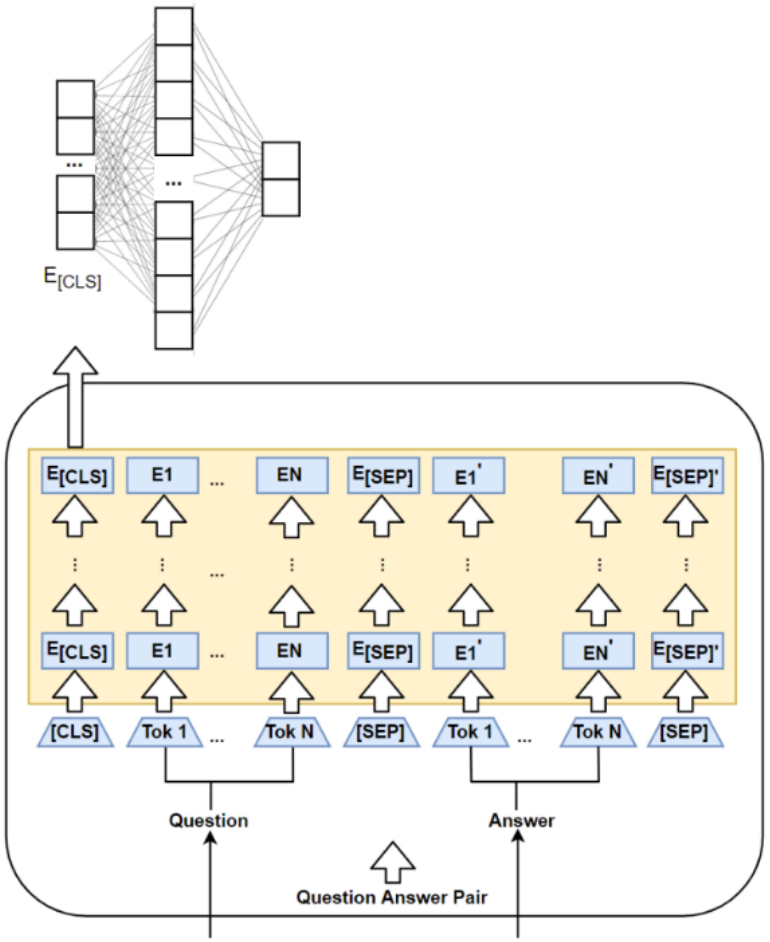
In these equations, Q is the set of questions, mj is the number of relevant answers to qj, Rjk is a list of candidate answers that contains top k relevant answers, Precision function is a function that measures the ratio of the number of relevant answers to the total candidate answers, rj is the inverse of the first relevant answer rank for qj.

For Example :
Assume A question has 7 candidates, and after prediction the order of candidates after sorting is

Predicted Order *:* C2 , C4 , C1 , C7, C5, C3, C6
Real Answer Candidates *:* C1, C3, C4
So AP for this Question is (½ (C4) + ⅓ (C1) + ⅙ (C3)) * ⅓
And Inverse Rank is ½ as C4 is the nearest correct answer and its rank is 2
MAP and **MRR** are average for all questions

Papers

BERT Baseline Model :
In this method, the classification method proposed by Devlin et al. (Devlin et al., 2019) is employed. That is, the output of the [CLS] token, a vector of length 768, is passed as input to a fully connected neural network with a hidden layer of length 256. The output layer of the fully connected neural network consists of two elements the first indicates the correctness of the answer candidate, and the latter indicates the incorrectness of the answer candidate.



Bert Baseline	MAP	MRR
Train	78.8	72
Dev	73.6	70.7
Test	70.3	63.9

A Study on Efficiency, Accuracy and Document Structure for Answer Sentence Selection

This paper explains that the AS2 model should consider the below two points

- Lexical Overlap
- Global Structure

Lexical Overlap:
One of the most reliable features in AS2 datasets is the lexical overlap, i.e., whether words appear in both questions and answer candidates

We used the number of unique words in both the question and the candidates as a single feature, which is called word-overlap WO, to rank question-answer pairs

Global Structure:

Another relevant feature for the AS2 datasets is the global structure present in the original rank. The structure of the document provides an essential signal for AS2.

For Example, in the case of WikiQA, SQuAD, and Natural Questions, there is a high chance that the answer is contained in the first sentence/paragraph
we believe that there is an intrinsic correlation between the real-world distribution of questions and the structure of the Wikipedia document: encyclopaedic knowledge is usually organised such that more general information about a topic is summarised and organised at the beginning of the document.

In contrast, the signal is less present in other datasets, such as SQuAD, where annotators are asked to write questions after reading the whole paragraph, i.e., they target each part of the text by construction Thus, the answer distribution is less skewed

This Paper Addresses these two features of As2 by building **Cosinet** for **Lexical Overlap**, and **Recursive Network** to capture **Global Structure**.

Cosinet Network :
The Cosinet has three building blocks:
● a word-relatedness encoder that performs the cosine similarity between the word embeddings in the question and the answer (generating word relatedness features);
● similar to (Severyn and Moschitti, 2016), the relational features are concatenated to the word embeddings and given in input to one layer of CNN, to create a representation for the question and candidate pair; and the information of the question and the candidate is combined at classification stage, by concatenating the vectors. That is, we use the component-wise multiplication and difference between a question and answer vectors.

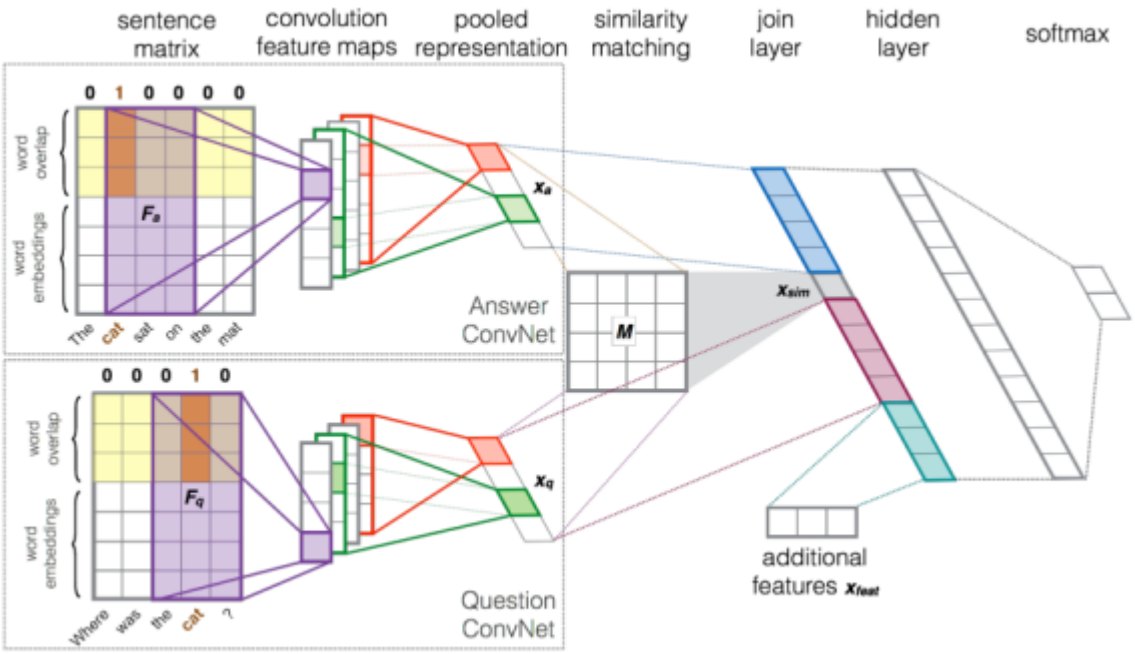
World-Relatedness Encoder :

- As mentioned in the paper, word relatedness encoder is used to capture the word overlap between the words of a question and all its candidates.
- To encode the word-relatedness information, we first map the words in the question and the answer to their respective word embeddings. We then perform comparisons between all the embeddings in the question w_{i}^{q} and all the embedding of the answer w_{j}^{c } using the cosine similarity

$$r_{i,j} = (w_i^q \cdot w_j^c)/(\|w_i^q\| \cdot \left\|w_j^c\right\|)$$

- For every word in the question we find which word in answer has the maximum cosine similarity, and we are gonna append that value to the 300 dimensional number-batch embeddings of that word in the question.
- Similarly for every word in answer, we find the maximum cosine similarity and will append that to the word-relatedness encoder.
- Note that after completing this process, the question embedding will not be the same for each candidate. It will be dependent on the candidate we are referring to, making the intuition of the question candidate pair.

Question-Candidate Encoder:
we encode the question and candidate independently using two single layers of CNN with a kernel size of 5 and global max pooling, producing two embeddings for the question qe and the candidate ce . These are then combined using their pointwise multiplication and their difference, i.e., qce = [qe ce ; qe − ce].



Global Optimization :

We use a Recurrent Neural Network applied on top of the qce representations for each ci of a given question q, to leverage the global structure of the rank. The resulting contextual representations are passed to the feed-forward network

MAP & MRR on WikiQA

Paper-1	MAP	MRR
RNN	66.08	60.17
BiRNN	66.3	60.20
LSTM	66.9	61.23
BiLSTM	67.4	61.69

Final Results:

Models	MAP	MRR
Baseline	70.3	63.9
RNN	60.08	60.17
BiRNN	66.3	60.20
LSTM	66.9	61.23
BiLSTM	67.4	61.69

Depending on the configuration, the model required roughly 1M parameters and outperformed earlier research that did not employ computationally costly pre-trained language models. On the WikiQA dataset, our model trains in 9.5 seconds, which is quite quick compared to the usual BERT-based fine-time tuning's of around 18 minutes.