MACQUARIE
University

A SHORT SURVEY OF PRE-TRAINED LANGUAGE MODELS FOR CONVERSATIONAL AI- A NEW AGE IN NLP

Munazza Zaib[1], Prof. Dr. Michael Sheng[1], Dr. Wei Zhang[2]

Macquarie University[1], The University of Adelaide[2]

Table of Contents

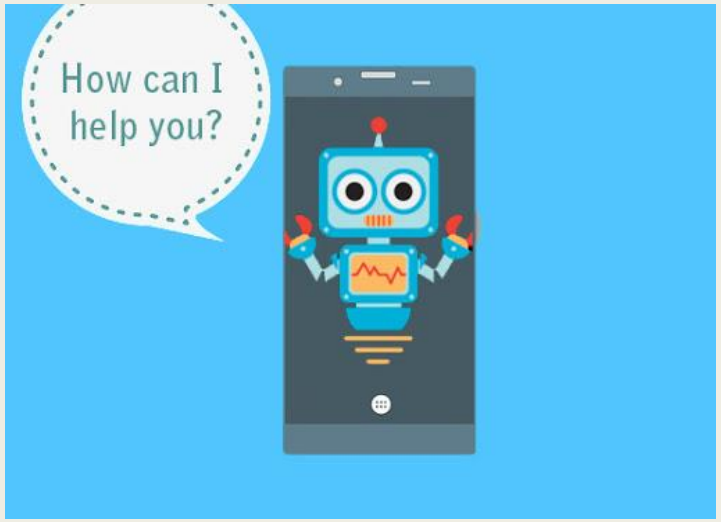
Introduction

- Building a dialogue system that can communicate naturally with humans is a challenging yet interesting problem of agent-based computing.
- The rapid growth in this area is usually hindered by the long-standing problem of data scarcity.
- The recently introduced pre-trained language models have the potential to address the issue of data scarcity and bring considerable advantages by generating contextualized word embeddings.

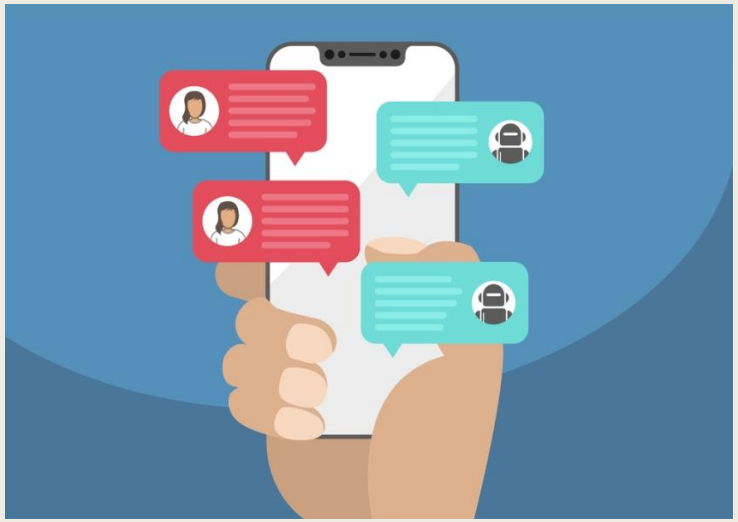
Introduction

- Based on functionality, conversational AI can be categorized into three categories:

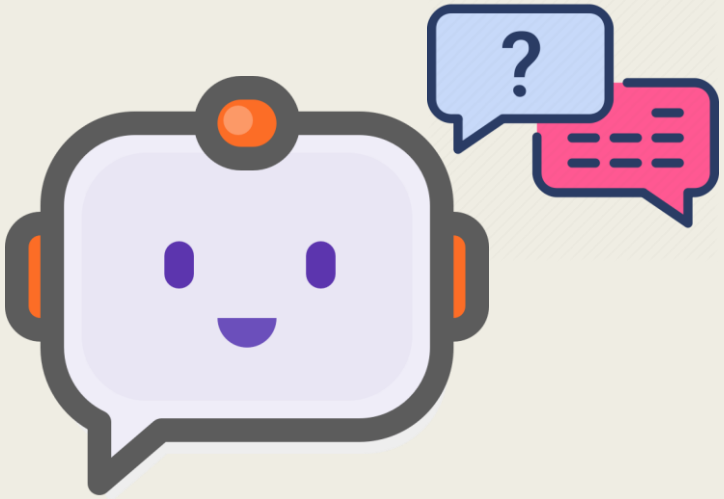
Task-oriented Dialog System



Chat-oriented Dialog System



QA Dialog System



Introduction

- Though, these systems have come a long way in terms of progress but conversing with such models for even a short amount of time quickly unveils the inconsistency in generated responses.
- A different number of strategies have been introduced over a period to address this issue such as word embeddings like GloVe or word2vec.
- However, these techniques failed to capture the correct context of the word used in a sentence.
- For example “an apple a day, keeps the doctor away" and “I own an Apple Macbook, two “apple" words refer to very different things but they would still share the same word embedding vector.

Pre-trained Language Modeling

- Pre-trained language modeling can be considered an equivalent of ImageNet in NLP and has achieved state-of-the-art results on various downstream NLP tasks such as sentiment analysis, data classification, and question answering, etc.
- The timeline of pre-trained language models can be divided into two categories:

Approaches Used

➢ Feature Based:

- The pre-trained word embeddings provides an edge to modern NLP systems over the embeddings learned from scratch and can be word level, sentence level, or paragraph level based on their granularity.
- These learned representations are also utilized as features in NLP downstream tasks.

Embeddings from Language Models (ELMo) revolutionized the concept of general word embeddings by proposing the idea of extracting the context-sensitive features from the language model.

Approaches Used

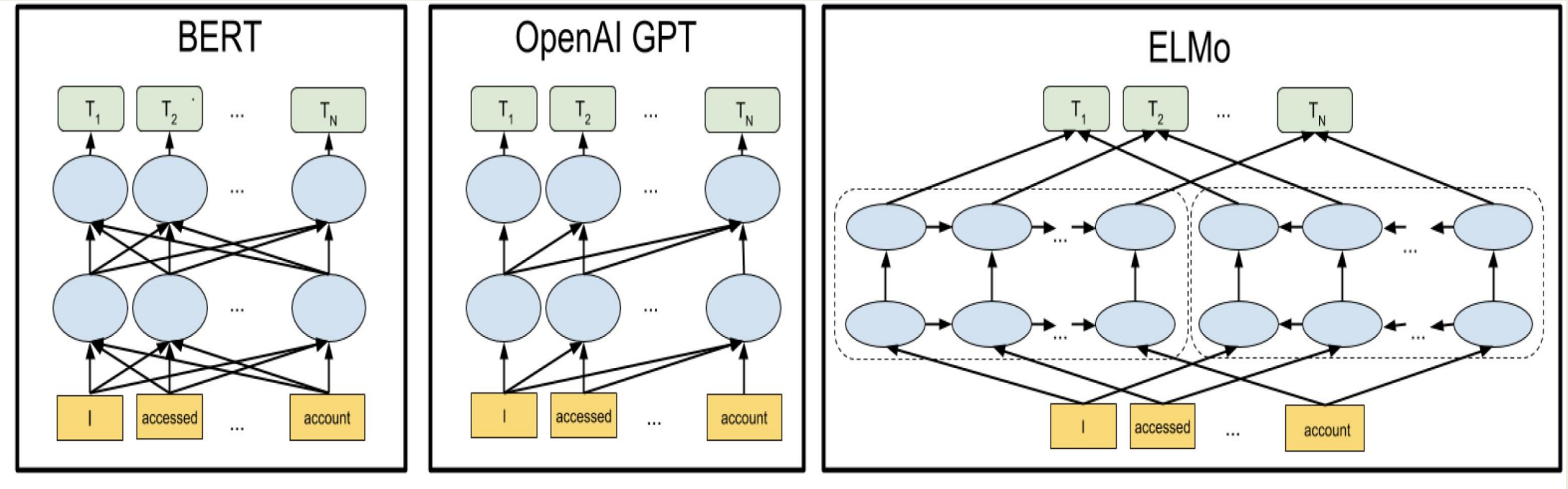
- **Fine-tuning Based:**
- There has been a trend of transfer learning from language models recently.
- The main idea is to pre-train a model on unsupervised corpora and then fine-tune the same model for the supervised downstream task.

These models are the adaptation of Google's Transformer model. First in the series is OpenAI's Generative pre-training Transformer (GPT). The model is trained to predict the words only from left-to-right hence, capture the context unidirectionally.

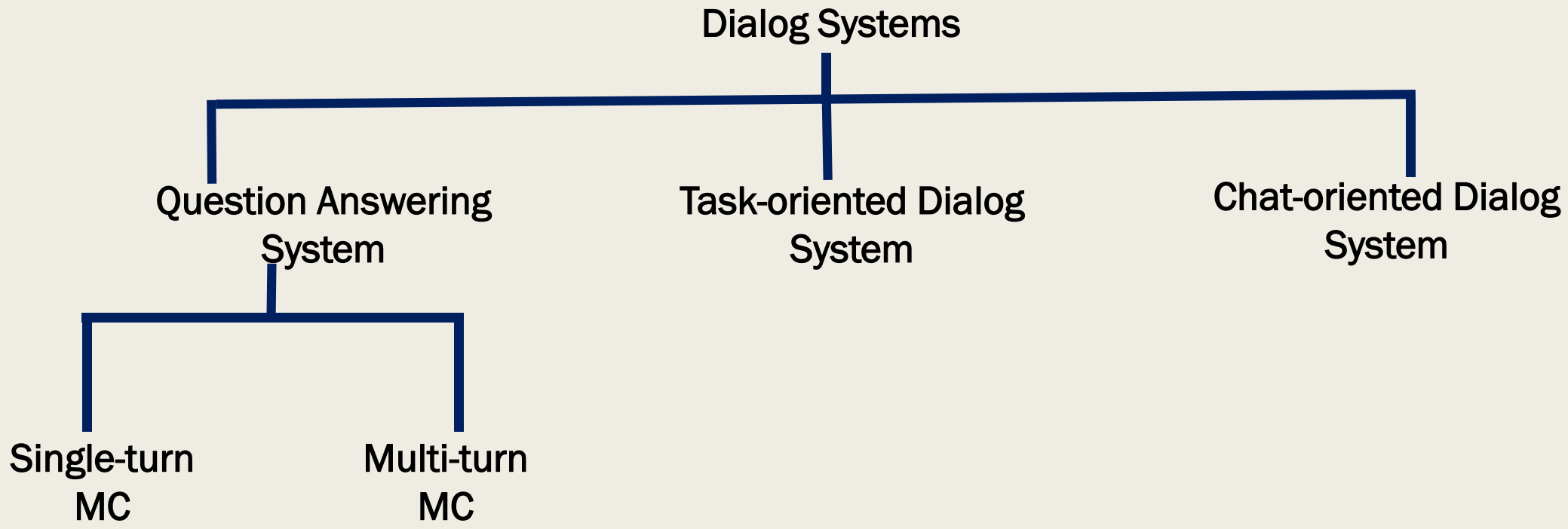
Bi-directional Encoder Representations from Transformers (BERT) addressed the unidirectional constraint by integrating the concept of 'Masked Language Model' in their model that randomly masks some of the input tokens and the goal is to predict the masked token based on the captured context from both directions.

Later, the predecessor of GPT, GPT2 was introduced. he model is bidirectionally trained on 8M web pages and has 1.5B parameters, 10 times greater than the original model. The model is based on the Transformer's decoder and is designed to generate language naturally.

Architectural Difference of Language Models



PRE-TRAINED LANGUAGE MODELING APPROACHES FOR DIALOGUE SYSTEMS



Question Answering Dialog Systems

➢ Single-turn MC:

- Benchmark dataset= SQuAD consisting of Wikipedia articles and questions posed on those articles by a group of co-workers.
- The system must select the right answer span for the question from all the possible answers in the given passage.

End of Conversation

Literature Survey

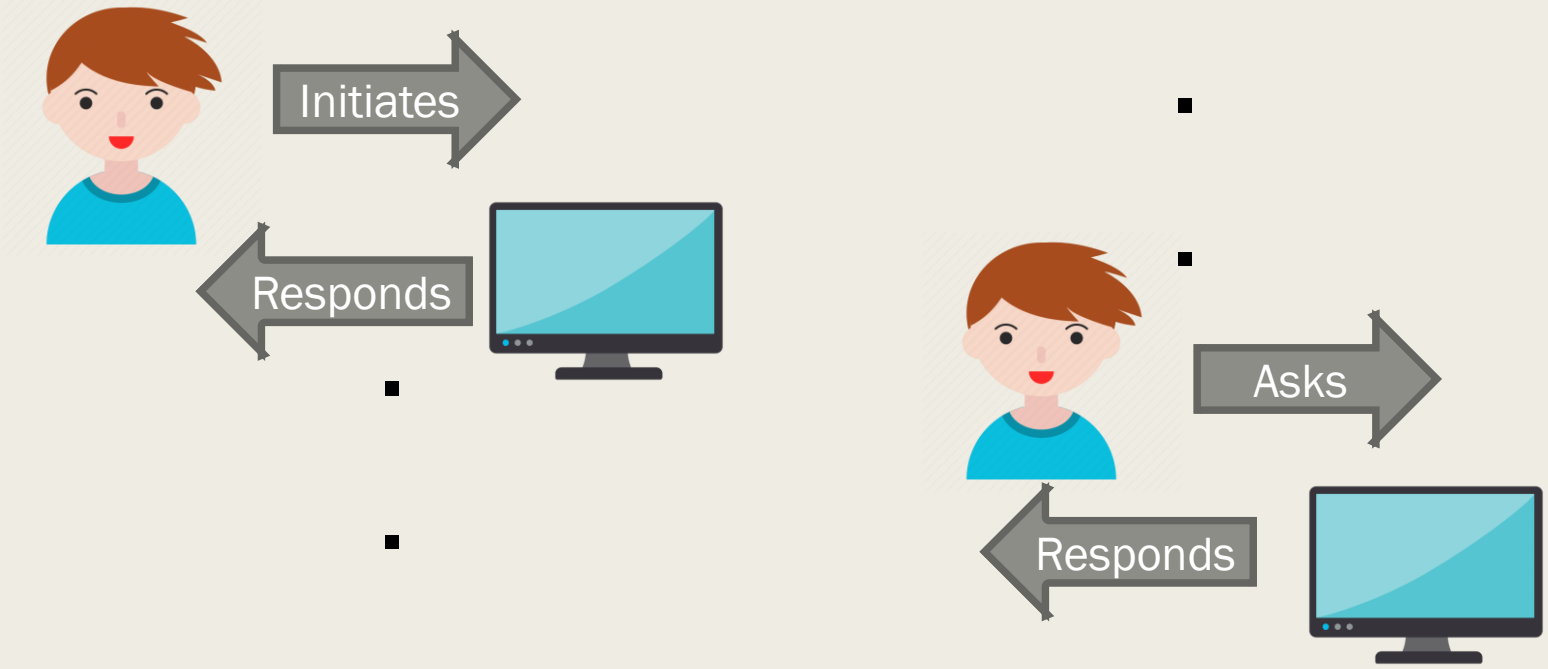
- Word embeddings from ELMo are used with BiLM and BiDAF setup to achieve better accuracy scores[1].
- When used in the architecture setting of BERT, the accuracy improved drastically and it broke all the records of language understanding in 11 NLP downstream tasks[2].
- Currently, the highest score on SQuAD leaderboard[3] is taken by model based on XLNet.

1. Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional Attention Flow for Machine Comprehension.
2. 2. Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019. Semantics-aware BERT for Language.
3. 3.https://rajpurkar.github.io/SQuAD-explorer/

Question Answering Dialog Systems

➢ Multi-turn MC:

- It combines the elements of chit-chat and question answering.
- High-quality conversational datasets such as QuAC [4] and CoQA [16] have provided the researchers a great source to work deeply in the field of CMC.

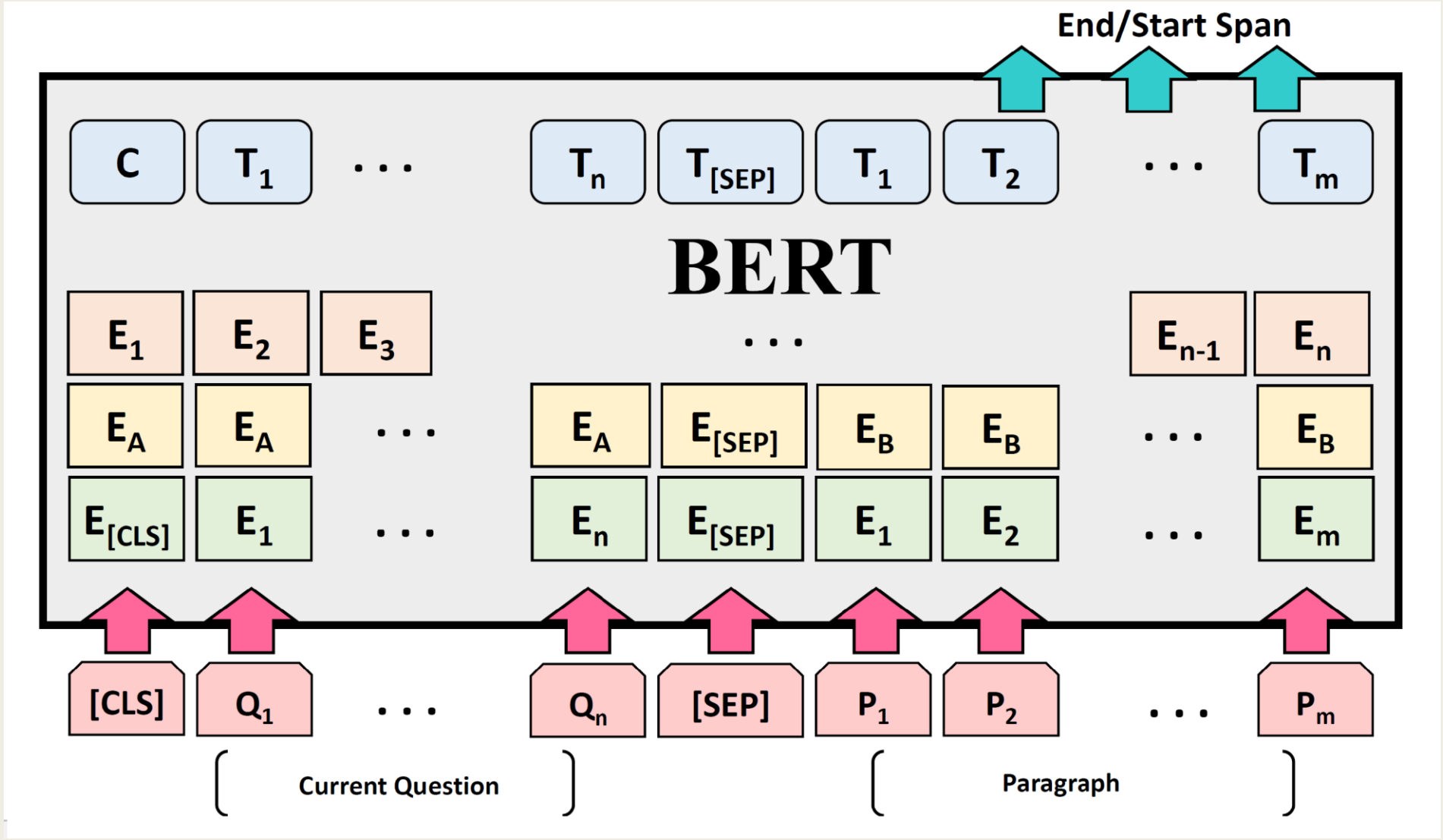


Literature Survey

- The first BERT based model for QuAC [4] was based on history answer embeddings to provide extra information to input tokens.
- Later, researchers [5] improved accuracy by introducing the last two contexts when answering the current question.
- Another research [6] introduced the reasoning process in BERT-based architecture.

4.https://quac.ai/
5. Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A Simple but Effective Method to Incorporate Multi-turn.
6. Yi Ting Yeh and Yun-Nung Chen. 2019. FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension.

Adapting BERT for Multi-turn QA Task

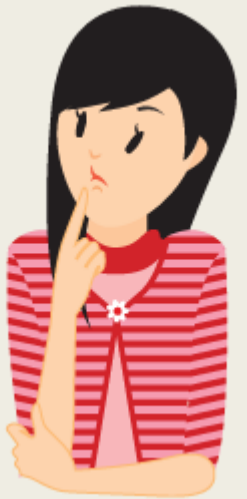


Other Dialog Systems

i. Task-oriented Dialog Systems:

- The goal of these systems is to assist the users by generating a valuable response.
- This response generation requires a considerable amount of labeled data for the training purpose.

Can we take advantage of transfer learning through pre-trained language models to enable the modeling of task-oriented systems?



Literature Survey

- The question has been addressed by [7] which introduces a GPT based framework to evaluate the ability to transfer the generation capability of GPT to task-specific multi-domains.
- Another work [8] recently utilized the strengths of BERT in improving the scalability of DST module. The DST module is use to maintain the state of user's intentions through out the dialogue.

7. Pawel Budzianowski and Ivan Vulic. 2019. Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented.
8. Guan-Lin Chao and Ian Lane. 2019. BERT-DST: Scalable End-to-End Dialogue State Tracking with Bidirectional Encoder Representatation from Transformer. In Proc. Interspeech 2019. https://doi.org/10.21437/interspeech.2019-1355

Literature Survey

➢ Chat-oriented Dialog Systems:

- Chat-oriented dialogue systems are known to have several issues such as they are often not very engaging and lack specificity.
- To address these problems, **TransferTransfo**, a persona-based model, [9] is introduced.
- They have extended the transfer learning from language understanding to generative tasks such as open-domain dialogue generation using GPT.
- They have addressed the above-mentioned issues by combining many linguistics aspects such as common-sense knowledge, co-reference resolution, and long-range dependency.

9. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Empathic Computing

Open Challenges

User Privacy

Lack of Inference Capability

Conclusion:

In recent years, the promising notion of pre-trained language models has gained widespread attention by researchers.

This paper is an effort to investigate the recent trends introduced in language models and their application to the dialogue systems.

Whilst these models tend to perform well, but, still there are some issues that needs to be addressed.


That's all Folks!