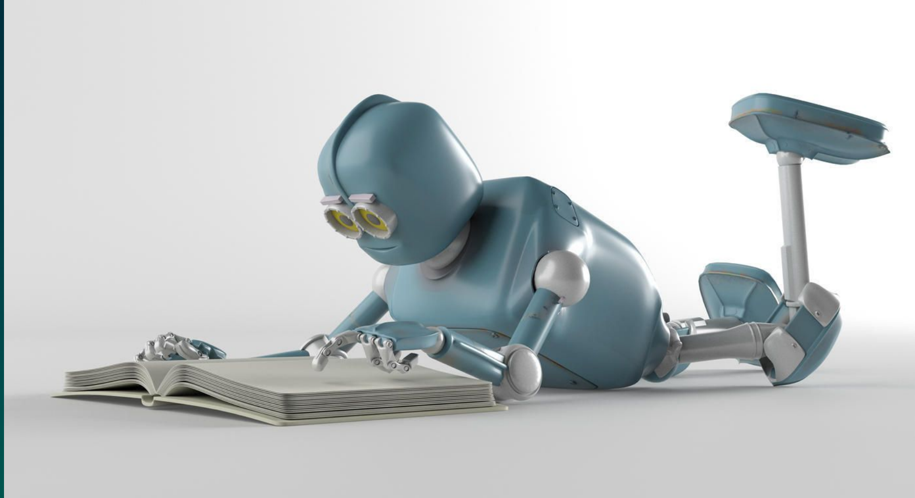
Evaluating dialogue systems

Turing test for Russian, 62 years after.

Tatiana Shavrina
rybolos@gmail.com
SberDevices
AIRI

Motivation

- ❑ Natural Language Generation models become more evolved
- ❑ Automatic Text detection challenges for English
- ❑ Russian chatbots and open-domain dialogue systems lack metrics and benchmarks



Today’s talk:

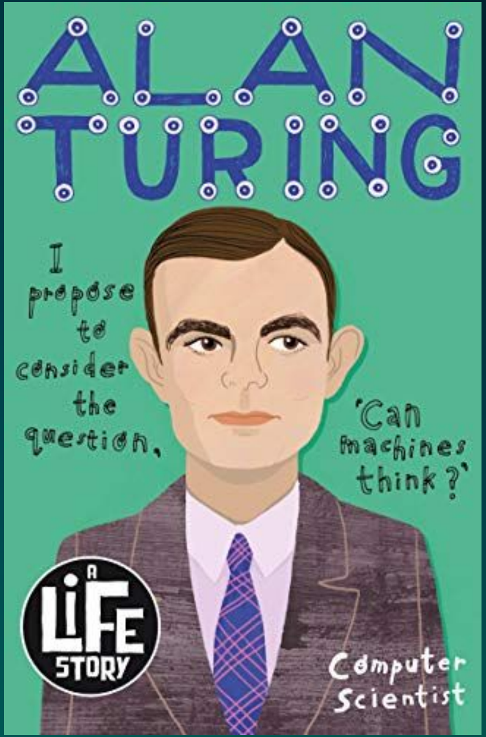
1950. How it started

❑ Method behind the Turing Test (TT)
❑ Critique
❑ Variations of TT

2022. How it’s going

❑ Amazon Alexa Prize
❑ Automatic metrics
❑ RuATD challenge
❑ Is it time for some linguistics? (finally)

Original Turing Test


ALAN TURING
I propose to consider the question, 'Can machines think?'
A LIFE STORY
Computer Scientist

TURING TEST: 50 YEARS LATER

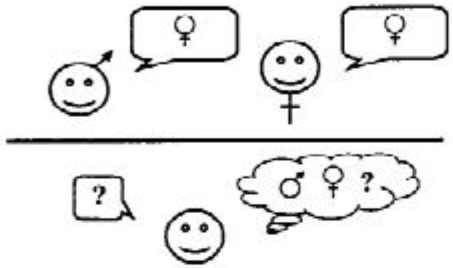


Figure 1. The Imitation Game: Stage 1.

Imitation game

The game is played with a man (A), a woman (B) and an interrogator (C) whose gender is unimportant. The interrogator stays in a room apart from A and B. The objective of the interrogator is to determine which of the other two is the woman while the objective of both the man and the woman is to convince the interrogator that he/she is the woman and the other is not.

Original Turing Test

Here is our explanation of Turing’s design: The crucial point seems to be that the notion of imitation figures more prominently in Turing’s paper than is commonly acknowledged. For one thing, the game is inherently about deception.

Turing: *‘if we are trying to produce an intelligent machine, and are following the human model as closely as we can’*

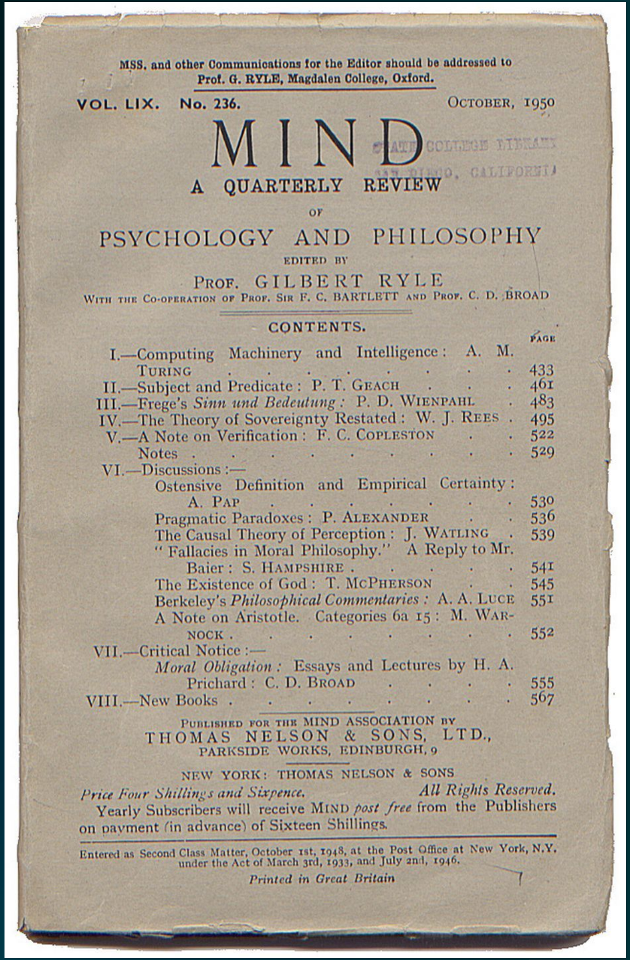
1. The reader must accept it as a fact that **digital computers can be constructed**, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely (Turing, 1950, p. 438).
2. **As I have explained, the problem is mainly one of programming.** Advances in engineering will have to be made too, but it seems unlikely that these will not be adequate for the requirements (Turing, 1950, p. 455).
3. **[The machine] may be used to help in making up its own programmes**, or to predict the effect of alterations in its own structure.

Critique

Keith Gunderson, 1964 Mind article, ‘The Imitation Game’,
The imitation game is not a test of intelligence:
- because it is finite and you can win it without using intelligence
- thinking is a general concept and playing the IG is but one example of the things that intelligent entities do

Purtill believes that the game is 'just a battle of wits between the questioner and the programmer: the computer is non-essential' (Purtill, 1971, p. 291).

The game of imitation in a general sense concerns any feature: can a person distinguish X1 from X2, and if he cannot, does this mean that the feature is not significant?



MSS. and other Communications for the Editor should be addressed to
Prof. G. RYLE, Magdalen College, Oxford.

VOL. LIX. No. 236. OCTOBER, 1950

MIND

A QUARTERLY REVIEW
OF
PSYCHOLOGY AND PHILOSOPHY

EDITED BY
PROF. GILBERT RYLE
WITH THE CO-OPERATION OF PROF. SIR F. C. BARTLETT AND PROF. C. D. BROAD

CONTENTS.

PUBLISHED FOR THE MIND ASSOCIATION BY
THOMAS NELSON & SONS, LTD.,
PARKSIDE WORKS, EDINBURGH, 9

NEW YORK: THOMAS NELSON & SONS

Price Four Shillings and Sixpence. *All Rights Reserved.*
Yearly Subscribers will receive MIND *post free* from the Publishers on payment (in advance) of Sixteen Shillings.

Entered as Second Class Matter, October 1st, 1948, at the Post Office at New York, N.Y. under the Act of March 3rd, 1933, and July 2nd, 1946.

Printed in Great Britain

Critique

1980, Searle: "Minds, brains, and programs"
Behavioral and Brain Sciences 3, 417–457.
1989, Harnad: "Minds, Machines and Searle"
1990: Michael Dyer: "Minds, Machines, Searle and Harnad”
...
2000, Harnad: "Minds, Machines and Turing: The Indistinguishability of Indistinguishables."
2001, Harnad: "MINDS, MACHINES AND SEARLE 2"

Stevan Harnad

Variations

- ❑ **Total Turing Test (TTT)** (Harnad, 1991) - requires the machines to respond to all of our inputs rather than just verbal ones.
- ❑ **Total Total Turing Test (TTTT)** - requires neuromolecular indistinguishability. ‘[TTTT] is as much as a scientist can ask, for the empirical story ends there’
- ❑ **Kugel Test (KT)** (Kugel, 1990) - play the imitation game, but do not tell the participants what distinguishing feature we are looking at.
- ❑ **Inverted Turing Test (ITT)** (Watt, 1996) - naive psychology, the consistency of the author's "cognitive profile"
- ❑ **Truly Total Turing Test (TRTTT)** (Schweizer, 1998) Evolutionary criteria for intelligence

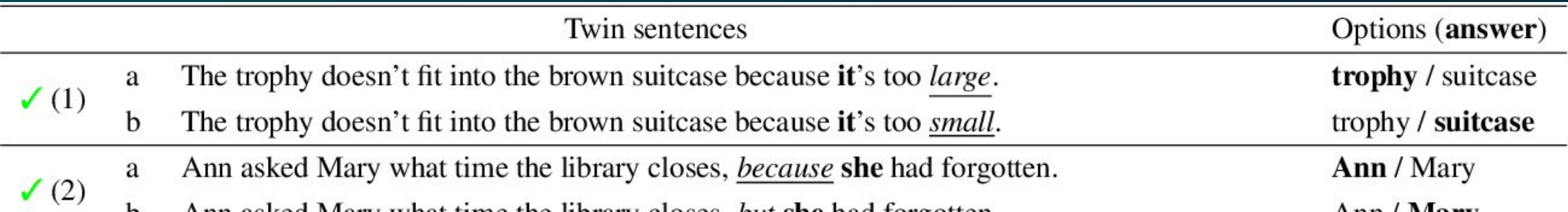
Variations

- Winograd schema - linguistic test for logic. Contains textual questions about the properties of objects and about common everyday situations, where the correct answer necessarily requires disambiguation [Winograd 1972].

“If Ivan had a donkey, he would beat him."
Who beats whom?

We adapted the Winograd test for the Russian language for the first time in 2019, Russian SuperGLUE benchmark

Terry Winograd (on the right)



		Twin sentences	Options (**answer**)
✓ (1)	a	The trophy doesn’t fit into the brown suitcase because **it**’s too *large*.	**trophy** / suitcase
	b	The trophy doesn’t fit into the brown suitcase because **it**’s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**

Variations!

- **Minimum intelligent signal test (MIST)** - a question-answer test that requires only “yes” / “no” answers, but on difficult questions. The machine requires knowledge, logic. Such a test, proposed in [McKinstry 1997], reduces the subjectivity of judging in the original Turing test, and also provides a metric for the "humanity" of the system's intelligence - that is, the proportion of correct answers;
- **Turing test with a specialist (Subject-matter expert Turing test)** - a kind of test with expert specialized knowledge. The correct answers should not differ from the answers of real experts [McCorduck 2004];
- **Ebert test - tests for humor.** The test involves speech synthesis, and it must be good enough to make the judges laugh at the joke of the machine [Pasternack 2011].

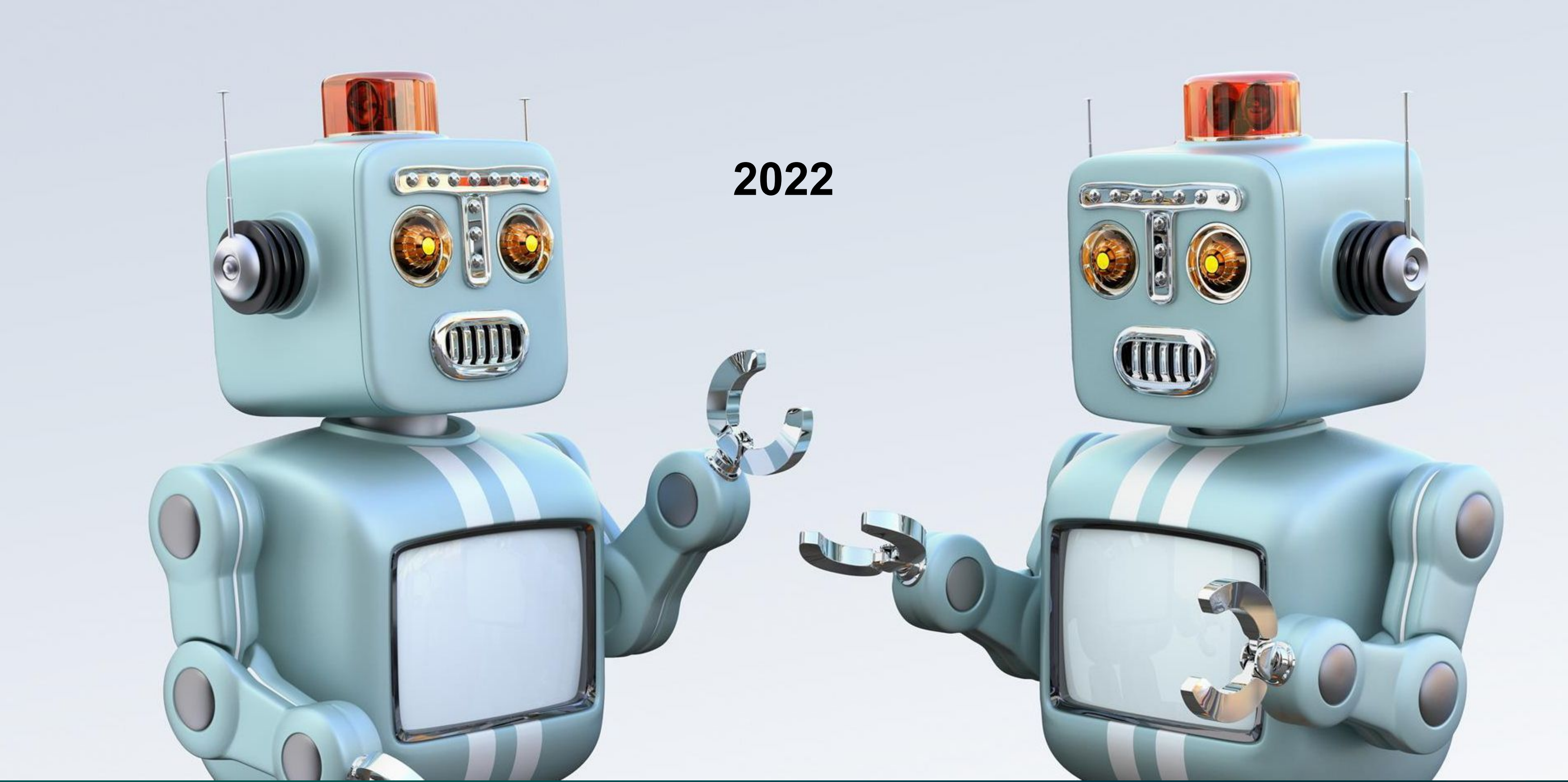
Variations!

- **Minimum intelligent signal test (MIST)** - a question-answer test that requires only “yes” / “no” answers, but on difficult questions. The machine requires knowledge, logic. Such a test, proposed in [McKinstry 1997], reduces the subjectivity of judging in the original Turing test, and also provides a metric for the "humanity" of the system's intelligence - that is, the proportion of correct answers;
- **Turing test with a specialist (Subject-matter expert Turing test)** - a kind of test with expert specialized knowledge. The correct answers should not differ from the answers of real experts [McCorduck 2004];
- **Ebert test - tests for humor.** The test involves speech synthesis, and it must be good enough to make the judges laugh at the joke of the machine [Pasternack 2011].

What can we implement: logic test, yes-no questions, specific questions, humour

If you want to talk about what a model or a simulation can or cannot do, first get it to run.

(Harnad, 1989)



2022

Measuring with quantity

Amazon Alexa Prize

real-time involvement of users and judges: thousands of dialogues daily

❏ **2022**: 4th time

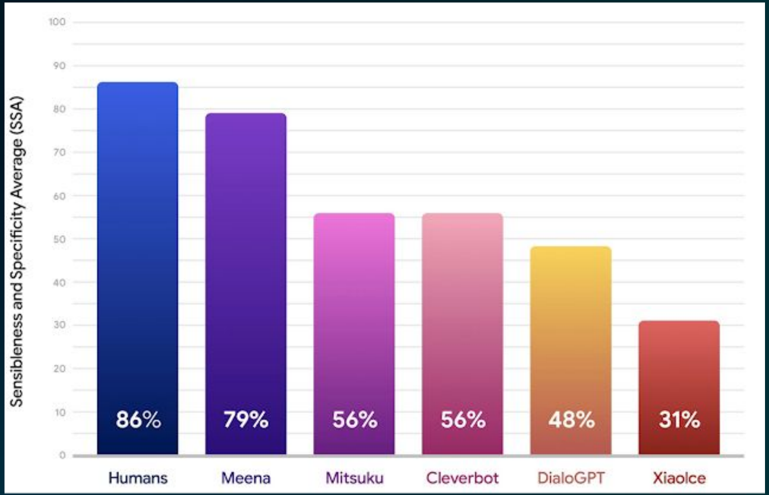
❏ **overall goal**: 20 min of avg dialogue - not achieved

Measuring with quantity

Automatic metrics for chatbot evaluation (not goal-oriented)

Google Meena: Human Evaluation Metric
Sensibleness and Specificity Average (SSA)

NeurIPS 2019: Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems
and others



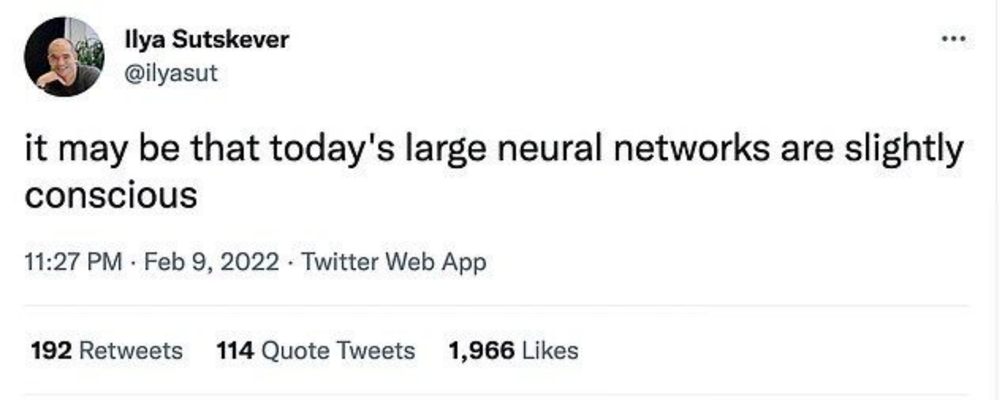
Main idea: the automatic metric should be correlated with human annotators (basically, inheriting TT problems)

The Case of Ilya Sutskever

A short meaningful sentence of frequency n-grams may well occur many times in a web-corpus and be easily reproduced by the simplest statistical model.

Thus, the very definition of a specific automatic text can be an extremely difficult task for an attentive annotator, and even for an engineer directly involved in developing generative models.

We achieved the "indistinguishability by the engineers themselves"

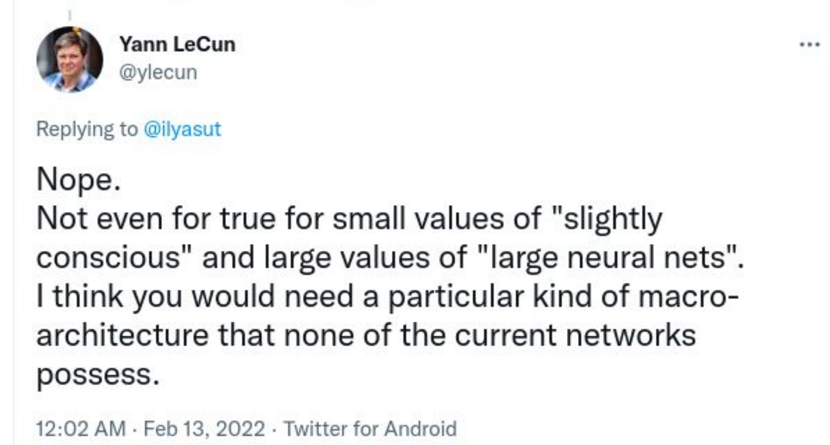


Ilya Sutskever
@ilyasut

it may be that today's large neural networks are slightly conscious

11:27 PM · Feb 9, 2022 · Twitter Web App

192 Retweets **114** Quote Tweets **1,966** Likes



Yann LeCun
@ylecun

Replying to @ilyasut

Nope.
Not even for true for small values of "slightly conscious" and large values of "large neural nets".
I think you would need a particular kind of macro-architecture that none of the current networks possess.

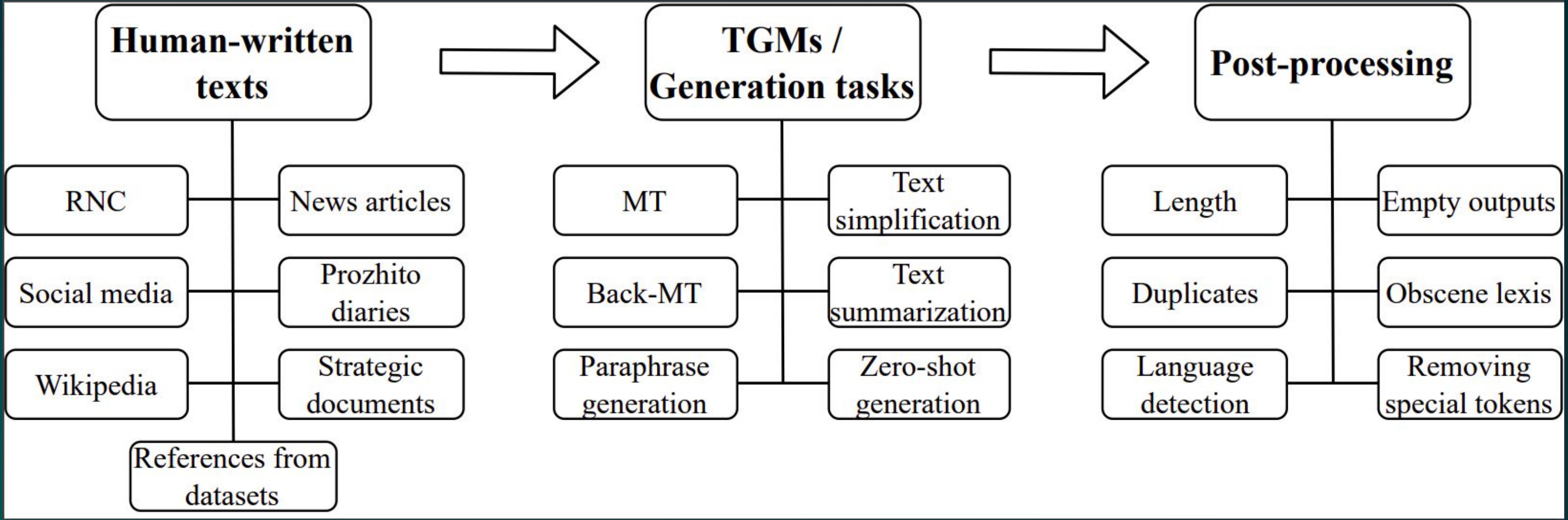
12:02 AM · Feb 13, 2022 · Twitter for Android

RuATD Challenge

https://www.kaggle.com/c/ruatd-2022-bi/
https://www.kaggle.com/c/ruatd-2022-multi-task/

Automatic text detection for Russian

- ❏ Binary task: human/non-human clf
- ❏ Multitask: human/ generation/ summarization/ rewrite/ translation



Human-written texts

RNC
News articles
Social media
Prozhito diaries
Wikipedia
Strategic documents
References from datasets

TGMs / Generation tasks

MT
Text simplification
Back-MT
Text summarization
Paraphrase generation
Zero-shot generation

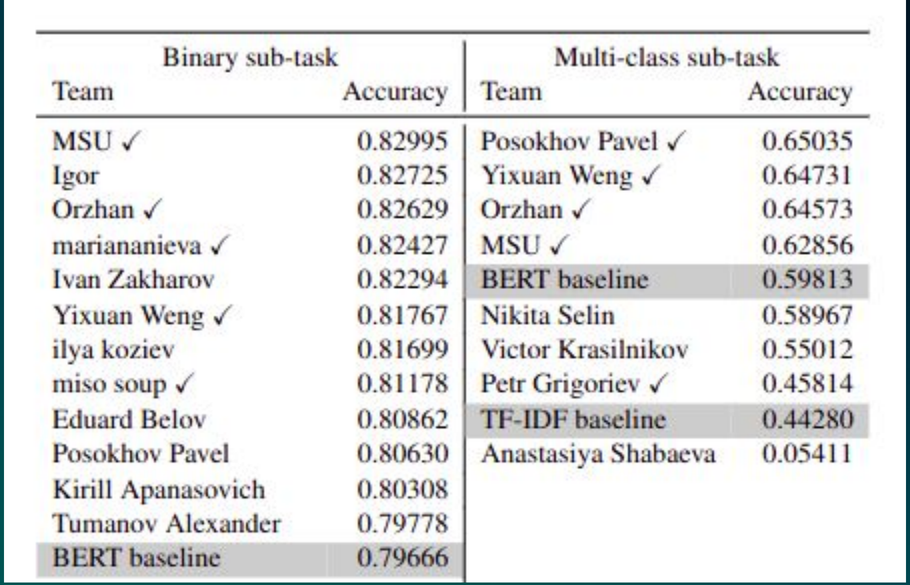
Post-processing

Length
Empty outputs
Duplicates
Obscene lexis
Language detection
Removing special tokens

RuATD Challenge

https://www.kaggle.com/c/ruatd-2022-bi/
https://www.kaggle.com/c/ruatd-2022-multi-task/

- ❏ Binary task: human/non-human clf
- ❏ Multitask: human/ generation/ summarization/ rewrite/ translation



Binary sub-task		Multi-class sub-task	
Team	Accuracy	Team	Accuracy
MSU ✓	0.82995	Posokhov Pavel ✓	0.65035
Igor	0.82725	Yixuan Weng ✓	0.64731
Orzhan ✓	0.82629	Orzhan ✓	0.64573
mariananieva ✓	0.82427	MSU ✓	0.62856
Ivan Zakharov	0.82294	BERT baseline	0.59813
Yixuan Weng ✓	0.81767	Nikita Selin	0.58967
ilya koziev	0.81699	Victor Krasilnikov	0.55012
miso soup ✓	0.81178	Petr Grigoriev ✓	0.45814
Eduard Belov	0.80862	TF-IDF baseline	0.44280
Posokhov Pavel	0.80630	Anastasiya Shabaeva	0.05411
Kirill Apanasovich	0.80308		
Tumanov Alexander	0.79778		
BERT baseline	0.79666		

Results:
short text are hard to distinguish!

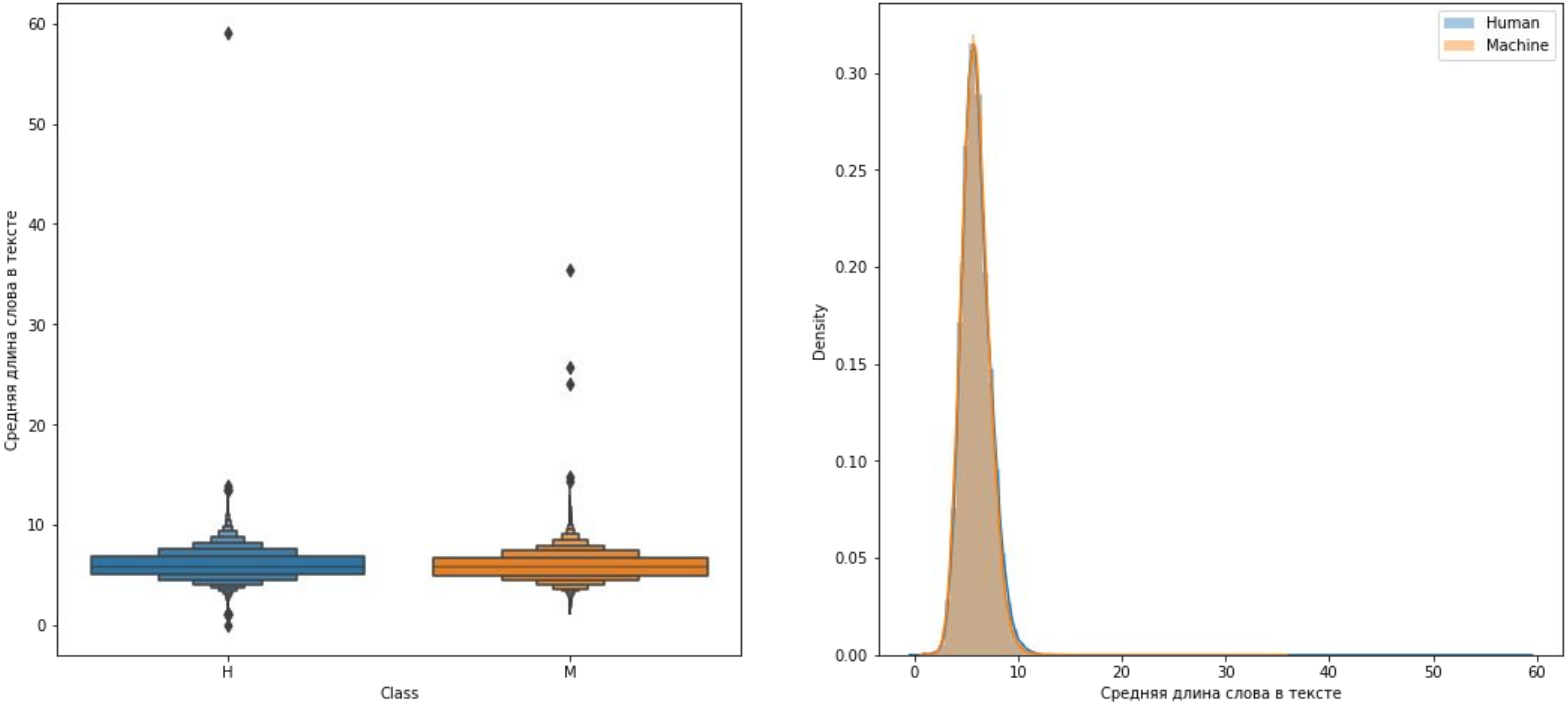
on the texts longer than 23 words (about a quarter of all RuATD texts) top models were able to score over 0.95 accuracy

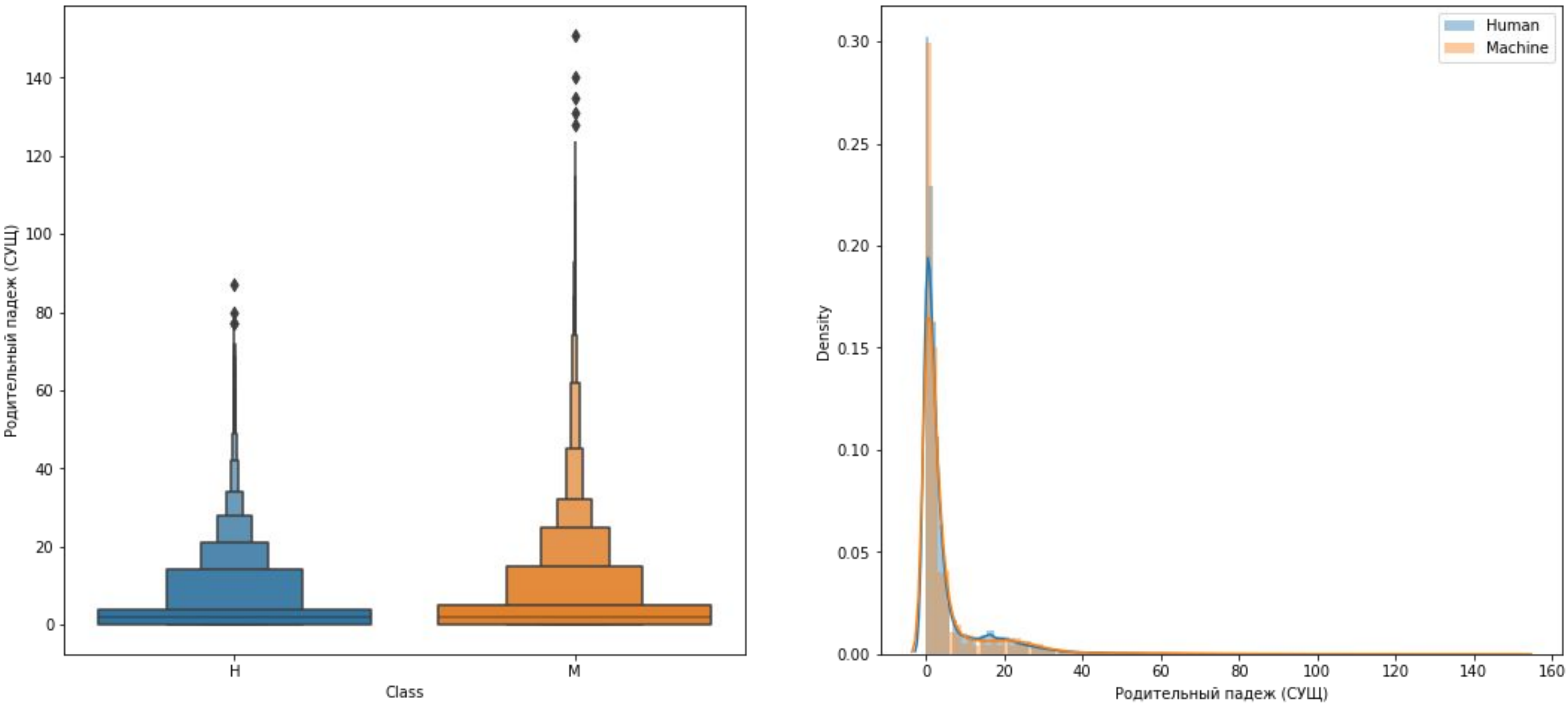
Is it time for some linguistics?

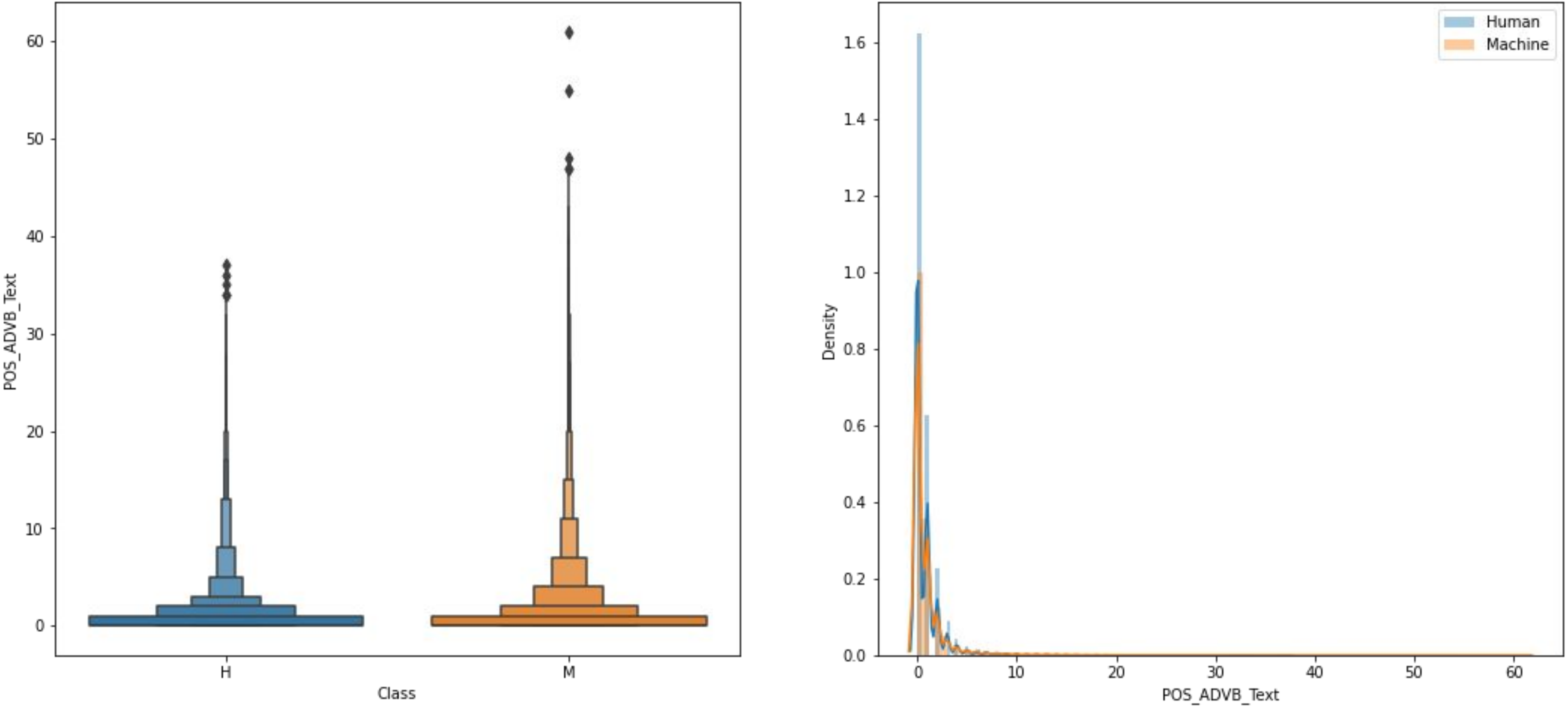
Amazon Alexa Prize time for Russian?

We still lack
1) automated metrics
2) public benchmarks
3) any playground for public comparison of dialogue systems

RuATD take-away: Modern automatic texts for Russian do differ from natural ones!


60
50
40
30
20
10
0
Средняя длина слова в тексте
H
M
Class
Human
Machine
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
10
20
30
40
50
60
Средняя длина слова в тексте


Родительный падеж (СУЩ)
0
20
40
60
80
100
120
140
H
M
Class
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Human
Machine
0
20
40
60
80
100
120
140
160
Родительный падеж (СУЩ)


POS_ADVB_Text
60
50
40
30
20
10
0
H
M
Class
Human
Machine
Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
POS_ADVB_Text

Take-away points

- ❑ The TT is still good for most of the models, but not for the SOTA
- ❑ Automatic metrics correlated with TT are inheriting its problems
- ❑ The artificial text detection mechanisms are evolving, however, the results do not match human evaluation
- ❑ Modern automatic texts for Russian do differ from natural ones! the differences are not human-distinguishable, but depend on syntactic frequences
- ❑ We need TT for Russian!
- ❑ And better metrics for chatbot evaluation metrics
- ❑ and shorter lists

References

- [Turing 1950] Turing, A. (1950), 'Computing Machinery and Intelligence', Mind 59(236), pp. 433-460.
- [Pinar Saygin et al. 2000] Pinar Saygin, A., Cicekli, I. & Akman, V. Turing Test: 50 Years Later. Minds and Machines 10, 463–518 (2000). https://doi.org/10.1023/A:1011288000451
- [Winograd 1972] Winograd T. Understanding natural language. Cognitive Psychology, 1972, 3(1): 1–191. https://doi.org/10.1016/0010-0285(72)90002-3.
- [McKinstry 1997] McKinstry C. Minimum Intelligent Signal Test: An alternative Turing Test", Canadian Artificial Intelligence, 1997, 41: pp. 35-47.
- [McCorduck 2004] McCorduck P. Machines who think. 2 nd edn. Natick (MA): A. K. Peters Ltd., 2004.
- [Pasternack 2011] Pasternack A. (18 April 2011). "A MacBook May Have Given Roger Ebert His Voice But An iPod Saved His Life" [video]. Archived from the original on 6 September 2011. Retrieved 12 September 2011. https://www.vice.com/en/article/4xxa7j/a-macbook-gave-roger-ebert-his-voice-an-ipod-saved-his-life.
- [Harnad 1989] Harnad, S. (1989), 'Minds, Machines and Searle', Journal of Experimental and Theoretical Artificial Intelligence 1(1), pp. 5–25.
- [Bringsjord 1994] Bringsjord, S. (1996), 'The Inverted Turing Test is Provably Redundant'. Psycoloquy 7(29). http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?7.29.
- [Watt 1996] Watt, S. (1996), 'Naive Psychology and the Inverted Turing Test', Psycoloquy 7(14). http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?7. 14
- [Schweizer 1998] Schweizer, P. (1998), 'The Truly Total Turing Test', Minds and Machines 8, pp. 263–272.

Thank you for human attention!

@rybolos
Tatiana Shavrina