NLP metrics & evaluation

@altsoph

plan

- **metrics and losses**
- **task specific metrics**
  - LM
  - NMT
  - style transfer
  - dialog systems
  - NLU
- **manual evaluation**

metrics

- numerical value
- function of the system
- correlates with some aspect
- dataset specific
- task specific
- interpretation may differ

metric VS loss

- evaluate
- optimize
- compare (intra/inter)

metric VS loss

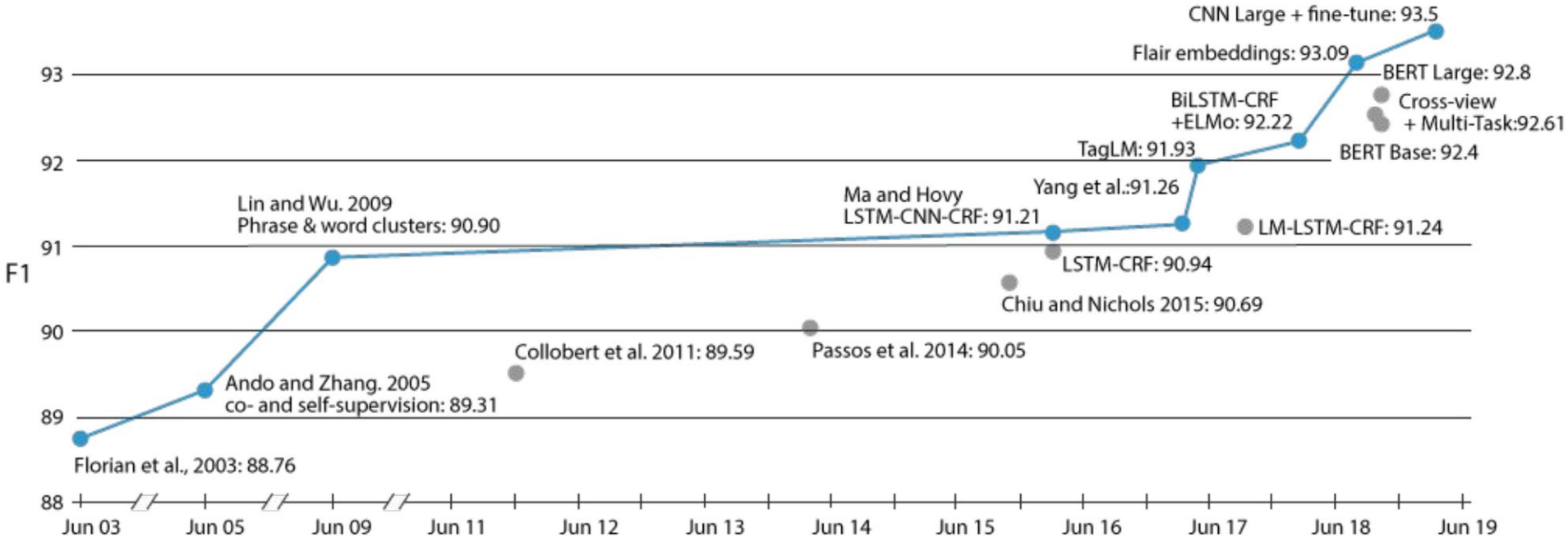
- cost
- scale
- meaning

- optimize what you can
- check what you want

metrics are task specific

- classic NLP
  - POS
  - NER
  - segmentation
  - classification
  - sentiment analysis
  - parsing
- NLG
  - LM
  - NMT
  - style transfer
  - dialog systems
  - NLU

progress in time



Performance on Named Entity Recognition (NER) on CoNLL-2003 (English) over time.

metrics are task specific

- classic NLP
  - POS
  - NER
  - segmentation
  - classification
  - sentiment analysis
  - parsing

BORING

- NLG
  - LM
  - NMT
  - style transfer
  - dialog systems
  - NLU

autoregressive language modeling

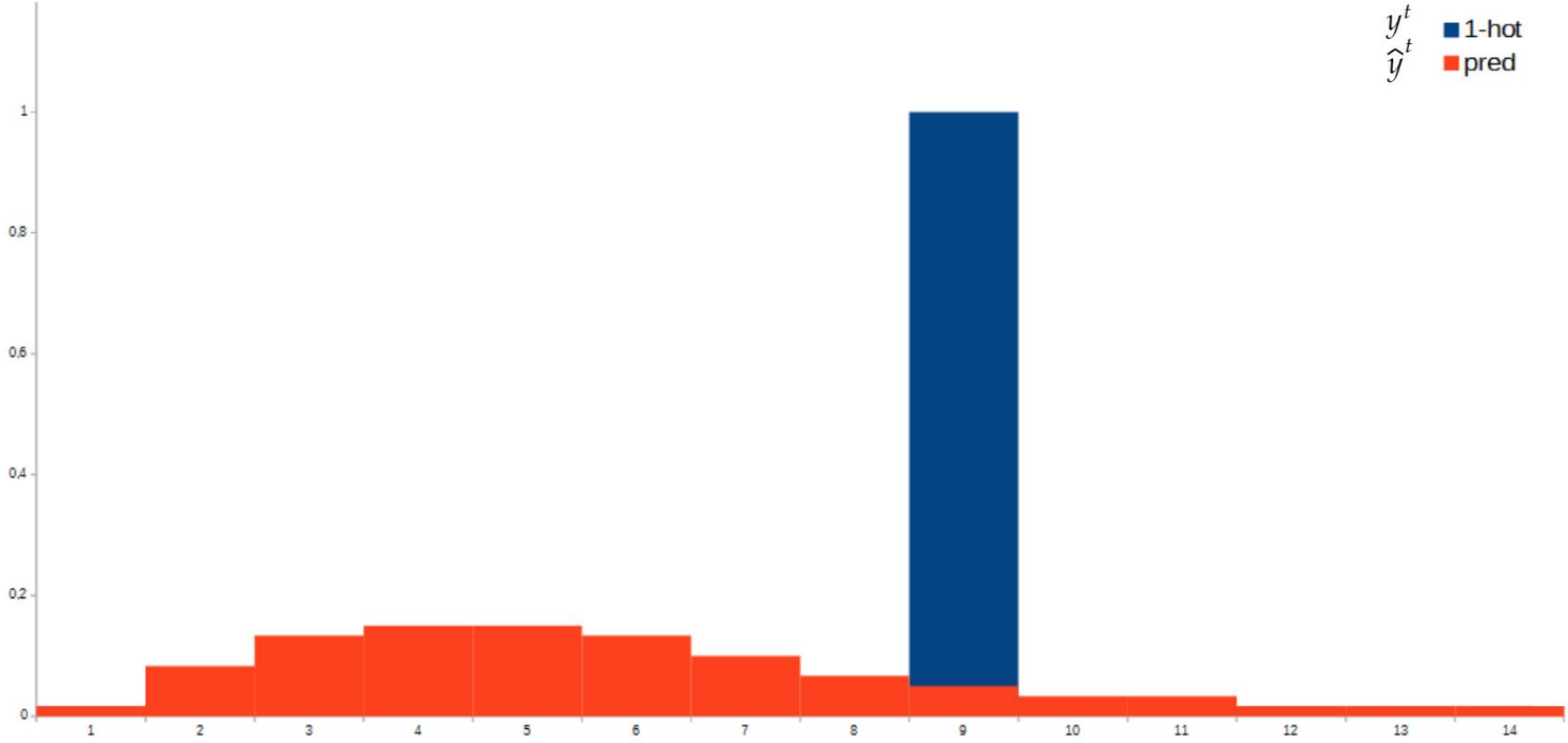
$$P(\text{I}, \text{am}, \text{good}) = P(\text{good} \mid \text{I}, \text{am})\, P(\text{am} \mid \text{I})\, P(\text{I})$$

autoregressive language modeling

Problem:

- **loss = *f*(token)**
- **quality = *f*(text)**

cross-entropy



cross-entropy & perplexity

$$L^t(\theta) = CE\left(y^t, \hat{y}^t\right) = -\sum_{i=1}^{|V|} y_i^t * log\, \hat{y}_i^t = -log\, \hat{y}_i^t = log\, PP\left(y^t, \hat{y}^t\right)$$

WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, b=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

Sampling, t=0.9

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Top-k, k=640

Pumping Station #3 shut down due to construction damage Find more at:
www.abc.net.au/environment/species-worry/
in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-k, k=40, t=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

Nucleus, p=0.95

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

WebText

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

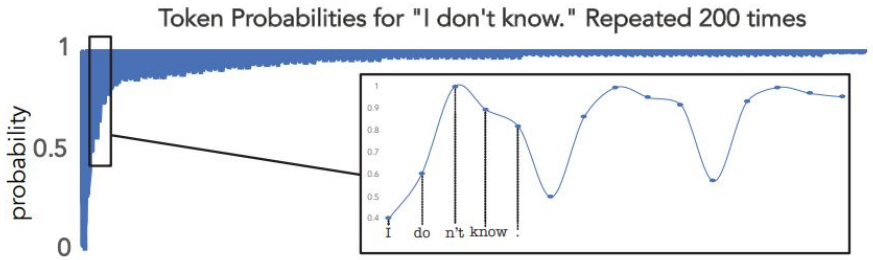
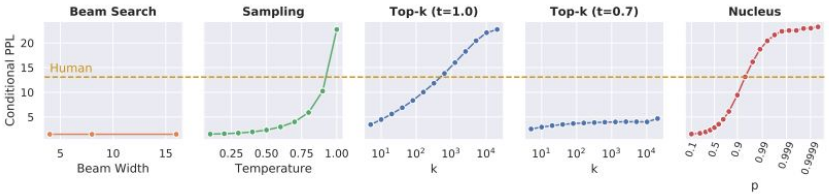
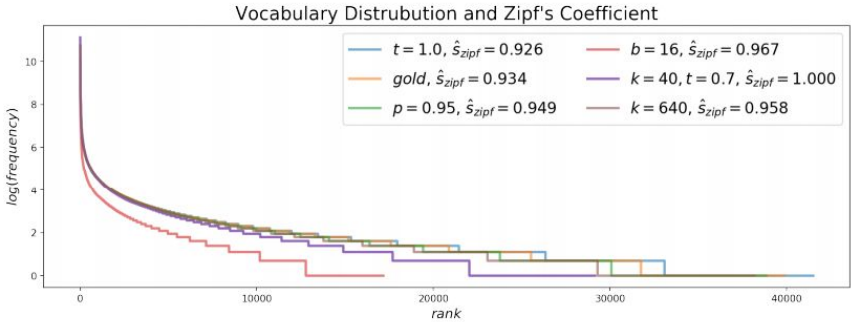


Figure 4: The probability of a repeated phrase increases with each repetition, creating a positive feedback loop. We found this effect to hold for the vast majority of phrases we tested, regardless of phrase length or if the phrases were sampled randomly rather than taken from human text.





[The Curious Case of Neural Text Degeneration, ICLR 2020]

autoregressive language modeling

- **optimize what you can**
  - CE is ***f(token)***

- **check what you want**
  - repetitions
  - fluency
  - … = ***f(text)***

NMT

NMT

- source text
- candidate text
- reference

NMT

- source text
- candidate text
- reference 1
  reference 2
  …

NMT: BLEU

- source text
- candidate text
- reference 1
  reference 2
  …

=> BLEU, 2002

NMT: BLEU

- **BLEU is n-gram precision**

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

NMT: BLEU

- BLEU is n-gram precision
- **with sentence brevity penalty**

Candidate: the .

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Unigram Precision = 1

NMT: BLEU

- BLEU is **modified** n-gram precision
- with sentence brevity penalty

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Unigram Precision $= 2/7$.[3]

NMT: BLEU

- BLEU**n** is modified n-gram precision
- with sentence brevity penalty
- **BLEUn is geometric mean of BLEU1, BLEU2, BLEU3, BLEU4**

NMT: BLEU

- BLEUn is modified n-gram precision
- with sentence brevity penalty
- BLEUn is geometric mean of BLEU1, BLEU2, BLEU3, BLEU4
- **and with weights**

NMT: BLEU

- BLEUn is modified n-gram precision
- with sentence brevity penalty
- BLEUn is geometric mean of BLEU1, BLEU2, BLEU3, BLEU4
- and with weights
- **and with smoothing**

what's wrong with BLEU

- not about meaning or structure or fluency
- not good for morphologically rich languages
- low correlation with human judgements
- needs reference[s]

- reference preprocessing matters
- many parameters

=> low reproducibility

BLEU family

BLEU family

- **sacreBLEU** - standardized BLEU implementation

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score
- **NIST** - BLEU reweighted by n-gram rareness

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score
- **NIST** - BLEU reweighted by n-gram rareness
- **ROUGE** - recall instead of precision

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score
- **NIST** - BLEU reweighted by n-gram rareness
- **ROUGE** - recall instead of precision
- **METEOR** - WN-synonyms, recall, precision, order penalty

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score
- **NIST** - BLEU reweighted by n-gram rareness
- **ROUGE** - recall instead of precision
- **METEOR** - WN-synonyms, recall, precision, order penalty
- **GLEU** - google BLEU, 1-4 grams, min(prec,rec)

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score
- **NIST** - BLEU reweighted by n-gram rareness
- **ROUGE** - recall instead of precision
- **METEOR** - WN-synonyms, recall, precision, order penalty
- **GLEU** - google BLEU, 1-4 grams, min(prec,rec)
- **RIBES** - uses rang correlation btw n-gram matches

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score
- **NIST** - BLEU reweighted by n-gram rareness
- **ROUGE** - recall instead of precision
- **METEOR** - WN-synonyms, recall, precision, order penalty
- **GLEU** - google BLEU, 1-4 grams, min(prec,rec)
- **RIBES** - uses rang correlation btw n-gram matches
- **[w]mpF** - F-score on word, morpheme and POS ngrams

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score
- **NIST** - BLEU reweighted by n-gram rareness
- **ROUGE** - recall instead of precision
- **METEOR** - WN-synonyms, recall, precision, order penalty
- **GLEU** - google BLEU, 1-4 grams, min(prec,rec)
- **RIBES** - uses rang correlation btw n-gram matches
- **[w]mpF** - F-score on word, morpheme and POS ngrams
- **[h]LEPOR** - recall and precision, lemmas, positions, ...

LEPOR, for example

LEPOR: automatic machine translation evaluation metric considering the enhanced Length Penalty, n-gram Position difference Penalty and Recall

—

In our evaluation, we used hLEPOR$_A$ v.3.1:

- $hLEPOR = Harmonic(w_{LP}LP, w_{NPosPenal}NPosPenal, w_{HPR}HPR)$
$= \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{Factor_i}} = \frac{w_{LP}+w_{NPosPenal}+w_{HPR}}{\frac{w_{LP}}{LP}+\frac{w_{NPosPenal}}{NPosPenal}+\frac{w_{HPR}}{HPR}}$

- $\overline{hLEPOR_A} = \frac{1}{SentNum}\sum_{i=1}^{SentNum} hLEPOR_{ithSent}$

—

(best metric from ACL-WMT 2013 contest)

LIKE BLEU, BUT BETTER

how to choose?

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
n	16	12	11	11	11	14	15
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	0.906	**0.993**	0.952	0.986	0.947	0.915	0.942
BERTr	**0.926**	0.984	0.938	0.990	0.948	**0.971**	0.974
BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899
CDER	0.890	**0.988**	0.876	0.967	**0.975**	0.892	0.917
CHARACTER	0.898	**0.990**	0.922	0.953	0.955	0.923	0.943
CHRF	**0.917**	**0.992**	0.955	0.978	0.940	0.945	0.956
CHRF+	**0.916**	**0.992**	0.947	0.976	0.940	0.945	0.956
EED	0.903	**0.994**	0.976	0.980	0.929	0.950	0.949
ESIM	**0.941**	0.971	0.885	0.986	**0.989**	**0.968**	**0.988**
HLEPORA_BASELINE	–	–	–	0.975	–	–	0.947
HLEPORB_BASELINE	–	–	–	0.975	0.906	–	0.947
METEOR++_2.0(SYNTAX)	0.887	**0.995**	0.909	0.974	0.928	**0.950**	0.948
METEOR++_2.0(SYNTAX+COPY)	0.896	**0.995**	0.900	0.971	0.927	**0.952**	0.952
NIST	0.813	0.986	0.930	0.942	0.944	0.925	0.921
PER	0.883	**0.991**	0.910	0.737	0.947	0.922	0.952
PREP	0.575	0.614	0.773	0.776	0.494	0.782	0.592
SACREBLEU.BLEU	0.813	0.985	0.834	0.946	0.955	0.873	0.903
SACREBLEU.CHRF	0.910	**0.990**	0.952	0.969	0.935	0.919	0.955
TER	0.874	**0.984**	0.890	0.799	0.960	0.917	0.840
WER	0.863	0.983	0.861	0.793	0.961	0.911	0.820
WMDO	0.872	**0.987**	0.983	**0.998**	0.900	0.942	0.943
YISI-0	0.902	**0.993**	**0.993**	0.991	0.927	**0.958**	0.937
YISI-1	**0.949**	**0.989**	0.924	0.994	0.981	**0.979**	**0.979**
YISI-1_SRL	**0.950**	**0.989**	0.918	0.994	**0.983**	**0.978**	0.977
QE as a Metric:							
IBM1-MORPHEME	0.345	0.740	–	–	0.487	–	–
IBM1-POS4GRAM	0.339	–	–	–	–	–	–
LASIM	0.247	–	–	–	–	0.310	–
LP	0.474	–	–	–	–	0.488	–
UNI	0.846	0.930	–	–	–	0.805	–
UNI+	0.850	0.924	–	–	–	0.808	–
YISI-2	0.796	0.642	0.566	0.324	0.442	0.339	0.940
YISI-2_SRL	0.804	–	–	–	–	–	0.947
	newstest2019						

Table 3: Absolute Pearson correlation of to-English system-level metrics with DA human assessment in newstest2019; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

[Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges]

more NMT metrics

- **distance based metrics:**
  - **CER / WER / TER / TER-plus / CDER**
  - **STM** - subtree metric
  - **...**

more NMT metrics

- **distance based metrics**
- **embedding distance metrics**
  - avg/max w2v/glove/fasttext
  - ELMo
  - BERT-score
  - ...

more NMT metrics

- distance based metrics
- embedding distance metrics
- learnable metrics:
  - ROSE / VERTa / MEWR / SIMILE
  - BLEURT
  - ...

NMT

- **optimize what you can**

- **check what you want**
  - “similarity” to references
  - meaning preservation
  - target language fluency
  - *...*

text style transfer

text style transfer

- text =?= style + meaning

text style transfer

- **text =?= style + meaning**
- **style domain examples**
  - formality
  - politeness
  - *sentiment*
  - author's style

text style transfer

- **text =?= style + meaning**
- **style domain examples**
  - formality
  - politeness
  - sentiment
  - author's style
- **style definition**
  - discrete / continuous
  - explicit corpora / rule based / classifier based

style transfer =?= NMT

- **no good parallel corpora :(**

- **check what you want**
  - target style matching
  - language fluency
  - *content / meaning preservation*

style transfer criteria

- **style matching**
  - cross-entropy / perplexity
  - classification problem
- **language fluency**
  - cross-entropy / perplexity
  - classification problem
- **content / meaning preservation**
  - self-NMT metrics
  - embedding based metrics
  - back translation tricks

style matching

- **cross-entropy under stylized LM**

Model $G(A_i)$/ author	Shakespeare	Poe	Carroll	Wilde	Marley	Nirvana	MUSE
Generated-Shakespeare	**19.0****	21.6	**18.5***	19.9	21.8	22.0	22.4
Generated-Poe	22.0	**20.4****	21.2	**19.0***	26.0	25.4	26.0
Generated-Carroll	22.2	23.6	**18.9***	22.5	22.4	**21.8****	23.8
Generated-Wilde	21.2	20.9	**20.5****	**18.4***	24.5	24.8	26.4
Generated-Marley	24.1	26.5	22.0	27.0	**15.5***	**15.7****	16.0
Generated-Nirvana	23.7	26.2	20.0	26.6	19.3	**18.3***	**19.1****
Generated-MUSE	21.1	23.9	18.5	23.4	17.4	**16.0****	**14.6***
Uniform Random	103.1	103.0	103.0	103.0	103.5	103.3	103.6
Weighted Random	68.6	68.8	67.4	68.5	68.5	68.0	68.0
SELF	23.4	21.8	25.1	27.3	20.8	17.8	13.3

Table 3. Sample cross entropy between generated texts $\{T_i^G|A_i\}$ and actual texts for different authors.

- **classification**

truth \ pred	Brodskiy	Pushkin	Esenin	Pasternak	Tsvetaeva	Mayakovskiy	Akhmatova	Tyutchev	Mandelshtam	Lermontov
Brodskiy	77,2%	1,7%	2,3%	4,3%	2,3%	1,5%	4,0%	1,3%	3,6%	1,7%
Pushkin	1,1%	77,0%	8,0%	0,3%	0,0%	0,3%	1,9%	3,3%	0,6%	7,5%
Esenin	3,9%	4,9%	73,8%	3,0%	1,3%	1,6%	5,9%	0,7%	1,6%	3,3%
Pasternak	16,3%	2,6%	10,7%	54,9%	2,1%	1,7%	3,9%	1,3%	6,0%	0,4%
Tsvetaeva	9,1%	2,8%	5,1%	4,0%	51,1%	1,7%	18,2%	1,1%	5,7%	1,1%
Mayakovskiy	8,2%	2,9%	11,7%	5,8%	3,5%	59,1%	0,6%	1,2%	7,0%	0,0%
Akhmatova	4,5%	4,5%	17,0%	3,4%	3,4%	0,0%	59,7%	1,1%	1,7%	4,5%
Tyutchev	3,0%	14,1%	3,7%	3,0%	0,7%	0,7%	5,9%	55,6%	2,2%	11,1%
Mandelshtam	9,2%	6,6%	9,2%	11,8%	1,3%	5,3%	15,8%	1,3%	35,5%	3,9%
Lermontov	2,6%	15,8%	9,2%	0,0%	2,6%	0,0%	9,2%	9,2%	2,6%	48,7%

fluency

- heuristics
- cross-entropy with general corpus
- hard classification problem

content preservation

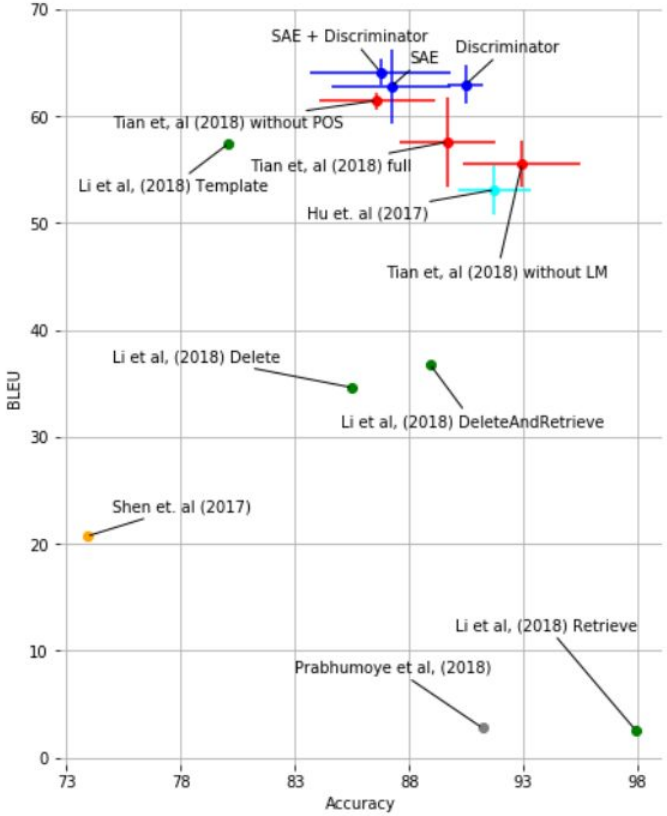
- **self-NMT metrics**
  - ...
- **embedding based metrics**
  - avg/max w2v/glove/fasttext
  - ELMo
  - BERT-score
  - ...
- **tricks**
  - double translation
  - ...

content preservation is hard

	POS-distance	Word overlap	chrF	Word2Vec	FastText	WMD	ELMO L2	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Meteor	BERT score	Human score
POS-distance	1,00	0,73	0,71	0,45	0,44	0,69	0,66	0,71	0,72	0,71	0,68	0,74	0,82	0,72
Word overlap	0,73	1,00	0,98	0,80	0,84	0,86	0,92	0,99	0,91	0,98	0,92	0,99	0,95	0,80
chrF	0,71	0,98	1,00	0,79	0,83	0,89	0,93	0,97	0,89	0,99	0,92	0,99	0,93	0,83
Word2Vec	0,45	0,80	0,79	1,00	0,98	0,87	0,88	0,78	0,79	0,78	0,82	0,77	0,73	0,64
FastText	0,44	0,84	0,83	0,98	1,00	0,86	0,90	0,83	0,81	0,83	0,85	0,81	0,76	0,65
WMD	0,69	0,86	0,89	0,87	0,86	1,00	0,96	0,86	0,92	0,89	0,92	0,86	0,85	0,89
ELMO L2	0,66	0,92	0,93	0,88	0,90	0,96	1,00	0,92	0,92	0,94	0,96	0,92	0,87	0,86
ROUGE-1	0,71	0,99	0,97	0,78	0,83	0,86	0,92	1,00	0,93	0,98	0,93	0,98	0,94	0,82
ROUGE-2	0,72	0,91	0,89	0,79	0,81	0,92	0,92	0,93	1,00	0,91	0,96	0,90	0,87	0,81
ROUGE-L	0,71	0,98	0,99	0,78	0,83	0,89	0,94	0,98	0,91	1,00	0,94	0,99	0,94	0,83
BLEU	0,68	0,92	0,92	0,82	0,85	0,92	0,96	0,93	0,96	0,94	1,00	0,92	0,87	0,84
Meteor	0,74	0,99	0,99	0,77	0,81	0,86	0,92	0,98	0,90	0,99	0,92	1,00	0,95	0,80
BERT score	0,82	0,95	0,93	0,73	0,76	0,85	0,87	0,94	0,87	0,94	0,87	0,95	1,00	0,82
Human score	0,72	0,80	0,83	0,64	0,65	0,89	0,86	0,82	0,81	0,83	0,84	0,80	0,82	1,00

Figure 1: Pairwise correlations of the orders induced by the metrics of semantic similarity.

content vs style trade-off



text style transfer

- **optimize trade-off of losses**
  - style acc
  - reconstruction loss
  - double translation acc
  - aux losses
- **check what you want**
  - target style matching
  - language fluency
  - content / meaning preservation

dialog systems

dialog systems

- **check what you want**
  - language fluency
  - ???

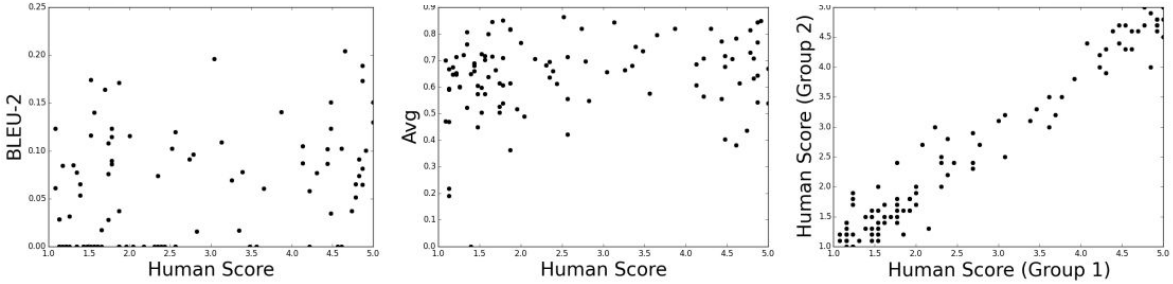
- **hard even for manual evaluation**
  satisfaction / comprehension / naturalness / specificity / sensibleness / engagingness / interestingness / ...

dialog system metrics

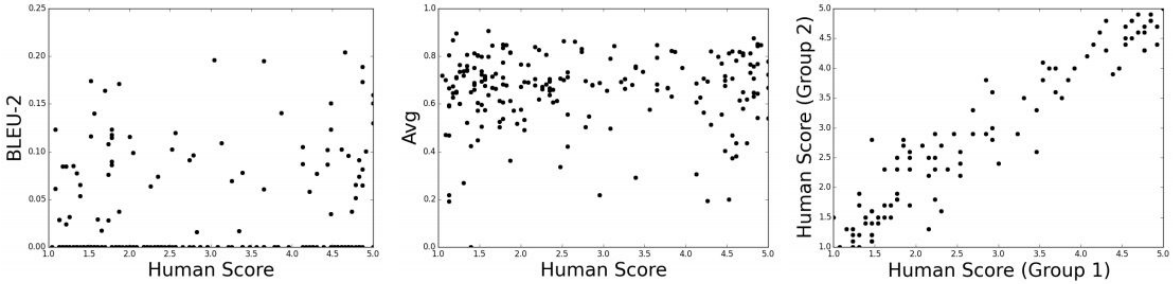
- **Q-A level**
  - NMT approach
  - style transfer approach

- **dialog level**
  - **goal-oriented**
    - success rate / time to success / funnel metrics / …
  - **general purpose chat**
    - inter-utterance coherence / memory / ...

dialog system metrics



(a) Twitter

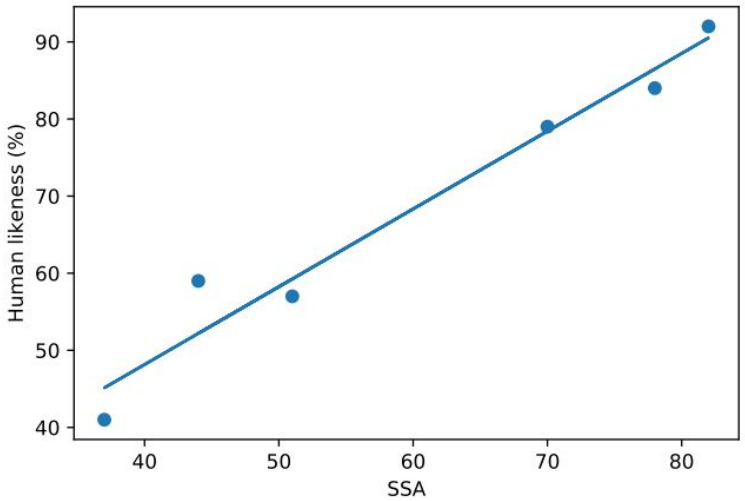


(b) Ubuntu

Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

[arXiv:1603.08023]

dialog system metrics



[arXiv:2001.09977]

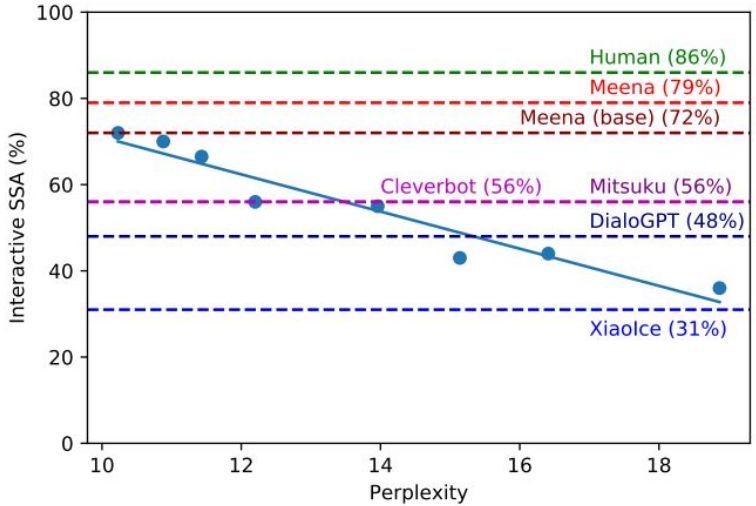


Figure 1: Interactive SSA vs Perplexity. Each point is a different version of the Meena model.

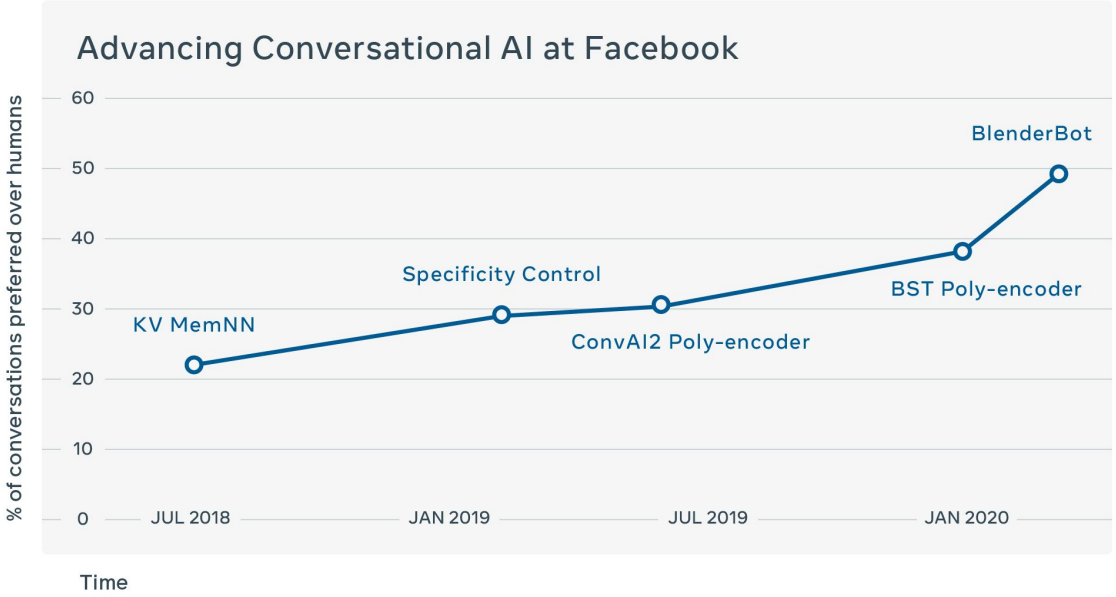
SSA = sensibleness and specificity average
!sensible -> !specific
manual labeling, closed simple questions
sensibleness agreement 76±3%, specificity 66±2%

dialog system metrics

Blended Skill Talk (BST) for training and evaluating these desirable skills. BST consists of the following skills, leveraging our previous research:

- Engaging use of personality (PersonaChat)
- Engaging use of knowledge (Wizard of Wikipedia)
- Display of empathy (Empathetic Dialogues)
- Ability to blend all three seamlessly (BST)

[arXiv:2004.08449]



NLU

current benchmarks

GLUE:

- **The Corpus of Linguistic Acceptability**
- **The Stanford Sentiment Treebank**
- **Microsoft Research Paraphrase Corpus**
- **Semantic Textual Similarity Benchmark**
- **Quora Question Pairs**
- **MultiNLI**
- **Question NLI**
- **Recognizing Textual Entailment**
- **Winograd NLI**

current benchmarks

SuperGLUE:

- CommitmentBank
- Choice of Plausible Alternatives
- Multi-Sentence Reading Comprehension
- Recognizing Textual Entailment
- Words in Context
- The Winograd Schema Challenge
- BoolQ
- Reading Comprehension with Commonsense Reasoning

current challenges

SemEval:
(some of 2020)

- **Task 5: Modelling Causal Reasoning in Language: Detecting Counterfactuals**
- **Task 6: DeftEval: Extracting Definitions from Free Text in Textbooks**
- **Task 7: Assessing Humor in Edited News Headlines**
- **Task 11: Detection of Propaganda Techniques in News Articles**
- **Task 12: OffensEval 2: Multilingual Offensive Language Identification in Social Media**

a bit more about manual evaluation

- **pairwise**
  - easy to set up
  - O(n^2)
  - hard to interpret
- **score based**
  - use instructions
  - O(n)
  - no hard tasks

[arXiv:1909.03087]

Question	Choice 1	Agrm.
Engagingness (PersonaChat)		
Which speaker is more engaging to talk to?	Speaker 1 is more engaging	82.5%
Who would you prefer to talk to for a long conversation?	I would prefer to talk to Speaker 1	***87.5%**
Which speaker do you think is more captivating?	Speaker 1 is more captivating than Speaker 2	84.2%
Interestingness (PersonaChat)		
If you had to say one of these speakers is interesting and one is boring, who would you say is more interesting?	Speaker 1 is more interesting	***86.7%**
Which speaker is more interesting to talk to?	Speaker 1 is more interesting	*81.5%
Which speaker is more boring to talk to?	Speaker 1 is more boring	69.6%
Who would you rather talk to for fun?	Speaker 1 is more fun	70.8%
Humanness (PersonaChat)		
Which speaker sounds more human?	Speaker 1 sounds more human	***76.9%**
If you had to guess that one speaker is human and one is a bot, which do you think is human?	Speaker 1 sounds human	71.4%
Which speaker sounds more like a real person?	Speaker 1 sounds more like a real person	76.9%
Knowledgeable (Wizard of Wikipedia)		
Which speaker is more knowledgeable?	Speaker 1 is more knowledgeable	*88.9%
If you had to say that one speaker is more knowledgeable and one is more ignorant, who is more knowledgeable?	Speaker 1 is more knowledgeable	***100%**
Which speaker is more well-informed?	Speaker 1 is more well-informed	*85.0%

Table 1: **Optimizing questions**: we measure the agreement rates for the most chosen response for different phrasings of questions, and choose the most agreed upon versions. Starred agreements indicate statistical significance (binomial test, $p < .05$), and bold agreements indicate the question was used in future trials.

recap

- **cross-entropy / perplexity**
- **syntactic similarity**
  - BLEU family and more
- **semantic similarity via pretrainded embeddings**
  - sent2vec, USE, bert_score, ...
- **“style” discriminators**
- **complex goals / human labeling**
- **NLU challenges**
- **learnable metrics on human labels**
  - BLEURT and others
- **blackbox testing**
  - CheckList, biases, etc

thanks for attention!

altsoph@gmail.com