ATTENTIVE HISTORY SELECTION FOR CONVERSATIONAL QUESTION ANSWERING

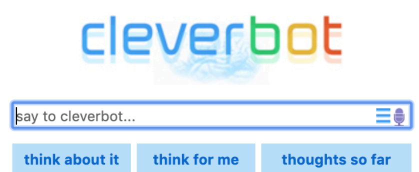
Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, Mohit Iyyer

Presented by - Vedanshi Kataria (20774266)

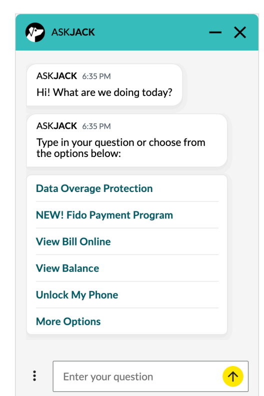
CONTENTS

CONVERSATIONAL AGENTS

- Can be of multiple types:
  - Open Domain : General conversation, Natural Dialogues. Example:



  - Closed Domain : Task(/s) specific conversation, Conversational Search



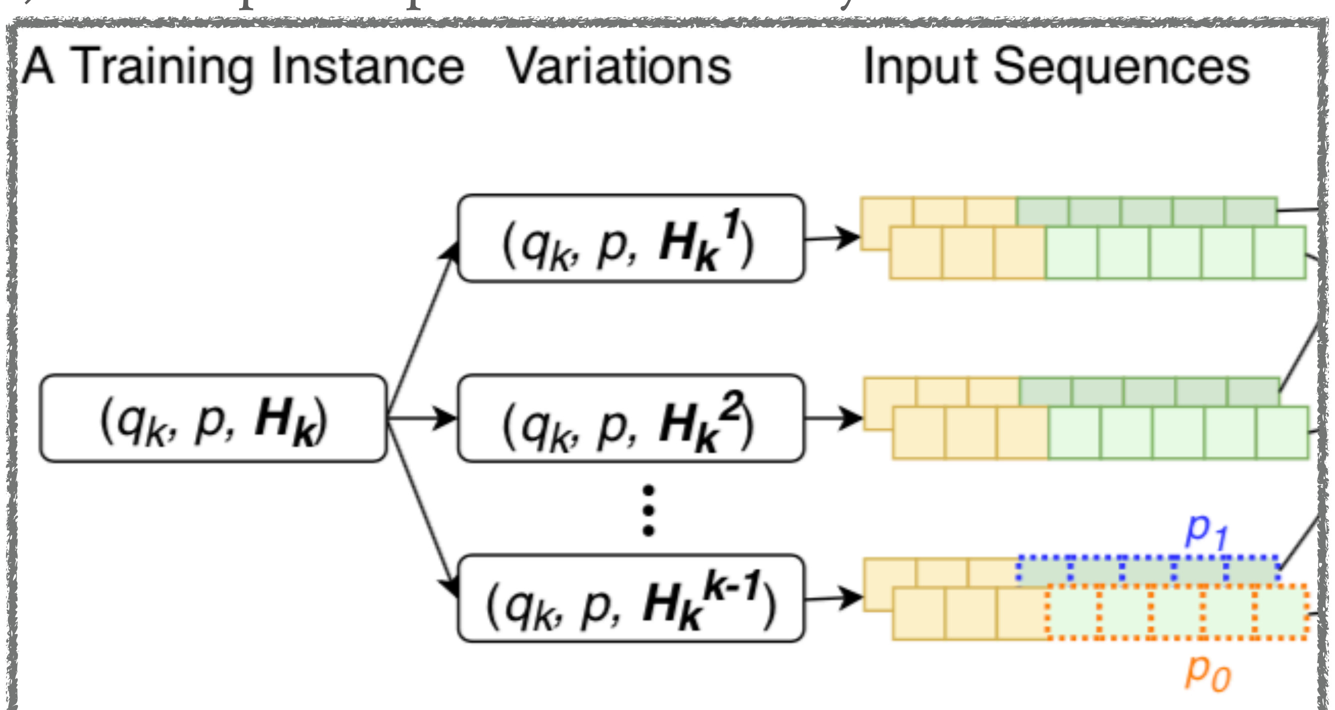
- Early Conversational Agents involved Intent Detection, Slot Filling, Information Retrieval Model, NLU module
- Siri and Google Assistant can be looked at as an example of a combination of both these types.

MOTIVATION

- Information Retrieval in the form of general conversational Question Answering (ConvQA) requires the system to remember old conversation as well.
- Existing systems only use the current question to find an answer from the context provided.
- No existing work that focuses on learning to select or re-weight conversational history turns.
- There may be three different types of conversation turns:
  - **Drill Down** : the current question is a request for more information about a topic being discussed
  - T**opic Shift** : the current question is not immediately relevant to something previously discussed
  - **Topic Return** : the current question is asking about a topic again after it had previously been shifted away from

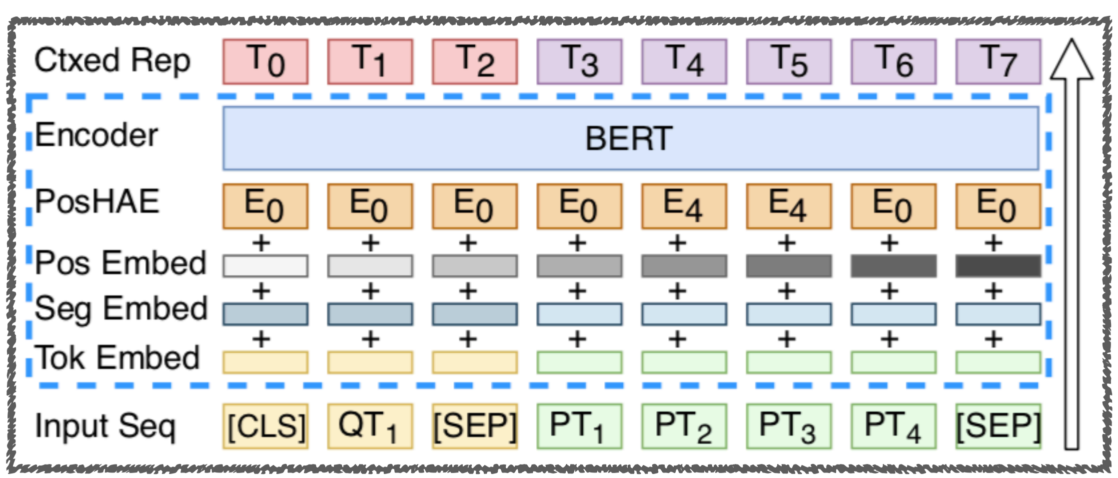
BERT ENCODER

- Encodes question q_k, paragraph p (context), and conversational histories H_k into contextualised representations.
- Input : (q_k, p, H_k). This input is used to generate $(k-1)$ variations of the instance where each variation contains the same question and passage, with only one turn of conversation history.
- If the context paragraph is too long, a sliding window is used to split it. Suppose the paragraph is split into n pieces, the training instance (q_k, p, H_k) will generate $n(k-1)$ input sequences.
- Generates contextualised token-level representations based on the embeddings for tokens, segments, positions, and a special positional history answer embedding (PosHAE)



PROPOSED METHOD 1 - POSITIONAL HISTORY ANSWER EMBEDDINGS

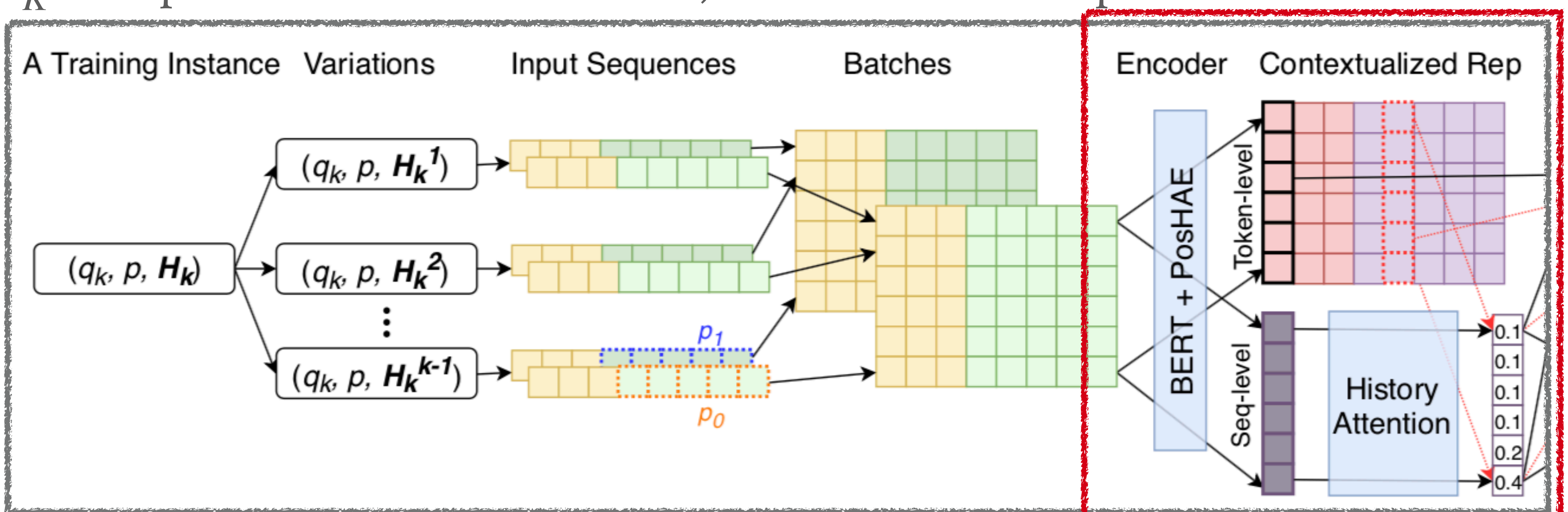
- **Intuition** behind adding Positional Embeddings: Utility of a historical utterance could be related to its position.
- Previous works have been simply appending "n" previous answers to the question.
- Observed Benefits: Enables the ConvQA model to capture the spatial patterns of history answers in context.



Encoder with PosHAE

PROPOSED METHOD 2 – HISTORY ATTENTION MECHANISM

- Inputs: Generated token-level and sequence-level representations for all variations
- A single layer feed forward network is used to learn the attention weights.
- Attention Vector $D \in R^h$ is learnt to compute attention weight for each sentence presentation s_k^i using $w_i = \frac{e^{D \cdot s_k^i}}{\sum_{i'=1}^{I} e^{D \cdot s_k^{i'}}}$
- *Fine-grained history attention*: Instead of using sequence level representation S_K as input for attention network, use token level representation



PROPOSED METHOD 3 – MULTI TASK LEARNING (1)

- Answer Span Prediction : For each token, predict the probability of being BEGIN token as well as END token i.e. learn *begin vector* B and *end vector* E.

- The probability for token being *begin token* and *end token* is $p_m^B = \frac{e^{\mathrm{B}\cdot\hat{\mathrm{t}}_k(m)}}{\sum_{m'=1}^{M} e^{\mathrm{B}\cdot\hat{\mathrm{t}}_k(m')}}$, $p_m^E = \frac{e^{\mathrm{E}\cdot\hat{\mathrm{t}}_k(m)}}{\sum_{m'=1}^{M} e^{\mathrm{E}\cdot\hat{\mathrm{t}}_k(m')}}$ respectively, where B and E are the learnt vectors and $t_k(m)$ is the token representation for the m^{th} token in the k^{th} sequence.

- Cross Entropy loss is computed for both, B and E as:

$$\mathcal{L}_B = -\sum_{M} \mathbb{1}\{m = m_B\} \log p_m^B \quad , \quad \mathcal{L}_E = -\sum_{M} \mathbb{1}\{m = m_E\} \log p_m^E$$

- The final loss is $L_{ans} = \frac{1}{2}(L_B + L_E)$.

- Invalid predictions are discarded at testing time. Examples:
  - predicted span overlaps with the question part of the sequence
  - end token comes before the begin token

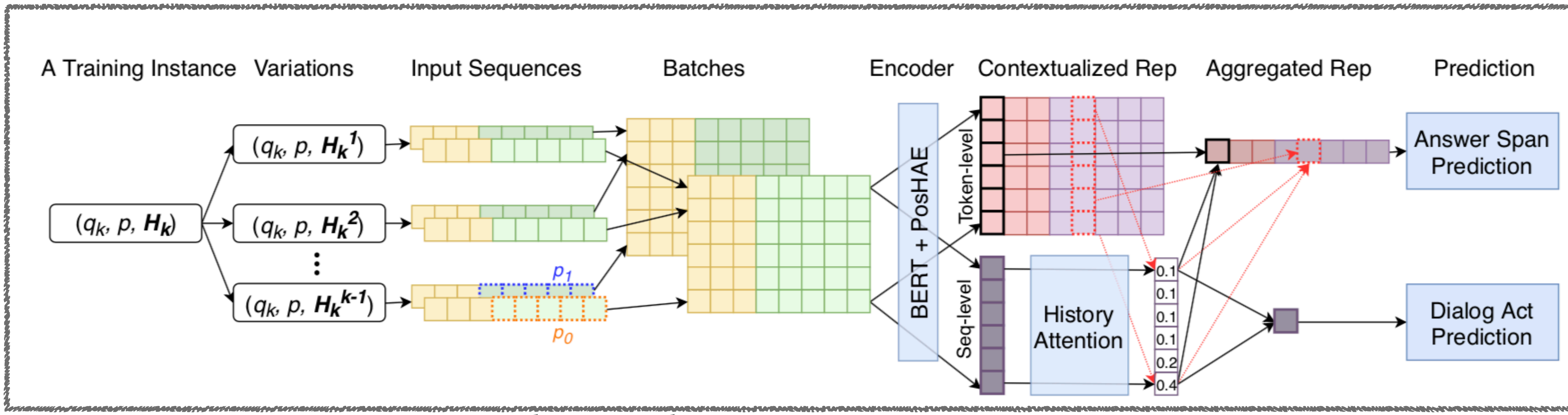
PROPOSED METHOD 3 – MULTI TASK LEARNING (2)

- Dialog Act Prediction: Two sets of parameters $A \in R^{|V_a| \times h}$ and $C \in R^{|V_a| \times h}$ are learnt predict the dialog act of *affirmation* and *confirmation* respectively. $|V_a|$ and $|V_c|$ denote number of classes.
- Affirmation Classes: Yes, No, Cannot Say
- Confirmation Classes: Drill Down, Topic Shift, Topic Return
- This is an independent predictor that does not consider conversation history.
- We calculate cross entropy loss for both *Affirmation* and *Confirmation* as L_A and L_C.

TRAINING

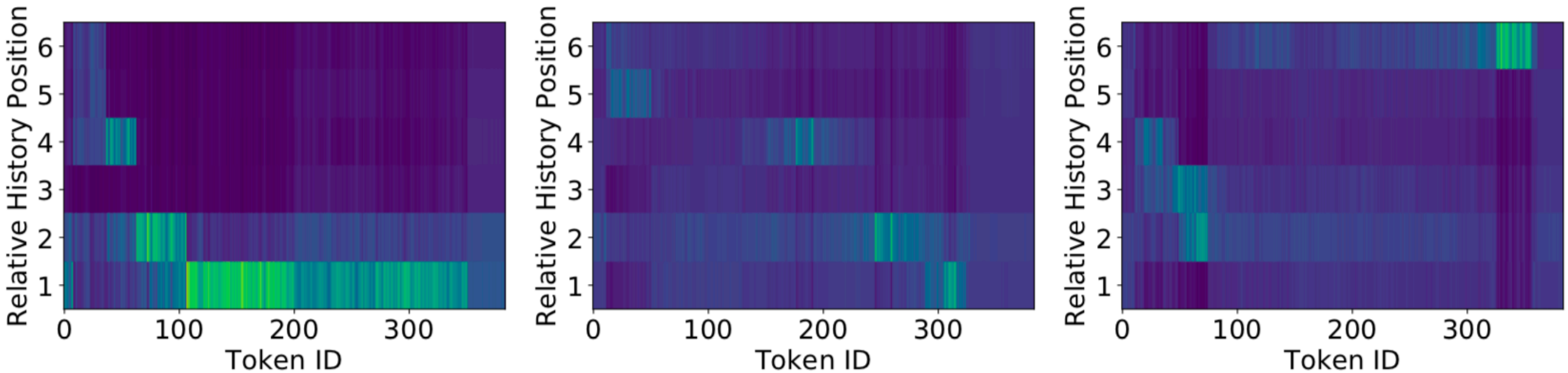
- Hyper parameters λ and μ are used combine the losses of both the tasks: $L = \mu L_{ans} + \lambda L_A + \lambda L_C$
- Advantages:
  - Two tasks provide more supervising signals to fine-tune the encoder.
  - Representation learning benefits from regularisation effect by optimising for multiple tasks.

COMBINED MODEL REPRESENTATION



End to End System Representation

ATTENTION VISUALIZATION



(a) Drill down **(b) Topic shift** **(c) Topic return**

- Brighter spots mean higher attention weights.
- Token ID refers to the token position in an input sequence. A sequence contains 384 tokens.
- Relative history position refers to the difference of the current turn # with a history turn #. The selected examples are all in the 7th turn.
- Dialog Acts (Confirmation):
  - **Drill Down** : the current question is a request for more information about a topic being discussed
  - **Topic Shift** : the current question is not immediately relevant to something previously discussed
  - **Topic Return** : the current question is asking about a topic again after it had previously been shifted away from

EXPERIMENTATION & EVALUATION

- Data: QuAC (Question Answering in Context) dataset
  - Designed for modelling and understanding information-seeking conversations
  - Contains interactive dialogs between an information-seeker and an information provider
  - Information-seeker tries to learn about a hidden Wikipedia passage by asking a sequence of freeform questions
  - Dialog data contains dialog act information
  - Questions are more open-ended, unanswerable, or only meaningful within the dialog context

Items	Train	Validation
# Dialogs	11,567	1,000
# Questions	83,568	7,354
# Average Tokens Per Passage	396.8	440.0
# Average Tokens Per Question	6.5	6.5
# Average Tokens Per Answer	15.1	12.3
# Average Questions Per Dialog	7.2	7.4
# Min/Avg/Med/Max History Turns Per Question	0/3.4/3/11	0/3.5/3/11

EXPERIMENTATION & EVALUATION

- **Key take-aways:**
  - Bert+PosHAE has better training efficiency and performance that FlowQA
  - HAM performs better than BERT + PosHAE
  - Applying BERT-Large to HAM substantially improves answer-span prediction. A more powerful encoder can boost the performance.

Models	F1	HEQ-Q	HEQ-D	Yes/No	Follow up
BiDAF++	51.8 / 50.2	45.3 / 43.3	2.0 / 2.2	86.4 / 85.4	59.7 / 59.0
BiDAF++ w/ 2-C	60.6 / 60.1	55.7 / 54.8	5.3 / 4.0	86.6 / 85.7	61.6 / 61.3
BERT + HAE	63.9 / 62.4	59.7 / 57.8	5.9 / 5.1	N/A	N/A
FlowQA	64.6 / 64.1	– / 59.6	– / 5.8	N/A	N/A
BERT + PosHAE	64.7 / –	60.7 / –	6.0 / –	N/A	N/A
HAM	65.7‡ / 64.4	62.1 / 60.2	7.3 / 6.1	**88.3** / **88.4**	62.3 / **61.7**
HAM (BERT-Large)	**66.7**‡ / **65.4**	**63.3** / **61.8**	**9.5** / **6.7**	88.2 / 88.2	**62.4** / 61.0

ABLATION ANALYSIS

- **Performance Drop:**
  - By replacing fine-grained history attention with sequence- level history attention
  - By disabling the history attention module, performance drops dramatically for 4.6% and 3.8%
  - Disabling history attention also hurts the performance for dialog act prediction
  - Removing the answer span prediction task, a relatively large performance drop for dialog act prediction is observed

- **Performance Increase:**
  - Removal of the dialog act prediction task results in a slight and insignificant increase in the performance for answer span prediction.
  - The encoder benefits from a regularisation effect because it is optimised for two different tasks and thus alleviates overfitting.

REFERENCES

- Qu, Chen, et al. "Attentive History Selection for Conversational Question Answering." Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019.
- Qu, Chen, et al. "BERT with History Answer Embedding for Conversational Question Answering." Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019.
- C. Zhu, M. Zeng, and X. Huang. SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering. CoRR, 2018.
- Choi, Eunsol, et al. "Quac: Question answering in context." arXiv preprint arXiv: 1808.07036 (2018)
- P. Rajpurkar, R. Jia, and P. Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In ACL, 2018.
- C. Qu, L. Yang, W. B. Croft, J. R. Trippas, Y. Zhang, and M. Qiu. Analyzing and Characterizing User Intent in Information-seeking Conversations. In SIGIR, 2018.
- H.-Y. Huang, E. Choi, and W. Yih. FlowQA: Grasping Flow in History for Conversational Machine Comprehension. CoRR, 2018.
- Tuason, Ramon, Daniel Grazian, and Genki Kondo. "Bidaf model for question answering." Table III EVALUATION ON MRC MODELS (TEST SET). Search Zhidao All (2017).

THANK YOU