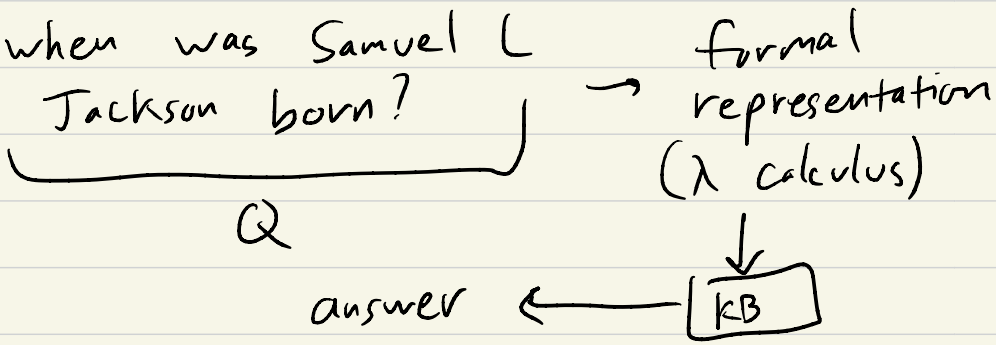
CS 378 Lecture 22

Today

- Reading comprehension
- Span-based QA
- Baseline methods
- Attentive reader

Recap Knowledge base QA

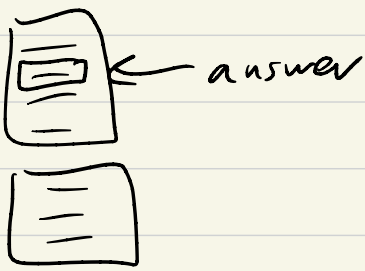


Announcements

- A5 due today
- FP out today

Reading Comprehension

What temperature should I cook chicken to?

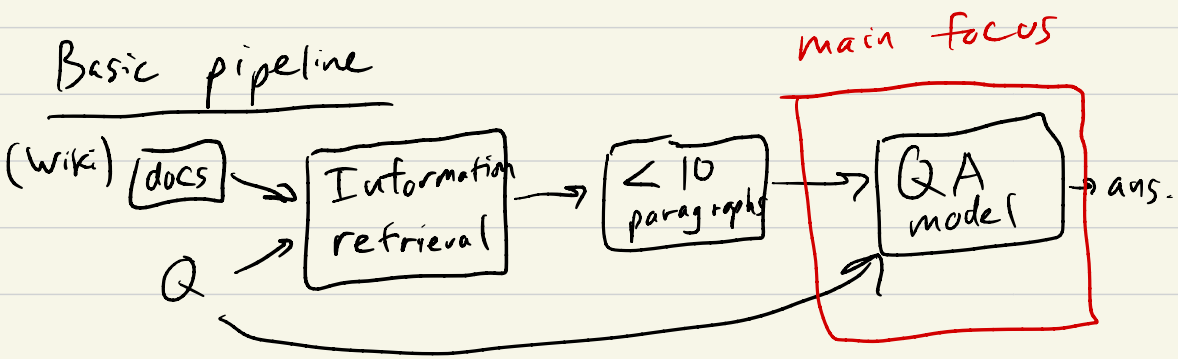


What event led to the start of WW1?

answer (span)

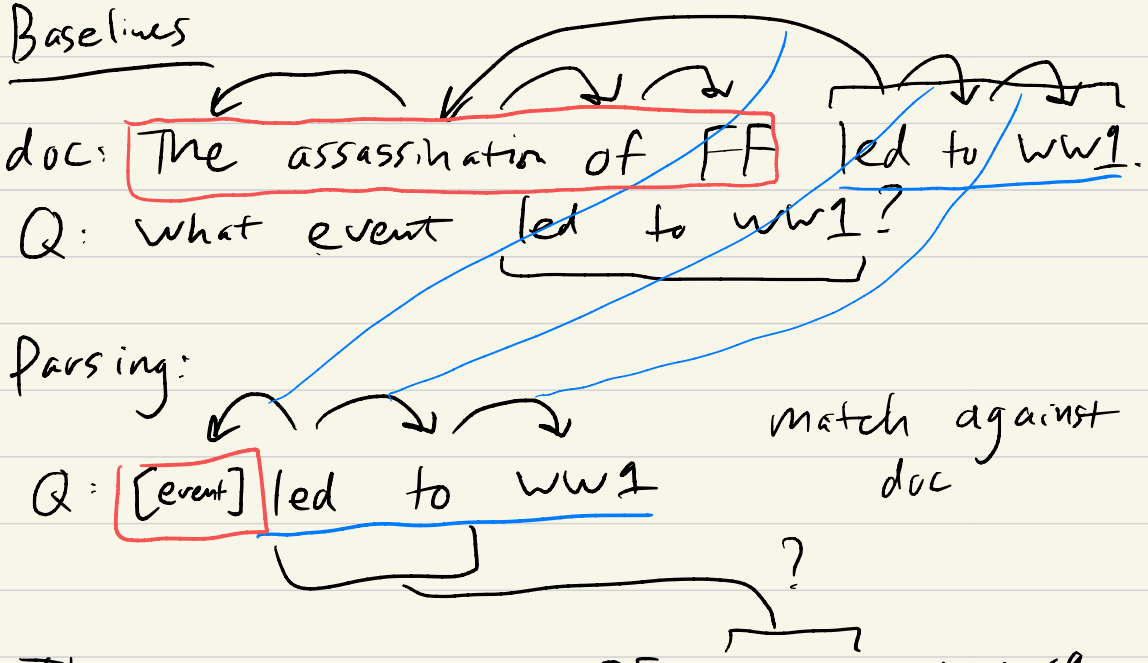
... The assassination of Franz Ferdinand took place on ... It led to WW1.

Basic pipeline



Span Extraction

Assume: answer can be identified as a span of the docs we look at



Too fragile!

NN Where is Paris?

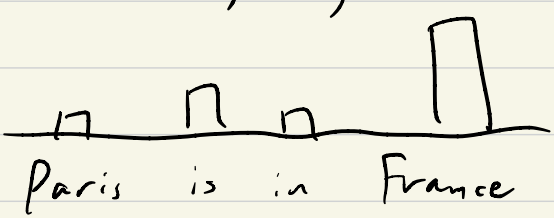
doc/passage Paris is in France . . .

Answer format

Picking a single word:

$p_1, \ldots, p_n$ words

Decision is a distribution over the indices $1, \ldots, n$



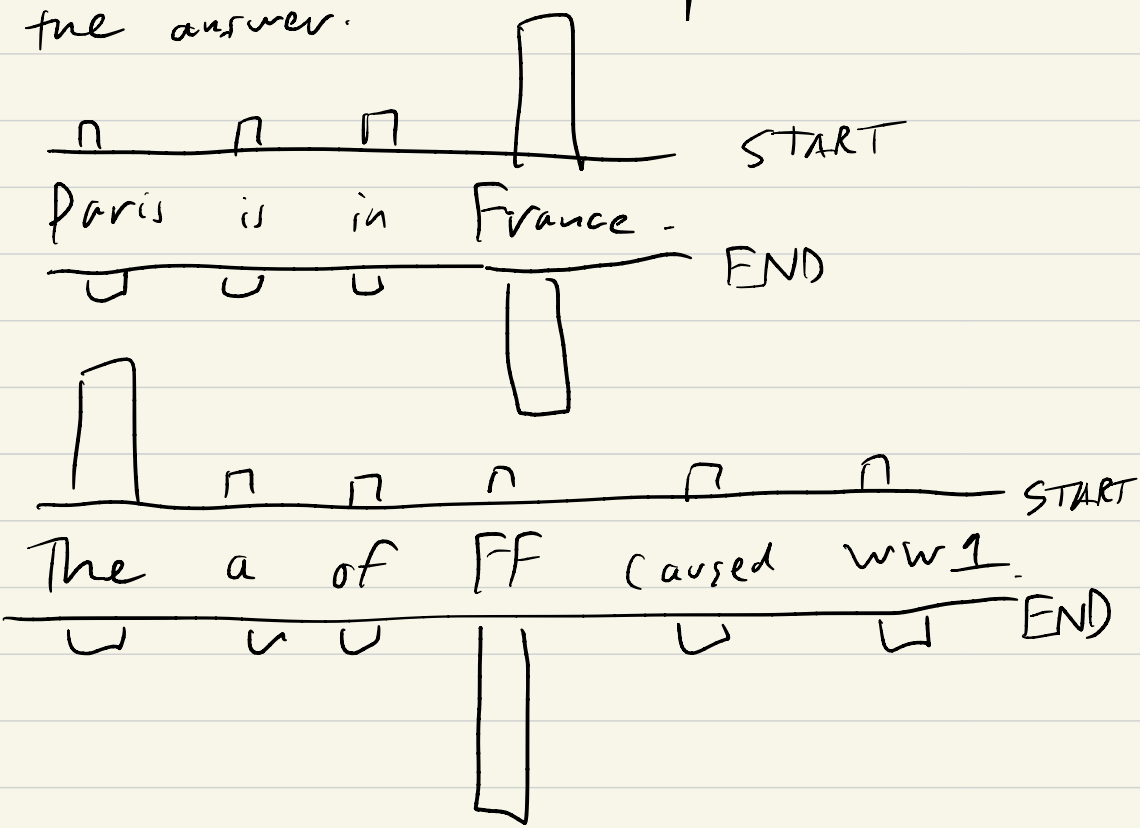
Picking a span:

O O O B-ANS O O

Paris is in France . . .

CRF / HMM

Pick start and end point for the answer.



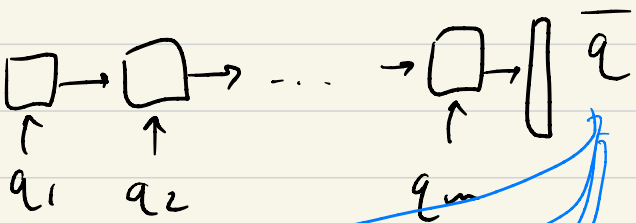
Final step: pull out answer which has highest P(start) · P(end)

Loop over all legal spans of fewer than 15 words

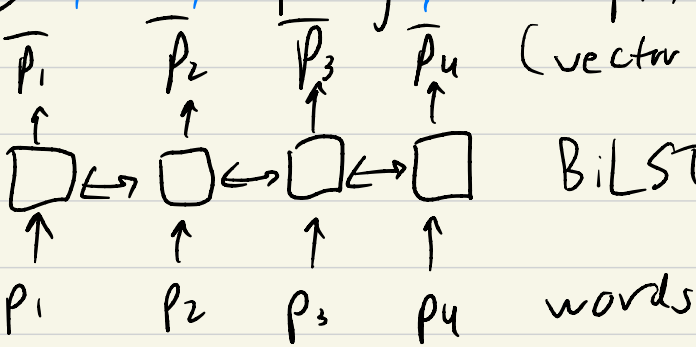
(Stanford)

Attentive Reader

① encode question into vector with LSTM



② encode passage into $\bar{p}_1, \ldots, \bar{p}_n$ (vector for each word)



③ Compute START dist:

$$\alpha = \text{softmax}_i\left(\bar{q}^T W^{START} \bar{p}_i\right)$$

dist over tokens in p

(4) Compute END dist

$$\beta = \text{softmax}_i\left(\bar{q}^T W^{END} \bar{p}_i\right)$$

dist over tokens in p

Parameters (1) 1x LSTM

(2) 1x BiLSTM

(3+4) W^{START}, W^{END}

Training Suppose we have docs, qs, ans

$$\left(p^{(i)}, q^{(i)}, \underbrace{\alpha^{(i)}, \beta^{(i)}}\right)$$

gold start + end points

Loss: $\log P(\alpha = \alpha^{(i)} \mid p^{(i)}, q^{(i)})$

$+ \log P(\beta = \beta^{(i)} \mid p^{(i)}, q^{(i)})$

Two extensions (next time):

① Additional attention between question + context to capture interactions

② More complex question encoder

Final project: These models do okay on their training data

But they don't generalize

SQuAD (training): Wiki articles

BioASQ, biomedical / NewsQA, news] models do badly here

Adversarial data: (see Jia+Liang)