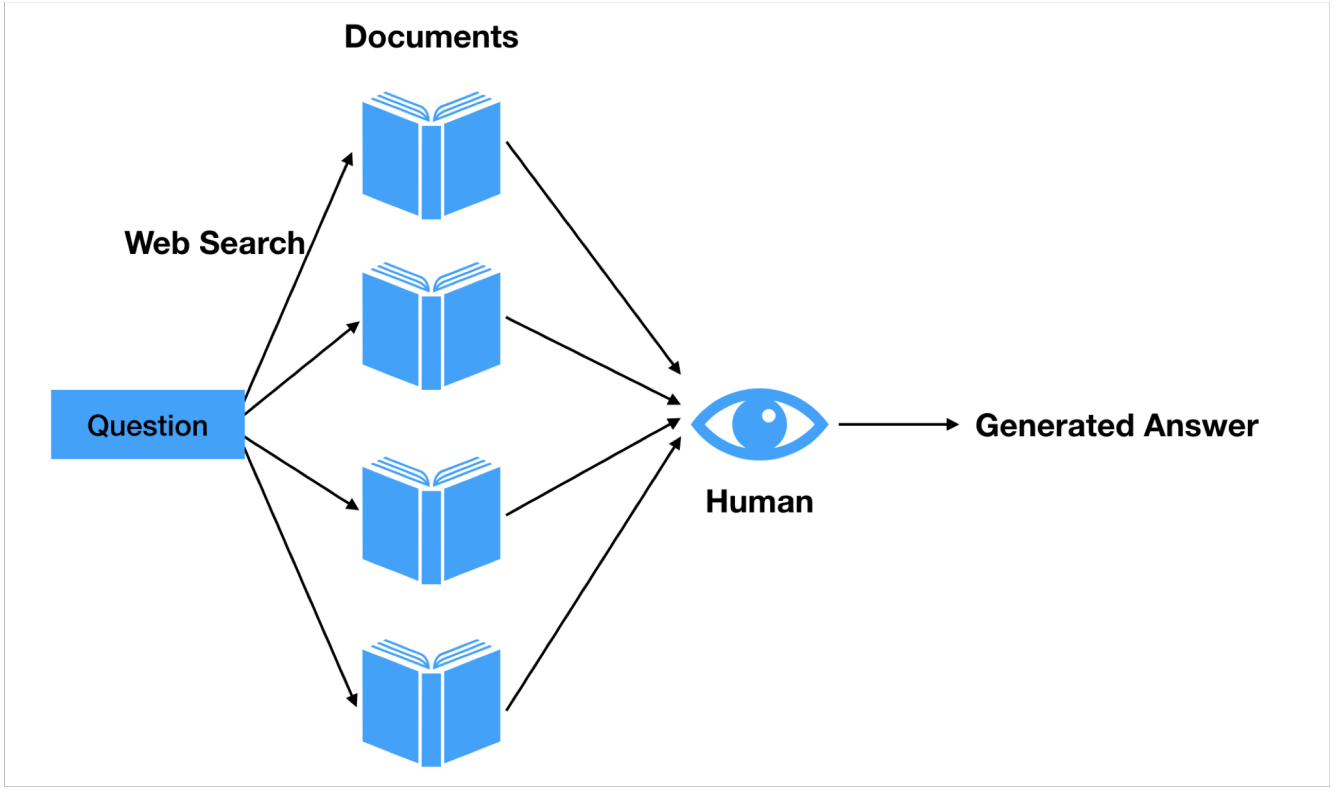
Project updates for 2/18

AngryFrog
Zihua Liu, Hui Luo, Huiming Jin, Yihui Peng

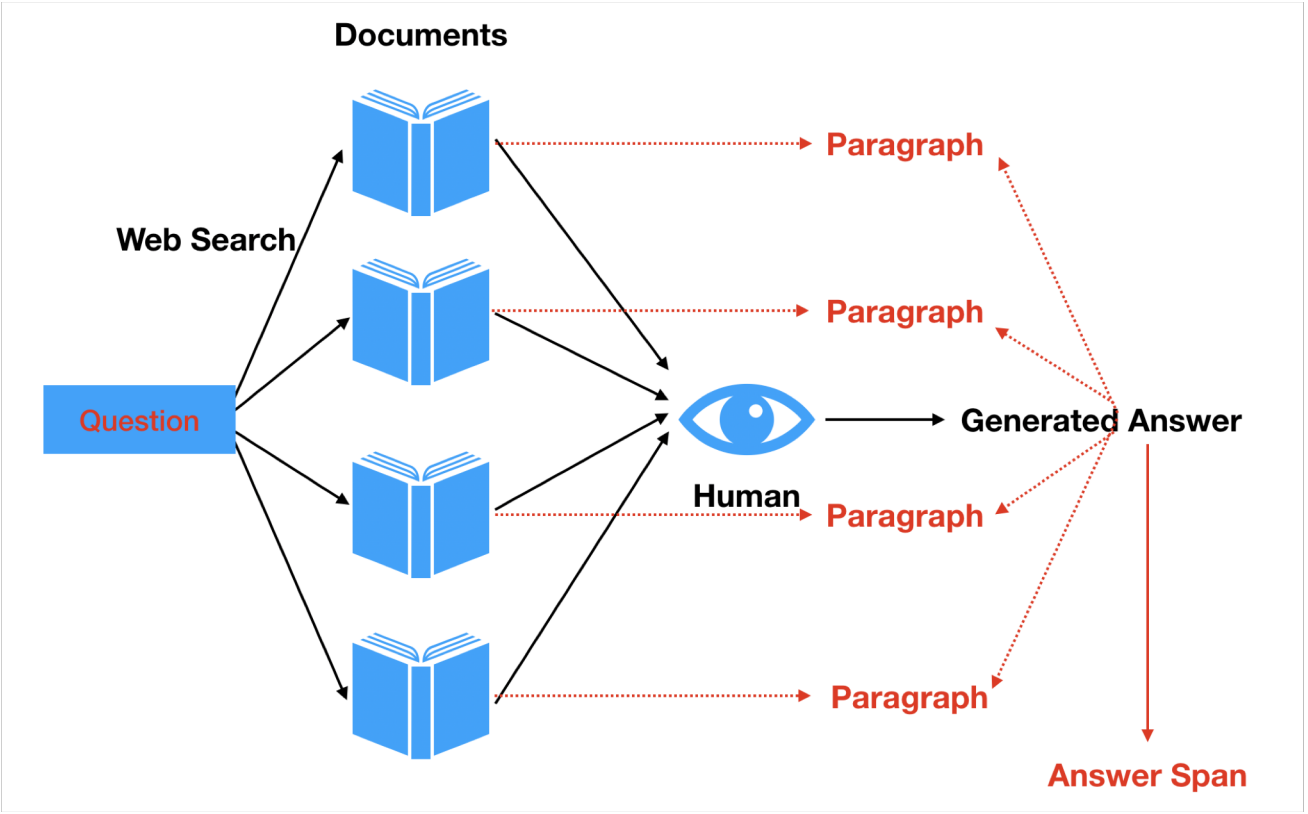
Dataset Structure



Updates for DuReader Dataset

- Chinese word segmentation.
- Answer paragraph targeting.
- Answer span locating.
- Word embedding initializing.

Dataset Preprocess



Chinese Word Segmentation

- Chinese sentences do not have space between words.
- Segment all questions, answers, document titles and paragraphs into Chinese words.
- Randomly initialize the word embedding with a length of 300

```
"question": "上海迪士尼可以带吃的进去吗",
"segmented_question": ["上海", "迪士尼", "可以", "带", "吃的", "进去", "吗"],
```

Answer paragraph targeting

- In the DuReader dataset, each question has up to five related document, and each document has 394 words on average.
- Find the most related paragraph according to the highest recall of the question.

Answer paragraph targeting using TF-IDF and BM25

- Construct sentence vectors based on TF-IDF scores.
- Select the most related paragraph according to the highest similarity between the question and paragraph in each document.

	TF-IDF vs Recall	BM25 vs Recall	TF-IDF vs BM25
Overlap	48%	52%	73%

Locating answer span

- Match real answer with the most related paragraph of each document.
- Search the substring with maximum F1-score of the real answers, and use the span of substring as the candidate answer span.

Updates for MS-MARCO Dataset

- Explored MS-MARCO dataset
- Explored BiDAF model
- Trained and applied BiDAF model as baseline
- Acquired prediction results
- Performed evaluation

Baseline Model: BiDAF

- Character Embed Layer: CNN
- Word Embed Layer: pre-trained embedding
- Contextual Embed Layer: BiLSTM
- Attention Flow Layer: Context2Query & Query2Context
- Modeling Layer: BiLSTM
- Output Layer: BiLSTM & softmax->index

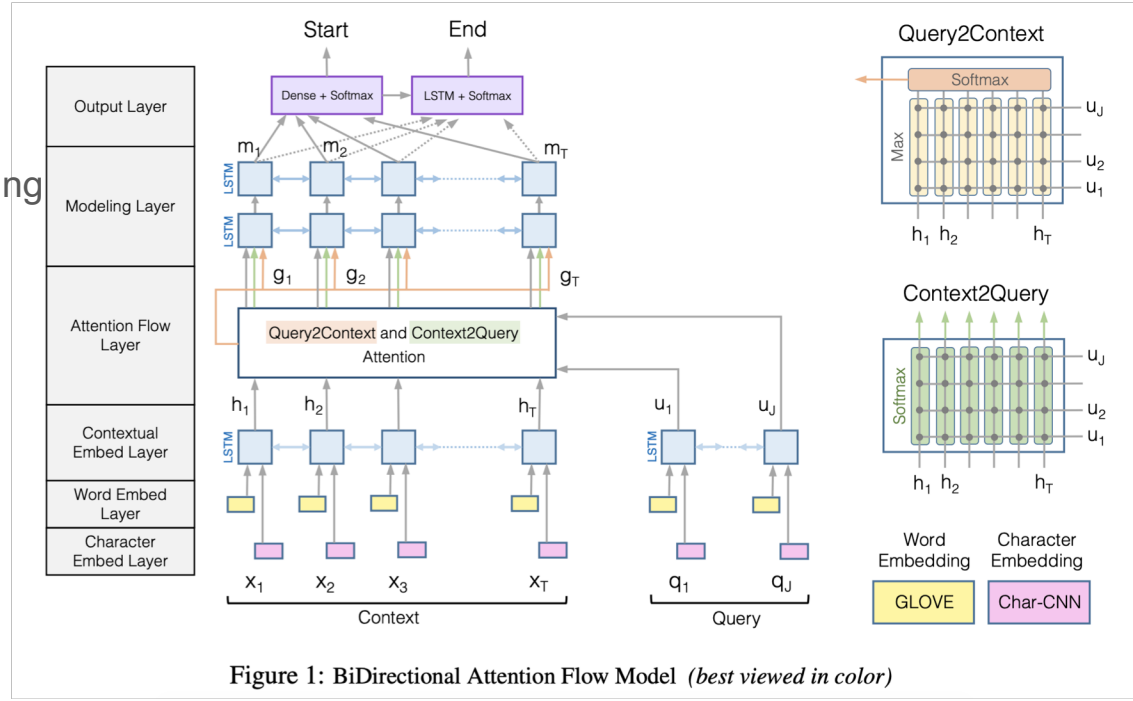


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

BiDAF Highlights

- Character embed layer & word embed layer & contextual embed layer: preserve information of character, word and sentence
- Attention Mechanism:

  Reduces the information loss caused by early summarization.

  Focus on learning the attention between the query and the context.

  Context2Query and Query2Context provide complimentary information to each other.

Training details

- Trained BiDAF on MS-MARCO
  - Dropout: 0.2
  - Highway layers: 2
  - LSTM layers: 2
  - Hidden units: 100
  - Embedding dimension: 300
  - Pre-trained embeddings: glove.840B.300d
  - Batchsize: 30
  - Learning rate: 0.001

Prediction results and evaluation

Evaluation on the official dev set where the answer is a span

Epoch	1	2	3
Semantic Similarity	0.806	0.808	0.812
BLEU_1	0.589	0.571	0.583
BLEU_2	0.553	0.537	0.550
BLEU_3	0.526	0.511	0.523
BLEU_4	0.502	0.489	0.500
ROUGE_L	0.591	0.597	0.607

Prediction examples

- **query**: how many calories in average apple
  output: 80 calories
  gold: 80 calories

- **query**: temperature and depth in earth
  output: 25 c per km
  gold: Geothermal gradient is the rate of increasing temperature with respect to increasing depth in the Earth's interior. Away from tectonic plate boundaries, it is about 25 °C per km of depth (1 °F per 70 feet of depth) near the surface in most of the world.

- **query**: weight capacity of gooseneck
  output: 30 000 lbs
  answer: The CURT over-bed folding ball gooseneck hitch offers a gross trailer weight capacity of 30,000 lbs. and a tongue we ight capacity of 7,500 lbs.

Next steps

- Apply baseline models on DuReader dataset.
- Compare the performace of different paragraph targeting algorithms.
- Explore Match-LSTM model
- Apply Match-LSTM model on MS-MARCO

Thanks!