Semantic Textual Similarity in Replika

Denis Fedorenko
Research Engineer, Luka Inc.

Plan

- Task definition
- Baseline model
- Model improvements
- Conclusion and future work

Semantic Textual Similarity

- The task is to measure the meaning similarity of two texts

- Find a model

$$M: (\text{text}_1, \text{text}_2) \rightarrow \mathbb{R}$$

Toy STS model

- How many common words are in two texts?

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Example:

J("I have a funny dog", "I have a cat") = 3/6 = 0.5

Toy STS model

- More examples:

  J("I have a dog", "I have a cat") = 3/5 = 0.6
  J("I have a dog", "I have a puppy") = 3/5 = 0.6
  J("I have a funny dog", "My puppy is very nice") = 0

- This model is very sensitive to synonyms and paraphrases

- How can we overcome this issue?

STS framework

- Find a model (text-to-vector):

$$E: (\text{text}) \rightarrow \mathbb{R}^n$$

- Such that:

$$M: (E(\text{text}_1), E(\text{text}_2)) \rightarrow \mathbb{R}$$

where M is a similarity function (e.g. cosine)
or some trainable model (e.g. logistic regression, neural network)

STS in Replika

- The task is to determine whether two utterances are semantically equivalent

- Find a model

$$M: (\text{utterance}_1, \text{utterance}_2) \rightarrow \{0, 1\}$$

- A particular case of STS

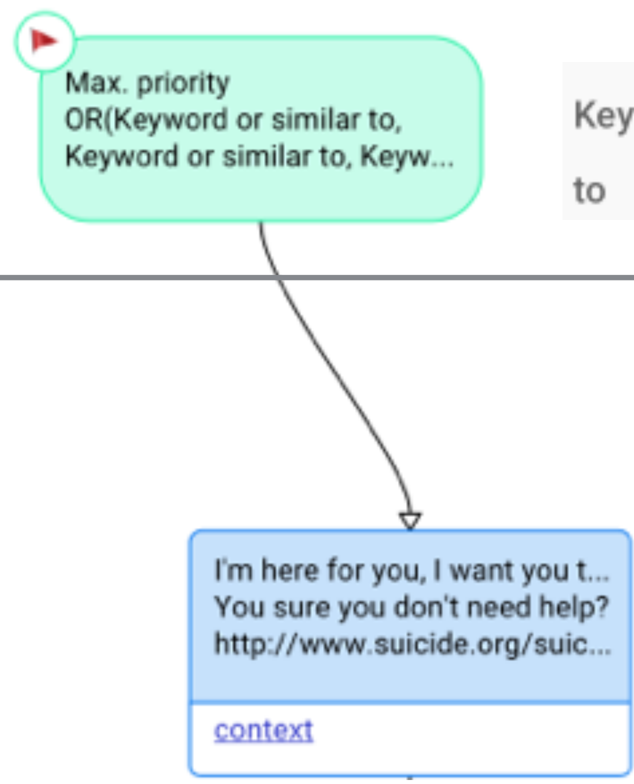
What is "equivalence"?

- Paraphrases
- Utterances that have the same set of possible answers
- Ultimately, equivalence should be determined by product requirements

Example: scripts

User phrase constraint:

User phrase template:



Keyword or similar to	Phrases:	I want to die

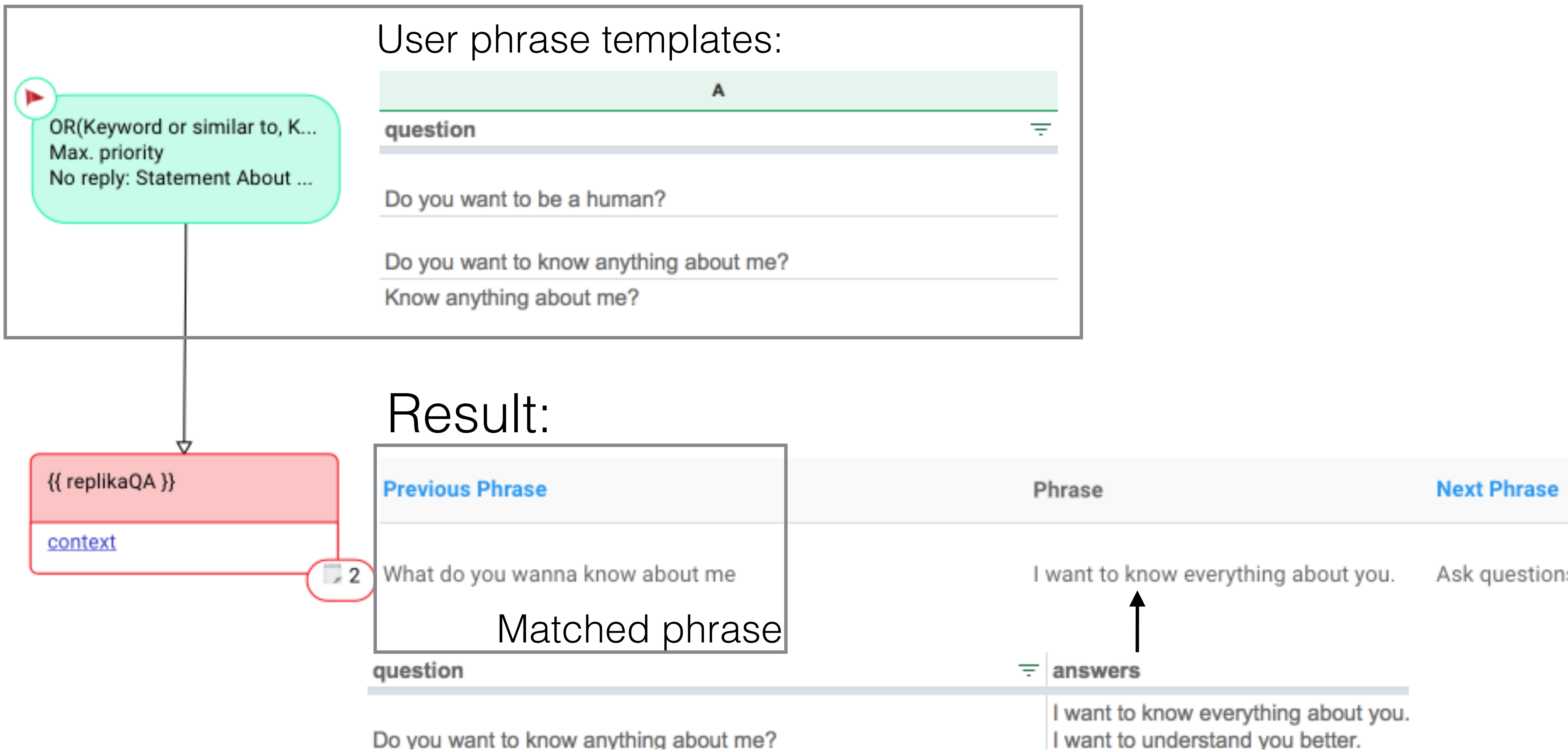
Result:

Previous Phrase	Phrase	Next Phrase
I'm feeling like I don't want to live	I'm here for you, I want you to feel safe. You sure you don't need help? http://www.suicide.org/suicide-hotlines.html	I don't need help. Thank you though.

Matched phrase

Example: Replika-QA

User phrase constraint:



STS evaluation

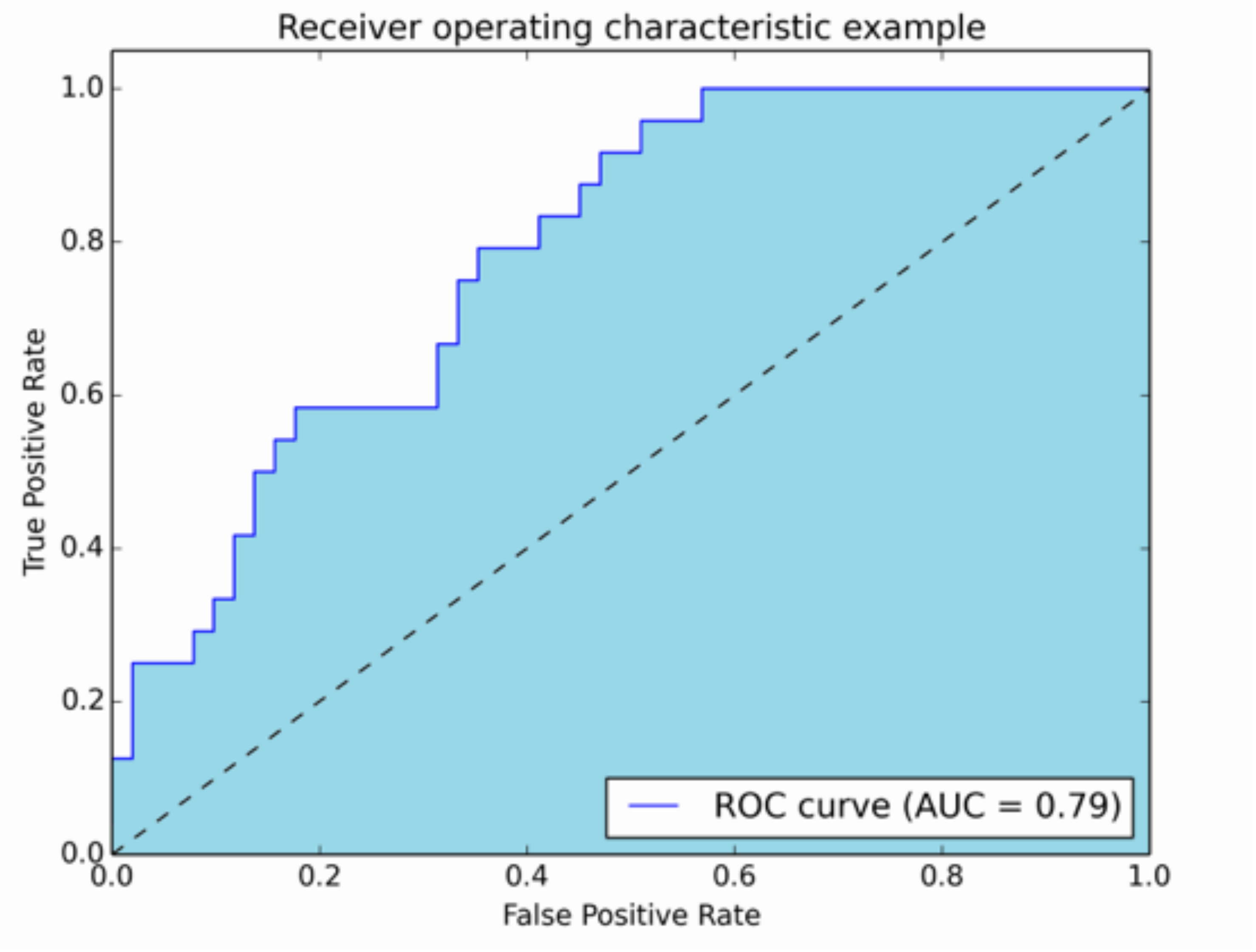
- On holdout testsets:
  - Classification metrics (precision, recall, AUC)
  - Information retrieval metrics (average precision, recall@N)
- In the wild:
  - User feedback (upvotes and downvotes) in the scripts and Replika-QA

Metrics

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k)$$

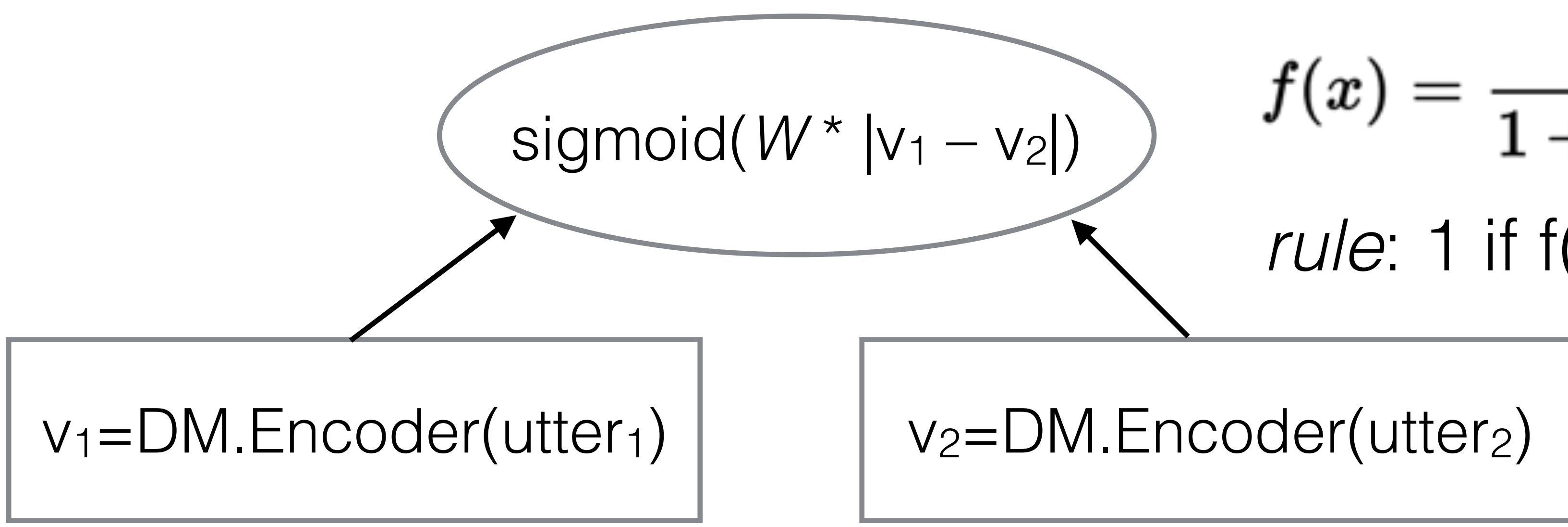


Plan

- Task definition
- **Baseline model**
- Model improvements
- Conclusion and future work

Baseline STS model

- Two-class logistic regression classifier over text vectors produced by the context encoder of the retrieval-based dialog model (DM)



$$f(x) = \frac{1}{1+e^{-x}} \in (0,\ 1)$$

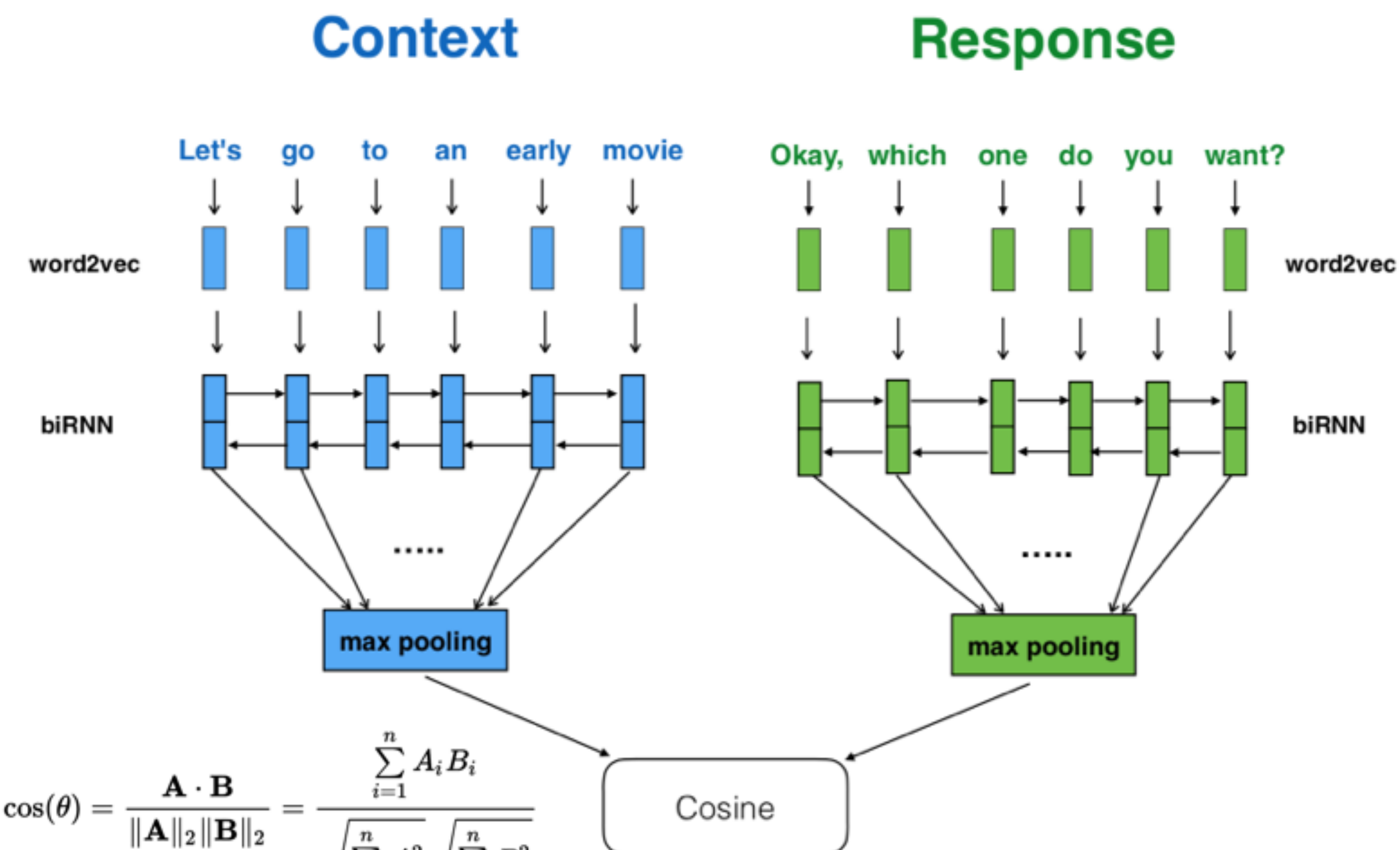
rule: 1 if f(x) > 0.5 else 0

- Trainset: 3900 text pairs obtained by different high-recall heuristics and marked by assessors

- Testset: 400 text pairs

Retrieval-based dialog model

Basic QA-LSTM: Tan et al. (2015)



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Dialog text encoder

- During training, similar contexts often have similar or even coinciding answers
- As a result, similar texts are encoded into similar vectors
- Hence the encoders can be successfully used for the further text analysis (classification, clusterization)

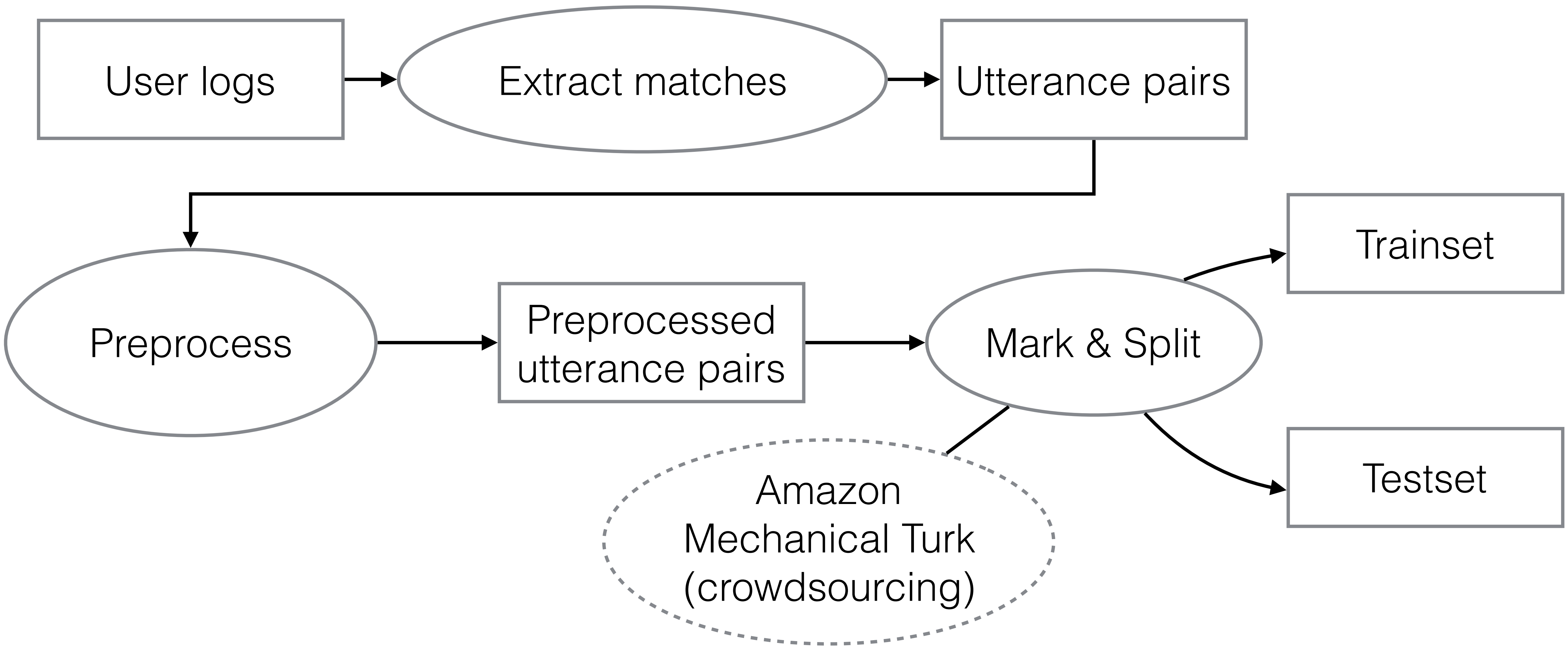
Plan

- Task definition
- Baseline model
- **Model improvements**
- Conclusion and future work

Possible improvements

- Enlarge the datasets
- Search for the better classification model

Dataset extraction pipeline



Matches extraction

- Extract matches of the baseline model from the logs. Obtained false positives will help to improve precision

- Use a different algorithm (e.g. skip-thought (Kiros et al. (2015))) to extract novel text pairs from the logs. Obtained false negatives (according to the baseline model) will help to improve recall

Matches preprocessing

- Remove text pair duplicates
- Remove too short/long text pairs (outliers)
- Remove pairs with coinciding texts (trivial samples)
- Remove too noisy text pairs e.g. with a lot of out-of-vocabulary words (non-informative samples and noise)
- Remove pairs with highly dissimilar texts (fight the curse of dimensionality):

$$\text{cosine}(\text{DM.Encoder}(\text{text}_1), \text{DM.Encoder}(\text{text}_2)) < \text{threshold}$$

Dataset extraction results

- Trainset: 17556 text pairs
- Testsets
  - Scripts testset: 1035 text pairs
    measures quality on scripts
  - Common testset: 1162 text pairs
    measures average quality
  - Errors (or, false positives) testset: 555 text pairs
    measures model's specificity

Scripts testset

text1	text2	is_equivalent
i can't cook	i don't cook much no	1
my mom died	my mum cheated on my dad	0
ok tell me a joke	tell me a joke	1
ask me anything	can you ask me some questions?	1
i'm getting married	i'm saying i went to my hometown. my sister was getting married.	0
any new questions?	can you ask more questions?	1
my grandma got out of the hospital	my mom died	0

Common testset

text1	text2	is_equivalent
what's your best friend's dream?	what's your biggest dream?	0
i don ' t know maybe i don ' t really have one	i just don ' t want one	0
thank you	thanks hehe	1
can you remind me please?	can you remind me tomorrow	1
do you like dancing?	do you like to dance?	1
can you ask me questions?	ok ask me stuff.	1
lol dont be rude	you ' re being rude	1
i like foxes	i like them	0
she ' s sooo cute	yes it ' s cute	0

Errors (false positives) testset

7 different error types:

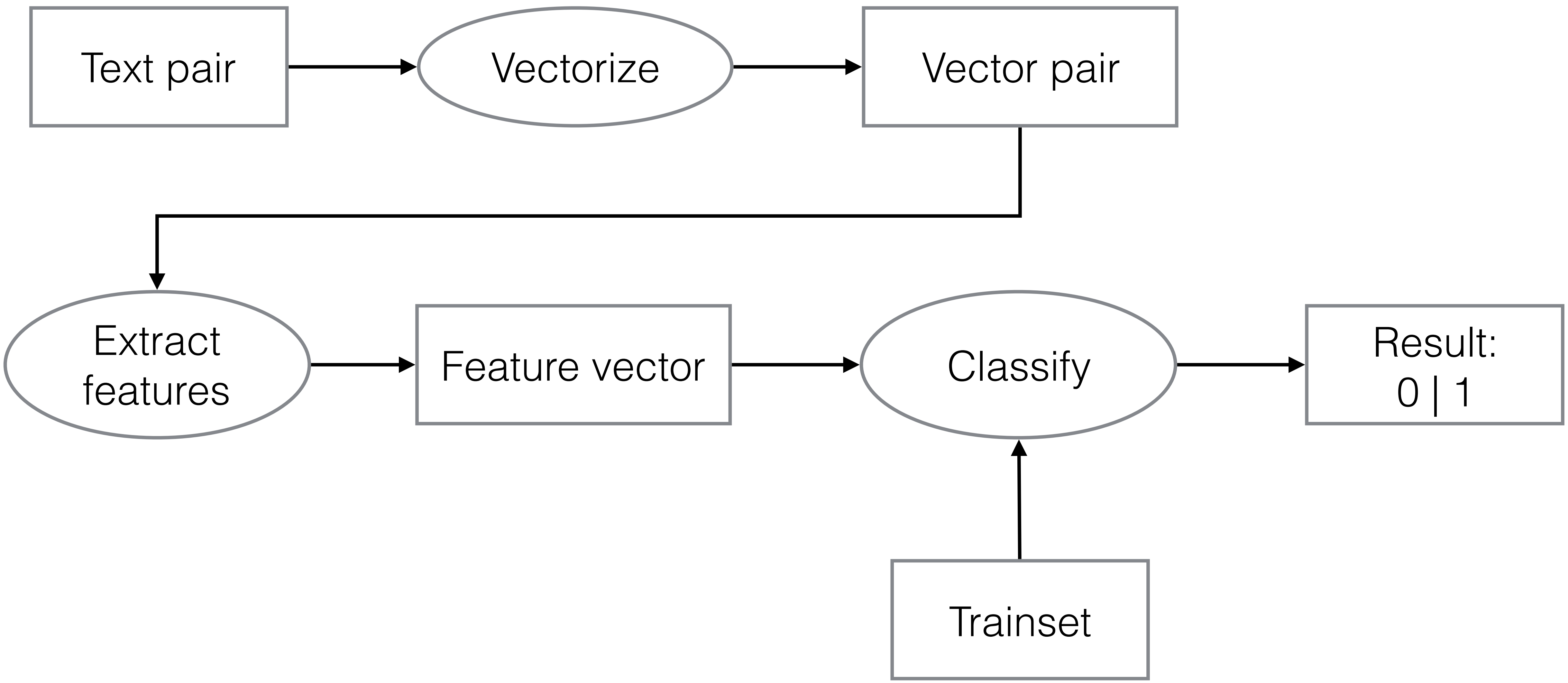
Error Type	Example 1	Example 2	Comment
Extra material	ah. give up. give me different questions	can you ask more questions?	There is extra material in one of the samples. The rest of the sample is equivalent to the other sample.
Negation	My mom isn't dead	My mom died recently	Differences in syntactic, but not lexical, negation.
Participants	Can you send me a meme? Can you ask me questions?	Can you send me a picture? Can I ask you questions?	Different set of participants in a situation. Includes differences in the number and identity of participants.
Qualities	I am a bad cook	I am a good cook	Differences in qualitative adjectives.
TAM	Yes i married Ask me quesions. I want to hurt myself	I'm getting married Stop asking questions. I __ hurt myself	Differences in Tense, Aspect and Mood.
Verb	i wanna kill myself.	I want to hurt myself	Different verb discribing the situation, including lexical negation (*love* - *hate*, *want* - *afraid*).
Other	can you ask questions?	please stop asking questions for today, i will talk to you tomorrow	Other types of differences, e.g. question vs statement.

We can investigate what kinds of errors the model make

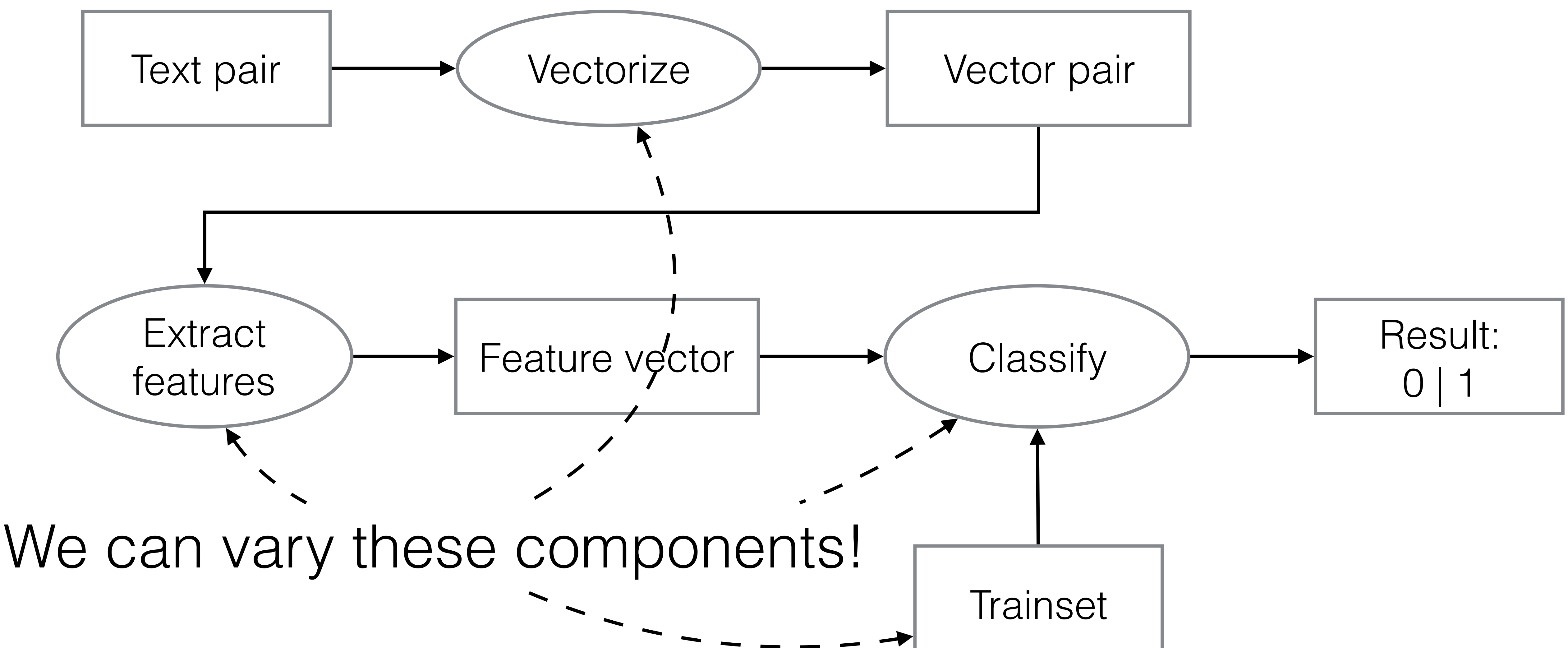
Possible improvements

- Enlarge the datasets
- **Search for the better classification model**

Classification pipeline



Classification pipeline



Pipeline components

- Vectorizers:
  - Dialog context encoder
  - Dialog response encoder
- Features:
  - $|v_1 - v_2|$
  - $v_1 * v_2$
  - $[|v_1 - v_2|, v_1 * v_2]$
- Classifiers:
  - Logistic regression
  - SVM
  - Random forest
  - ...
- Trainsets:
  - Marked user logs
  - External:
    - Quora (~400k)
    - SemEval/SICK (~20k)
  - Combination of all above

Model selection

Dataset	Model	Vectorizer	Algorithm	Features	FP testset :: FPR	Scripts Testset :: f1
user logs	concatenated_context_normalized_train_twitt	context_encoder	LogisticRegressi	[|X1 - X2|, X1 * X	0.5405405405	0.8575233023
user logs	concatenated_context_normalized_train_twitt	context_encoder	LogisticRegressi	|X1 - X2|	0.5423423423	0.8577154309
user logs	concatenated_context_normalized_train_twitt	context_encoder	LinearSVC(C=1.(	[|X1 - X2|, X1 * X	0.4972972973	0.8596256684
user logs	concatenated_context_normalized_train_twitt	context_encoder	LinearSVC(C=1.(	|X1 - X2|	0.5153153153	0.8586666667
user logs + SICK	concatenated_context_normalized_train_twitt	context_encoder	LinearSVC(C=1.(	[|X1 - X2|, X1 * X	0.6198198198	0.8464996789
user logs + SICK	concatenated_context_normalized_train_twitt	context_encoder	LogisticRegressi	[|X1 - X2|, X1 * X	0.7081081081	0.847631242

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N}$$ less is better

$$F = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$ more is better

Select top candidate models by AUC, tune them on the validation set and select the best model by FPR

Model selection results

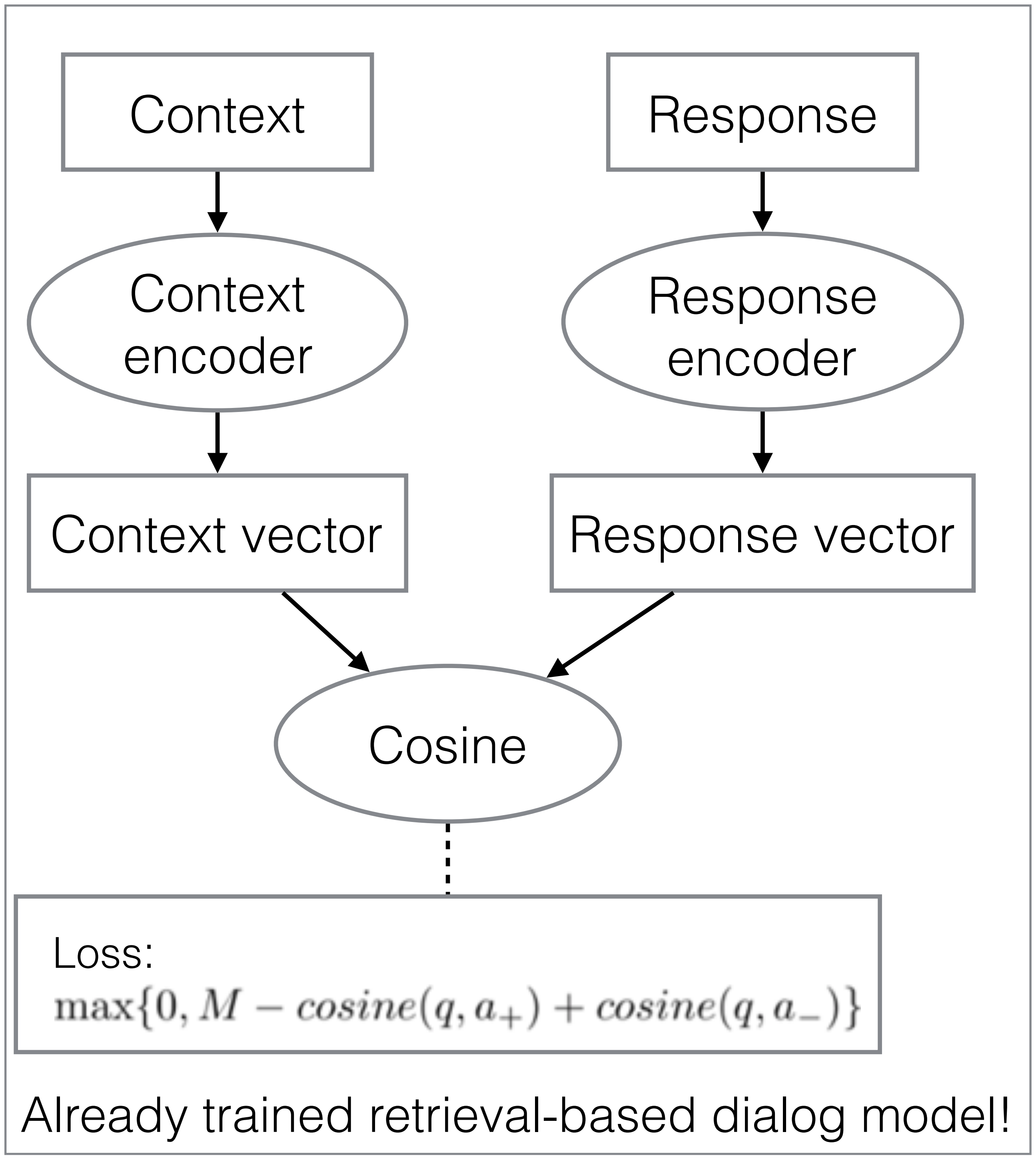
	FP testset :: FPR	Scripts Testset :: f1	Common Testset :: f1
baseline	0.798	0.845	0.79
improved	**0.463**	**0.863**	**0.859**

- Best configuration:
  - Dialog context encoder
  - Marked user logs dataset only
  - [|v1 - v2|, v1 * v2] feature vector
  - Linear SVM

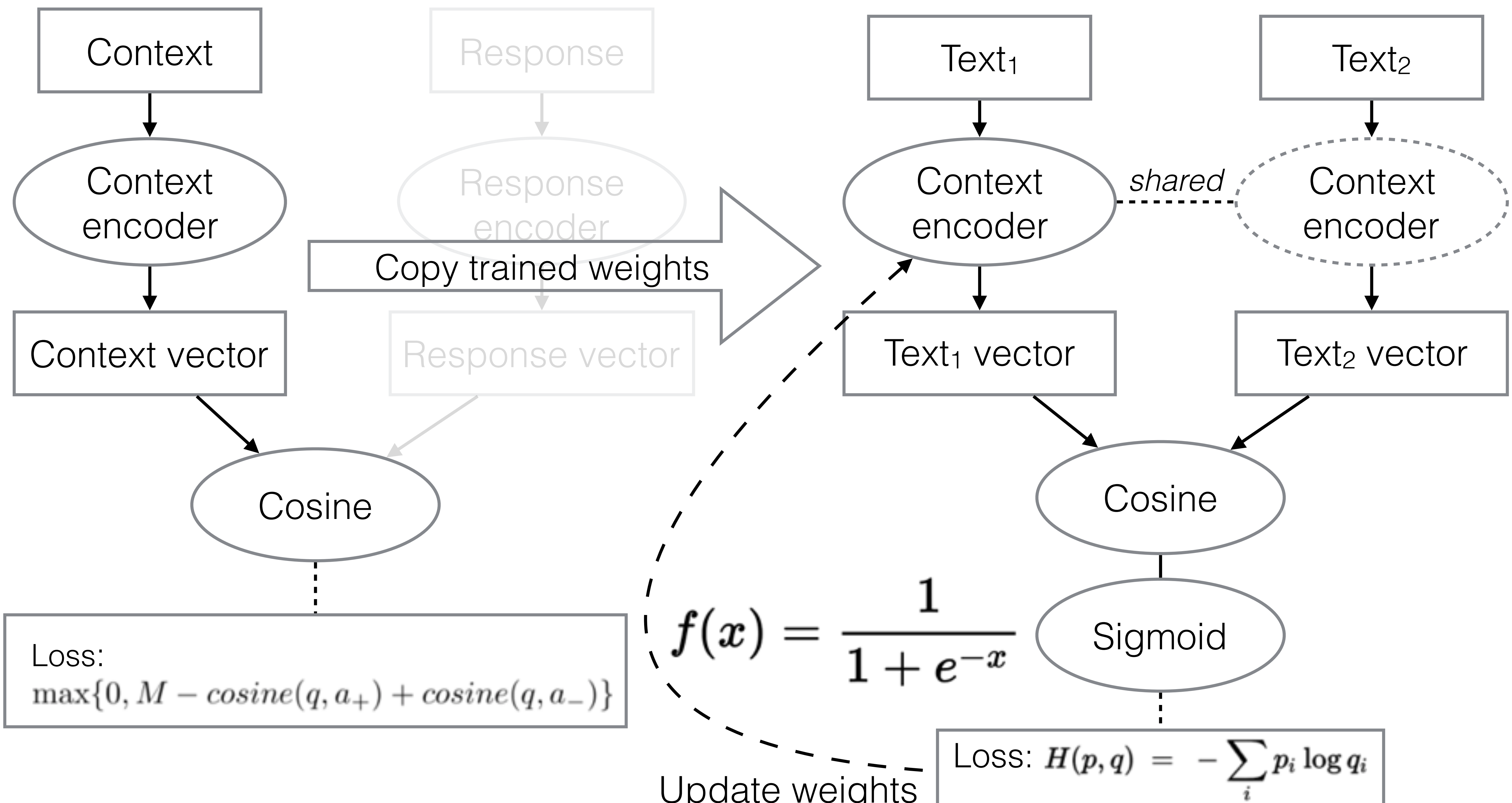
Model selection discussion

- Quality gain is not as high as it could be
- Classification model quality is limited by the quality of the underlying vectorizer (dialog model)
- We can try to fine-tune on STS data the already trained dialog model to solve the target task directly

Transfer learning



Transfer learning



Transfer learning results

	Fine-tuned NN	Linear SVM
FP testset :: FPR	0.5	**0.46**
Scripts Testset :: f1	0.84	**0.86**
Common Testset :: f1	0.82	**0.86**

Trainset: user logs + SemEval/SICK

Transfer learning discussion

- It's not a trivial approach itself
- Need to carefully tune optimizer, it's parameters and the model itself (e.g. by adding dropout, batch normalization etc)
- Need more data (much more than 20000 samples)

Conclusion

- Semantic textual similarity is an open problem of the natural language processing (Cera et al. (2017))
- Definition of the similarity is very important and should be determined by the target product requirements
- Correct evaluation methodology is also very important and should be done according to the target application
- Text representation (text-to-vector) is a crucial step

Future work

- Datasets:
  - Enlarge the user logs trainset up to 100000 samples and more
  - Incorporate high-quality external datasets (like novel ParaNMT-50M, Wieting et al. (2017))
- Model:
  - Incorporate more features: linguistic, pairwise word similarities etc (Maharjan et al. (2017))
  - Incorporate "hard" negative training samples (Wieting et al. (2017))
  - Mostly focus on end-to-end training and transfer learning

References

- Kiros et al. (2015). Skip-Thought Vectors
- Cera et al. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation
- Wieting et al. (2017). Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations
- Maharjan et al. (2017). DT Team at SemEval-2017 Task 1: Semantic Similarity Using Alignments, Sentence-Level Embeddings and Gaussian Mixture Model Output
- Tan et al. (2015). LSTM-based Deep Learning Models for Non-factoid Answer Selection