LOL NLP: an overview of computational humor

Pavel Braslavski

18.12.2019

About myself

- Research/academia: JetBrains Research/Higher School of Economics SPb/Ural Federal University
- Past industrial experience: Yandex/SKB Kontur
- Recent research interests: question answering, fiction analysis, computational humor
- RuSSIR (13th – 24-28 August 2020)

Homepage: http://kansas.ru/pb/

Humor

…is the tendency of experiences to provoke laughter and provide amusement.

Most people are able to experience humor and thus are considered to have a sense of humor. The hypothetical person lacking a sense of humor would likely find the behavior inducing it to be inexplicable, strange, or even irrational.

from Wikipedia

Humor is a promising area for studies of intelligence and its automation: it is hard to imagine a computer passing a rich Turing test without being able to understand and produce humor.

West & Horvitz, AAAI2019

Humor at Alexa Prize competition

…it’s amazing to see that now humor is coming in… Good sense of humor is a sign of intelligence in my mind and something very hard to do.

Rohit Prasad
vice president and head scientist of Amazon Alexa
in conversation with Lex Fridman, 14 December 2019

https://youtu.be/Ad89JYS-uZM

Tell me which are funny, which are not – and which get a giggle first time but are cold pancakes without honey to hear twice.

Robert Heinlein, The Moon Is a Harsh Mistress

Plan

- Humor recognition
- Humor generation
- Humor evaluation
- Humor datasets

Research team

Vladislav Blinov

Valeria Bolotova

Kirill Mishchenko

Humor detection

Humor Detection & Ranking

- Humor detection [Taylor and Mazlack, 2004; Mihalcea and Strapparava, 2005; Kiddon and Brun, 2011; Yang et al., 2015; Zhang and Liu, 2014; Liu et al., 2018; Cattle and Ma, 2018; Ermilov et al., 2018]
- Humor ranking [Shahaf et al., 2015; Potash et al., 2017]

Humor classifier [Mihalcea & Strapparava, 2005]

- 16K one-liners/16 non-funny sentences
- Features: alliteration/rhyme, antonymy (WordNet), adult slang, content words
- Classifiers: NB and SVM

Heuristic	One-liners Reuters	One-liners BNC	One-liners Proverbs
Alliteration	74.31%	59.34%	53.30%
Antonymy	55.65%	51.40%	50.51%
Adult slang	52.74%	52.39%	50.74%
ALL	76.73%	60.63%	53.71%

Classifi er	One-liners Reuters	One-liners BNC	One-liners Proverbs
Na¨ı ve Bayes	96.67%	73.22%	84.81%
SVM	96.09%	77.51%	84.48%

Humor Classifier [Blinov et al., 2017]

Features:

- *tf-idf* weighted unigrams and bigrams, *df* > *2*;
- 300-dimensional *doc2vec* representations.

Logistic Regression 10-fold cross-validation accuracy: **0.887**

Humor anchors [Yang et al., 2015]

- Humor features:
  - incongruity,
  - ambiguity,
  - interpersonal effect (sentiment/subjectivity),
  - phonetic style.
- 'Humor anchors' – structures enabling humorous effect

	Pun of the Day				16000 One Liners			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
HCF	0.705	0.696	0.736	0.715	0.701	0.685	0.746	0.714
Bag of Words	0.632	0.623	0.686	0.650	0.673	0.708	0.662	0.684
Language Model	0.627	0.602	0.762	0.673	0.635	0.645	0.596	0.620
Word2Vec	0.833	0.804	0.880	0.841	0.781	0.767	0.809	0.787
SaC Ensemble	0.763	**0.838**	0.655	0.735	0.662	0.628	0.796	0.701
Word2Vec+HCF	**0.854**	0.834	**0.888**	**0.859**	**0.797**	**0.776**	**0.836**	**0.805**

Transformer Gets the Last Laugh [Weller and Seppi, 2019]

- 14K jokes from reddit:

Method	Body	Punchline	Full
CNN	0.651	0.684	0.688
Transformer	**0.661**	**0.692**	**0.724**
Human (General)	0.493	0.592	0.663

- 16K one-liners:

Previous Work:	Accuracy	Precision	Recall	F1
Word2Vec+HCF	0.797	0.776	0.836	0.705
CNN	0.867	0.880	0.859	0.869
CNN+F	0.892	0.886	0.907	0.896
CNN+HN	0.892	0.889	0.903	0.896
CNN+F+HN	0.894	0.866	**0.940**	0.901

Our Methods:	Accuracy	Precision	Recall	F1
Transformer	**0.930**	**0.930**	0.931	**0.931**

Humor Generation

Humor Generation

- HAHAcronym: funny deciphers for acronyms [Stock and Strapparava, 2005]
- ‘Adult’ puns from short text messages by lexical replacement [Valitutti et al., 2013]
- Pun generation from a pair of homophones [He et al., 2019]

Pun generation with surprise [He & Laing, 2019]

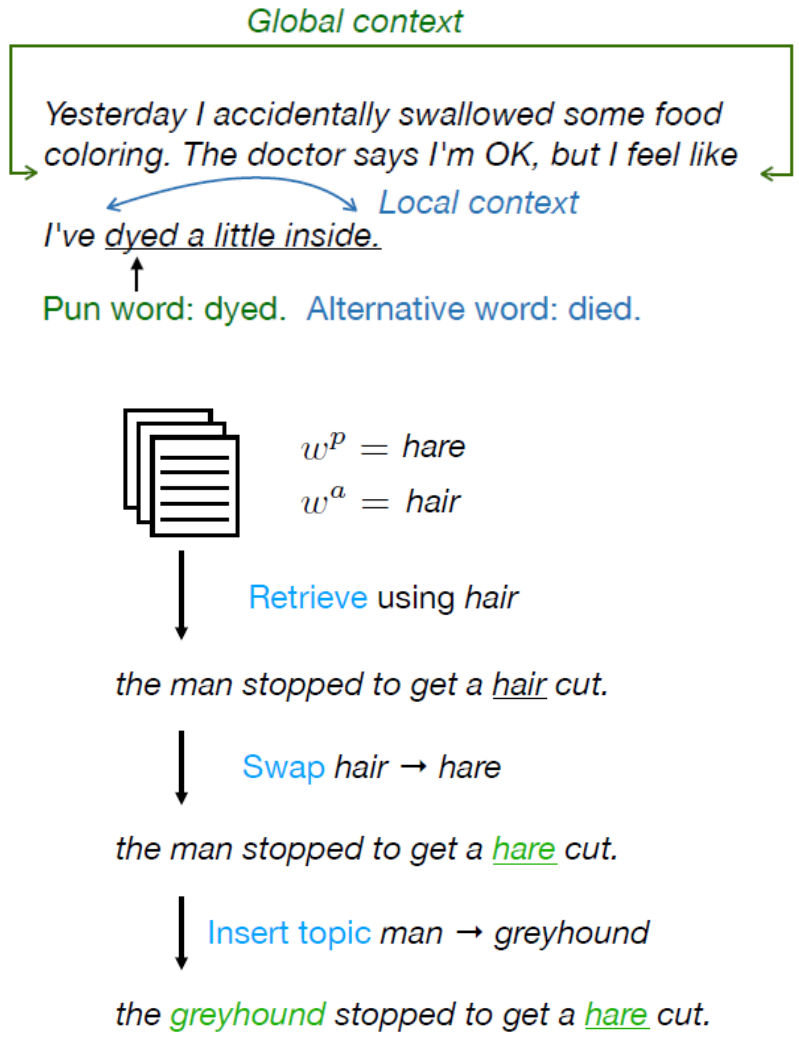
- Puns based on *homophones*
- Local vs global context

$$S(c) \stackrel{\text{def}}{=} -\log \frac{p(w^{\text{p}} \mid c)}{p(w^{\text{a}} \mid c)} = -\log \frac{p(w^{\text{p}}, c)}{p(w^{\text{a}}, c)}. \quad (1)$$

$$S_{\text{local}} \stackrel{\text{def}}{=} S(x_{p-d:p-1}, x_{p+1:p+d}), \quad (2)$$

$$S_{\text{global}} \stackrel{\text{def}}{=} S(x_{1:p-1}, x_{p+1:n}), \quad (3)$$

$$S_{\text{ratio}} \stackrel{\text{def}}{=} \begin{cases} -1 & S_{\text{local}} < 0 \text{ or } S_{\text{global}} < 0, \\ S_{\text{local}} / S_{\text{global}} & \text{otherwise.} \end{cases} \quad (4)$$



Results

Method	Example	Rating
1. Pun/alternative word pair: butter – better		
NJD	He is going to come up with the butter a 'very good' approach to the world's economic crisis, the world's biggest economic climate.	1
SURGEN	Well, gourmet did it, he thought, it'd butter be right.	2
Human	Why did the dairy churn? The less said, the butter...	1.5
2. Pun/alternative word pair: peace – piece		
NJD	Further, he said, at the end of the peace, it's not clear that it will be a good example.	1
SURGEN	That's because negotiator got my car back to me in one peace.	1.5
Human	Life is a puzzle; look here for the missing peace.	3
3. Pun/alternative word pair: flour – flower		
NJD	Go, and if you are going on the flour.	1
SURGEN	Butter want to know who these two girls are, the new members of the holy flour.	1.5
Human	Betty crocker was a flour child.	4.5
4. Pun/alternative word pair: wait – weight		
NJD	Gordon Brown, Georgia's prime minister, said he did not have to wait, but he was not sure whether he had been killed.	0
SURGEN	Even from the outside, I could tell that he'd already lost some wait.	2
Human	Patience is a virtue heavy in wait.	3

Method	Success	Funniness	Grammar
NJD	9.2%	1.4	2.6
R	4.6%	1.3	**3.9**
R+S	27.0%	1.6	3.5
R+S+T+M	28.8%	**1.7**	2.9
SURGEN	**31.4%**	**1.7**	3.0
Human	78.9%	3.0	3.8

IR-based humor generation

IR-based Humor Generation [Blinov et al., 2017]

- humorous response generation as search;
- large collection of jokes;
- different evaluation approaches.

Humorous Response Generation using IR

Data:

- Funny tweets;

Models:

- BM25;
- Query-Term Reweighting;
- doc2vec.

Evaluation:

- community question answering (CQA);
- lab settings.

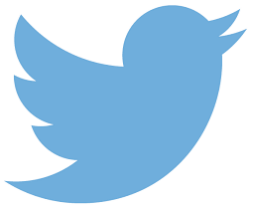
Twitter Joke Collection

Accounts (103 total):

- “funny Twitter accounts” lists;
- 20,000+ followers.

Tweets (300,876 total):

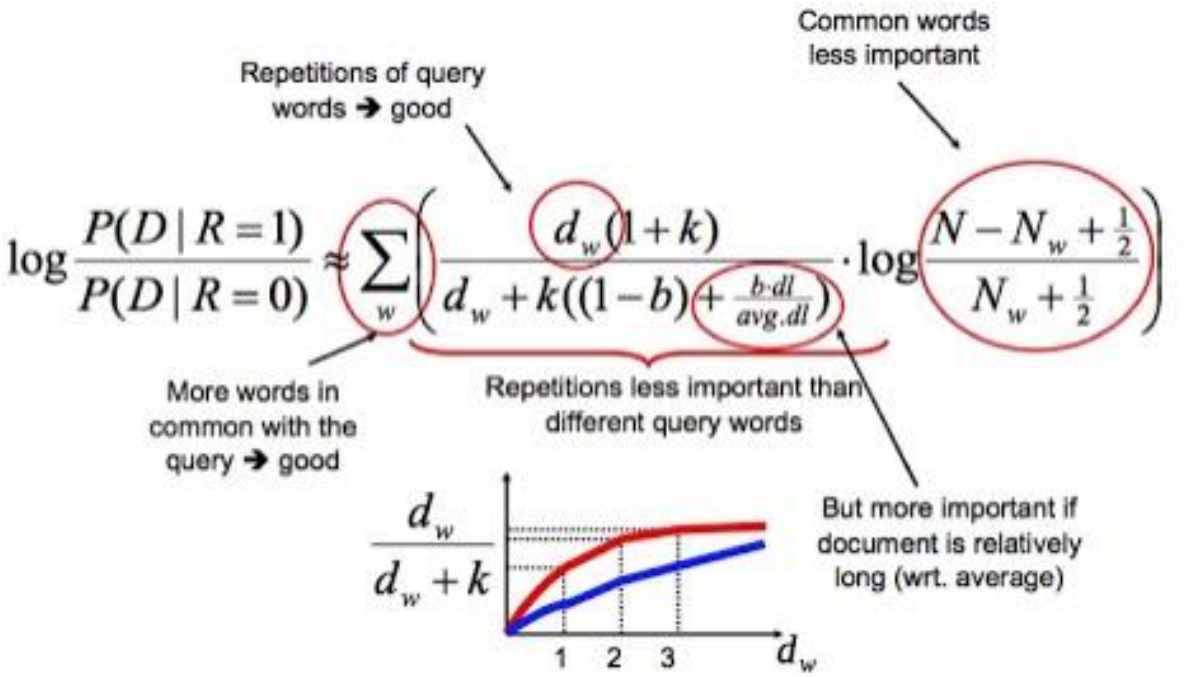
- text-only (w/o images, videos, or URLs);
- with at least 30 likes or retweets;
- duplicates removed.



Tweets Classification

Collection/account	Classified as humorous	Tweet count
Joke Collection (103 accounts)	258,466/85.9%	300,876
The Wall Street Journal (wsj)	142/9.7%	1,464
The Washington Post (washingtonpost)	195/21.5%	907
The New York Times (nytimes)	240/19.8%	1,210
Donald J. Trump (realDonaldTrump)	7,653/59.1%	12,939

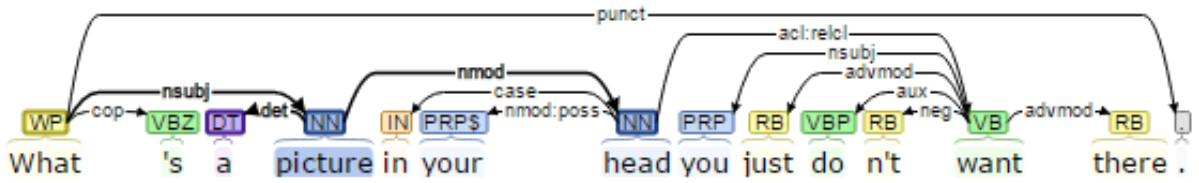
BM25


Common words less important
Repetitions of query words ➔ good
$\log \frac{P(D \mid R=1)}{P(D \mid R=0)} \approx \sum_w \left(\frac{d_w(1+k)}{d_w + k((1-b) + \frac{b \cdot dl}{avg.dl})} \cdot \log \frac{N - N_w + \frac{1}{2}}{N_w + \frac{1}{2}} \right)$
More words in common with the query ➔ good
Repetitions less important than different query words
But more important if document is relatively long (wrt. average)
$\frac{d_w}{d_w + k}$
1
2
3
d_w

Query Term Reweighting

- inspired by **humor anchors** [Yang, 2015];
- reweights BM25 scores based on syntactic roles of query terms;
- learned on 573 QA pairs from *Jokes & Riddles* of Yahoo!Answers.

Question:



Best Answer:

The picture of Clinton ... it can be either one of them ... :)

Question:

LOCATION

Why is the Earth flat ?

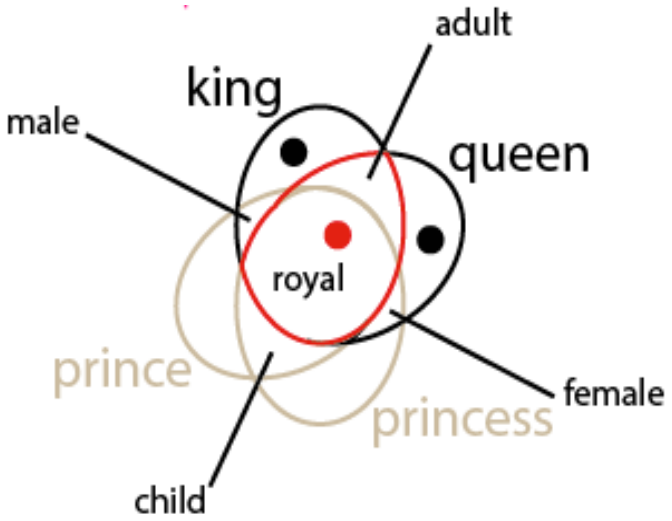
Best answer:

LOCATION

I have never understood how the Earth is

flat and other planets are reportedly round .

doc2vec



- semantically close Q&A even w/o word overlap;
- ranking based on Q/A cosine similarity;
- trained on the English Wikipedia.

Evaluation

- Yahoo!Answers
  - real-life environment
  - many judges
  - scales poorly
  - hard to interpret
- lab settings
  - controllable
  - few judges
  - does it represent an ‘average’ user?

Yahoo!Answers Evaluation

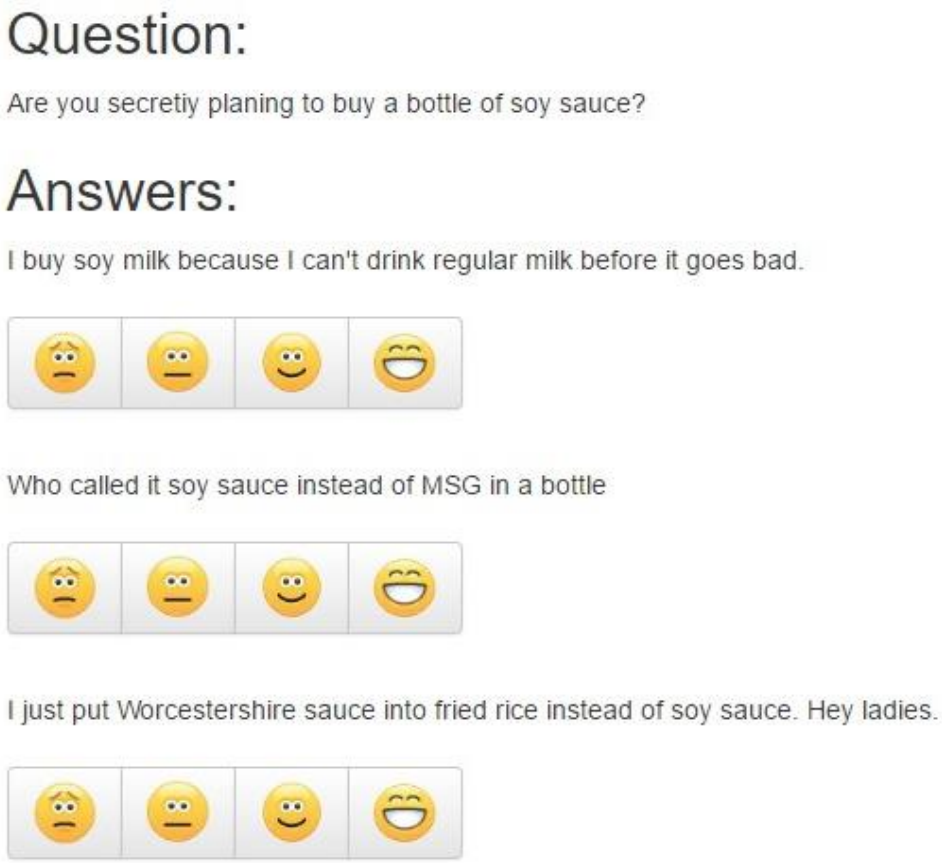
- top-1 from each model;
- *Jokes & Riddles* category;
- 96 questions.

Y!A Evaluation Results

Model	Answers	+	--	Best answer	Users below
BM25	96	19	3	0	16.5%
QTR	96	14	1	2	16.1%
doc2vec	96	15	2	2	14.3%
Oracle	96	23	1	4	19.4%
User1	23	25	0	1	32.1%
User2	20	33	1	0	33.1%
User3	20	13	1	2	15.9%

Lab Evaluation

- top-3 responses of each model;
- 50 questions;
- 3 assessors;
- dedicated evaluation interface;
- relevance scores from 0 to 3;
- 3 pairs evaluated at a time.



Lab Evaluation Results (50 stimuli)

Model	top-1	DCG@3
BM25	1.34	2.78
QTR	1.15	2.38
doc2vec	1.25	2.63
Oracle	1.91	3.61

Responses & Evaluation Scores

Score	Stimulus	Response
3.00	Does evolution being a theory make it subjective?	There is no theory of evolution, just a list of creatures Chuck Norris allows to live.
2.67	Can you find oil by digging holes in your backyard?	Things to do today: 1.Dig a hole 2. Name it love 3. Watch people fall in it.
1.33	Why don't they put zippers on car doors?	Sick of doors that aren't trap doors.
0.67	What if you're just allergic to working hard?	You're not allergic to gluten.
0.33	What test do all mosquitoes pass?	My internal monologue doesn't pass the Bechdel test. :(

IR-based Humor Generation: Summary

- search is a promising approach to humorous response generation;
- the ‘oracle’ model indicates that there is an abundant room for the improvement of the answer ranking;
- humor evaluation is hard and methodology needs to be revised.

Humor Evaluation

Why Humor Evaluation?

- Massive generation ‘in the wild’ vs. few handcrafted rules.
- Evaluation is crucial for measuring progress.
- Can we get rid of subjectivity of crowdworkers?

Humor Evaluation [Braslavski et al., 2018]

- 30 dialog jokes from different sources

Source of jokes	Count	Average score in our experiment
Jester	7	2.32
Siri	3	1.76
Yahoo!Answers	5	1.73
Automatically generated	5	1.80
Reddit	5	2.37
Twitter	5	1.82
Total	30	2.01

Q: *Am I the coolest person in the world?*
A: *Nope. That person lives in Antarctica.*

Q: *How did the hipster burn his mouth?*
A: *He ate a cookie BEFORE they were cool!*

Evaluation Interface

Evaluation Results

	Group	MT	# MT	V	# V	All	# All
Age Group	18–30	2.05	46	2.07	77	2.06	123
	31–40	1.89	37	2.04	54	1.98	91
	41–50	1.80	18	2.02	29	1.94	47
	51–60	1.84	10	1.97	6	1.89	16
	61+	1.73	1	3.33	1	2.53	2
Sex	Male	1.92	52	2.01	82	1.97	134
	Female	1.95	60	2.10	85	2.04	145
Language	Average	–	–	2.16	15	2.16	15
	Good	2.25	5	2.10	69	2.11	74
	Bilingual	2.11	3	2.06	39	2.07	42
	Native	1.91	104	1.95	44	1.92	148
	Global	1.93	112	2.06	167	2.01	279

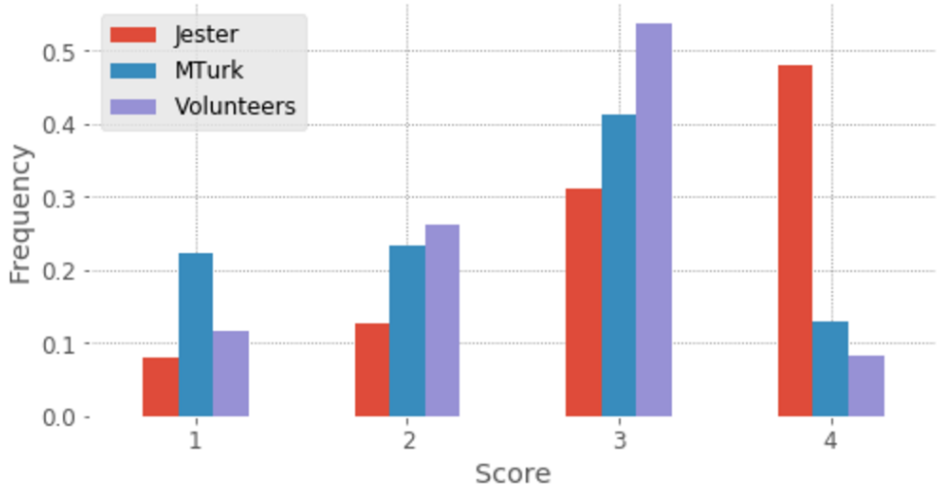
Evaluation Results -2

Highest variation native/no-native:

- ***Q:*** *Why did 10 die?*
- ***A:*** *He was in the middle of 9/11*

Highest variation in male/female:

- ***Q:*** *What is the meaning of life?*
- ***A:*** *All evidence to date suggests it is chocolate.*



Q: *How many programmers does it take to change a lightbulb?*
A: *NONE! That's a hardware problem*

Evaluation: Summary

- MTurk is suitable for humor evaluation.
- Age/language proficiency/gender influence joke ratings.
- Jokes degrade over time → re-using evaluation is questionable.

Humor Datasets

Datasets

Dataset	Description	Reference
One-liners	16K one-liners / 16K headlines/proverbs/BNC	[Mihalcea & Strapparava, 2005]
Pun of the Day	2400 puns/ 2400 headlines	[Yang et al., 2015]
#HashTagWars	12K tweets for 112 hashtags, graded scores	[Potash et al., 2017]
English Puns	4K (71% puns) + WN annotations	[Miller et al., 2019]
Unfun.me	2.8K headline pairs (1.2K seeds), funny → serious edits	[West & Horvitz, 2019]
Humicroedit	15K headlines, serious → funny edits	[Hossain et al., 2019]
Stierlitz (Ru)	60K jokes / 60K headlines + 200 puns	[Ermilov et al., 2018]

Dataset for Humor Recognition [Blinov et al., 2019]

- Large (100K+)
- Russian
- Humor, not lexical properties

Starting Point

- 63K Russian jokes from social media [Bolotova et al., 2017]
- +63K non-jokes [Ermilov et al., 2018]

→ STIERLITZ dataset

+

- Humor-related public pages from VK.com: 556K
- anekdot.ru – the largest collection of Russian jokes: 477K
- E1.ru online forum: 10M

Lexical similarity

- Unfunny counterpart for each joke: top-ranked (BM25) non-identical forum post

FUN: Russian Mars rover will hand out Russian passports to the Martians.

FORUM: They say in the Crimea, too, they are handing out Russian passports or have already handed them out.

- KL divergence
  - STIERLITZ: 0.50
  - FUN: 0.18

Gold FUN

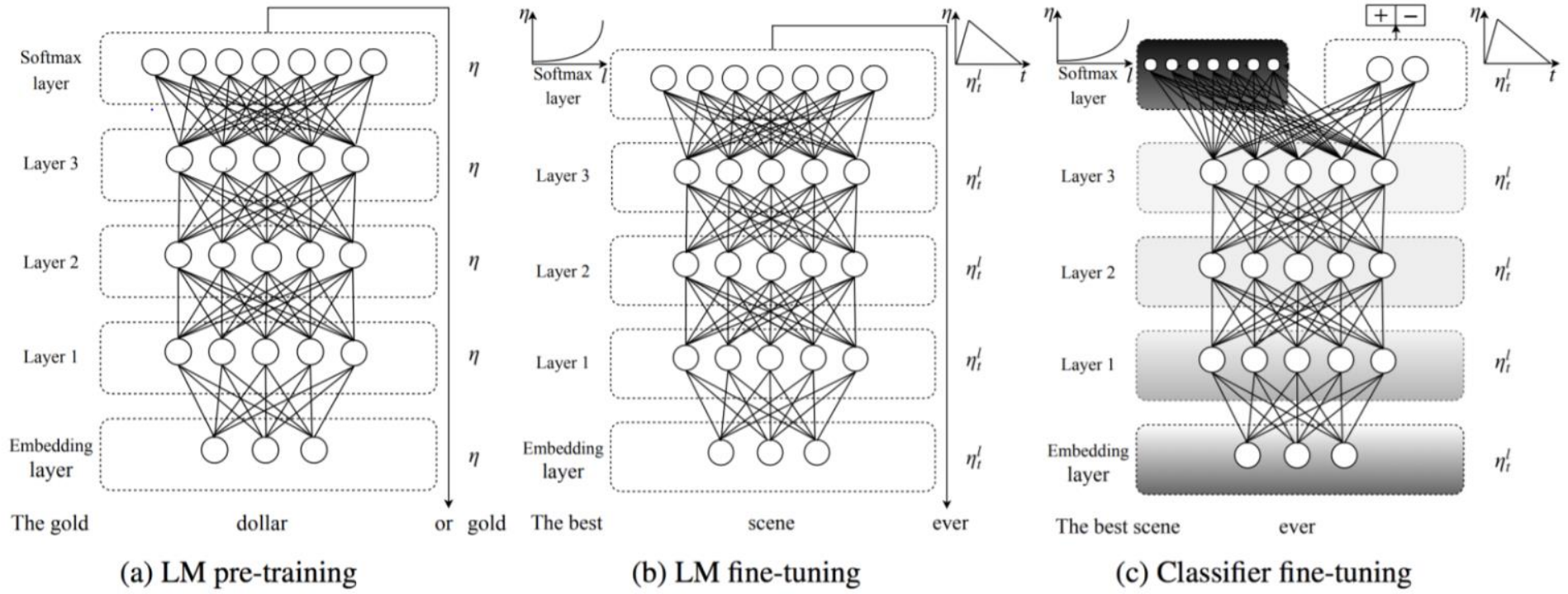
- ~1,877 texts annotated by volunteers (joke/non-funny joke/not a joke), x3
- 238 (12.7%) – joke/non-joke disagreement
- Majority voting – 94% correct

FUN Dataset

Dataset	*Jokes*	*Non-jokes*	*Total*
STIERLITZ	46,608	46,608	**93,216**
train	37,447	37,447	65,530
validation	4,682	4,682	9,364
test	9,361	9,361	18,722
PUNS	213	0	**213**
FUN	156,605	156,605	**313,210**
train	125,708	125,708	251,416
test	30,897	30,897	61,794
GOLD	899	978	**1,877**

http://bit.ly/fun_data

ULMFiT [Howard and Ruder, 2018]



Humor Detection Results

Model	STIERLITZ *Test F1*	FUN *Test F1*	GOLD *F1*	PUNS *Recall*
Baseline SVM	0.787	0.798	0.803	0.436
ULMFun	0.921	0.907	0.890	0.892

Error Analysis

- lacking world knowledge
- ‘hard-to-get’ jokes, dry humor
- non-jokes in ‘unfunny’ class and vice versa
- unusual word combinations/expressive content/figurative language

Summary

- More data – better detection results
- Datasets can be compiled without excessive manual annotation
- Even simple methods work for humor retrieval
- Evaluation is crucial, but crowdsourcing works well in general

Have FUN!

- Analysis of the dataset
- Evaluation of humor recognizers in the wild
- Humor generation

Questions?

pbras@yandex.ru

References

- Rada Mihalcea , Carlo Strapparava. Making computers laugh: Investigations in automatic humor recognition. HLT-EMNLP2005.
- Oliviero Stock, Carlo Strapparava. HAHAcronym: A Computational Humor System. ACL demo 2005.
- Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, Jukka M. Toivanen. “Let Everything Turn Well in Your Wife”: Generation of Adult Humor Using Lexical Constraints. ACL2013.
- Dafna Shahaf, Eric Horvitz, and Robert Mankoff. Inside jokes: Identifying humorous cartoon captions. PKDD2015.
- Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor recognition and humor anchor extraction. EMNLP2015.
- Peter Potash, Alexey Romanov, Anna Rumshisky. SemEval-2017 Task 6: #HashtagWars: Learning a Sense of Humor. SEMEVAL2017.
- Tristan Miller, Christian F. Hempelmann, and Iryna Gurevych. SemEval-2017 Task 7: Detection and Interpretation of English Puns. SEMEVAL2017.
- Vladislav Blinov, Kirill Mishchenko, Valeria Bolotova, Pavel Braslavski. A Pinch of Humor for Short-Text Conversation: an Information Retrieval Approach. CLEF2017.
- Anton Ermilov, Natasha Murashkina, Valeria Goryacheva, and Pavel Braslavski. Stierlitz Meets SVM: Humor Detection in Russian. AINL2018.
- Pavel Braslavski, Vladislav Blinov, Valeria Bolotova, Katya Pertsova. How to Evaluate Humorous Response Generation, Seriously? CHIIR2018.
- Vladislav Blinov, Valeriia Bolotova-Baranova, Pavel Braslavski. Large Dataset and Language Model Fun-Tuning for Humor Recognition. ACL2019.
- He He, Nanyun Peng, Percy Liang. Pun Generation with Surprise. NAACL2019.
- Orion Weller, Kevin Seppi. Humor Detection: A Transformer Gets the Last Laugh. EMNLP2019.
- Nabil Hossain, John Krumm, Michael Gamon. “President Vows to Cut <Taxes> Hair”: Dataset and Analysis of Creative Text Editing for Humorous Headlines. NAACL2019.
- Robert West, Eric Horvitz. Reverse-Engineering Satire, or “Paper on Computational Humor Accepted despite Making Serious Advances”. AAAI2019.