Reasoning Systems

Introduction

OBJECTIVE

- Addresses one of the major challenge in AI i.e. to describe long term dependencies in data
- To develop a system that can generate conclusion from available knowledge
- In terms of question answering tasks

TASK 1

→ Use of end to end memory networks to simulate the reasoning tasks on bABI dataset

→ Different experiments on bABI dataset with different parameters.

TASK 2

→ Fine tuning BERT model for question answering tasks on bABI dataset as well as real word dataset.

TASK 1

END TO END MEMORY NETWORK

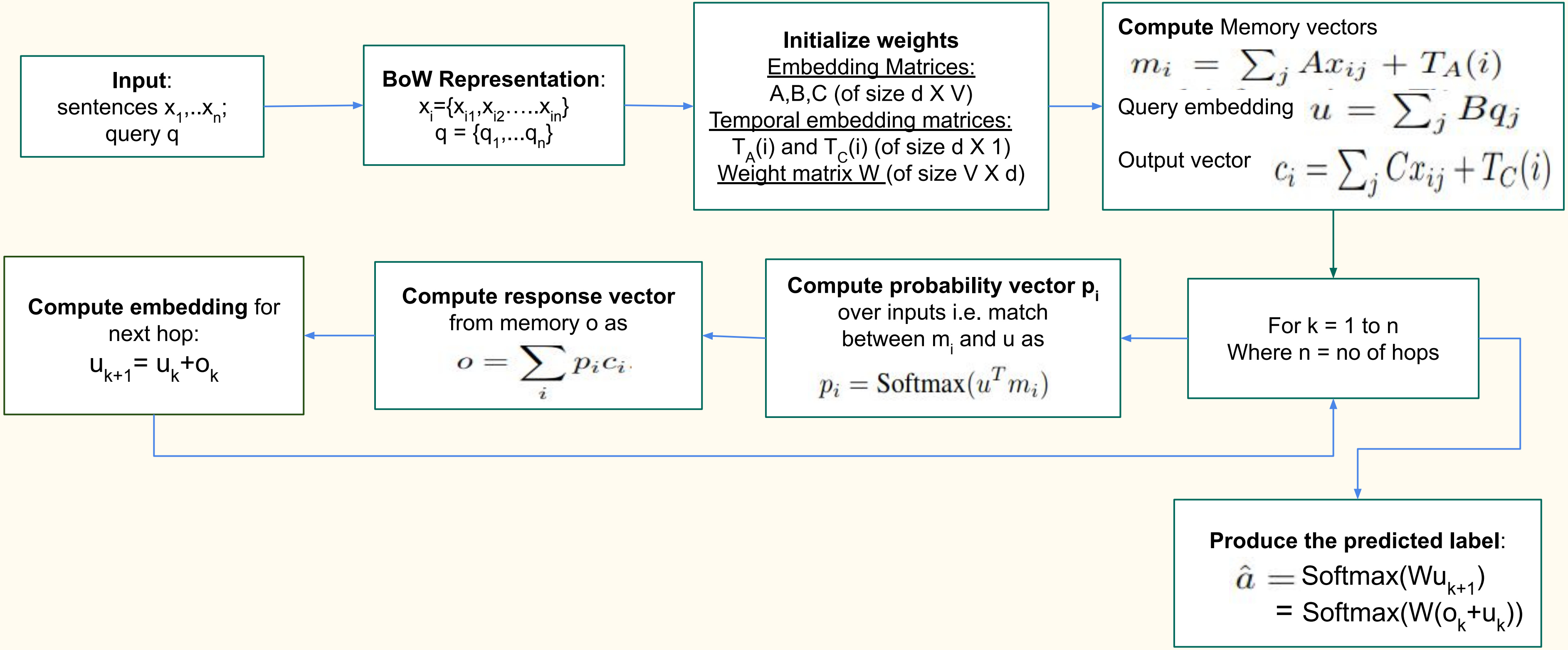
bABI Dataset

→ QA tasks related to bABI project of Facebook AI Research designed for testing text understanding and reasoning

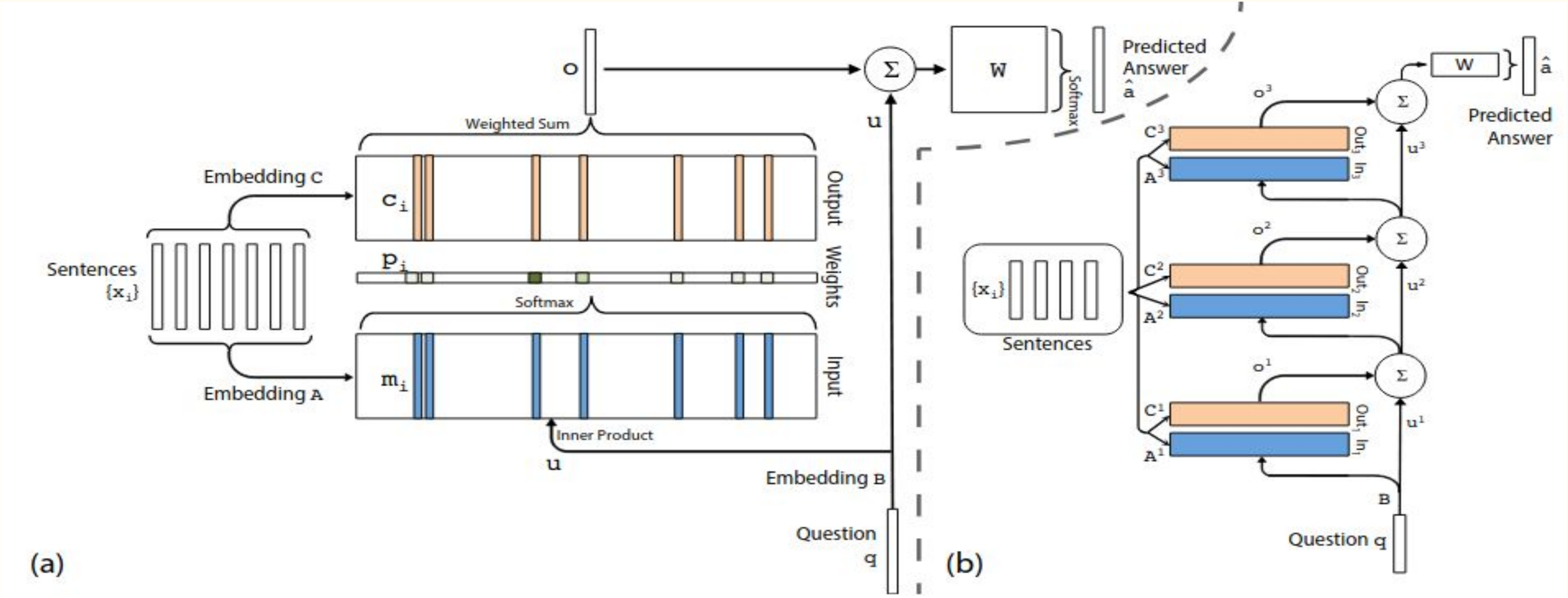
→ A given QA task consists of a set of statements, followed by a question whose answer is typically a single word which is available to the model at training time, but must be predicted at test time

Sam walks into the kitchen.	Brian is a lion.	Mary journeyed to the den.
Sam picks up an apple.	Julius is a lion.	Mary went back to the kitchen.
Sam walks into the bedroom.	Julius is white.	John journeyed to the bedroom.
Sam drops the apple.	Bernhard is green.	Mary discarded the milk.
Q: Where is the apple?	Q: What color is Brian?	Q: Where was the milk before the den?
A. Bedroom	A. White	A. Hallway

End to end Memory Network – Flowchart



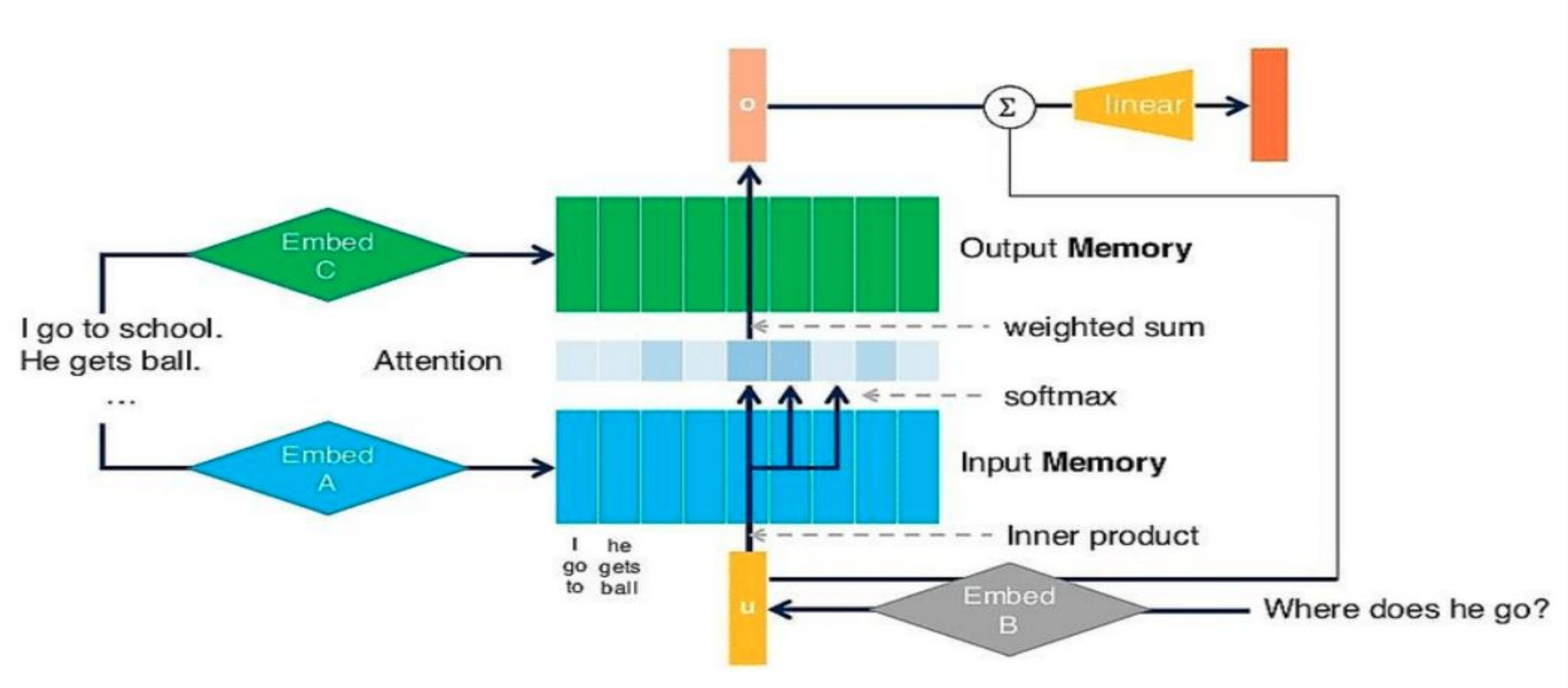
End to end Memory Network – Architecture



(a): A single layer version of our model. (b): A three layer version of our model.

Ref: https://arxiv.org/pdf/1503.08895.pdf

End to end Memory Network – Example



Ref: https://www.researchgate.net/figure/A-high-level-overview-of-end-to-end-memory-networks-in-the-context-of-question-answering_fig4_313374979

End to end Memory Network

TRAINING DETAILS

- Initial learning rate eta = 0.01 which anneals every 25 epochs by eta/2
- Weights initialized randomly with Gaussian distribution N ~ (0,0.1)
- Memory dimension d = 100 for joint training and d = 20 for independent training

OBSERVATIONS AND RESULTS

→ As training rate is decreased, accuracy keeps on improving.

→ Increasing number of hops and memory dimension d, improves accuracy over joint tasks.

Accuracy on joint training	n=3	n=6
d = 20	0.67	0.66
d = 50	0.69	0.67
d = 100	0.65	0.71

TASK 2

FINE TUNED BERT

BERT

- BERT - Language representation model that stands for Bidirectional Encoder Representation from Transformers
- designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context
- can be fine-tuned to create state-of-the-art models for a wide range of tasks
- alleviates unidirectionality constraint by using "masked language model" (MLM) that randomly masks some tokens from input, and then predict the masked word based only on its context

<u>BERT ARCHITECTURE</u>

- Multi-layer bidirectional transformer encoder where transformer is built using stacked self-attention and position wise FFN as

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

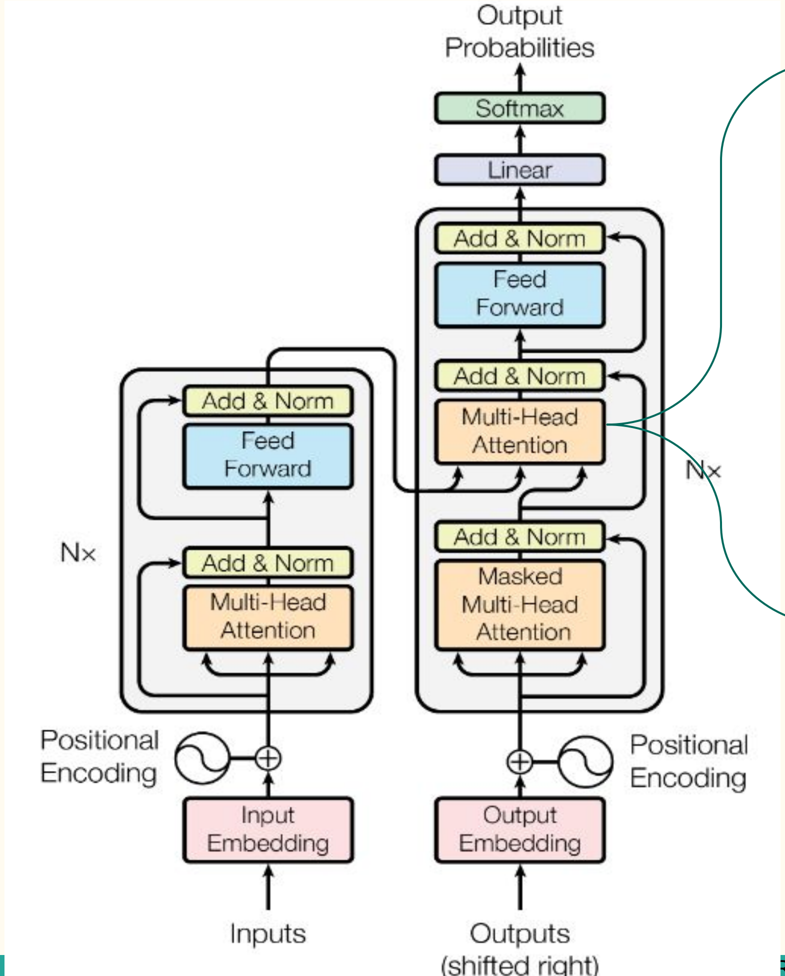
- Attention function maps query and key-value pairs to an output i.e. weighted sum assigned by compatibility function

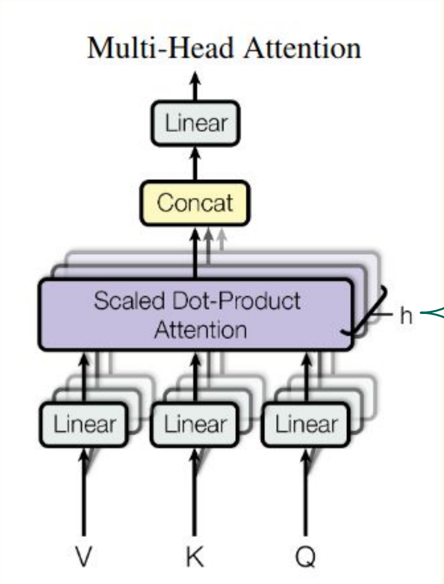
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

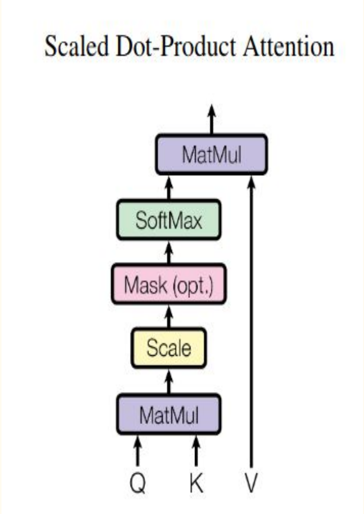
- Multi-head attention allows model to jointly attend information at different positions

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

TRANSFORMER ARCHITECTURE

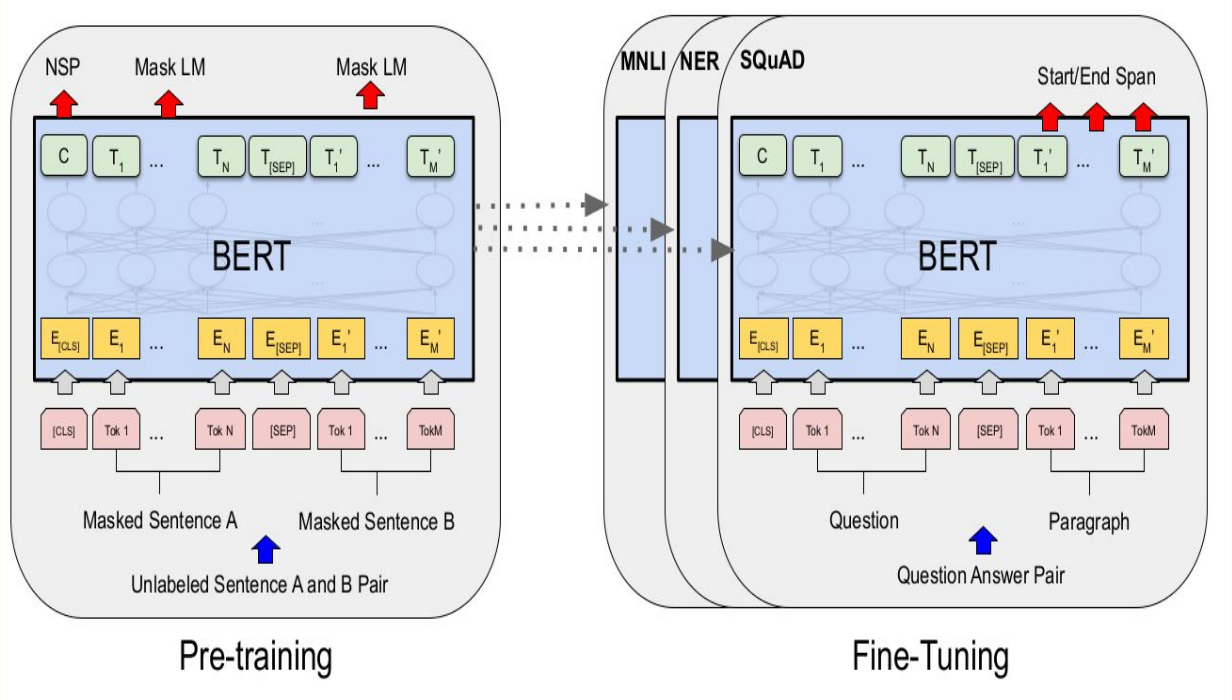






Ref: https://arxiv.org/pdf/1706.03762.pdf

BERT ARCHITECTURE



- Pre-trained with MLM (Masked Language Model) and NSP (Next Sentence Prediction)
- Bidirectional training helps to work in multiple context i.e. from left-to-right and right-to-left
- Self-attention mechanism allows BERT to model many downstream tasks by swapping out appropriate inputs and outputs

Ref: https://arxiv.org/pdf/1810.04805.pdf

BERT FINE-TUNING

METHODOLOGY

- Input question and text paragraph for context reference.
- Convert input to features in form of token encodings, positional and segment encodings along with padding information.
- Load data in form of batches using DataLoader.
- Predict answer for the question using features computed in previous steps.

OBSERVATIONS AND RESULTS

→ Fine tuned BERT performed much better than end to end memory network on bABI dataset with an accuracy ~0.80.

→ BERT performed really better on even real time questions given the context as

```
Enter story: Hey! I am Rhea. I love dancing. I am a resident of Singapore. I have one brother and one sister. My father is a busine
ssman. My mother is a teacher. I love my family.
Enter question: What is Rhea's hobby?
Answer is  dancing
Ask more: y/n y
Enter question: What is Rhea's mother occupation?
Answer is  teacher
Ask more: y/n y
Enter question: What do Rhea's father do?
Answer is  businessman
Ask more: y/n y
Enter question: How many siblings she has?
Answer is  one brother and one sister
Ask more: y/n n
```

Thank you

Mehak
2018MCS2143