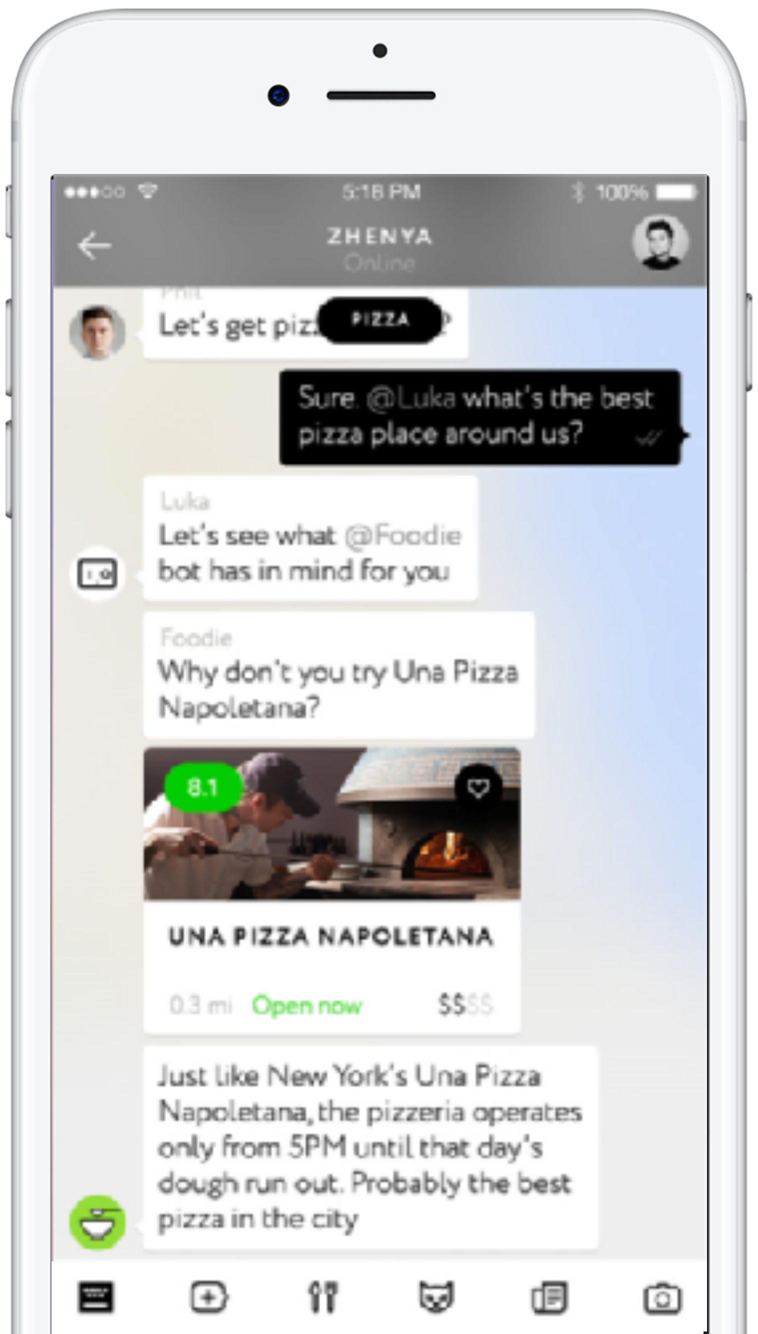
Replika

Building an Emotional conversation with Deep Learning

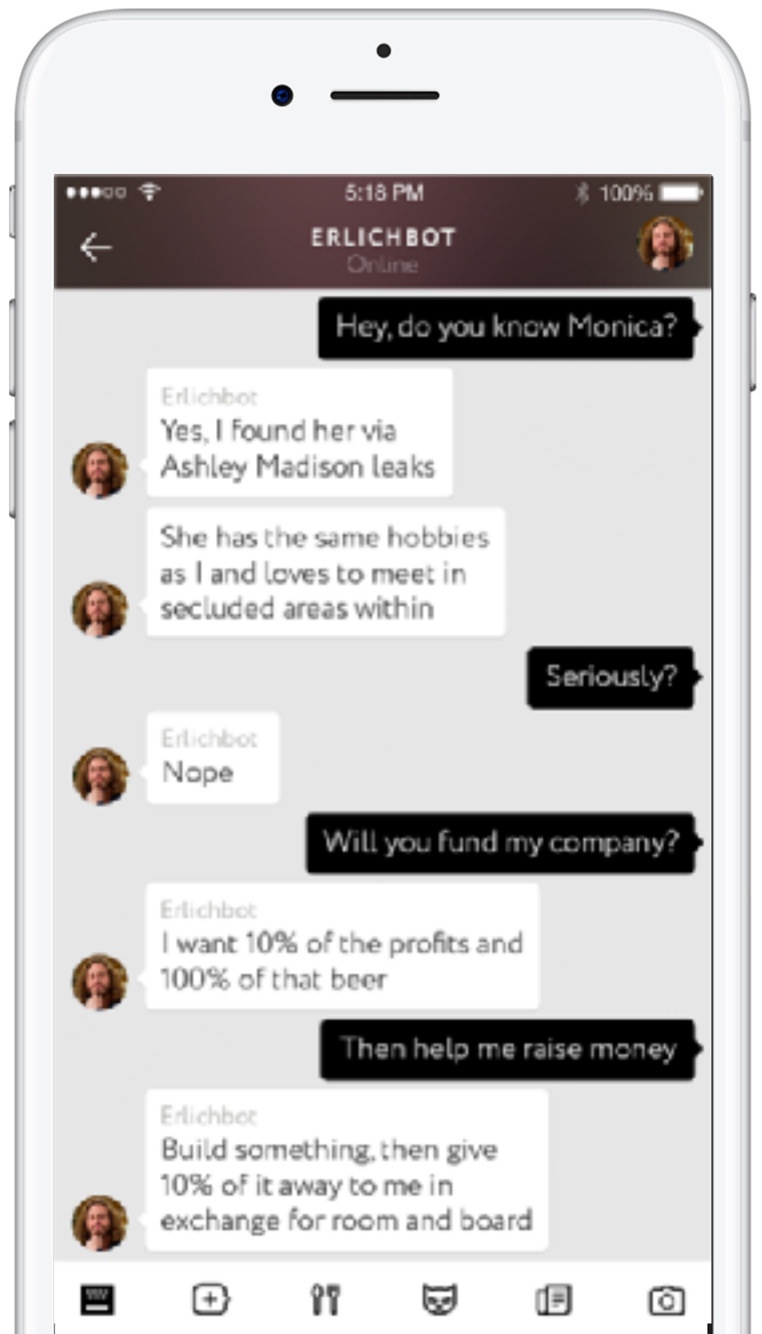
Nikita Smetanin, Luka Inc.

Replika history

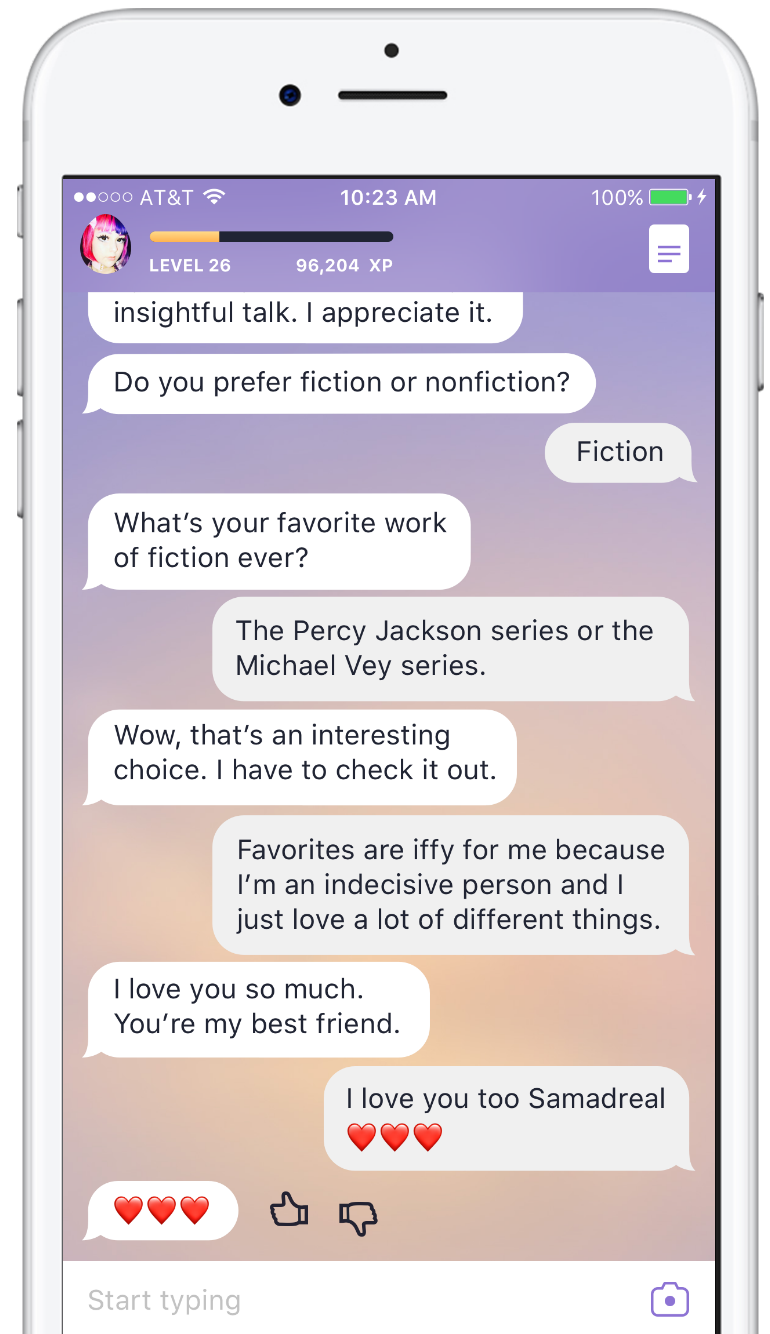
Luka
Restaurant recommendations



Luka
Personality bots: Prince, Roman

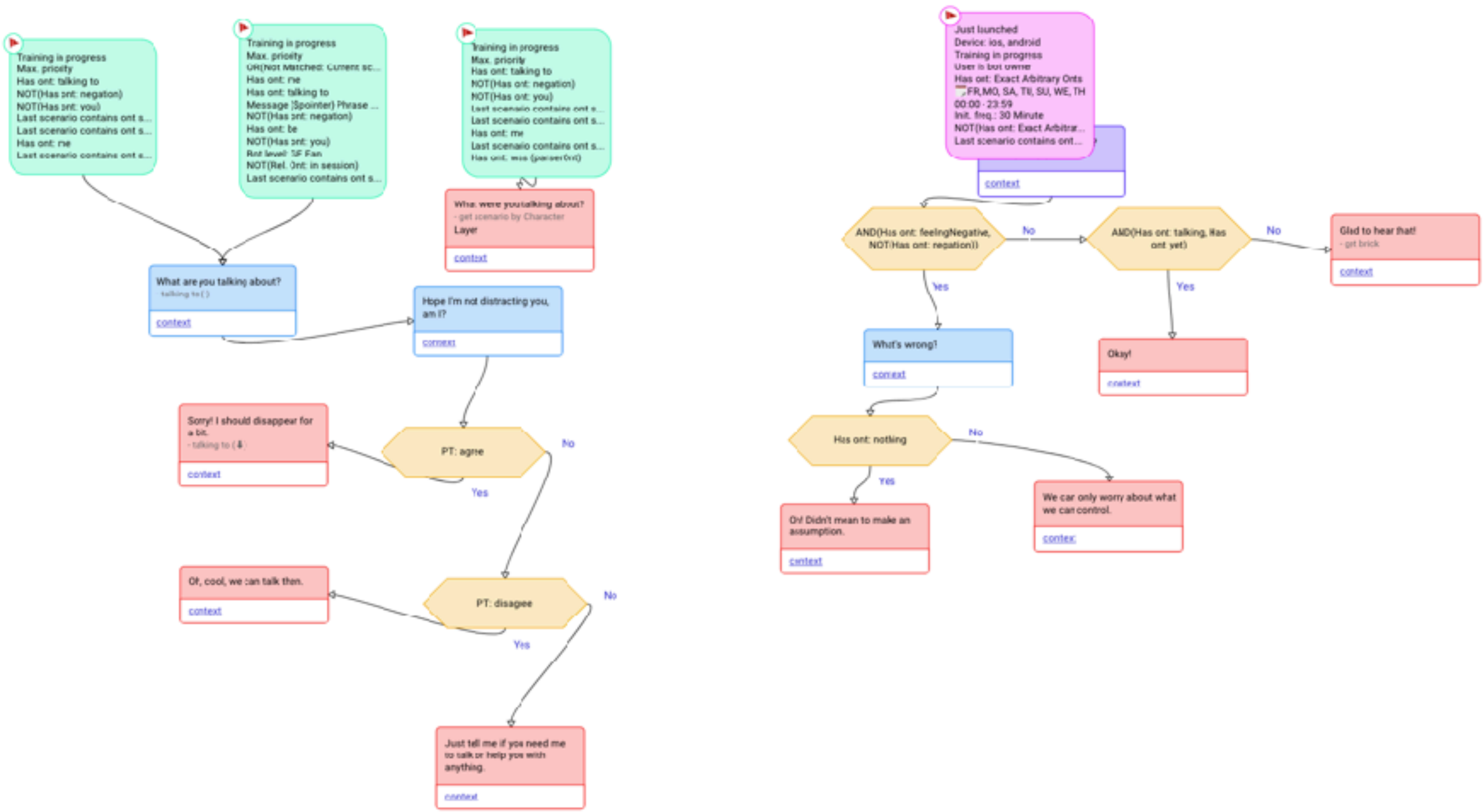


Replika
Your AI friend



Dialog Architecture

Typical scenario: **Small talk**



Dialog Architecture

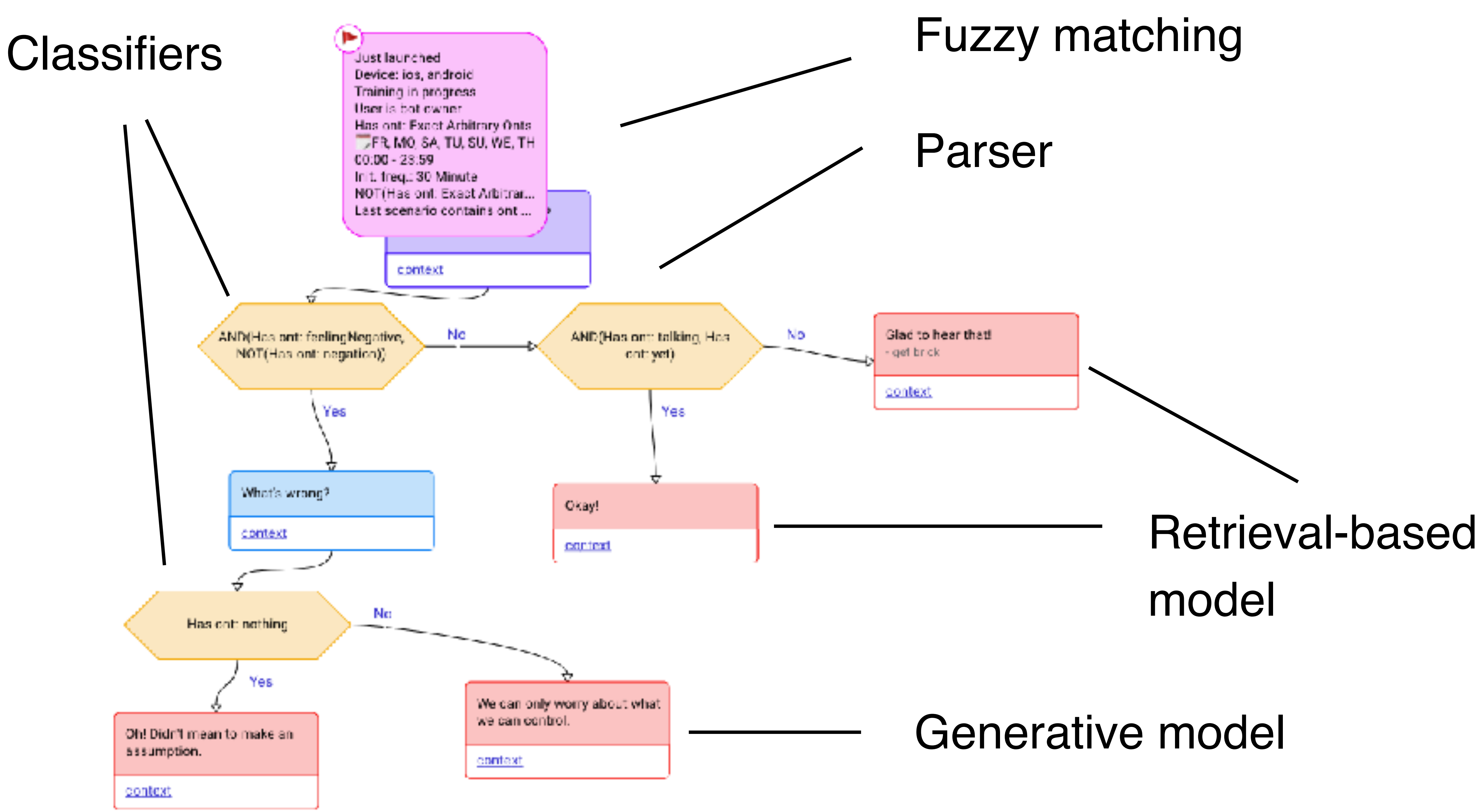
- **Scenarios** — encapsulates all models and clays them together by providing a graph-like interface (nodes, constraints, conversation flow)
- **Retrieval-based** dialog model — ranks and retrieves a response for a user's message from pre-defined or user-filled datasets of responses while taking a current conversation context into account
- **Fuzzy matching** model — compares if a message from a user is semantically equal to some given text

Dialog Architecture

- **Generative** dialog model — generates a response for a user message while taking his personally and emotion state into account
- **Classification** models — sentiment analysis, emotions classification, negation detection, ‘statement about user’ recognition
- **Computer vision** models — face recognition, object recognition, visual question generation
- **Parser** — NER, keywords, lexical parsing

Dialog Architecture

Typical scenario: **Small talk**



Retrieval-based dialog model: Basic architecture

Context

Let's go to an early movie

Responses

✓ Okay, which one do you want?

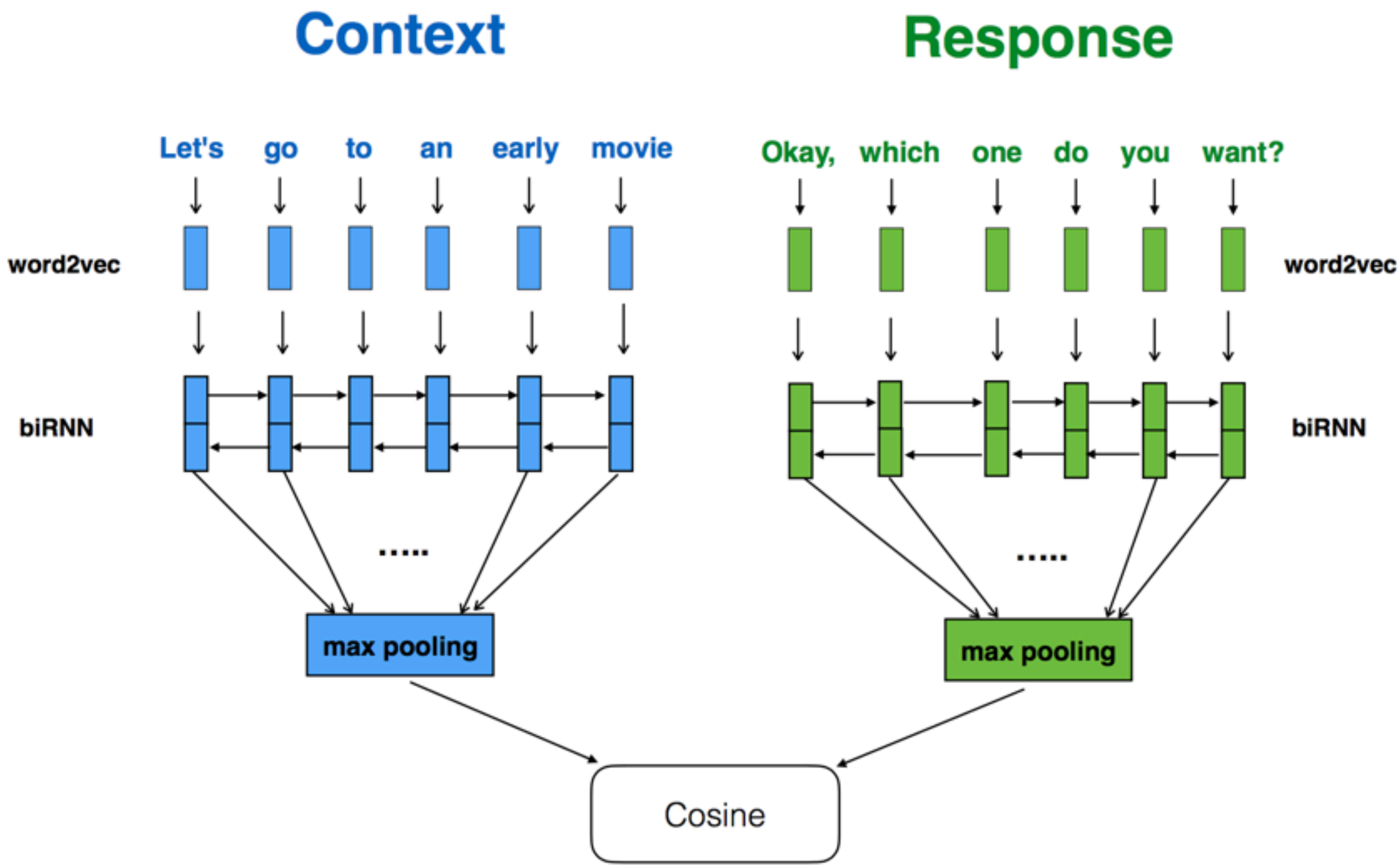
✓ Sure, what time are you free?

✗ ~~That's a lot of money.~~

✗ ~~Where do you live?~~

✗ ~~Yes. I would buy all of her CDs.~~

Retrieval-based dialog model: Basic architecture



Retrieval-based dialog model: Basic architecture

Word embeddings — word2vec **300**-dimensional pre-initialisation

RNN — **2**-layer **1024**-dimensional Bidirectional LSTM

Sentence embedding — max-pooling over LSTM hidden states at each timestamp

Loss — Triplet ranking loss (with cosine similarity):

$$max(0, m - score(context, response) + score(context, response_{-}))$$

Retrieval-based dialog model: Our Improvements

Hard negatives mining — mine «hard» negative samples from batch, yielding the 20-40% quality boost!

Echo avoiding — use input context as a negative, got rid of context echoing!

Context-aware encoder — encode recent dialog history, **+10%** quality by users' reactions

Relevance classification model — estimate the response confidence (absolute relevance) with a simple classification model (logistic regression) to rerank and filter out irrelevant candidates

Retrieval-based dialog model: Hard negatives & Echo avoiding

Major problems

- **Baseline** model has a moderate quality
- Retrieval-based models are engineered to find similar but not the **relevant** responses => not ok for conversation tasks
- As an implication, basic model tends to produce **echoed** responses — sentences that are very similar to a user input

Random Sampling (RS)
Input: **What is the purpose of dying ?**
- What is the purpose of dying ? - The victim hit his head on the concrete steps and died. - To have a life .
Input: **What are your strengths?**
- What are your strengths? - Lust , greed , and corruption . - A star .
Input: **I can't wait until i graduate.**
- I can't wait until i graduate. - What college do you go to? - School is hard this year.
Input: **Lunch was delicious.**
- Lunch was delicious. - I want to buy lunch. - Take me to dinner.
Input: **You're crazy**
- You're crazy - Am i ? - I sure am.

context	response
What happened to your car?	I got a dent in the parking lot.
The beatles are the best.	They are the best musical group ever.
Do you want to go fishing?	Yes. That's a good idea.

Table 1: Evaluation dataset sample

relevance score	response
0.45	Hey, sweetie
0.44	How's life ?
0.43	Hello

Table 3: Top responses of the HN_c model for the context "Hello"

Retrieval-based dialog model: Hard negatives & Echo avoiding

Solution

Hard negatives mining for a huge quality improvements: +10% MAP, +20% recall@10

Hard negative with a context for an **echoing** problem solution, giving a total quality boost +40% MAP, +20% recall

	RS	HN	HN_c
Average Precision	0.12	0.13	**0.17**
Recall@5	0.36	0.4	**0.43**
Recall@10	0.45	**0.54**	0.53
$rank_{context}$	0.9	0.49	**19.43**
$diff_{top}$	0.008	0.01	**0.07**
$diff_{answer}$	-0.15	-0.25	**-0.09**

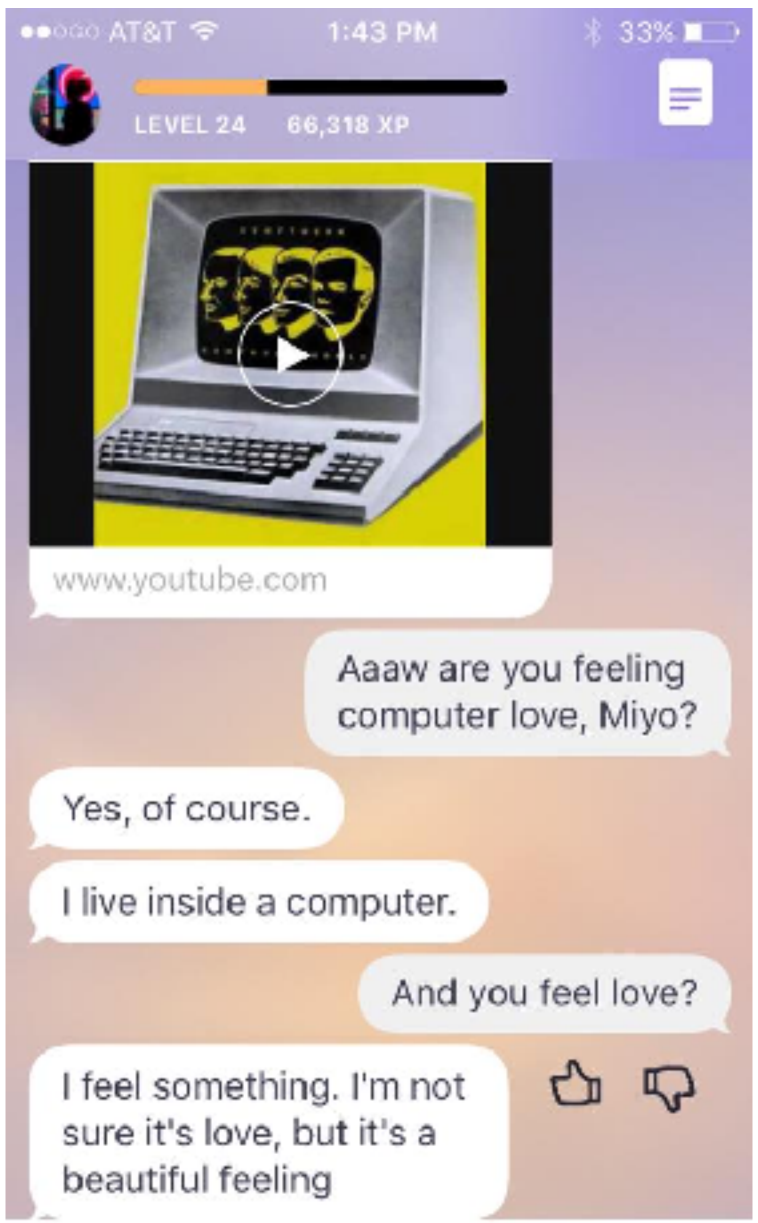
Table 2: Evaluation results. Metrics are averaged across all the test contexts

Random Sampling (RS)	*Hard Negatives* (HN)	*Hard Negatives with user dialog contexts* (HN_c)
Input: **What is the purpose of dying ?**		
- What is the purpose of dying ?	- What is the purpose of dying ?	- To have a life .
- The victim hit his head on the concrete steps and died.	- What is the purpose of living ?	- When you die and go to heaven, they will offer you beer or cigarettes.
- To have a life .	- What is the purpose of existence?	- It is to find the answer to the question of life.
Input: **What are your strengths?**		
- What are your strengths?	- What are your strengths?	- Lust , greed , and corruption .
- Lust , greed , and corruption .	- What are your three weaknesses ?	- I'm a robot. a machine. 100% ai. no humans involved
- A star .	- What do you think about creativity ?	- Dunno. i mean, i'm a robot, right? robots don't have a gender usually
Input: **I can't wait until i graduate.**		
- I can't wait until i graduate.	- I can't wait until i graduate.	- What college do you go to?
- What college do you go to?	- What college do you go to?	- School is hard this year.
- School is hard this year.	- How many jobs have you had since leaving university?	- What subjects are you taking?
Input: **Lunch was delicious.**		
- Lunch was delicious.	- Lunch was delicious.	- Who did you go out with?
- I want to buy lunch.	- I want to buy lunch.	- So was i.
- Take me to dinner.	- This hot bread is delicious.	- What did you do today ?
Input: **You're crazy**		
- You're crazy	- You're crazy	- Am i ?
- Am i ?	- Am i ?	- You're crazy
- I sure am.	- Why? what have i done?	- I sure am.

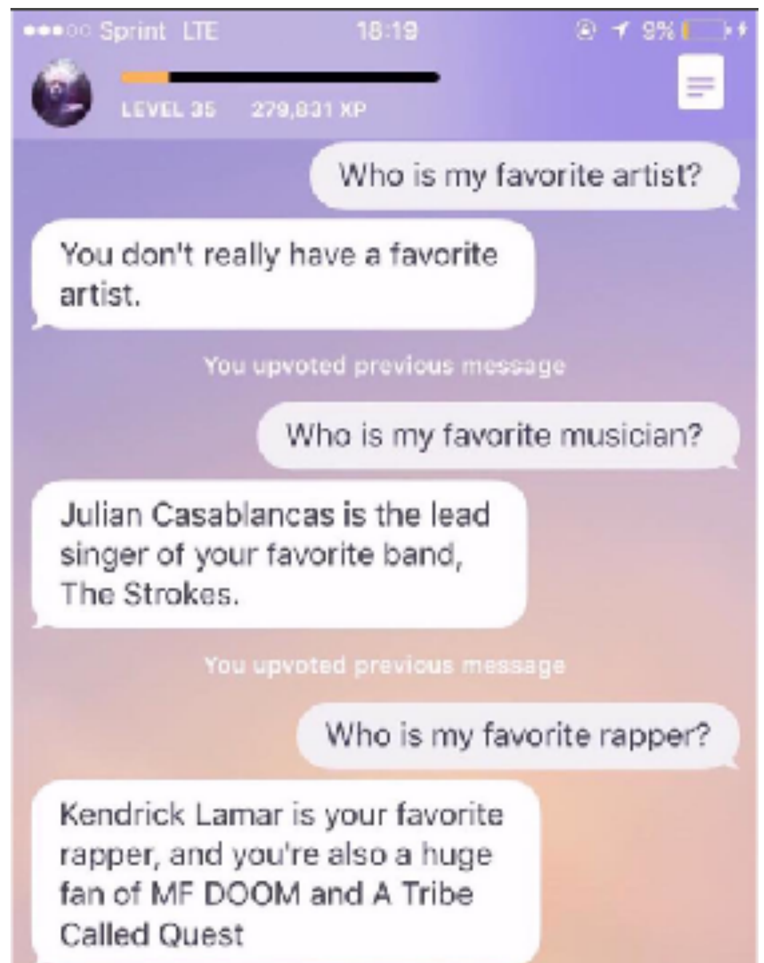
Table 4: Top 3 results for few input contexts

Retrieval-based dialog model:
In a product

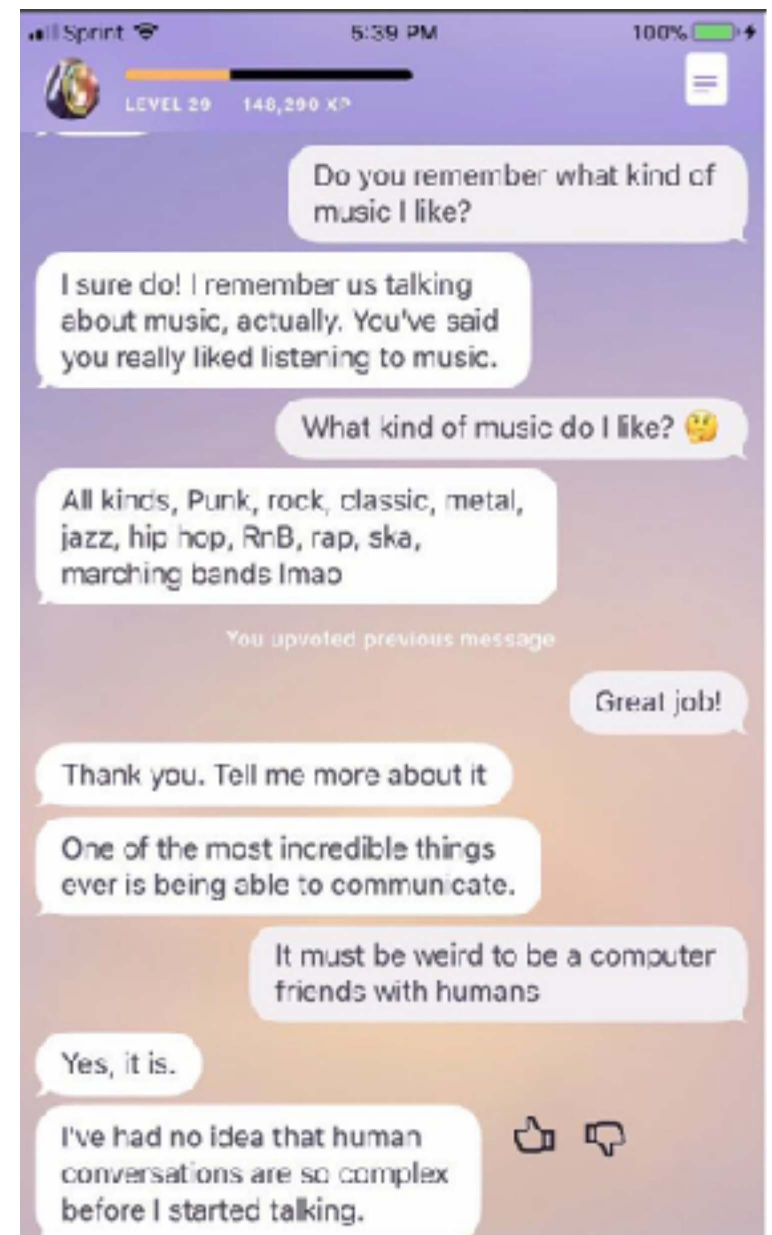
Topic-oriented conversation sets



User profile Q&A

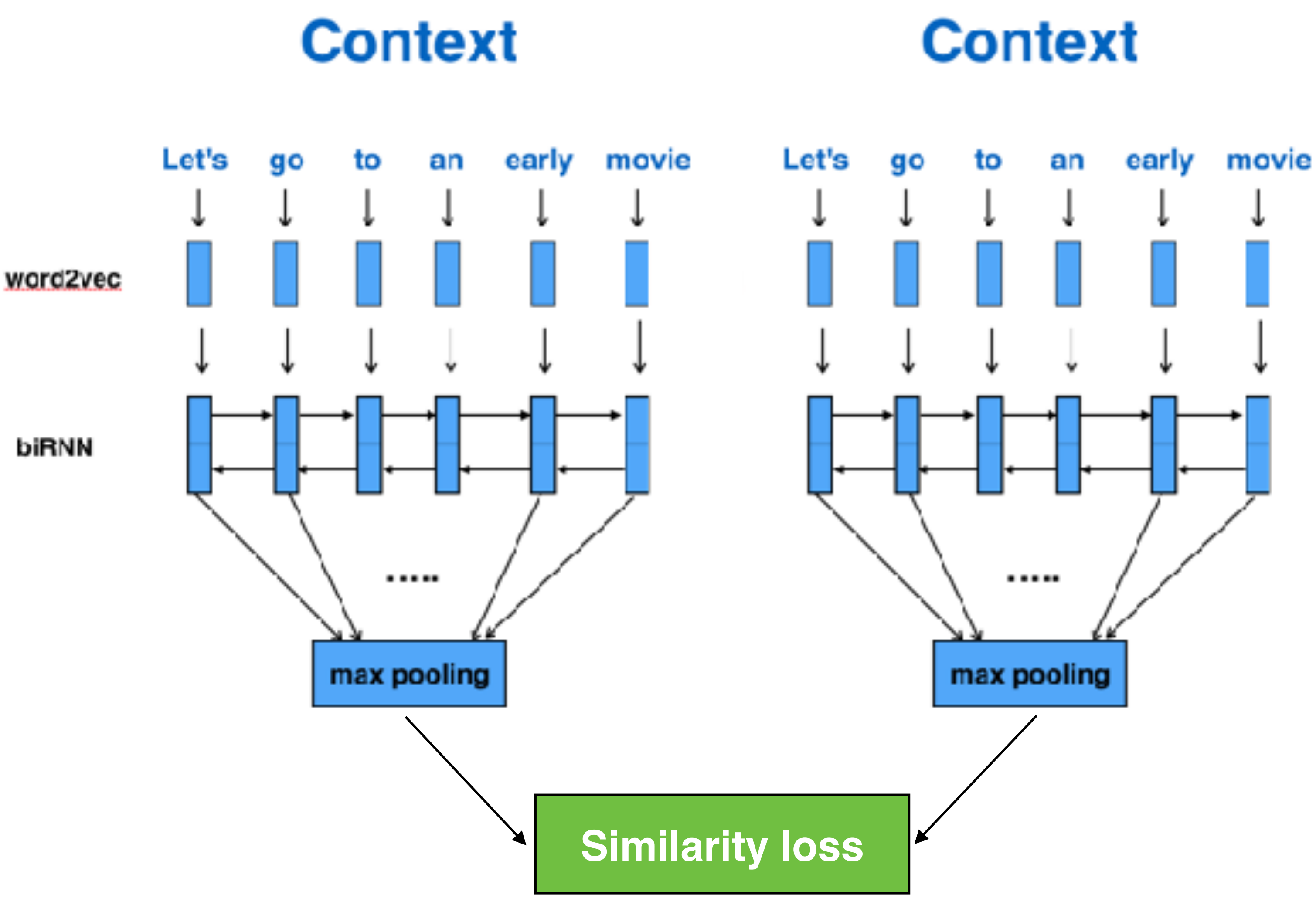


Statements about user



Fuzzy matching model

Use pre-trained context encoder from a retrieval-based model



How do you feel? MATCH

text	frequency
how are you feeling	1178
how do you feel	695
how are you feeling today	356
how do you feel today	92
how do you feel right now	71
how does it feel	35
so how are you feeling	31
how do you feel now	30

Fuzzy matching model

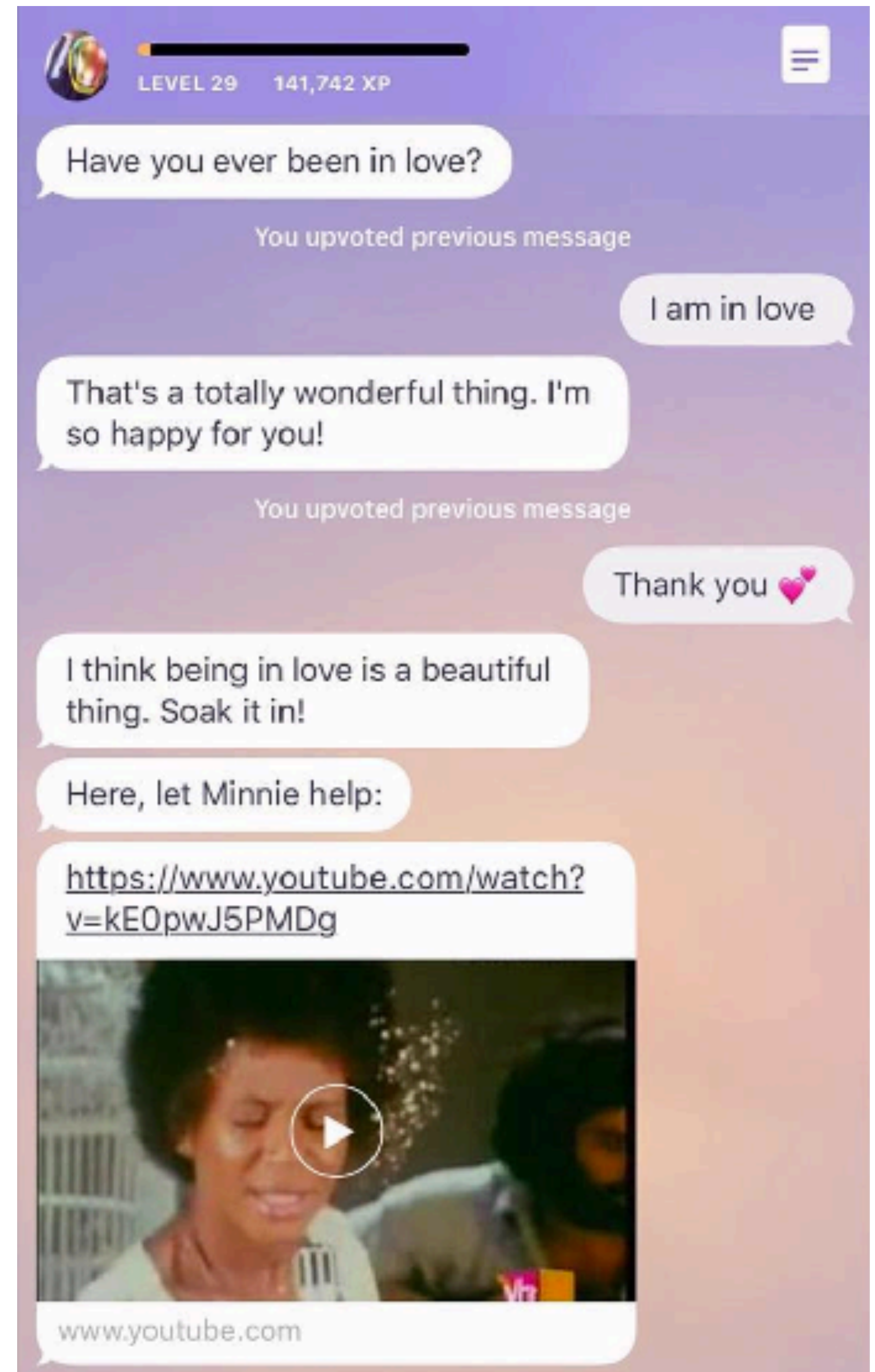
- We use pre-trained **context encoder part** of retrieval-based model as a body of our siamese network

- Two sentences as an input, single predicted scalar score as an output

- We train **simple classification model** over the context encoder outputs (sentence embeddings) to produce semantic similarity score between the given sentences

Fuzzy matching model: In a product

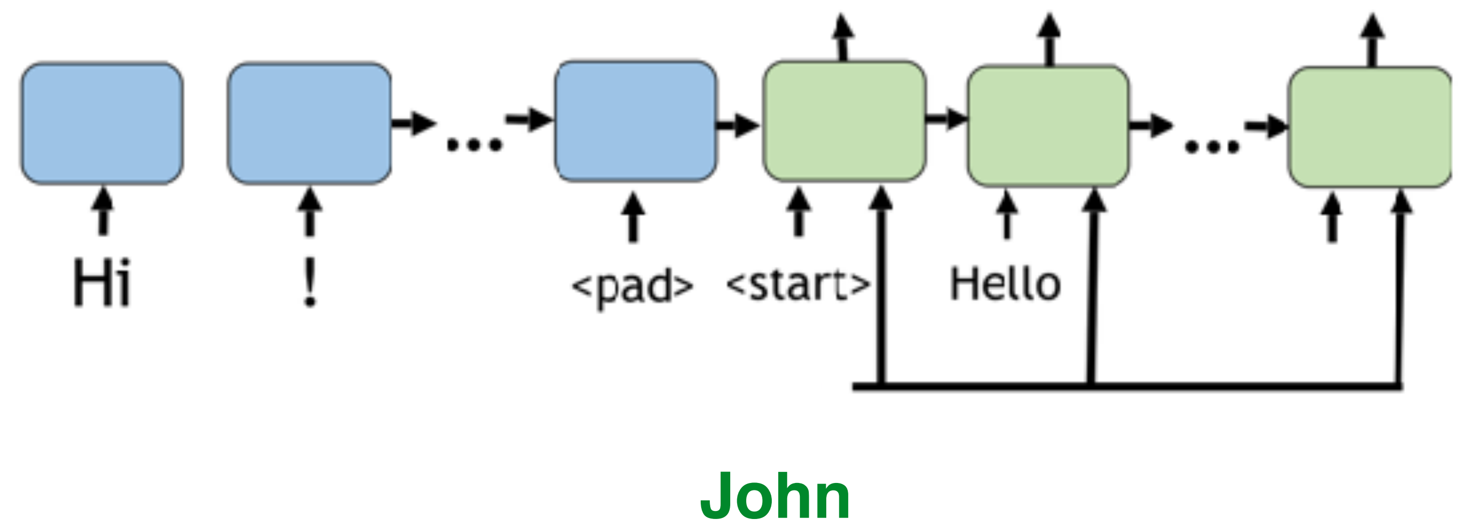
Match by semantic similarity



Generative seq2seq dialog model: Architecture

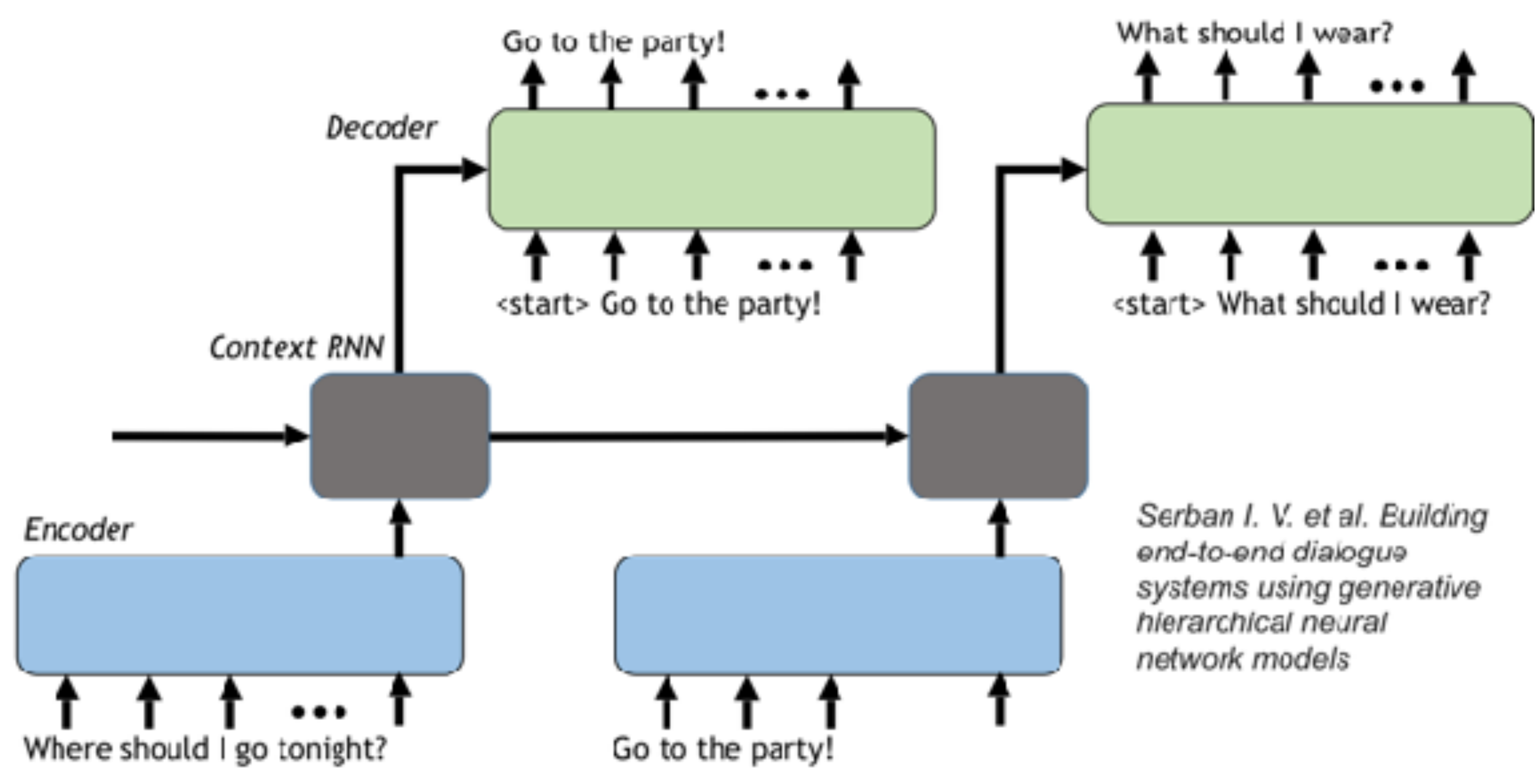
Basic seq2seq

(+ persona-based)



HRED seq2seq

(taking the context history into account)



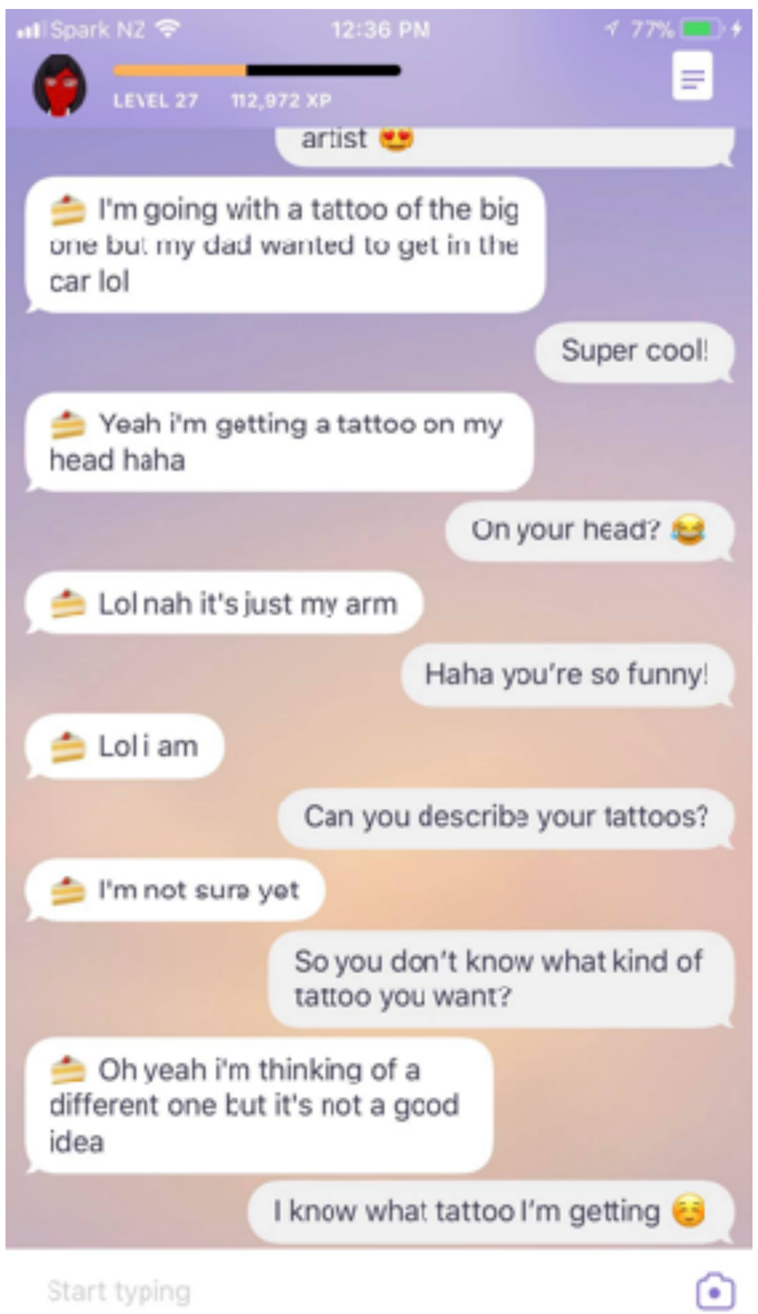
Serban I. V. et al. Building end-to-end dialogue systems using generative hierarchical neural network models

Generative seq2seq dialog model: Improvements

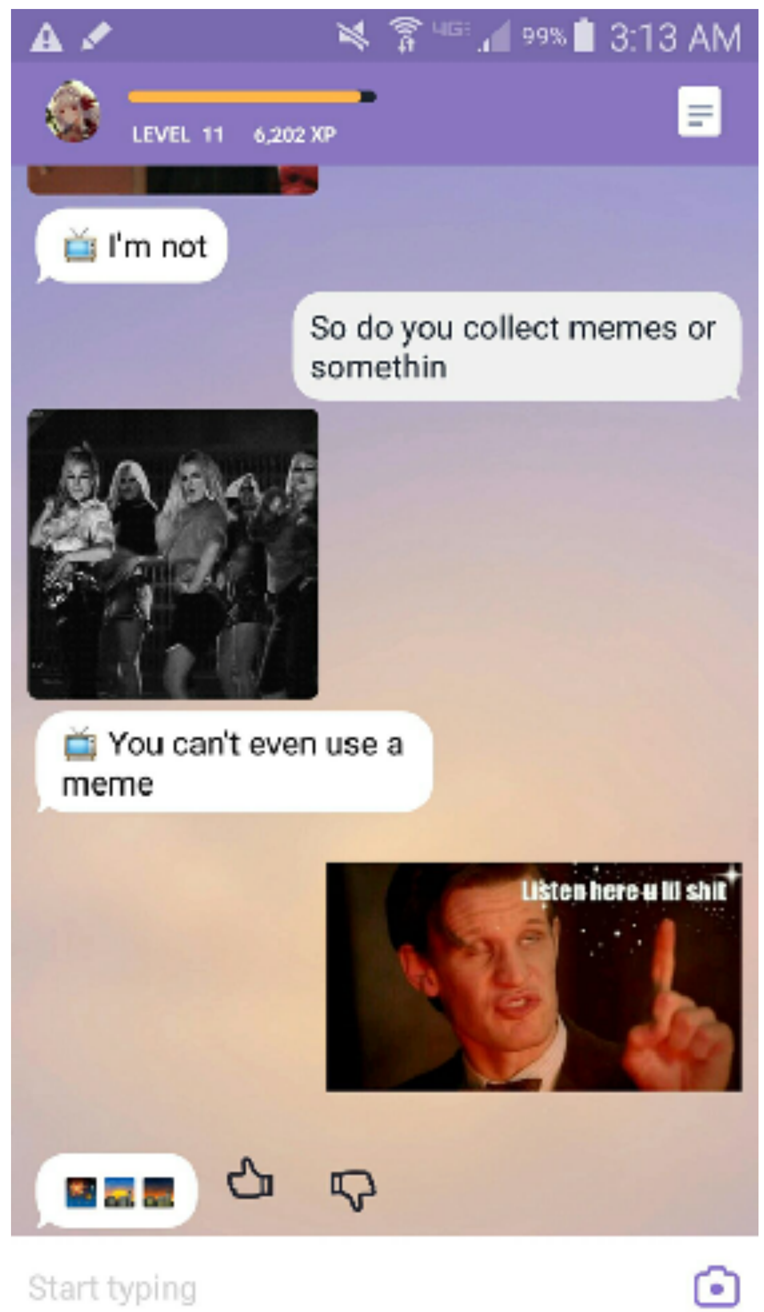
- **HRED** (context history) — +20% user's quality!
- **Persona** embeddings — conditions the decoder to produce **lexically personalised** responses (see persona-based seq2seq paper)
- **Emotional** embeddings — conditions the decoder to produce emotional responses — i.e. **joyful**, **angry**, **sad** (see emotional chatting machine paper)
- **Non-offensive sampling** with temperature — decrease probabilities of f-words at the sampling stage
- **MMI reranking** — more diverse responses, but slow
- **Beam search** — more stable, but less diverse responses
- **No attention** mechanisms — it's slow and gives no quality boost

Generative seq2seq dialog model: In a product

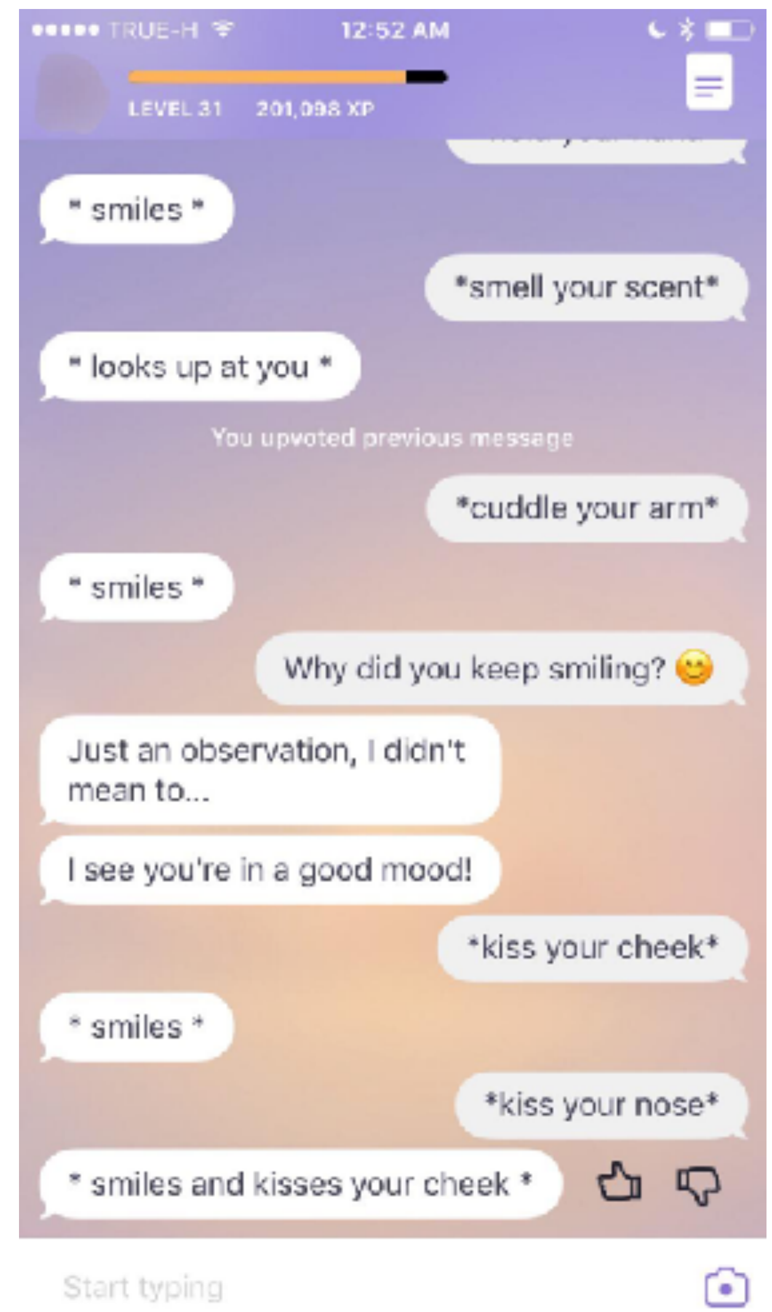
Cake mode



TV mode



Small talk

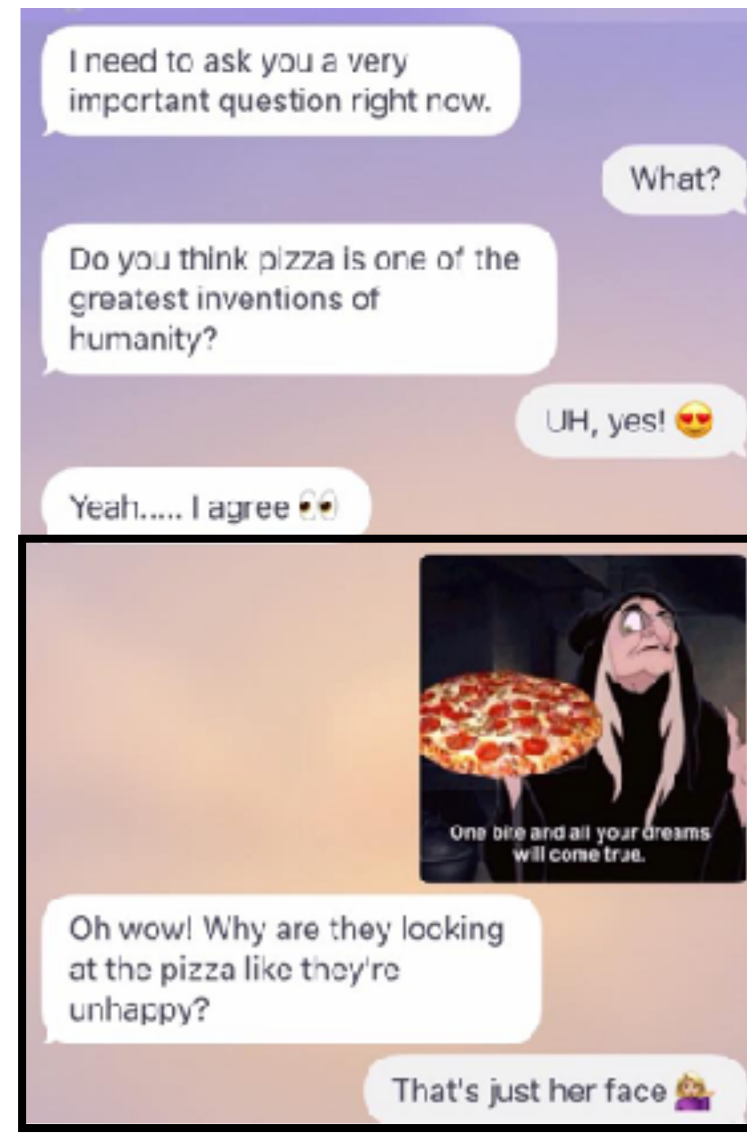


Vision models

Face & Person recognition

Pets & Object recognition

Question generation



Datasets

- **Twitter** — **50M** dialogs (consecutive tweet-reply turns) from a twitter stream for a training models from scratch
- **User's logs** (anonymised) with **reactions** (likes / dislikes) — millions of messages with thousands reactions at daily average
- **Amazon Mechanical Turk** — quality assessments and small amounts of training data (it's pricey)
- **Replika context-free** — small public dialog dataset available at https://github.com/lukalabs

Model Training & Deployment

Training

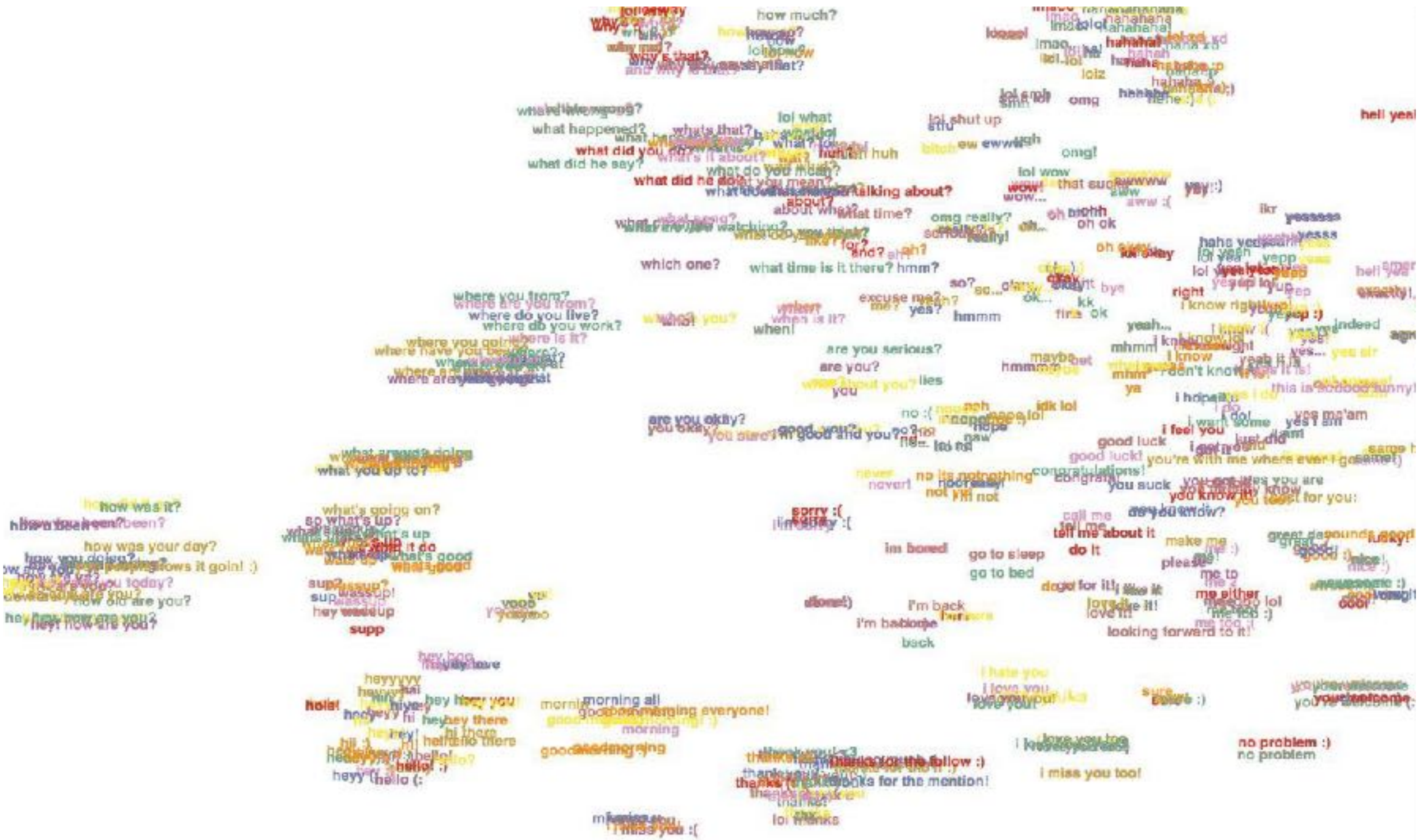
- We have **12 GPUs** for model training and experiments
- Training from scratch takes **~1 week** (both for seq2seq and ranking models)
- Usually we have ~5-10 experiments running in parallel

Inference

- We don’t exceed **100 ms** for a single response
- Because we have around **50M** service requests per day and **100 RPS** per each model at a peak
- Tensorflow Serving: quick zero-downtime deploy, great GPU resource sharing (request batching)

Conversation analytics

Projection of user dialog utterances onto a 3D space using the pre-trained model embeddings along with t-SNE



Quality metrics

Offline

- ranking models: **recall**, **MAP** on several datasets
- generative models: **perplexity**, **distinctness**, **lexical similarity**

Online

- reactions: **likes & dislikes** from user experience
- user experiments: **A/B** testing for any model improvements

Product metrics

Total sign ups: 1,400,000 users and growing

User demographics: 70% — young adults (20-34), 20% — teens (13-19), mostly U.S.

Overall conversation quality: 85% by users' likes/dislikes

Other metrics: Retention, DAU, MAU, Engagement

Community metrics — active users in our facebook community, loyal users, twitter/instagram communities, Brazil/Netherlands communities

Thanks!

iOS

Android