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

Recap Knowledge base QA

Q: when was Samuel L Jackson born?

formal representation
(λ calculus)

answer ← KB

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

(Wiki) docs \rightarrow \text{Information retrieval} \rightarrow \text{< 10 paragraphs} \rightarrow \text{QA model} \rightarrow \text{ans.}
Span Extraction

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

Baselines

doc: The assassination of FF led to WW1.

Q: what event led to WW1?

Parsing:

Q: [event] led to WW1

The assassination of FF caused WW1.

Too fragile!
NN Where is Paris?

doc/ Paris is in France.

passage

Answer format

Picking a single word:

P1,...,pn words

Decision is a distribution over the indices 1,...,n

Paris is in France

Picking a span:

O O O O B-ANS O O

Paris is in France

CRF/HMM

Not common
Pick start and end point for the answer.

```
START
Paris is in France.
END
```

```
START
The a of FF caused WW1.
END
```

Final step: pull out answer which has highest \( P(\text{start}) \cdot P(\text{end}) \)

Loop over all legal spans of fewer than 15 words
(stanford)

**Attentive Reader**

1. Encode question into vector with LSTM

   ![Diagram of encoding question]

   - q1
   - q2
   - qn

   Compare q to each \( p_i \)

2. Encode passage into \( p_1, \ldots, p_n \)

   - \( \bar{p}_1 \)
   - \( \bar{p}_2 \)
   - \( \bar{p}_3 \)
   - \( \bar{p}_4 \)

   (vector for each word)

   ![Diagram of encoding passage using BiLSTM]

   - \( p_1 \)
   - \( p_2 \)
   - \( p_3 \)
   - \( p_4 \) words

3. Compute START dist:

   \[
   \alpha = \text{softmax}(\bar{q}^T W^{\text{START}} \bar{p}_i) \\
   \text{dist over tokens in } p
   \]
④ Compute $\text{END dist}$

\[ \beta = \text{softmax}_i \left( \frac{q^T W_{\text{END}} p_i}{\text{dist over tokens in } p} \right) \]

**Parameters**

1. $1 \times \text{LSTM}$
2. $1 \times \text{BiLSTM}$
3. $W_{\text{START}}, W_{\text{END}}$

**Training**

Suppose we have docs, $qs, ans$

\[
(\rho^{(i)}, q^{(i)}, \alpha^{(i)}, \beta^{(i)}), \quad \text{gold start+end points}
\]

**Loss:**

\[
\log P(\alpha = \alpha^{(i)} | \rho^{(i)}, q^{(i)}) + \log P(\beta = \beta^{(i)} | \rho^{(i)}, q^{(i)})
\]
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, News QA ] models do poorly here
Adversarial data: (see Jia & Liang)