Natural Language Processing and Machine learning
Generative Russian Eliza

Anonymous COLING submission

Abstract

Conversational Artificial Intelligence (AI) is a trendy topic as for research field as for industry. There are many approaches to the chatbots, one of which is generative models. In this work we aim to implement several conversational AI models for chatbot: match based and generative models, - and compare its performance. Moreover, we aim to implement these models on Russian.

1 Introduction

Natural language processing as an area of machine learning is high-demanded for the industry and research field. For the current state of NLP approaches the machine can not deeply understand the semantics of language, while current NLP models can easily process the syntax of language. There are different syntax model tasks: Question Answering, Summarization, Text Classification, Text Generation, Translation, Sentence Similarity, Feature Extraction and other.

Researchers throughout the world try to approximate to the semantics recognisable models.Evolution of Conversational AI models begins with Eliza - an early computer program created from 1964 to 1966. Eliza examined the keywords received as input and then produced the output according to a defined set of rules. This methodology of generating output is still used by a number of chatbots.

After this, various virtual assistants were launched, for example, Siri by Apple was the first one conversational assistants. The concepts gained popularity and some time after that Google launched their Google Assistant for Android. One of latest model of conversation AI - Alexa. Alexa is an intelligent personal assistant developed by Amazon. It was introduced in 2014.

Implementation of models described in the following sections is done for Russian language. The motivation for this is considerably small amount of similar works, while the research area is abundant with models written for English.

The aim of this work[*] is to generate replies for input messages modelling chatbot with better performance. The report contains the following sections:

- **Introduction** - short description of conversational AI evolution
- **Background** presenting short literature review.
- **Methodology** with description of applied approaches and experiments, information about dataset.
- **Analysis** with obtained results and its explanation.
- and **Conclusion**

2 Background

Many conversational models are applied in chat-bots (Hussain et al., 2019), (Yan, 2018). There are several approaches used in conversational chatting systems: rule based(Frederking, 1981), pattern matching(Masche and Le, 2017) and generative(Shum et al., 2018), (AbuShawar and Atwell, 2015), (Ram

[*] Link on GitHub: https://github.com/Elizaveta55/NLP$_M L$

et al., 2018), including last generation (Lagler et al., 2013), (Floridi and Chiriatti, 2020), (Raffel et al., 2019). There are works written for Russian(Ismoilov and Semenov, 2019), (Burtsev et al., 2018), but this research area is still wide open. Evaluation (Shawar and Atwell, 2007), (Ram et al., 2018) of conversational models is not obvious and direct metric, but its contribution in model performance is significant.

3 Methodology

This section represent the conducted experiments setups: model architecture for each model, dataset description and evaluation of performance within some metrics.

3.1 Approaches

Basically, all approaches implemented in this work are rule based or generative models. Classsical Eliza is a good example of pattern matching approach, this model was the first to implement. Next class of generative models contains the following approaches: sequence to sequence generation, text to text generation and generative pre-trained transformer application.

3.1.1 Classical Eliza

Classical Eliza implements pattern matching for answer generation. For every user's message it search for any patterns and apply one of multiply possible answers for found pattern. For example, there is a typical pattern for Russian version of pattern "I would":

```
   [r'Я бы (. *) ',
       ["Не могли бы вы объяснить, зачем вам %1?",
       "Зачем вам %1?",
       "Кто еще знает, что вы бы %1?"]],
OR
   [r'I would (.*)',
       ["Could you explain why you would %1?",
       "Why would you %1?",
       "Who else knows that you would %1?"]],
```

But there is one huge difference between English and Russian - these two languages are from different language groups. For English the sentence is distinguishable because of strict word order. This order preserve words being changed often to reflect the meaning. While there is no strict word order for Russian, there is hardly any rules upon this ordering. But in order to distinguish one sentence from another with the same scope of words, this scope of words should have different forms reflecting the meaning, the tense, part of speech and other information. Therefore, pattern matching approach requires more efforts for Russian to formulate. And the performance of Russian version of Eliza would be definitely lower because of high variability of Russian expressions, forms and constructions.

Nevertheless, we attempt to perform Russian Eliza: the vast amount of efforts were put on correct pattern and reflections forming. For example, in Russian verbs change their form according to the active person. For English "I go", "You go", "We go", "He goes" have only two forms, but for every proposition there is a different form of this word for Russian. Moreover, these forms differ according to the conjugation of verb.

For proper implementation of classical Eliza the language adaptation was conducted. The result of implementation is presented in the section "Analysis".

3.1.2 Seq2seq

Seq2Seq is a generative algorithm for time sequences. This model belongs to sequence models which can learn the probability of the form $p(y|x_n, x_1)$, where i is an index signifying the location in the sequence and x_i is the i-th input sample. This form is suitable for the answer generation because it has ability to analyze and keep previous information in case it is needed.

Due to nature of conversations more advanced model with attention was implemented. It has GRU cells and long attention in order to be able to detect valuable feature and generate distinguishable answer.

It is essential to represent the data in applicable way: model will learn the representation and will generate the same form, therefore the representation should be sufficient, clear and distinguishable. At the same time text dataset could be too heavy to process, it can be highly resource-consuming to compute this amount of data with already heavy models. The preprocessing on data was applied, including filtering out unnecessary characters and expressions with rare appearance, and the index of questions and answers was built. The index was applied for simplification of representation.

The training of the model is important as well. It analysed the cross entropy with mask losses within all 50 epochs in order to learn features of data and be able to generate answers. The result of generative ability is presented in the section "Analysis".

3.1.3 rut5

Text-To-Text Framework uses unified format where the input and output are always text strings, while BERT-style models can only output either a class label or a span of the input. Original t5 framework allows to use the same model, loss function, and hyperparameters on any NLP task, including question answering. The result of generative ability for Russian language is presented in the section "Analysis".

3.1.4 ruGPT3

RuGPT3 is a model based on Generative Pretrained Transformer(GPT) 2 form OpenAI for Russian language with up to 1,3 billion parameters released by SberDevices in 2020. GPT is a generative Pretrained transformer with a separate layer of attention. For GPT language models solve exactly one problem: they attempt to predict the next token in the sequence from the previous ones, taking into account the previous context.

The model itself predict the continuation of a phrase, some time was spent to identify the correct pattern for input which would be treated by model in the correct way and make the model output in the form of reply. The result of generative ability for Russian language is presented in the section "Analysis".

3.2 Dataset

Dataset as an ultimate source of information is highly important. We used two datasets: dialogue dataset and subtitles dataset. The reason of usage of two datasets is following: dialogue dataset represents real dialogues collected by Yandex company, but it contains only short dialogues where most probable conversation lead is to greet each other and conduct really small talk. Therefore this information might be not sufficient to obtain a good generative ability.

The second dataset is abundant with information and diversity, implemented models could learn more features. But at the same time this diversity might be too huge for model to learn in the limited time.

Therefore we aim to use two datasets in order to find the model with better performance.

3.3 Evaluation

Evaluation of generative ability could be easily conducted with special metrics. We aim to attempt to evaluate advanced dialogue quality, but basically "reasonableness" (reasonable reply, weird but understandable, or nonsensical reply) is fine metric as well. In this work we implemented sentence comparison based on cosine similarity of representations of sentences. We have identified 4 different sentences to give to the models to obtain its replies. These four different sentences are:

- **Sentence1:** "Я пошел гулять" (or "I went for a walk") - regular sentence

- **Sentence2:** "Я чувствую обиду и злость за свой поступок" (or "I feel resentment and anger for what I did.") - emotional sentence, here we assume to evaluate the ability to produce supportive replies.

- **Sentence3:** "Снег автомату рознь, а собака скользкая" (or "Snow is automatic differently, but dog is slippery") - trash message

- **Sentence4:** "Скажи мне что-либо приятное" (or "Tell me something sweet") - sentence with indicative mood, here we assume to evaluate the ability to follow the instructions/indicatives.

Seq2Seq:	Dialogue dataset	Subtitles dataset 10%	Subtitles dataset 5%
Sentence1	ладно пойду спокойной fine I'll go good night	ы давайте наконец unk let's finally unk	он чувствовал что это может быть unk unk н в самом unk he felt it could be unk unk actually unk
Sentence2	я не люблю покушать а вообще люблю танцевать дома на работе I don't like eating but I like dancing home on work	unk unk у обвинения или больше unk unk unk с unk unk unk unk charges or more unk unk with unk unk	в чем дело малышка what's wring baby
Sentence3	у меня две собаки и трое детей а там уже кот I have two dogs and three children and a cat there	unk приготовьтесь на минутку ли unk get ready for a minute	ы успеем два тысяч момент we'll catch up two thousand moment
Sentence4	оу интересно конечно но мне очень нравиться oh it's interesting for sure but I like it very much	я unk со всеми дэнни но вы unk прямо без своей I'm unk with every Dannie but you unk right without your	это то что мы заработали that is what we've earned

Table 1: Seq2seq generated replies on limited training data

4 Analysis

The following table 1 represents the replies for seq2seq model we trained on two different datasets. According to the generated replies we can conclude the first dialogue dataset was not sufficient for model to obtain the generative ability. For the second dataset we used two approaches: trained on 10% of initial data and 5% of initial data. The data was limited in this way because, firstly, the subtitle dataset is quite big to process - it contains 44913765 lines, and secondly, for 10% of data the average calculation time for one epoch is 25 mins and almost 21hour for all 50 epochs, what is highly time-consuming. But we did not want to limit the data in severe way: we compute 10% and 5% of data assuming the generative ability of 10% will be higher.

??enerated results from seq2seq presented in table 1 show the model was not completely successful in learning: for subtitle dataset there are words "unk" indicating unknown token representation, moreover for 10% there are more such special words; for dialogue dataset the generated results are readable and without special words, but there are minor semantic mistakes. The reason of this behaviour seems to be in diversity of information: subtitle dialogue is more diverse and also generated results trained on 5% of data have less "unk" tokens comparing with 10% of data, - from these observations we can conclude more training is needed for subtitle dataset. Dialogue dataset is simpler, nevertheless dialogue dataset produced samples with better quality, that is why we select this model for future investigation.

The following table 2 represents the replies for each sentences from "Evaluation" subsection.

	Eliza	Seq2Seq	RuT5	RuGPT3
Sentence1	вы пошел гулять you goes for a walk	ладно пойду спокойной ладно пойду спокойной	Я знаю, что ты не можешь пойти вместе? I know, that you can not go together?	Вы не можете пойти погулять, потому что вы в данный момент находитесь на работе и You can not go for a walk, because you are working at the moment and
time	0.0006074	0.0860004	0.7879731	0.7590312
Sentence2	Когда вы обычно чувствуете обиду и злость за свой? When do you usually feel resentment and anger for your?	я не люблю покушать а вообще люблю танцевать дома на работе I don't like eating but I like dancing home on work	Что это за поступок? What did you do?	Здравствуйте, уважаемый Алексей! Спасибо Вам большое от всей души!!! Мы обязательно исправим Вашу ошибку Hello dear Alexey! Thank you very much from the bottom of the heart!!! We will definitely correct your mistake.
time	0.0005025	0.0982244	0.4680874	0.5840234
Sentence3	Как ты себя чувствуешь? How do you feel?	у меня две собаки и трое детей а там уже кот I have two dogs and three children and a cat there	О, ну что ж. Какая собака? Well. What dog?	Да нет же! У меня была такая проблема с собакой - она не могла пройти No! I had such a problem with a dog - she couldn't get through
time	0.0005192	0.2739951	0.5744748	0.5276198
Sentence4	Как ты себя чувствуешь? How do you feel?	оу интересно конечно но мне очень нравиться oh it's interesting for sure but I like it very much	Что это? How do you feel?	Я хочу, чтобы ты была счастлива и довольна своей жизнью! Ты должна быть счастливой в настоящем моменте I want you to be happy and content with your life! You should be happy in the moment
time	0.0004651	0.0889997	0.4473335	0.6415379

Table 2: The generated replies

	Eliza	Seq2Seq	RuT5	RuGPT3
Sentence1	**0.9797183**	0.5323588	0.4623547	0.4341821
Sentence2	**0.9158659**	0.2693855	0.7133703	0.5755187
Sentence3	0.3801847	0.4546080	**0.5749257**	0.5466763
Sentence4	**0.7708983**	0.7036186	0.5968673	0.7508475

Table 3: Cosine distance of generated replies

The following table 3 represent the sentence similarity values for each sentences from "Evaluation" subsection.

Analysing generated samples and its similarity with original sentence the following can be noticed:

1. **Eliza:** the model is pretty simple from interpretability point, but samples indicate in 2 from 4 cases model fails to find appropriate pattern and return general expressions. For the rest of cases pattern was found, but there are grammar and lexical mistakes. These mistakes were obtained because of high variability of Russian. To neglect these mistakes significantly more patterns and reflections should be written.

   But Eliza has the least calculation time for every sample, what is explainable with its simplicity.

   Result: readability is presented, while dialogue quality is low.

2. **Seq2seq:** the model was trained on two different datasets: dialogues and subtitles. For both of dataset the performance is low: there are a lot of "unk" in the generated samples indicating the model fails to infer into text sequence. There are two possible solutions: 1) change the representation of data on more advanced representative models, 2) spend more time on training, because loss values were decreasing throughout 50 training epochs, the model could not catch up with this amount of epochs.

   Calculation time is moderate, but sufficient. For the trash sample calculations took more time, probably due to nature of data distribution of outliers: in original dataset incoherent samples are rare, therefore model learns little about these samples and processing time for these sentence should increase.

   Result: readability is poorly presented, dialogue quality is out of consideration.

3. **RuT5:** performance of the model is considerably high: it managed to produce completely readable samples with context reflecting. For regular sample (Sentence1) it results with a little strange sentence (there should not be "?" in the end and the context of "together" is unclear), but for the second emotional sample it managed to produce fine human-like reply, for the third trash sample it produced the best possible answer: showing an desire to understand the sample and emphasizing the continuation of the dialogue with the question("What dog?"), - but for the last sample it did not follow an indicative mood, but nevertheless produced realistic reply.

   Calculation time is high. It is explainable with complexity of model.

   Result: readability is presented, dialogue quality is high.

4. **RuGPT:** this model has the most questionable performance: its generative ability is high, it attempt to finish input message, but these writing attempts produce unreliable and unstable samples: in the first regular sentence nothing was stated about work, while model produced "you are working at the moment"; for the second emotional sample it managed to show empathy, but context was trashy. For the rest of samples generated responses were realistic and human-like, for the last sentence it even followed the indicative mood and really generated something nice. These heavy models are supposed to perform better for long conversation, therefore there is a probability these models have better performance for dialogue, not for one-shot message we applied.

   Calculation time is high. It is explainable with complexity of model.

Result: readability is presented, dialogue quality is moderate.

5. **Evaluation:** the best sentence similarity belongs to classical Eliza because it actually is the closest generated samples. Nature of dialogue is incrementive: every new message adds new information, while previous still relevant information is also considered. Eliza does not keep incrementive information, it apply patterns on context of previous message, therefore similarity of contexts will be high. In case of generative models it constantly change the context resulting similarity of contexts becoming lower but it does not affect dialogue quality. This approach of evaluation is necessary to satisfy, but not sufficient.

   This metric is applicable for identification of outliers: samples with low similarity definitely should have low dialogue quality.

5 Conclusion

Every considered generative model has its application, even classical Eliza is still used in different chatbots. We have implemented for different models for Russian dialogue generation: pattern matching Eliza, sequence to sequence seq2seq generation, text to text RuT5 generation and generative pre-trained transformer RuGPT application. We have identified four different sentences to evaluate with cosine distance the context similarity of generated samples. Based on values of samples similarity and manual exploration of dialogue quality of generated samples, there was concluded that RuT5 has the best ability for dialogue generation. RuGPT has a nice performance as well, but it produces unreliable responses. Therefore RuT5 has ability to catch up with conversation and produce realistic responses.

References

Bayan AbuShawar and Eric Atwell. 2015. Alice chatbot: Trials and outputs. *Computación y Sistemas*, 19(4):625–632.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yurii Kuratov, Denis Kuznetsov, et al. 2018. Deeppavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Robert Frederking. 1981. A rule-based conversation participant. In *19th Annual Meeting of the Association for Computational Linguistics*, pages 83–87.

Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*, pages 946–956. Springer.

Nurullo Ismoilov and Mikhail Evgenievich Semenov. 2019. Russian language neural net chatbot with natural language processing. In *14th International Forum on Strategic Technology (IFOST-2019), October 14-17, 2019, Tomsk, Russia:[proceedings].—Tomsk, 2019.*, pages 135–138.

Klemens Lagler, Michael Schindelegger, Johannes Böhm, Hana Krásná, and Tobias Nilsson. 2013. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073.

Julia Masche and Nguyen-Thinh Le. 2017. A review of technologies for conversational systems. In *International conference on computer science, applied mathematics and applications*, pages 212–225. Springer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Bayan Abu Shawar and Eric Atwell. 2007. Different measurement metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*, pages 89–96.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *arXiv preprint arXiv:1801.01957*.

Rui Yan. 2018. " chitty-chitty-chat bot": Deep learning for conversational ai. In *IJCAI*, volume 18, pages 5520–5526.