Learning Recurrent Span Representations For Extractive Question Answering

Kenton Leey, Shimi Salant, Tom Kwiatkowksiz, Ankur Parikhz, Dipanjan Dasz, and Jonathan Berant

Introduction

Task

- **SQUAD** dataset, the **answer extraction** task

Motivation

- Previous approaches identify **answer spans** by labeling either individual words, or the start and end of the answer span, and are susceptible to search errors due to **greedy** training and decoding.
- So consider enumerating all possible answer spans

Challenge

- For enumerating all possible answer spans, a naive approach require a network **cubic** in size with respect to the passage length, and such a network would be **untrainable**.

Contributions

- Present a novel neural architecture called **RASOR**, **reusing recurrent computations for shared substructures**. (cubic -> quadratic)
- Directly classifying each of the competing spans, and training with global normalization over all possible spans, leads to an increase in performance.

Extractive Question Answering

Task Definition

- a question $\mathbf{q} = \{q_0, \ldots, q_n\}$, a passage of text $\mathbf{p} = \{p_0, \ldots, p_m\}$
- predict a single **answer span** $\mathbf{a} = \langle a_{start}, a_{end} \rangle$, represented as a pair of indices into p.
- Machine learned extractive question answering systems learn a predictor function $f(\mathbf{q}, \mathbf{p}) \rightarrow \mathbf{a}$ from a training dataset of $\langle \mathbf{q}, \mathbf{p}, \mathbf{a} \rangle$ triples.

Related Work

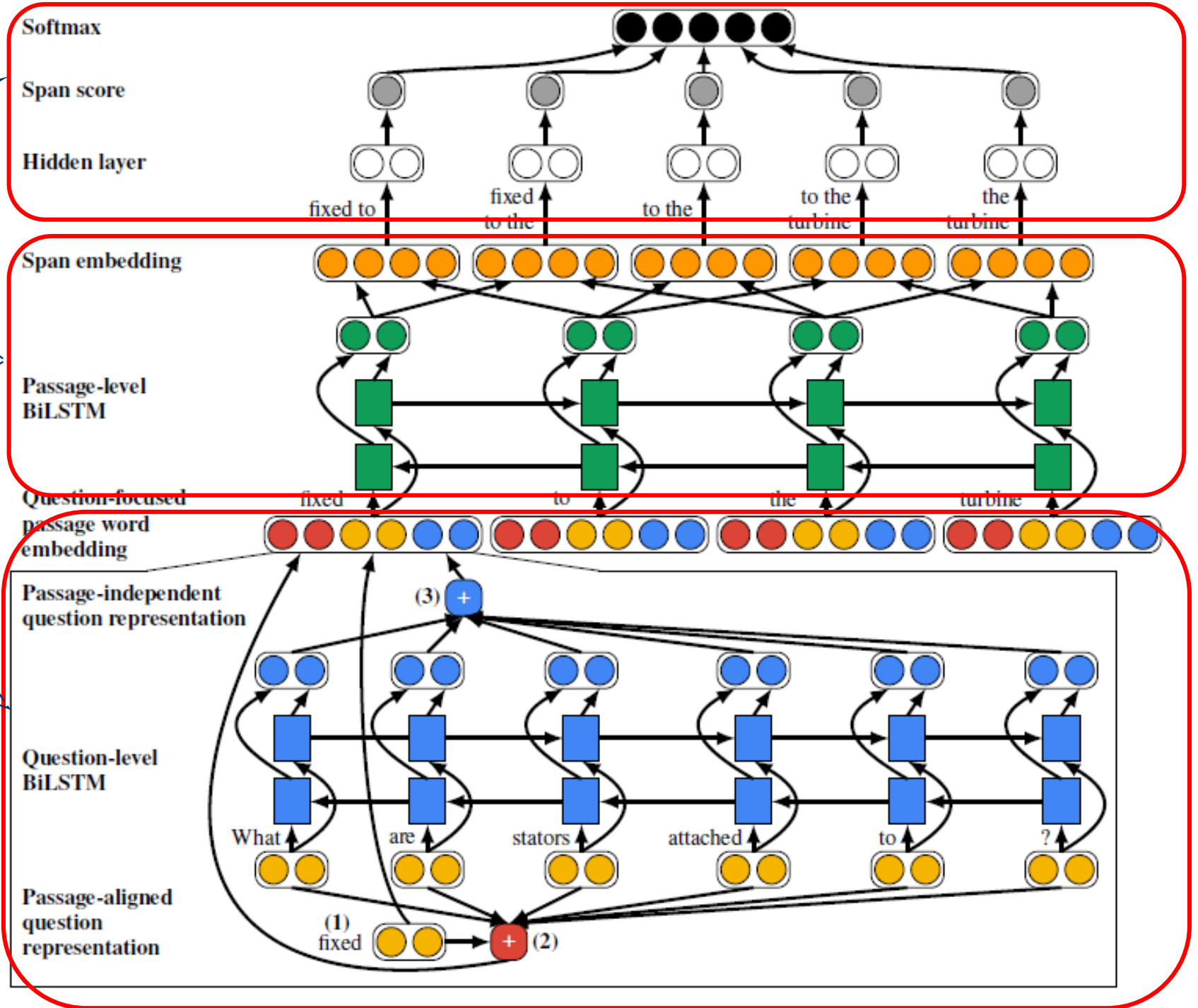
- Work on SQUAD dataset
  - lexical features: a linear model with sparse features
  - syntactic information: of dependency paths
  - Match-LSTM, pointer networks
- The Cloze task
- Extractive question answering on sentences
  - TREC, WikiQA

Model

Scoring Answer Spans

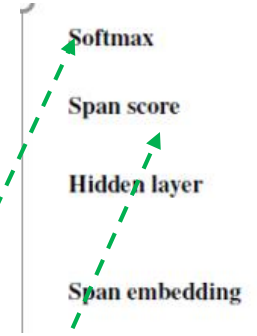
RASOR: Recurrent Span Representation

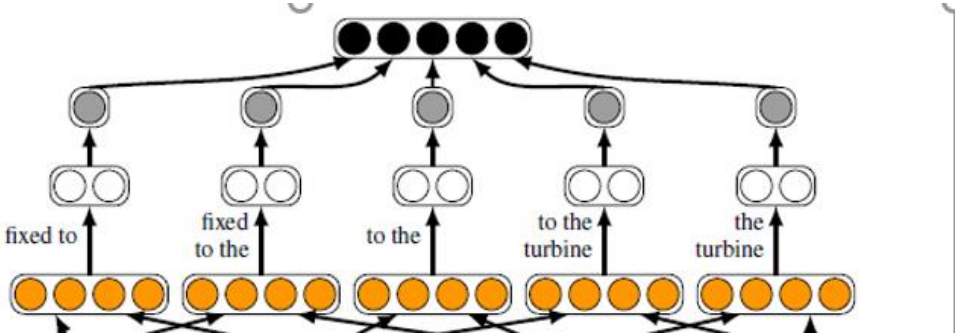
Question-focused Passage Word Embeddings



Model

Scoring Answer Spans





- predictor function

$$f(\mathbf{q}, \mathbf{p}) := \underset{\mathbf{a} \in \mathbf{A}(\mathbf{p})}{\operatorname{argmax}} P(\mathbf{a} \mid \mathbf{q}, \mathbf{p})$$

A(p): answer span candidate

- assume a fixed-length representation h_a, and then

$$s_{\mathbf{a}} = w_a \cdot \mathrm{FFNN}(h_{\mathbf{a}}) \qquad \mathbf{a} \in \mathbf{A}(\mathbf{p})$$

$$P(\mathbf{a} \mid \mathbf{q}, \mathbf{p}) = \frac{\exp(s_{\mathbf{a}})}{\sum_{\mathbf{a}' \in \mathbf{A}(\mathbf{p})} \exp(s_{\mathbf{a}'})} \qquad \mathbf{a} \in \mathbf{A}(\mathbf{p})$$

FFNN(): a fully connected feed-forward neural network
w_a: a learned vector for scoring the last layer of the feed-forward neural network

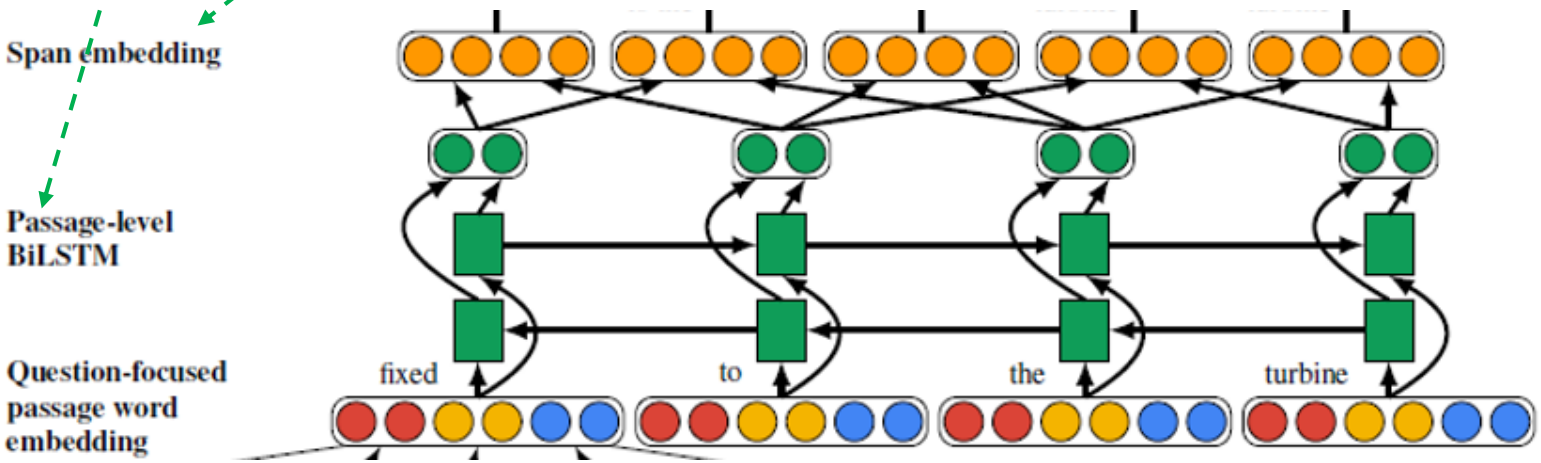
Model

RASOR: Recurrent Span Representation

- Computing h_a

$$\{p_1^{*\prime}, \ldots, p_m^{*\prime}\} = \text{BILSTM}(\{p_1^*, \ldots, p_m^*\})$$

$$h_{\mathbf{a}} = [p_{a_{start}}^{*\prime}, p_{a_{end}}^{*\prime}] \qquad \langle a_{start}, a_{end} \rangle \in \mathbf{A}(\mathbf{p})$$



Model

Question-focused Passage Word Embeddings

Question-independent passage word embedding

- The first component simply looks up the pretrained word embedding for the passage word, p_i.

Passage-aligned question representation

- Parikh et al.(2016), variant of neural attention

$$s_{ij} = \text{FFNN}(p_i) \cdot \text{FFNN}(q_j) \qquad 1 \le j \le n$$

$$a_{ij} = \frac{\exp(s_{ij})}{\sum_{k=1}^{n} \exp(s_{ik})} \qquad 1 \le j \le n$$

$$q_i^{align} = \sum_{j=1}^{n} a_{ij} q_j$$

s_{ij}: attention scores

Model

Question-focused Passage Word Embeddings

Passage-independent question representation

- Li et al. (2016)

$$\{q'_1, \ldots, q'_n\} = \text{BILSTM}(\mathbf{q})$$

$$s_j = w_q \cdot \text{FFNN}(q'_j) \qquad 1 \leq j \leq n$$

$$a_j = \frac{\exp(s_j)}{\sum_{k=1}^{n} \exp(s_k)} \qquad 1 \leq j \leq n$$

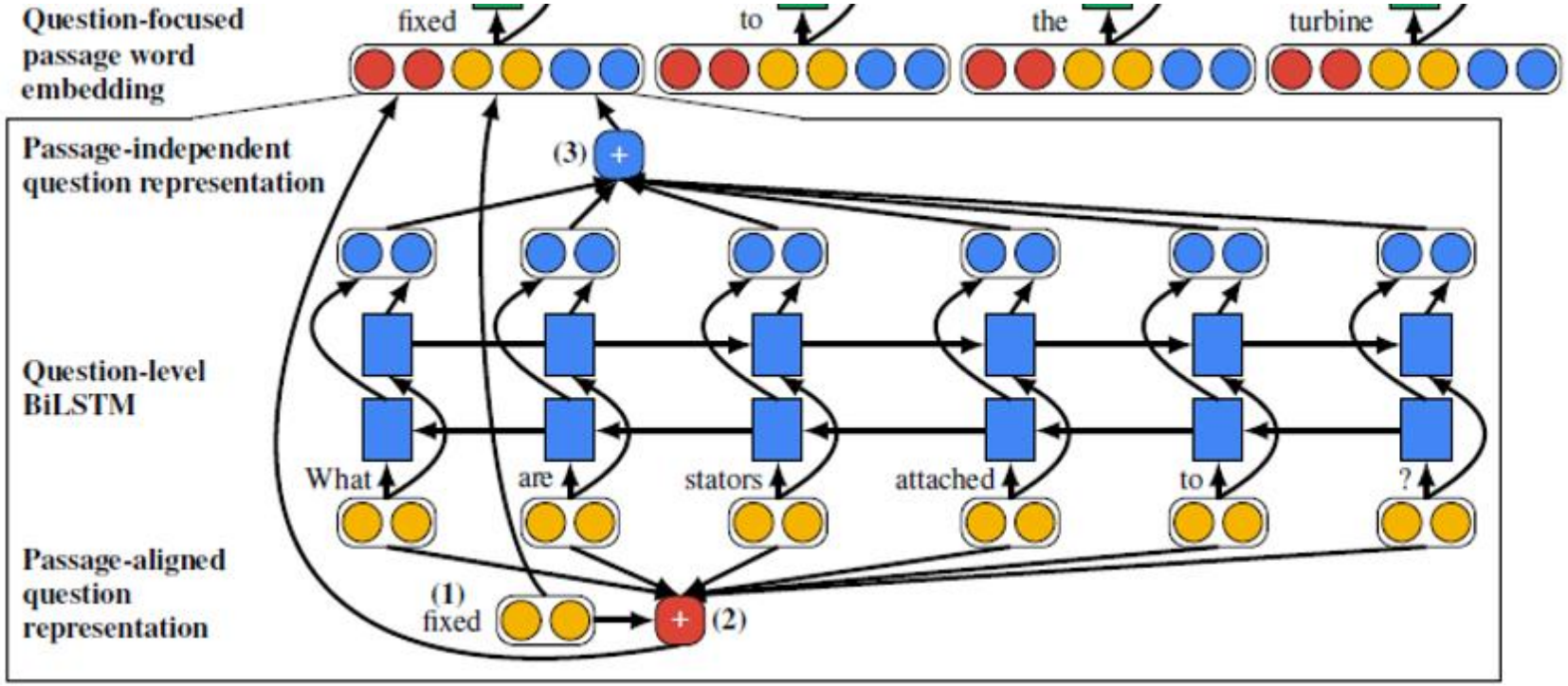
$$q^{indep} = \sum_{j=1}^{n} a_j q'_j$$

Concatenation

$$p_i^* = [p_i, q_i^{align}, q^{indep}].$$

Model

Question-focused Passage Word Embeddings



(1) the original passage word embedding
(2) a passage-aligned representation of the question
(3) a passage-independent representation of the question shared across all passage words

Model

Learning

- Maximize the loglikelihood of the correct answer candidates
- Backpropagate the errors end-to-end.

Experiments

Experimental Setup

- SQUAD dataset
- 300 dimensional GloVe embeddings trained on a corpus of 840bn words
- Grid searches over parameters

Experiments

Comparison to Other Work

	Dev		Test	
System	EM	F1	EM	F1
Logistic regression baseline	39.8	51.0	40.4	51.0
Match-LSTM (Sequence)	54.5	67.7	54.8	68.0
Match-LSTM (Boundary)	60.5	70.7	59.4	70.0
RASOR	66.4	74.9	67.4	75.5
Human	81.4	91.0	82.3	91.2

Table 1: Exact match (EM) and span F1 on SQUAD.

- Match-LSTM: both training and evaluation are greedy, making their system susceptible to search errors when decoding.
- RASOR: can efficiently and explicitly model the quadratic number of possible answers

Experiments

Model Variations

Question representation	EM	F1
Only passage-independent	48.7	56.6
Only passage-aligned	63.1	71.3
RASOR	66.4	74.9

(a) Ablation of question representations.

Learning objective	EM	F1
Membership prediction	57.9	69.7
BIO sequence prediction	63.9	73.0
Endpoints prediction	65.3	75.1
Span prediction w/ log loss	65.2	73.6

(b) Comparisons for different learning objectives given the same passage-level BiLSTM.

Table 2: Results for variations of the model architecture presented in Section 3.

- First, we observe general improvements when using labels that closely align with the task.
- Second, we observe the importance of allowing interactions between the endpoints using the span level.

Analysis

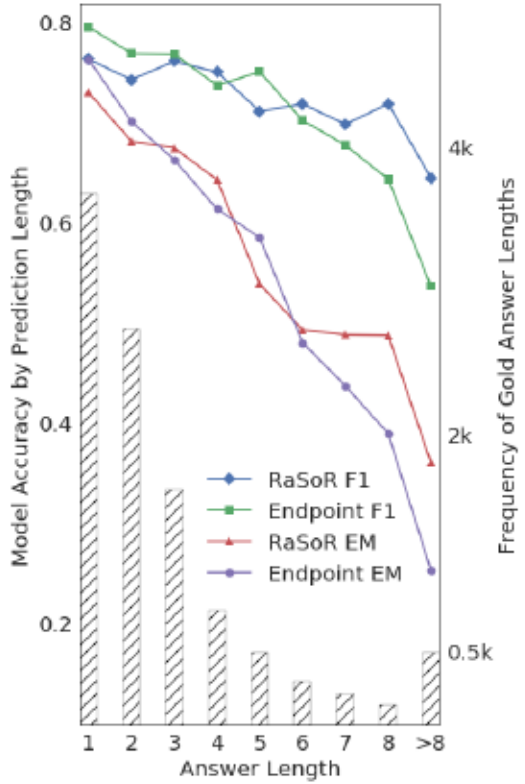


Figure 2: F1 and Exact Match (EM) accuracy of RASOR and the endpoint predictor baseline over different prediction lengths.

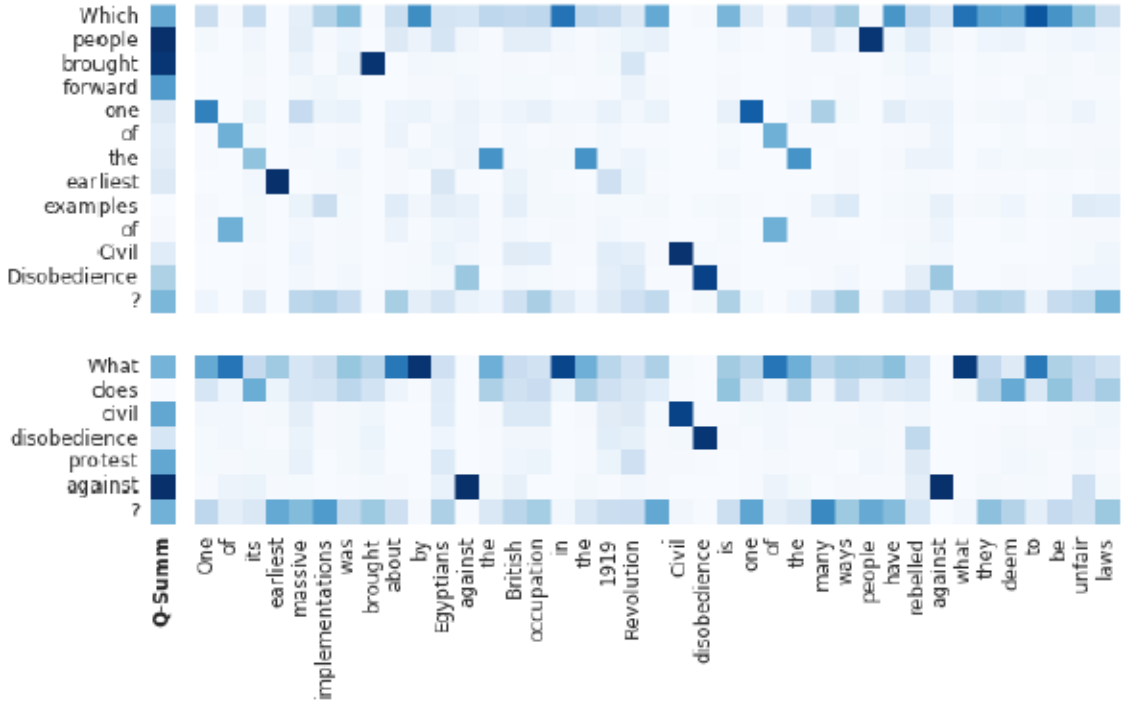


Figure 3: Attention masks from RASOR. Top predictions for the first example are 'Egyptians', 'Egyptians against the British', 'British'. Top predictions for the second are 'unjust laws', 'what they deem to be unjust laws', 'laws'.

THANKS

Ni Yihan