Dialogue system for classification of interlocutors as humans or bots

Agafonov Alexey
frizman04@gmail.com

Abstract

This work presents a system, which classifies interlocutors as humans or bots on human-to-bot conversations dataset. Proposed classifier was trained in end-to-end style, with the auc roc score 0.968 on the test dataset. The dataset was presented on hackathon, which took place in Moscow Institute of Physics and Technology (MIPT) in July 2017. Hackathon participants solved this classification task in feature engineering style with the auc roc score 0.966.

1 Introduction

In this work we have solved the task of classification an interlocutor as a human or a chatbot. Classification, which was proposed earlier, makes this decision with high accuracy, but uses a set of handmade features. This feature is specific for the proposed dataset, so classification system will not have a generalization ability.

2 Datasets

In this article, we use several datasets of english chatbot-human intersections:

The first dataset (D1) was created by a framework, that joins volunteers with a randomly selected chatbots (occasionally, with another volunteer) and allows them to have a conversation. Dialogue participants are asked to discuss short text from SQuAD dataset Rajpurkar et al. (2016). They contain extracts from Wikipedia articles accompanied by questions to these extracts. Created dataset was used to train human-bot recognition system and validate one during the training.

At the second stage, we used WOCHAT Shared Task datasets that contain human-bot intersections. Such datasets contain human dialogues with different chatbots. We chose all chatbots that have more than one hundred sessions (TickTock chatbot, IRIS chatbot, Joker chatbot), and used them to evaluate generalization ability of our classification system.

Table 1: Statistics for datasets

	D1	D1 cleanup	TikTok	IRIS	JOKER
Number of dialogues in dataset	4750	2759	206	163	112
Average number of phrases	10.6	12.8	26.5	34.9	40.0
Average number of words	63.8	76.4	189.0	182.7	154.2
Average number of characters	343.9	386.3	983.1	960.5	776.6

2.1 Describe and clean up datasets

First dataset was created on DeepHach.Turing hackathon in MIPT and included 4750 sessions of human-bot/human-human intersections. There where different type of noises, which were excluded:

- Monologs : $1125(\approx 24\%)$.
- Empty sessions : $272(\approx 6\%)$.
- Sessions where were used cyrillic symbols : $176(\approx 4\%)$.

Then variations of dialogues where minimized - from dataset were excluded all samples less then $5th$-percentile and more then $95th$-percentile for next features:

- Count of phrase in dialogue.
- Count of word in dialogue.
- Number of characters in dialogue.

Dataset without noise contains 2760 sessions of human-bot/human-human intersections.

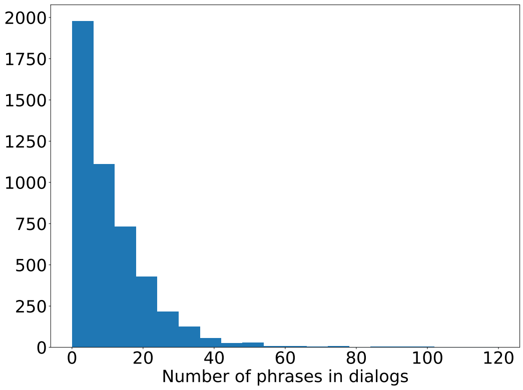

Figure 1: Dataset1 dialogues length

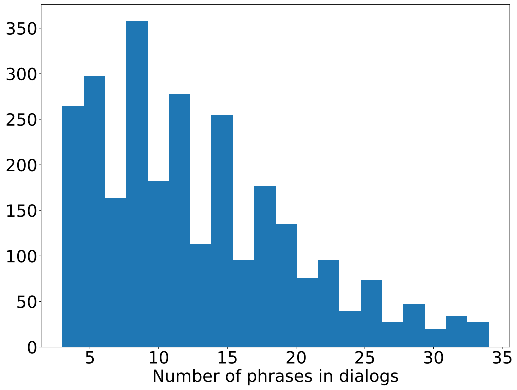

Figure 2: Dataset1 cleanup dialogues length

3 Model

We had human-chatbot conversations and human-human conversations, and is a interlocutor a chatbot or not label was predicted for each dialogue participant in such way:

3.1 Baseline

We used pre-trained word vectors, trained by fastText on Wikipedia, as word embeddings. These vectors in dimension 300 were obtained by means of the skip-gram model with default parameters described in Bojanowski et al. (2016). In order to get dialogue embeddings we averaged word embeddings for each phrase in dialogues. Then we use vanilla Long Short-Term Memory cell (LSTM) with 16 hidden units. All hidden states were subsequently flatted and pushed in a dense layer with 64 units. At the final stage of the work a dense layer with two outputs and the sigmoid activation function was obtained.

Baseline model was trained with Adam optimizer with standard hyperparameters, at 300 epochs. The results are shown in Table 2.

3.2 InferSent

Pre-trained sentence encoder InferSent (Conneau et al. 2017) was used in order to improve the baseline model. As far as it transforms a sentence to vector in dimension 4096, dialogues were presented as sequences of vectors $\in \mathbb{R}^{4096}$. Sentence embeddings were pushed in LSTM with 34 hidden units. Later, we have built our classifier on top of final cell state with two dense layers (first - 64 neurons and relu activation function, second - 2 neurons and sigmoid activation function).

Sentence embeddings approach significantly improved performance. The results are shown in Table 2.

3.3 Add extra features (InferSent+)

Finally, we have added extra features to input sentence embedding vector, that was stacked with information about the author of each phrase (first or second user).

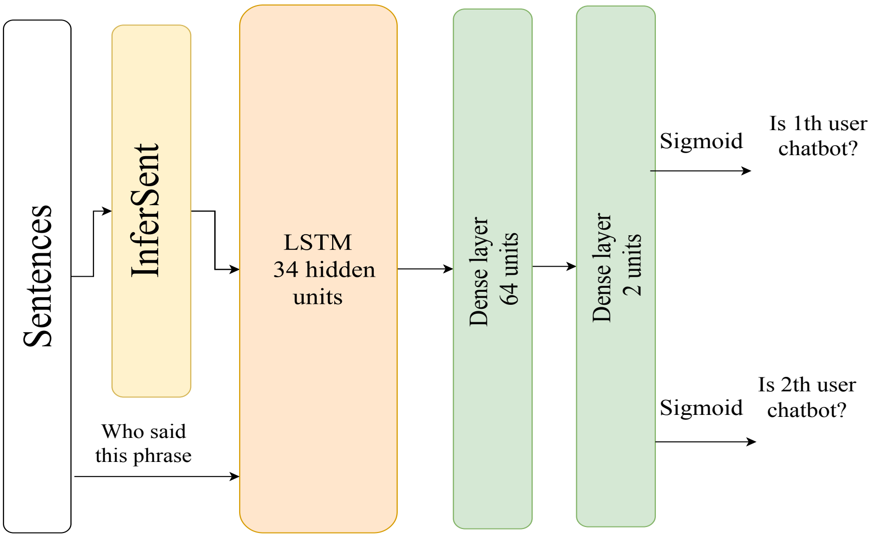

Figure 3: InferSent+ model

Model with InferSent sentence embeddings and extra features was trained with Adam too (first 200 epoch with learning rate = 10^{-3} and then 500 epoch with learning rate = 10^{-4}). The results are shown in Table 2.

Table 2: Model's quality

Model	AUC ROC	Accuracy
Baseline	0.767	0.780
InferSent	0.925	0.855
InferSent+	**0.968**	**0.904**

4 Generalization ability

In order to discover best model generalization ability, we tested it on human-bot dialogues for several chatbots in WOCHAT Shared Task datasets (TickTock chatbot, IRIS chatbot, Joker chatbot were chosen). Table 3 shows us the best model performance on different datasets (model was explained in section 3.3).

Table 3: Generalization ability

Dataset	AUC ROC	Accuracy
Train dataset (validation data)	0.968	0.904
TickTock chatbot	0.993	0.950
IRIS chatbot	0.745	0.500
Joker chatbot	0.993	0.656
WOCHAT bots together	0.788	0.735

From the Table 3 we can conclude, that this InferSent+ model (look at 3.3) unlike previous one has good performance on datasets. Thus, it can be inferred, that model has generalization ability.

Then we investigated an impact of phrase number in a dialogue on model quality. Classification system was tested on WOCHAT Shared Task datasets with diferent maximums dialogue length $lmax_i \in [1..180]$. If $lmax_i$ was smaller then length of current dialogue l_j , then part of dialogue was taken. If

$lmax_i > l_j$, then whole dialogue was classified. As shown in Fig. 4 AUC ROC score grows rapidly from 0 to 10.

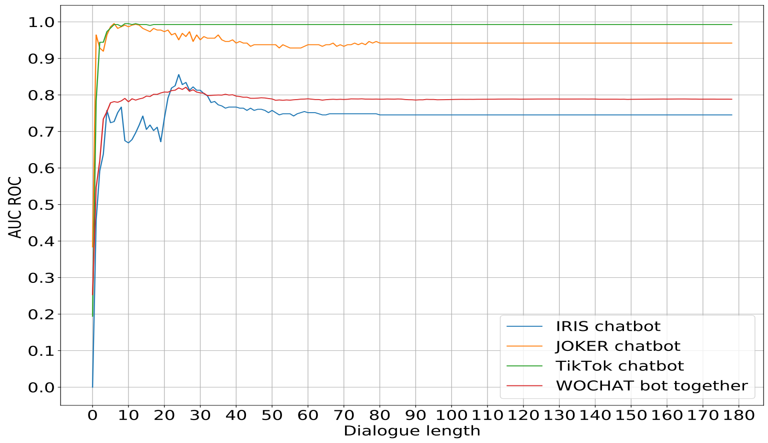

Figure 4: Impact dialogue length on AUC ROC score

Fig. 5 illustrates changing of accuracy depending on dialogue length. There is a sharp growth between 0 and 20 phrases in dialogue.

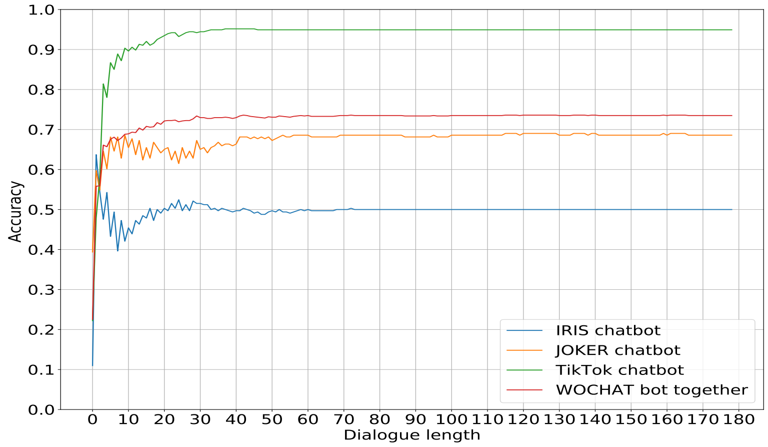

Figure 5: Impact dialogue length on Accuracy

5 Conclusion

In the work different deep learning models have been tasted in order to solve the task of classification an interlocutor as a human or a chatbot. In addition, classification system trained in end-to-end style ($AUC = 0.968$) was investigated. The last one seems to be better than model with feature engineering ($AUC = 0.966$) proposed on hackathon in Moscow Institute of Physics and Technology (MIPT) in July 2017.

References

Varvara Logacheva et al. 2017. *Automatic Quality Evaluation of Dialogues at ConvAI.* https://drive.google.com/file/d/0B3pvTVUKR3GGVFJLSE91WHc5cWM

Workshop on Chatbots and Conversational Agent Technologies. 2016. *WOCHAT Shared Task Datasets.* http://workshop.colips.org/wochat/data/

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang 2016. *SQuAD: 100,000+ Questions for Machine Comprehension of Text* Emnlp,(ii):23832392.

Bojanowski Piotr, Grave Edouard, Joulin Armand and Mikolov Tomas. 2016. *Enriching Word Vectors with Subword Information.* arXiv preprint arXiv:1607.04606

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loc Barrault and Antoine Bordes 2017. *Supervised learning of universal sentence representations from natural language inference data.* arXiv preprint arXiv:1705.02364

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas A.-Gontier, Yoshua Bengio and Joelle Pineau. 2017. *Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses.* arXiv preprint arXiv:1708.07149

Chia-Wei Liu1, Ryan Lowe1, Iulian V. Serban, Michael Noseworthy1, Laurent Charlin1, Joelle Pineau1 2016. *How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation.* arXiv preprint arXiv:1603.08023

Sepp Hochreiter and Jurgen Schmidhuber. 1997. *Long short-term memory.* Neural computation, 9(8):17351780

Ondrej Dusek, Jekaterina Novikova and Verena Rieser 2017. *Referenceless Quality Estimation for Natural Language Generation* arXiv preprint arXiv:1708.01759