Question Answering on SQuAD 2.0

by Boxiao Pan, Gaël Colas and Shervine Amidi

Stanford University

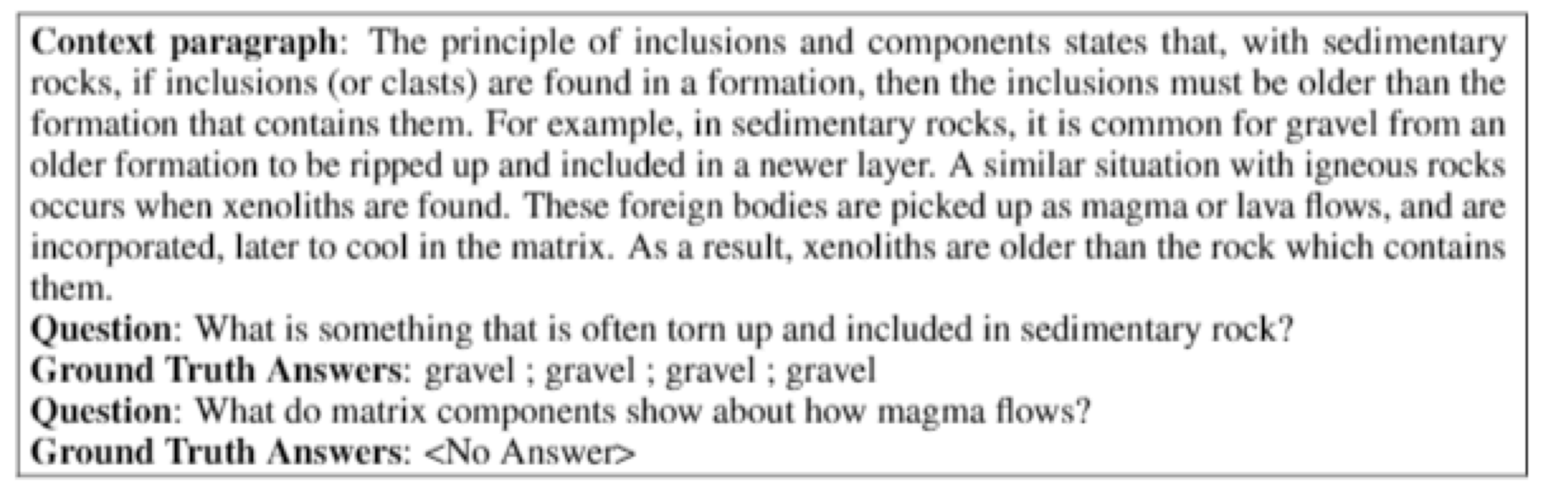
Introduction

Problem

- Task: Extractive question answering
- Dataset: SQuAD 2.0
- Challenge: Unanswerable questions

Current state-of-the-art:
BERT-based model from Google AI
EM = 86.7 ; F1 = 89.1

non-PCE category



Context paragraph: The principle of inclusions and components states that, with sedimentary rocks, if inclusions (or clasts) are found in a formation, then the inclusions must be older than the formation that contains them. For example, in sedimentary rocks, it is common for gravel from an older formation to be ripped up and included in a newer layer. A similar situation with igneous rocks occurs when xenoliths are found. These foreign bodies are picked up as magma or lava flows, and are incorporated, later to cool in the matrix. As a result, xenoliths are older than the rock which contains them.
Question: What is something that is often torn up and included in sedimentary rock?
Ground Truth Answers: gravel ; gravel ; gravel ; gravel
Question: What do matrix components show about how magma flows?
Ground Truth Answers: <No Answer>

Figure 3: SQuAD 2.0 dev set example: a context with two associated questions, one with an answer and one without

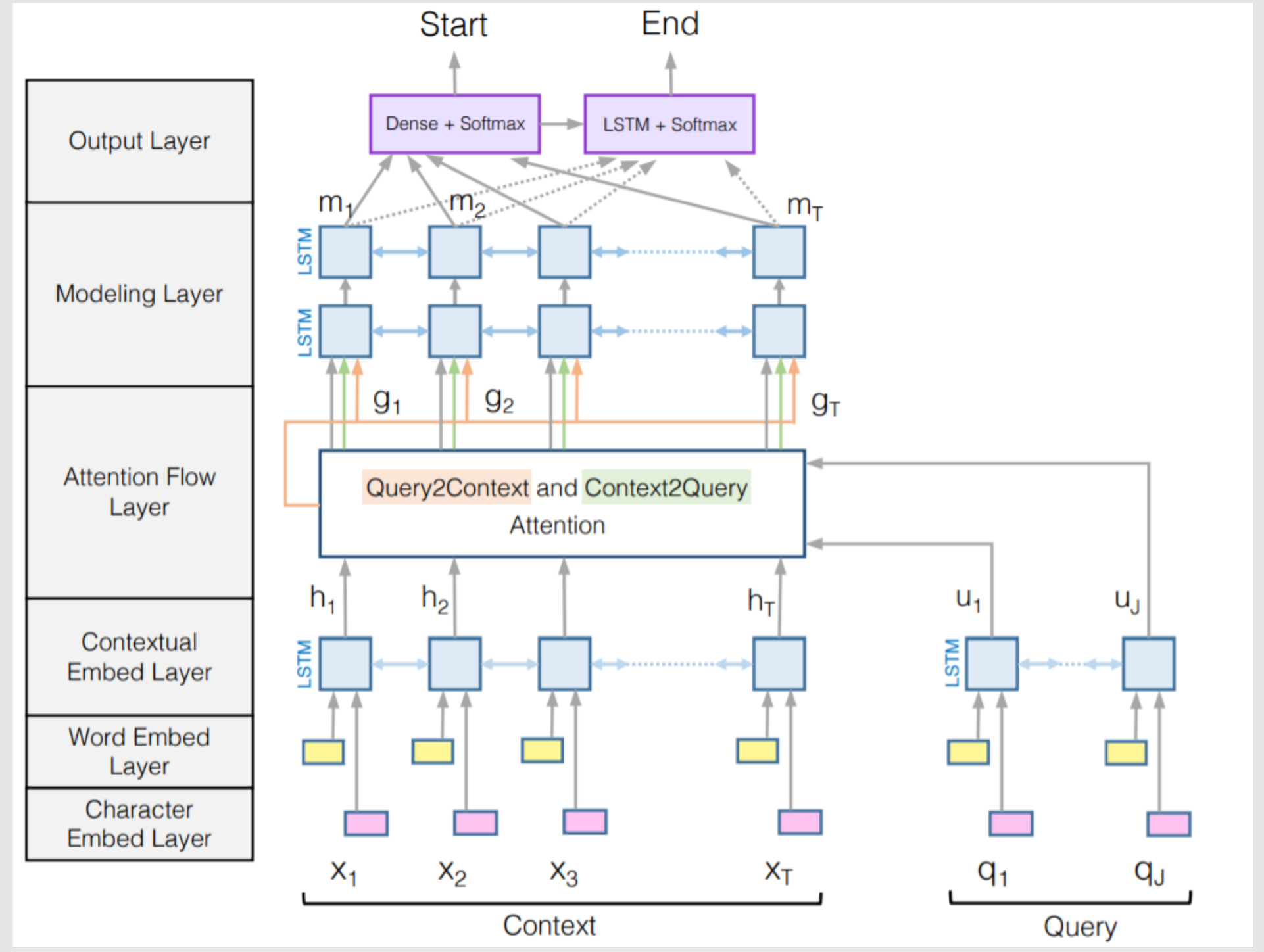
Data

SQuAD 2.0 split:

- train = 129,941 examples
- dev = 6078 examples
- test = 5915 examples

Baseline

BiDAF – BiDirectional Attention Flow

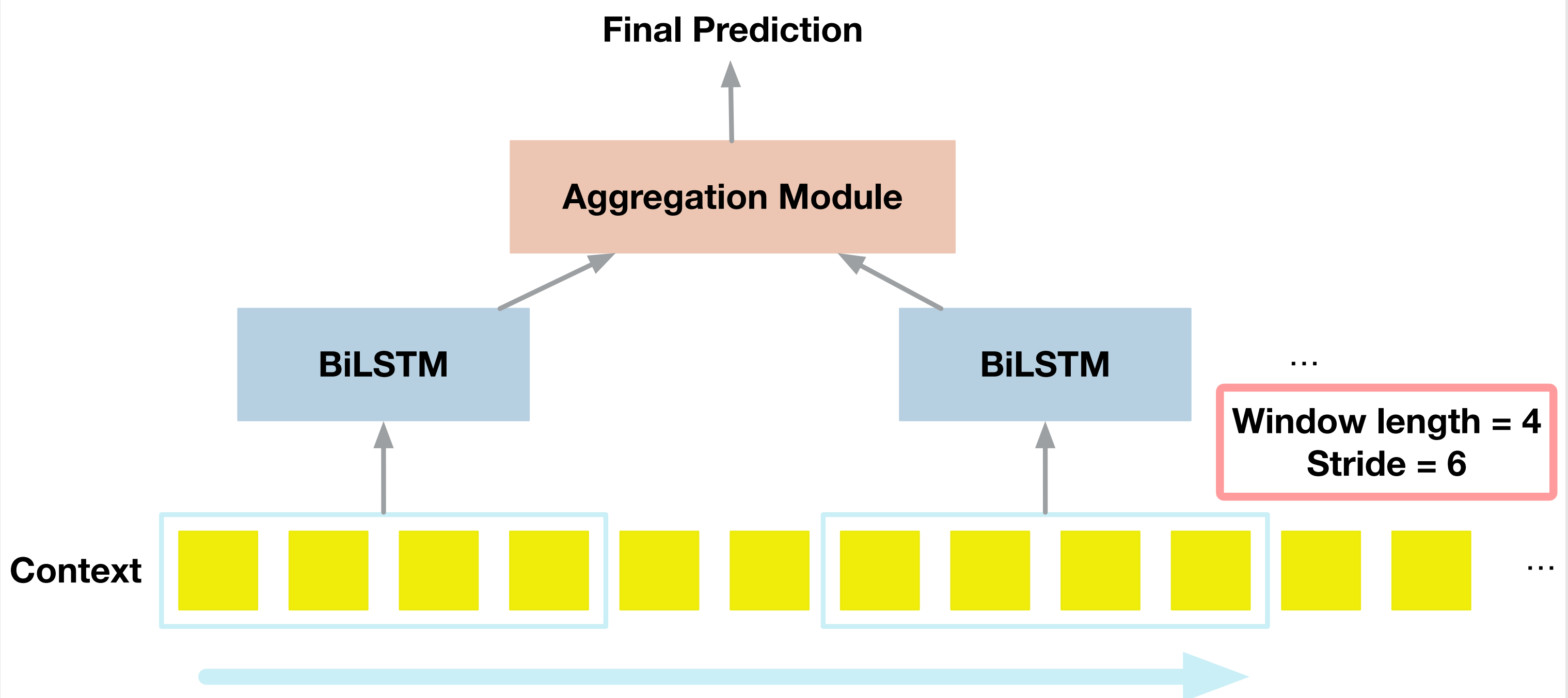


Approach

A more powerful Word Embedding

Word features	BiDAF word embedding	Character-level word embedding	Tag features: POS ; NER	Additional features
Embeddings	Pretrained GloVe embeddings	Trainable character embeddings	One-hot encoding Trainable tag classes embeddings	EM = context-question exact match TF*IDF = term-frequency

Segment-Based Aggregation (SBA)



Approach Overview

- Slide a window over context
- Predict for each window
- Aggregate over all windows

Hyper-param Setting

- Window length = 50
- stride = 20

Deep Dynamic Co-attention (DDCo)

Adopted from [3], this idea generates a co-attention score by looking at context and question simultaneously, thus being able to leverage useful mutual information. A dynamic decoder is used to iteratively predict the start and end index at each time step.

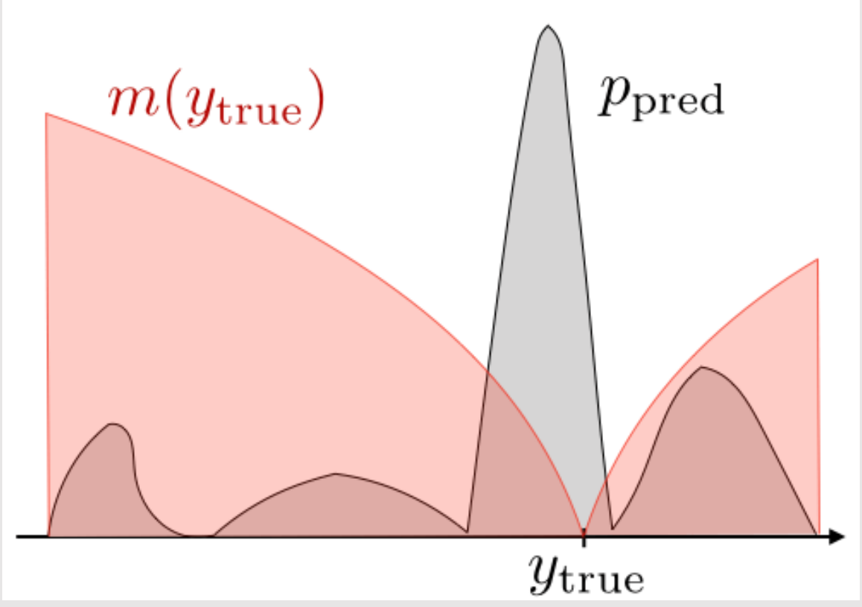
A new loss term

Distance loss

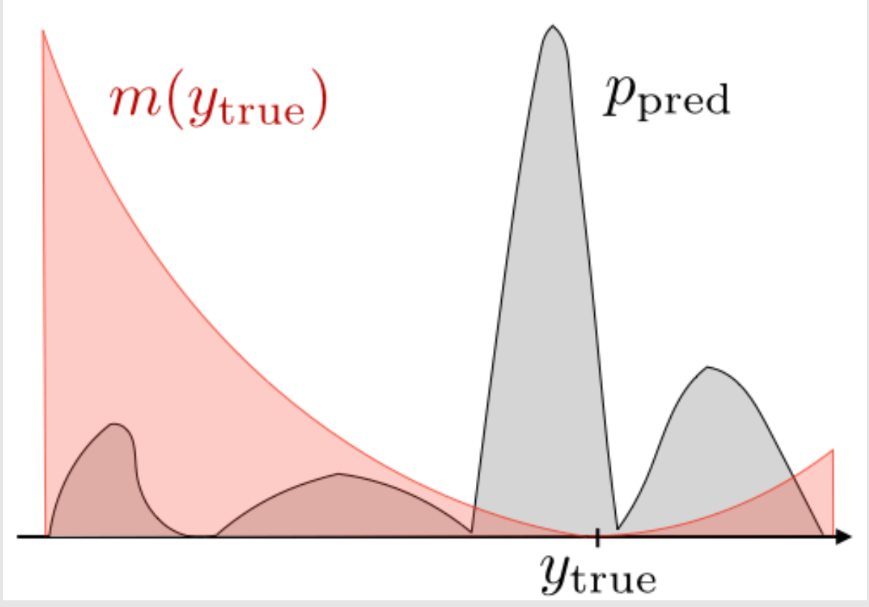
Goal: penalize more predictions far from the true label

Formula $\mathcal{L}_{\text{ind}} = \lambda \cdot \sum_i [m(y_{\text{true}}) \circ p_{\text{pred}}]_i$

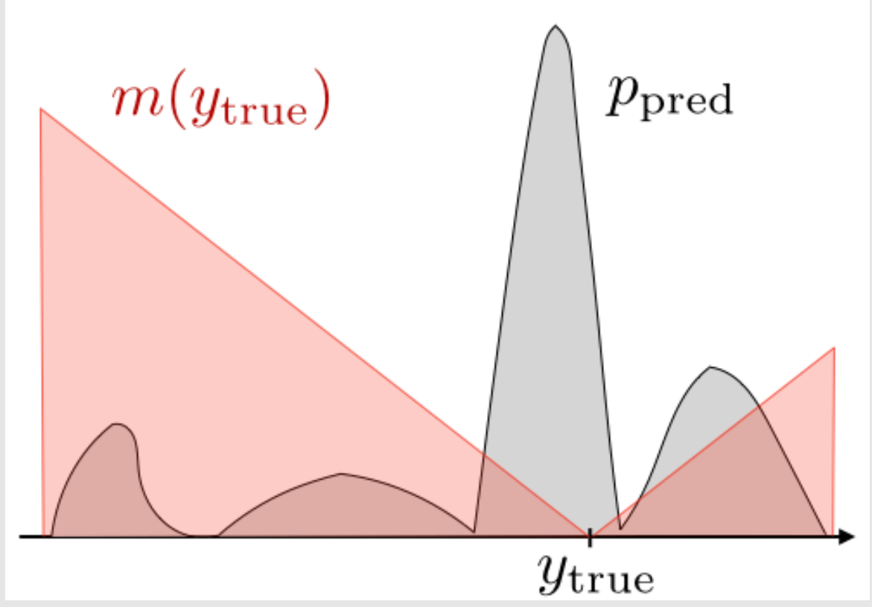
3 types of penalization tested



Square root (SR)



Linear (L)



Quadratic (Q)

Results (dev)

Model	EM	F1
BiDAF	55.1	58.2
BiDAF + SR	55.7	59.8
BiDAF + L	56.0	59.7
BiDAF + Q	56.6	60.5
BiDAF + DDC	57.3	61.7
BiDAF + SBA	58.4	62.1
BiDAF + Char (C-BiDAF)	60.6	64.0
C-BiDAF + Tags (Tag-BiDAF)	61.4	64.9
Tag-BiDAF + DDCo	60.7	62.5
Tag-BiDAF + SBA	**62.0**	**65.4**

Conclusion

- More powerful word embeddings significantly improved performance
- Segment-based approach can include more information
- A penalization term based on the distance of the distribution from the true label yields promising results

References

[1] Seo et al. Bidirectional attention flow for machine comprehension.
CoRR, abs/1611.01603, 2016.

[2] Chen et al. Reading Wikipedia to answer open-domain questions.
arXiv preprint arXiv:1704.00051, 2017

[3] Xiong et al. Dcn+: Mixed objective and deep residual coattention for question answering.
arXiv preprint arXiv:1711.00106, 2017