Department of Computer Science and Engineering (CSE)
IIT Hyderabad

September 6, 2022

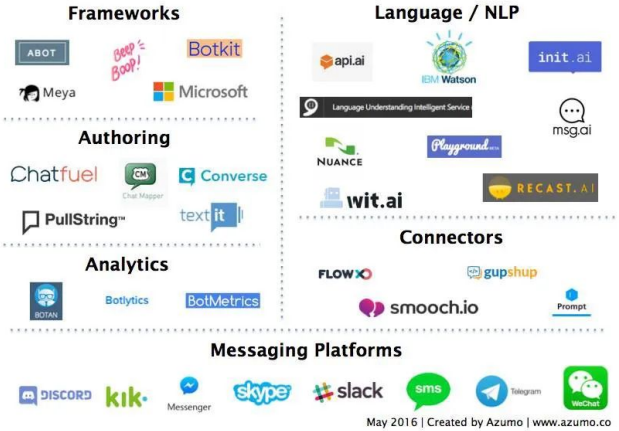
Conditional Natural Language Generation for Dialogue Systems and Recommendation Engine (Y4J Platform)

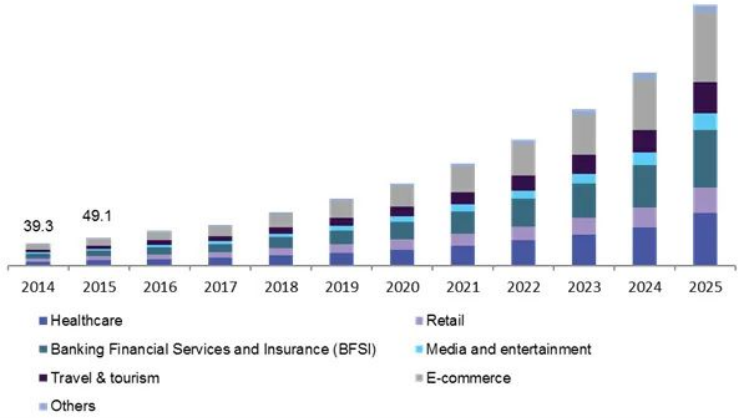
Thesis Stage-1 Presentation

Kamal Shrestha, cs21mtech16001

Thesis Supervisor: **Dr. Maunendra Sankar Desarkar**

The global chatbot market size of chatbots is projected to expand to **3,411 million dollars by 2030.**





Market	Chatbot Market
Market Size 2021	USD 521 Million
Market Forecast 2030	USD 3,411 Million
CAGR During 2022 - 2030	23.7%
Analysis Period	2018 - 2030
Base Year	2021
Forecast Data	2022 - 2030
Segments Covered	By Type, By Application, By Product Landscape, By End-User, And By Geography
Regional Scope	North America, Europe, Asia Pacific, Latin America, and Middle East & Africa
Key Companies Profiled	Facebook, Inc., Pandorabots, Inc., ToyTalk (PullString Inc.), Google, Inc., Yahoo Inc., Haptik, Inc., Helpshift, Kasisto Inc., Slack Technologies, Inc., Astute Solutions, Kiwi, Inc., and Microsoft Corporation.
Report Coverage	Market Trends, Drivers, Restraints, Competitive Analysis, Player Profiling, Regulation Analysis
Customization Scope	10 hrs of free customization and expert consultation

Conditional Natural Language Generation for Dialogue Systems

Dialogue History

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch GOT
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?

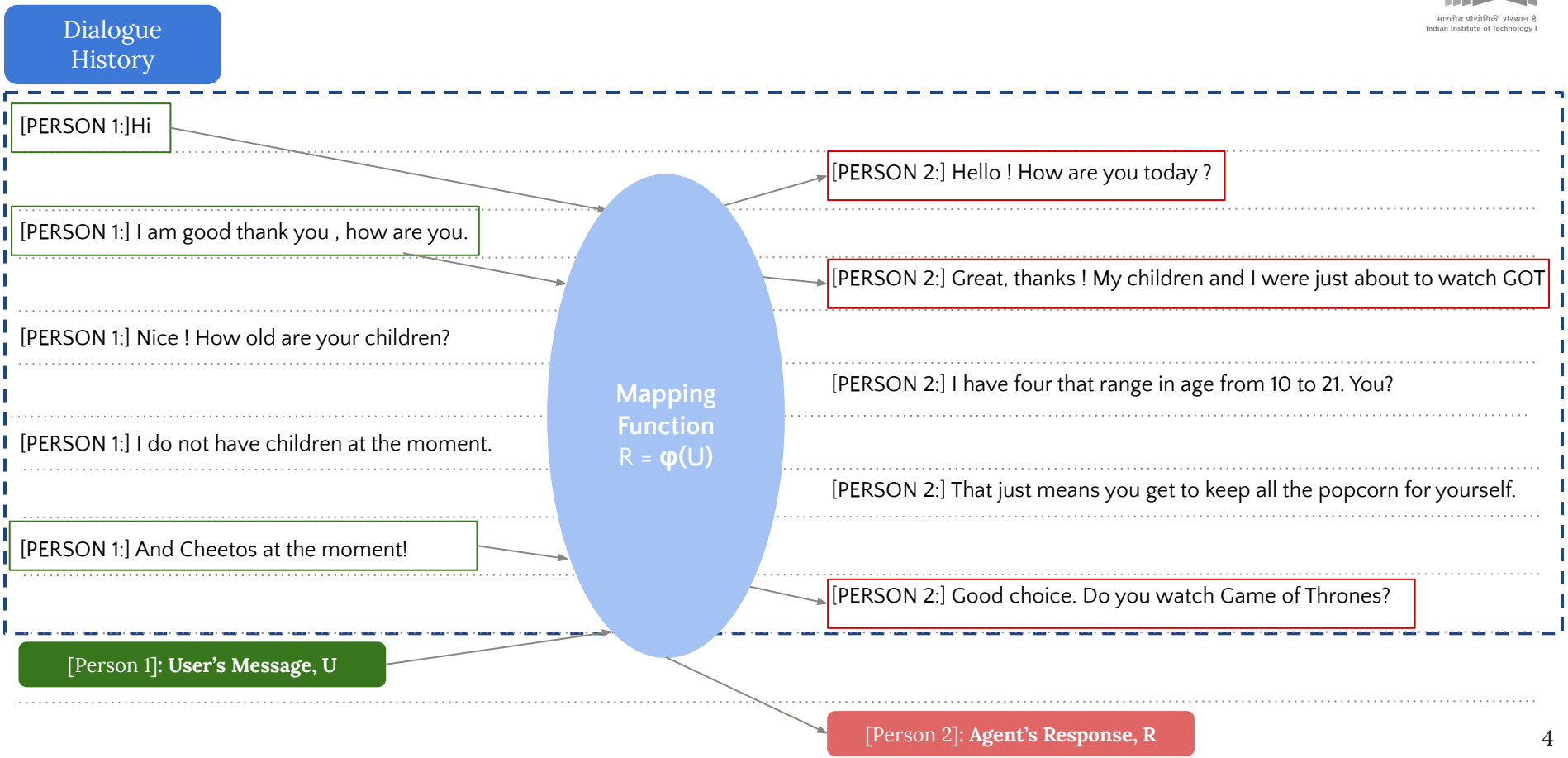
User's Message

[PERSON 1:] No, I do not have much time for TV.

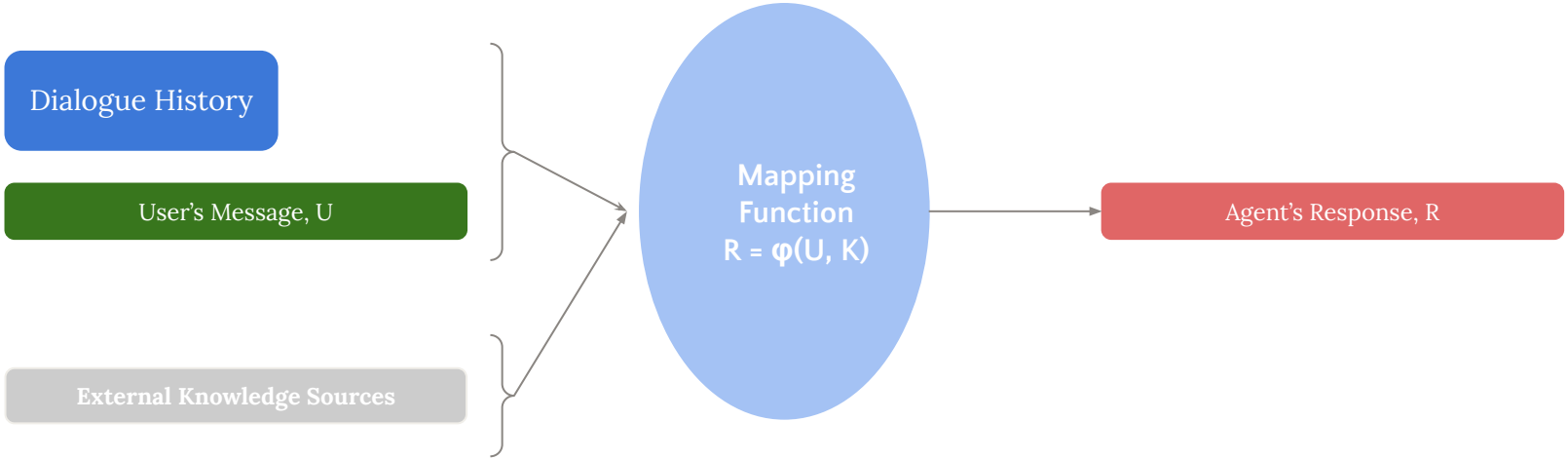
Agent's Response

[PERSON 2:] I usually spend my time painting: but, I love the show.

A dialogue system can be seen as a **mapping function, φ,** from **user's message U,** to **agent's response, R**



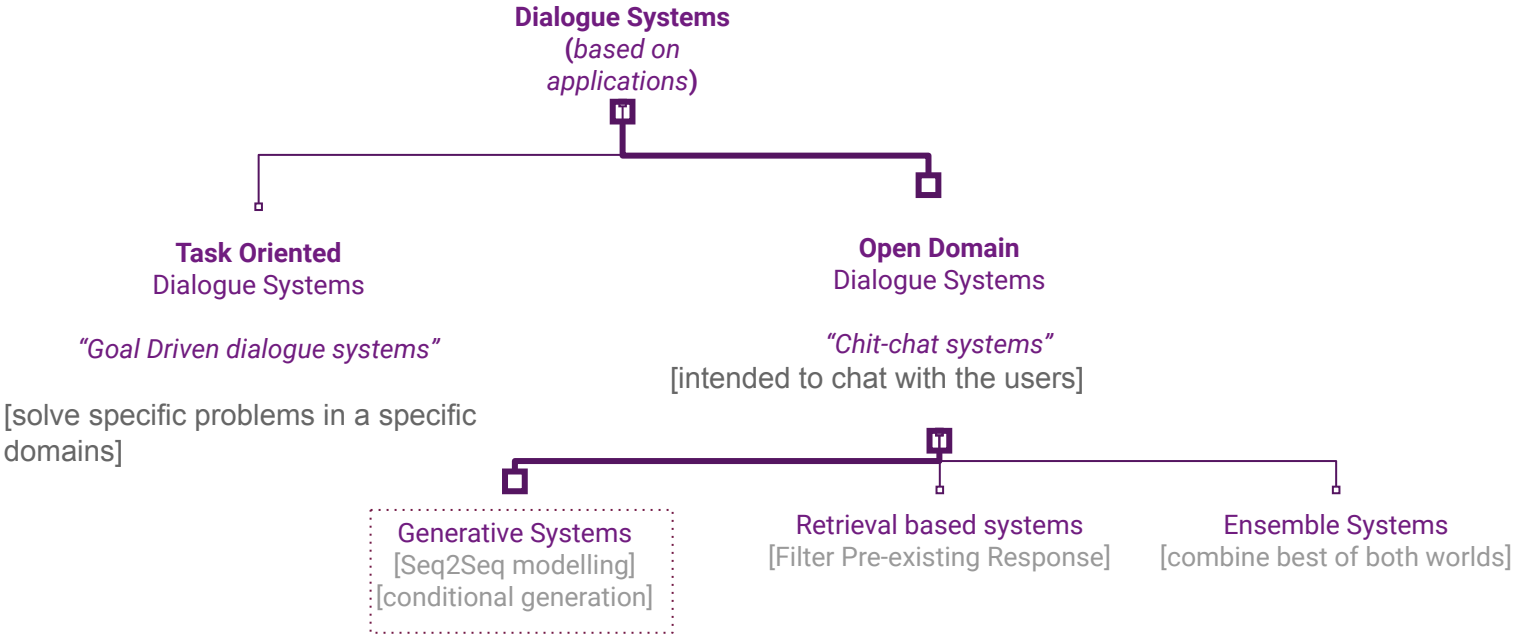
Knowledge Grounded Systems use an external knowledge such as common-sense knowledge as a significant source of information when organizing an utterance.



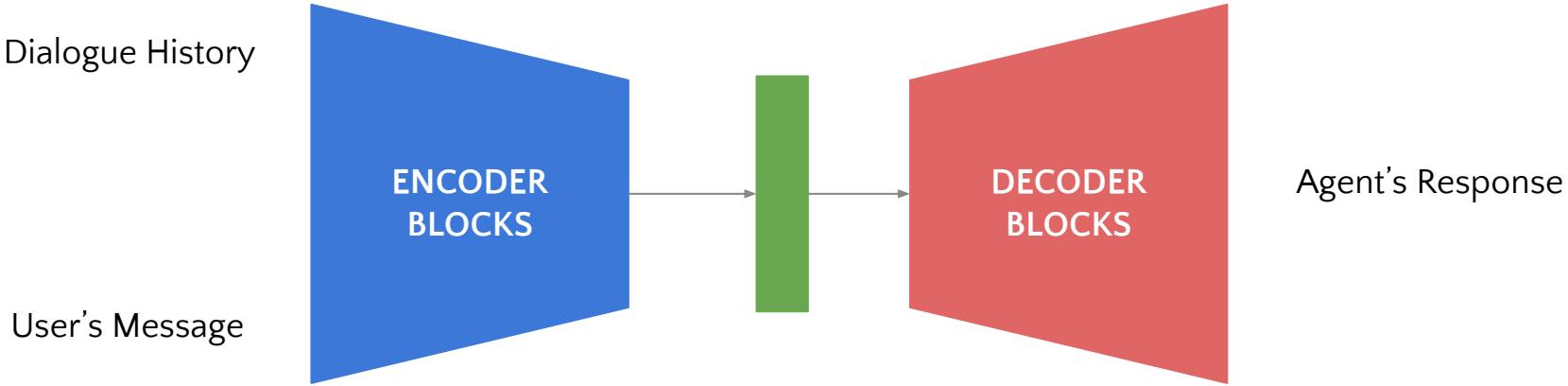
Dialogue Systems with external knowledge (k)

$$U = \{u^{(1)}, u^{(2)}, \ldots, u^{(i)}\},\ K \longrightarrow \mathrm{R} = \varphi(U, K) \longrightarrow \mathrm{R} = \{\mathrm{r}^{(1)}, r^{(2)}, \ldots, r^{(j)}\}$$

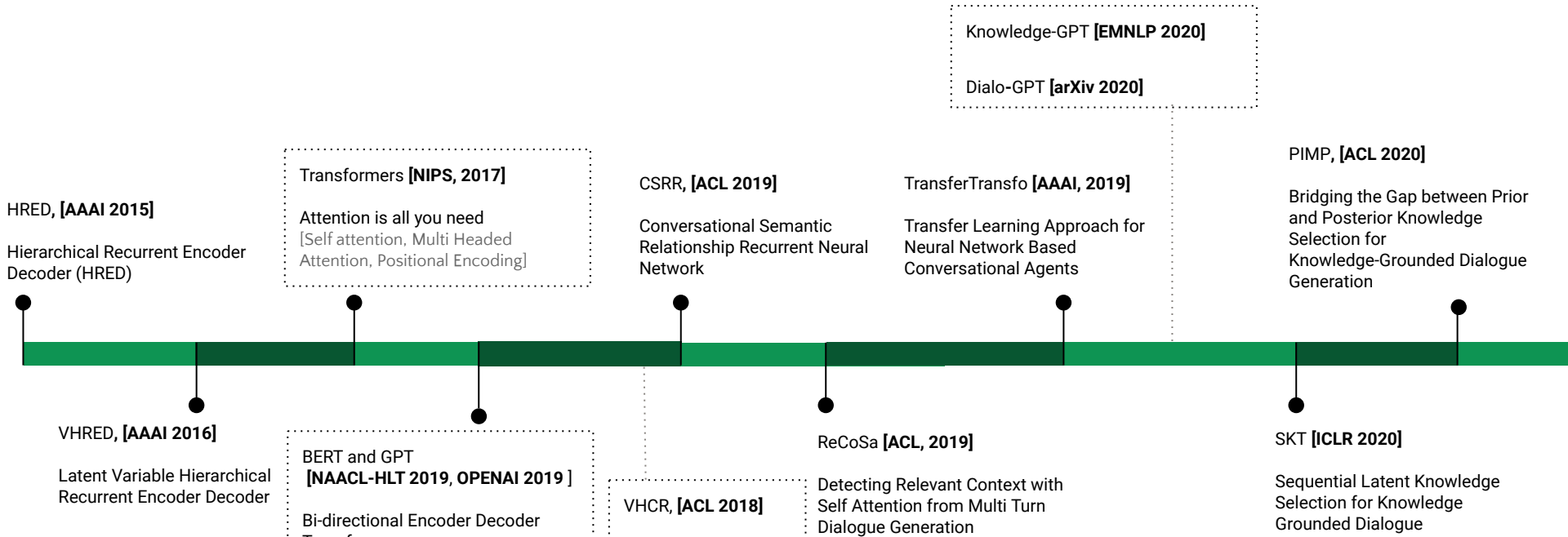
On the basis of **use cases and applications,** dialogue systems are divided into two types: **Task oriented and open domain.**



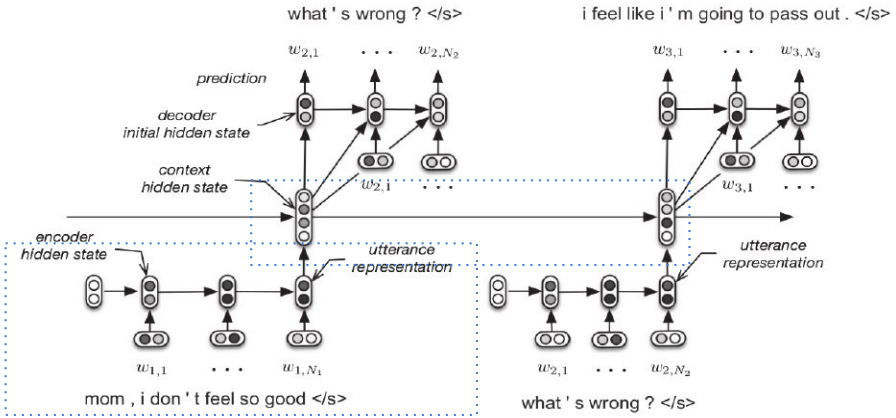
Encoder decoder architecture for conditional natural language generation



Literature Survey and SOTA results, discussions and implementations



In HRED, the context RNN allows the model to represent a form of common ground between speakers.



- <u>Preceding to HRED:</u> Traditional Dialogue system (RNNLM / LSTM based models) used single turn history to generate responses

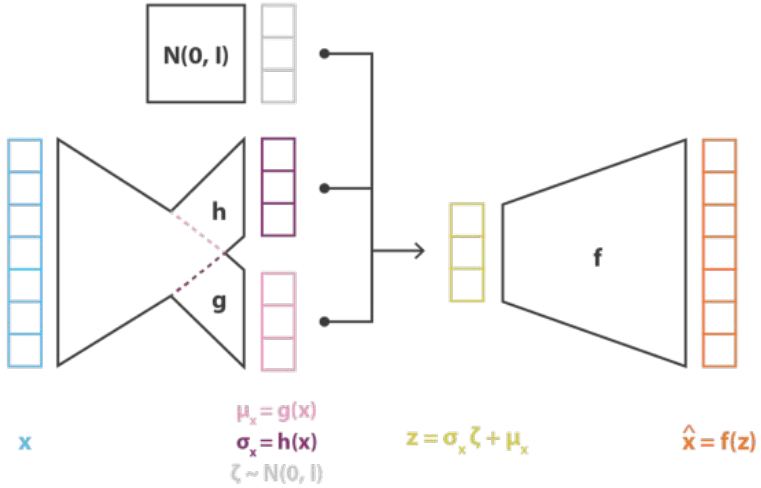
- **ENCODER:**
  The Dialogue Context awareness was generated at two levels: [RNN used in both the levels]
  - each utterance is encoded into a dense vector and then mapped into the dialogue context
  - higher-level context RNN keeps track of past utterances by processing iteratively each utterance vector [continuous-valued state of the dialogue system]

- **DECODER:**
  Greedy Search Decoding [Probability Distribution on |V| and Maximization of Log Likelihood]

- Superiority over standard RNN because the context RNN allows the model to represent a form of common ground between speakers, e.g. to represent topics and concepts shared between the speakers using a distributed vector representation.

Serban, I., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2015, March). Building end-to-end dialogue systems using generative hierarchical neural network models.

Models HRED and all preceding it suffered from **a deficient generation problem** of generating meaningful dialogue utterances.



Minimization of reconstruction loss and KL divergence

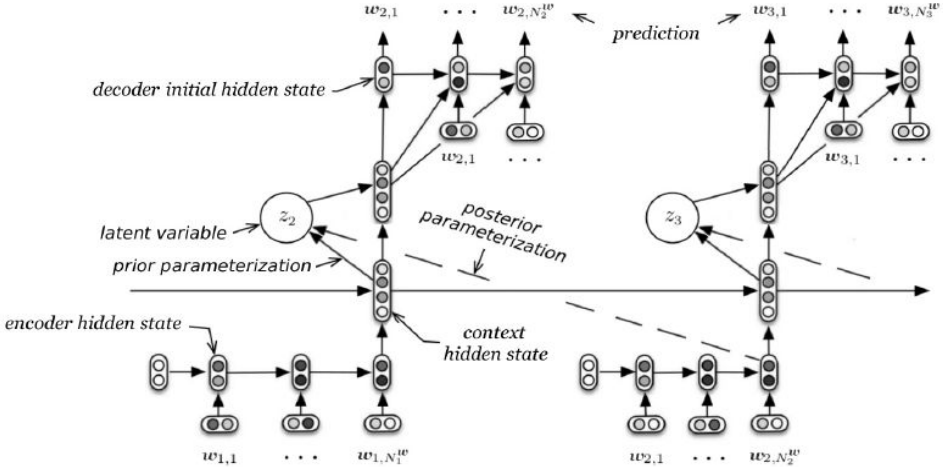
$$\text{loss} = C \| x - \hat{x} \|^2 + KL[\, N(\mu_x, \sigma_x), N(0, I) \,] = C \| x - f(z) \|^2 + KL[\, N(g(x), h(x)), N(0, I) \,]$$

Maximization of variational lower-bound

$$(f^*, g^*, h^*) = \underset{(f,g,h) \in F \times G \times H}{\arg\max} \left(\mathbb{E}_{z \sim q_x} \left(-\frac{||x - f(z)||^2}{2c} \right) - KL(q_x(z), p(z)) \right)$$

Likelihood Maximization — variational inference

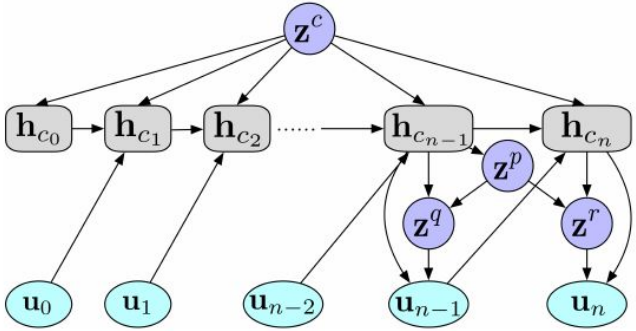
VHRED[2016]: models hierarchically-structured sequences **in a two-step generation process**—first sampling the latent variable, and then generating the output sequence



- Preceding to VHRED [in RNNLM and HRED]: **The Restricted Shallow Generation Process**
- Maximizing a variational lower-bound on the log-likelihood. [**Prior Distribution:** Concatenated (encoder + context) hidden state at M utterance], [**Posterior Distribution:** encoder hidden state at M+1 utterance]
- **DECODER:**
  - **Two-step generation process**—first sampling the latent variable, and then generating the output sequence
  - Greedy Search Decoding [Probability Distribution on |V| and Maximization of Log Likelihood]
- Randomness injected by the variable **z** corresponds to higher-level decisions, like topic or sentiment of conversation.

Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., & Bengio, Y. (2017, February). A hierarchical latent variable encoder-decoder model for generating dialogues.

CSRR[2019]: More variations are imported into utterance level to help generate more diverse responses.



- **Preceding to CSRR [HRED, VHRED]:** Do not model meaning of each utterance explicitly, rather summarize the meaning when needed with no guarantee that inferred meaning is adequate to the original utterance.

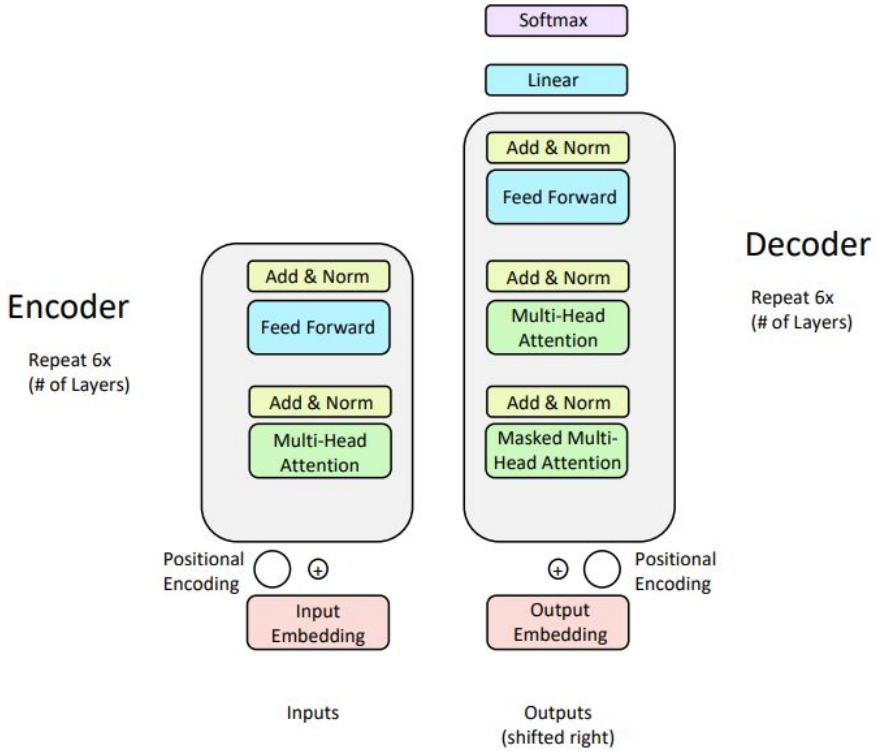
- Add more variations of utterance level for more better responses [not just on the decoding level]
  - **[Discourse Level]:**
    Models the background of the conversation. **{Zc}**
  - **[Pair Level]:**
    Models the **consistent semantics between query and response** [topic of conversation] with a common latent variable shared by the query and response pair. **{Zp}**
  - **[Utterance Level]:**
    Models the specific meaning of the query and the response with a certain latent variable for each of them to capture the content difference. **{Zq, Zr}**

- Significantly improves the quality of responses in terms of fluency, coherence and diversity

Brauweiler, C., & Yerim Sheva, A. (2019). CSR and CSRR in Kazakhstan: state and prospects. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu

Meanwhile,

- Transformers [2017]
  [Self attention, Multi Headed Attention, Positional Encoding]
- Embeddings from Language Model (ELMo) [2018]
  [Contextualized Word Embeddings]
- BERT [2018]
  [Bidirectional context encoding]
- GPT, GPT-2 [2018]
  [Generative Task based]

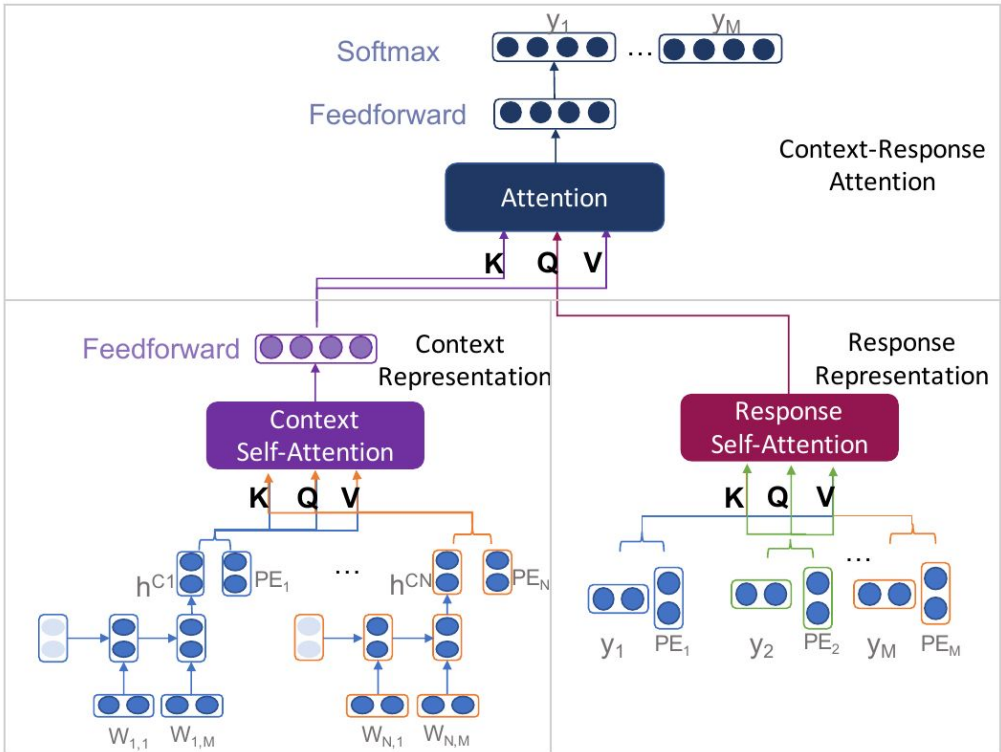


[ReCoSa,2019]: Ideally model should be able t**o detect these relevant contexts** and produce a suitable response accordingly

Preceding to ReCoSa [RNNLM, HRED, VHRED, CSRR]:
Processed all contexts in the dialogue history indiscriminately, meaning **did not discriminate between the relevant contexts history.**

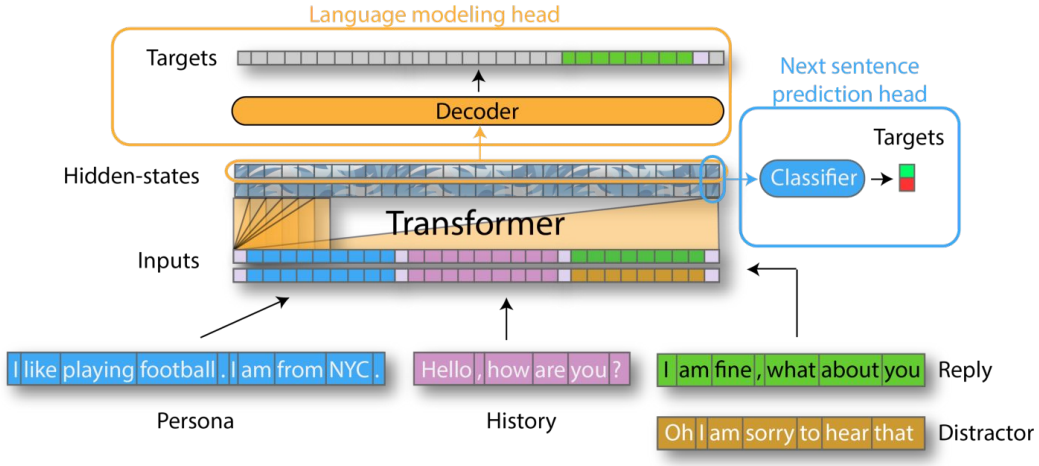
Inspired by Transformer's Self Attention:
How to effectively extract and use the relevant contexts for better encoding of the contexts.

1. Word-level LSTM uses **self-attention mechanism** to encode each context to content attentive representation [Concatenated with PE]
2. Similarly, we get masked response attentive representation
3. Calculate the **attention scores** between the context as **Key, Value** and response representations as **Query**.
4. Greedy Decoding with log likelihood maximization. [Encoder Decoder Attention for decoding]



Zhang, H., Lan, Y., Pang, L., Guo, J., & Cheng, X. (2019). Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation.

TransferTransfo, [2019]: An AI with **personality**

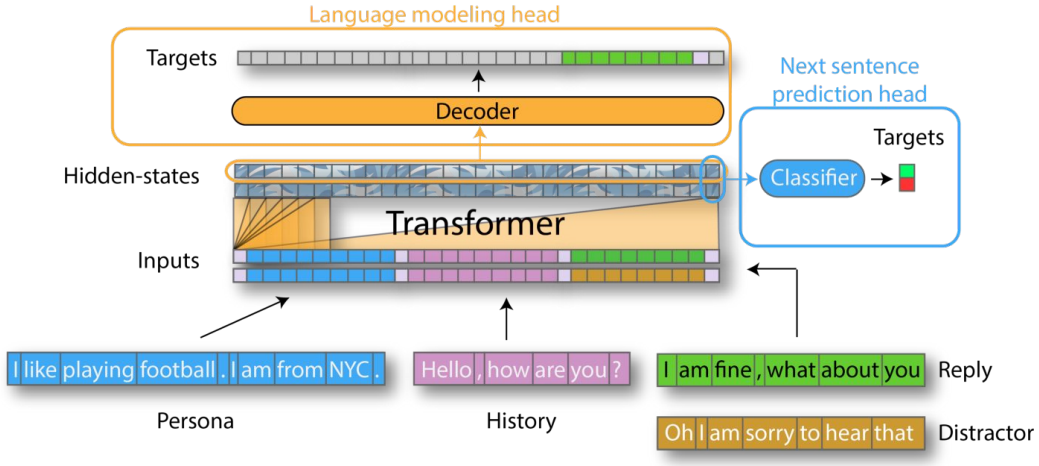


- **Preceding to TransferTransfo:**
  a. Wildly inconsistent outputs and the **lack of a consistent personality**
  b. Tendency to produce consensual and generic responses (e.g. I don't know) which are vague and not engaging for humans

- **Transfer Learning:** Fine Tuned on OpenAI GPT2 [concatenation with Special Tokens and Beam Search Decoding Technique]
- **Training Objectives:** Language Modelling Task [Perplexity], Next Utterance Retrieval Task [Classification], Generation Task [Human Evaluations]
- Dialogue Agent will have a **knowledge base** to store a few sentences describing its **personality traits** and a preceding dialogue history

Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents.

TransferTransfo, [2019]: An AI with **personality**



- <u>Preceding to TransferTransfo:</u>
  a. Wildly inconsistent outputs and the **lack of a consistent personality**
  b. Tendency to produce consensual and generic responses (e.g. I don't know) which are vague and not engaging for humans

- **Transfer Learning:** Fine Tuned on OpenAI GPT2 [concatenation with Special Tokens and Beam Search Decoding Technique]
- **Training Objectives:** Language Modelling Task [Perplexity], Next Utterance Retrieval Task [Classification], Generation Task [Human Evaluations]
- Dialogue Agent will have a **knowledge base** to store a few sentences describing its **personality traits** and a preceding dialogue history

Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents.

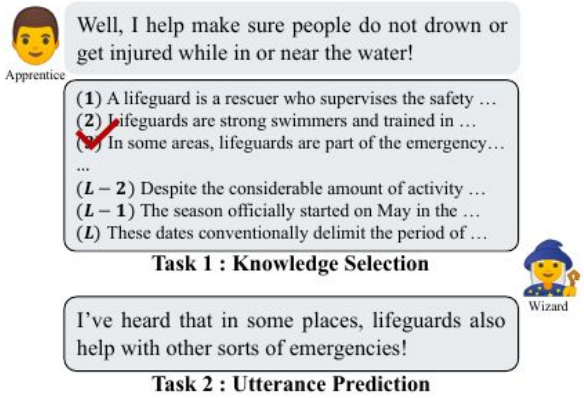
External knowledge such as common-sense knowledge is **a significant source of information** when organizing an utterance.

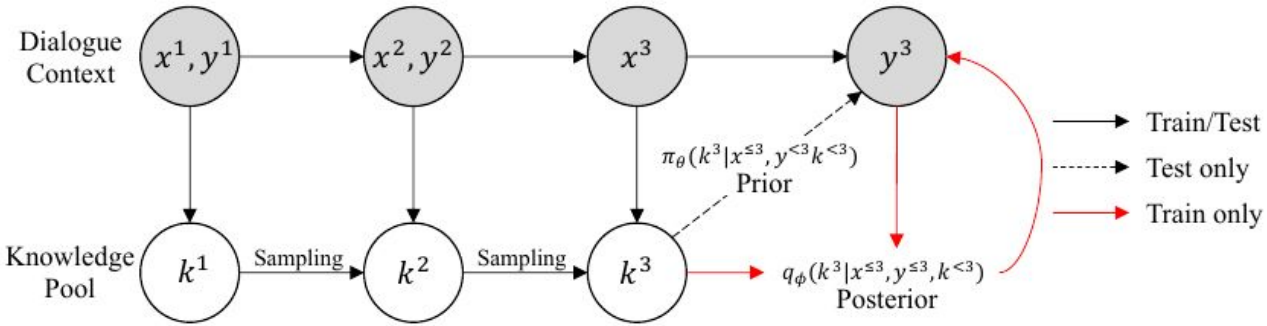
Dialogue Systems with external knowledge (k)

$U = \{u^{(1)}, u^{(2)}, \ldots, u^{(i)}\}$, $K \longrightarrow R = \varphi(U, K) \longrightarrow R = \{r^{(1)}, r^{(2)}, \ldots, r^{(j)}\}$

User message (U)	Agent response (R)	External Knowledge (K)
I need to find a nice restaurant in Madrid that serves expensive Thai food.	There is a restaurant called Bangkok City locating at 9 Red Ave.	restaurant database
I love the grilled fish so much!	Yeah. it's a famous Chinese dish.	knowledge graph

[SKT, 2021]: The diversity of knowledge selection in dialogue [multimodal in nature], is modeled it as **latent variables.**

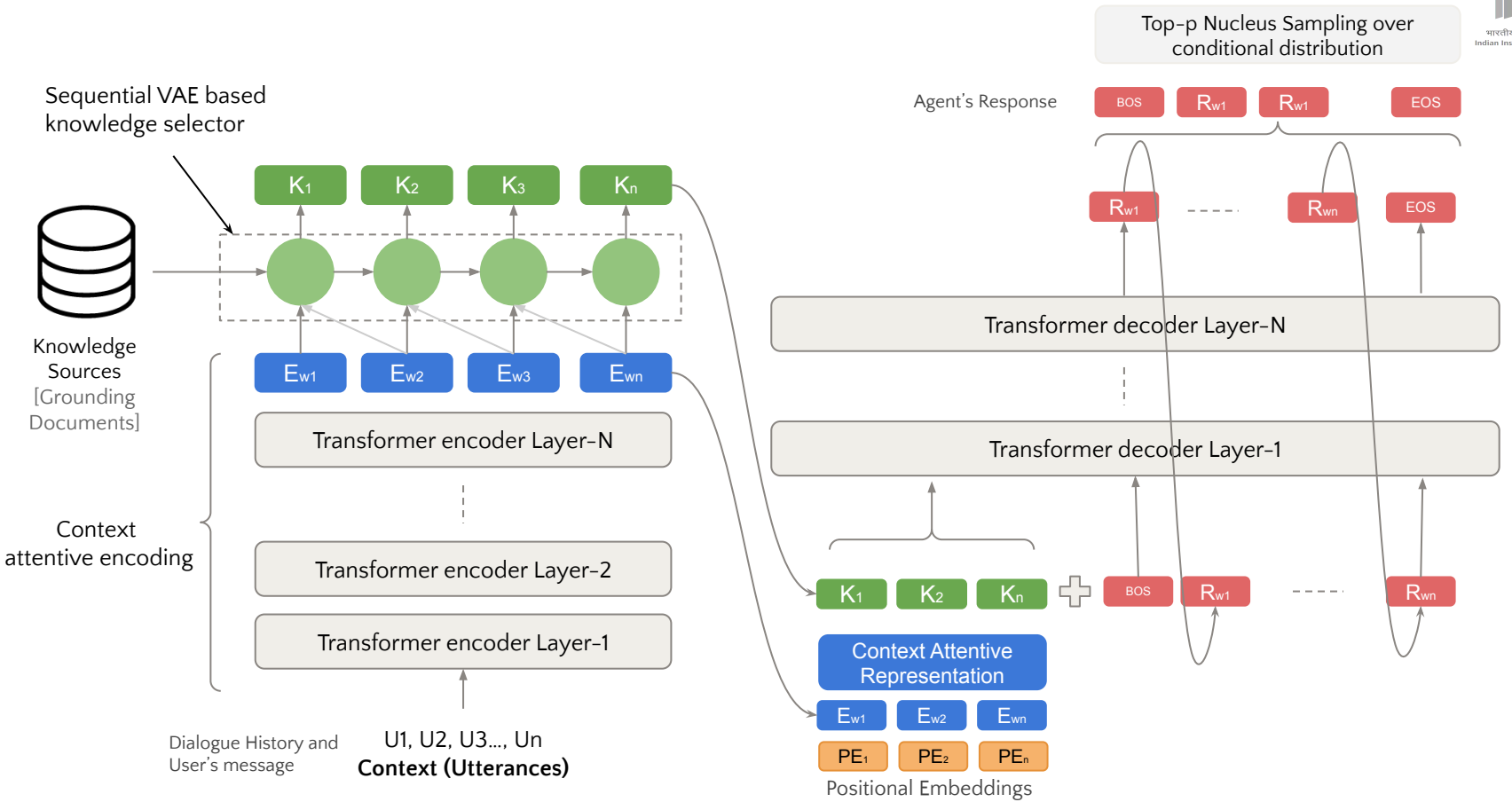




- **Motivation**: More engaging and accurate knowledge-based chit-chat
- **Knowledge Selection:** A sequential latent variable model for knowledge selection as a sequential decision process [continuous sampling] instead of a single-step decision process
- **Training Objective:** Maximization of variational lower bound and knowledge loss (cross entropy between true and predicted knowledge sentences) [GRU is used to encode history]
- If we can sequentially model the history of knowledge selection in previous turns, we can **reduce the scope of probable knowledge candidates** at current turn and generate more engaging responses.

Kim, B., Ahn, J., & Kim, G. (2021). Sequential latent knowledge selection for knowledge-grounded dialogue

Knowledge Grounded Dialogue Systems



Datasets used in previous papers.

भारतीय प्रौद्योगिकी संस्थान है
Indian Institute of Technology I

Wizards of Wikipedia

8, 430 training instances

1948 validation instances

1933 testing instances

First 10 sentences of the original Wikipedia page of the topic + top 7 articles from IR system
[67.5 sentences on average]

Holl-E

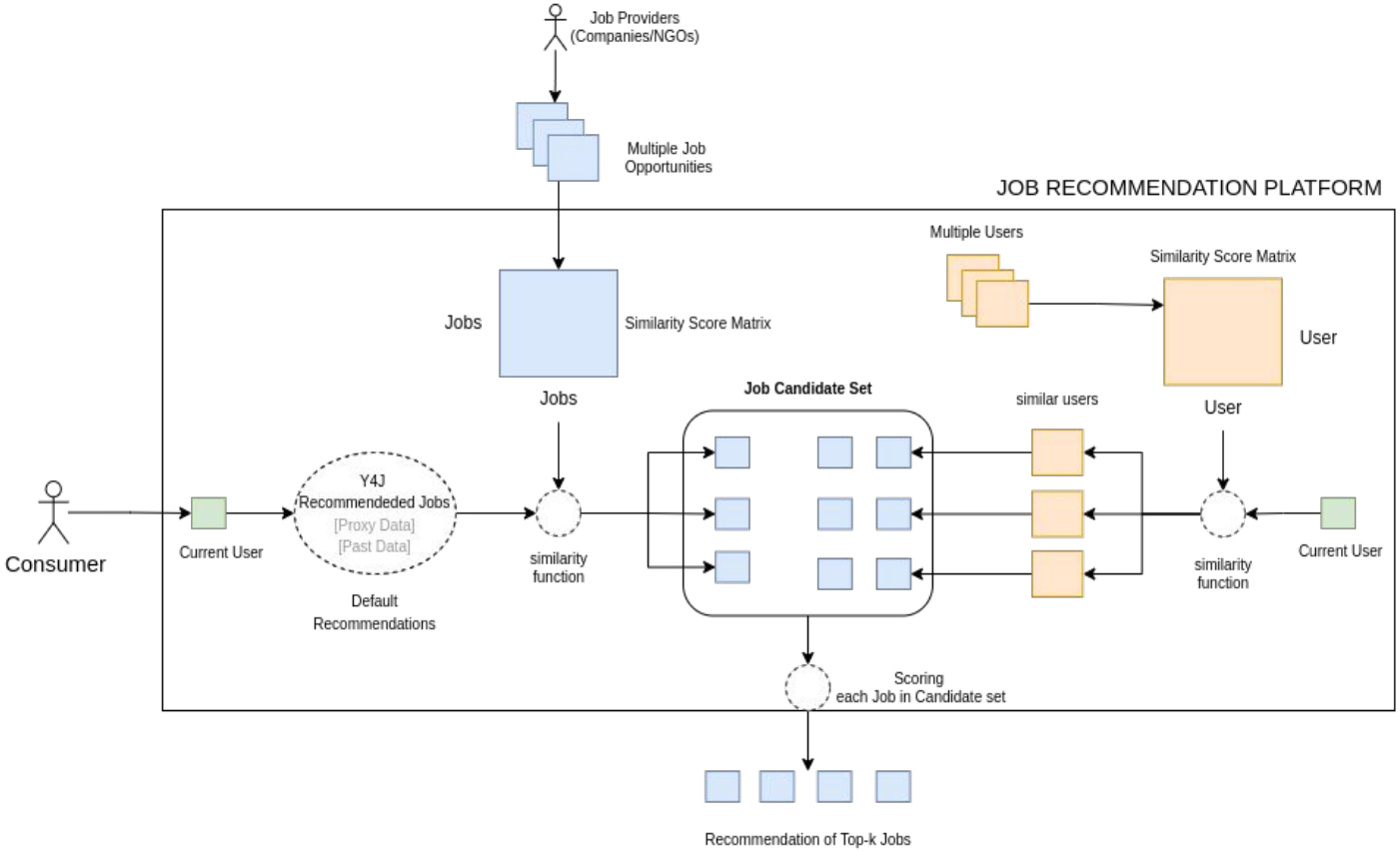
7, 228 training instances

930 validation instances

913 testing instances

Single Document Given per dialogue
[58-63 sentences on average]

A heuristic based filtering approach for job platform **recommender systems**



Thanks!

Any questions ?

You can find me at:

- cs21mtech16001@iith.ac.in

Academic and Professional Updates at:

- https://shresthakamal.github.io/home