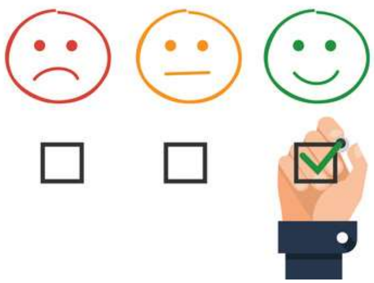
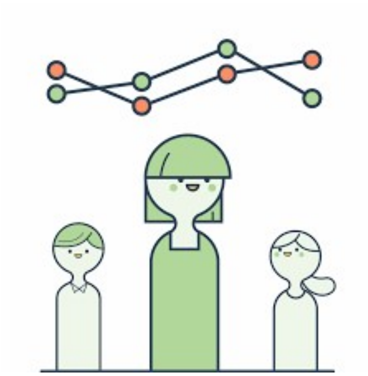
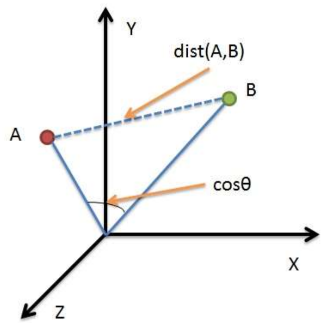
Evaluation and Benchmarks



Types of Evaluation Methods for Text Generation



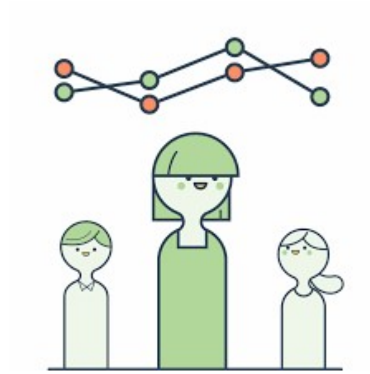
Human Evaluations



Un-trained Metrics

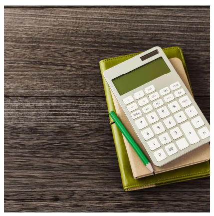
Trained Metrics

Human Evalutions



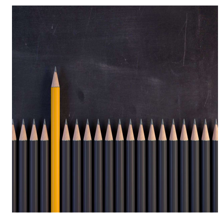
- Most important form of evalation for NLG systems
- Automatic metrics fall short of replicating human decisions
- Gold standard in developing new automatic metrics

Human Evalutions: Issues

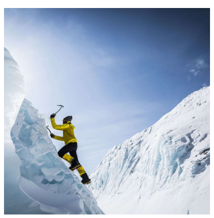


Expensive

Time Consuming



Quality Control



Challenging Criteria

Inconsistency in Evaluations

Inconsistency in reporting

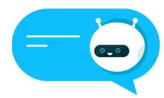
Intrinsic Human Evaluations

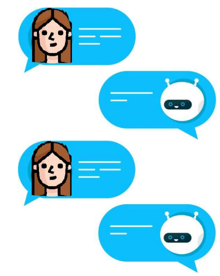
- Ask *humans* to evalute the quality of generated text
- Overall or along some specific dimension:
  - fluency
  - coherence
  - factuality and correctness
  - adequacy
  - commonsense
  - style / formality
  - grammaticality
  - typicality
  - redundancy

For details Celikyilmaz, Clark, Gao, 2020

Extrinsic Human Evalutions

- **Humans** evaluate a system's performance on the task for **which it was designed**
- For instance, **dialog systems** are typically evaluated extrinsically!





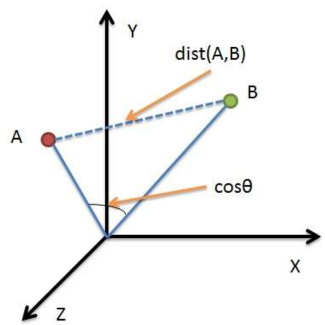
Turn Level	Dialog Level
▪ Interesting ▪ Engaging ▪ Generic/Specific ▪ Relevant ▪ Semantically appropriate ▪ Understandable ▪ Fluently Written ▪ Correct vs. Misunderstanding ▪ Overall Impression	• Coherent • Recovers from errors • Consistent • Diversity in its responses • Topic Depth • Likable (empathy, personality) • Understanding • Flexible and adaptable • Informative • Inquisitive • Overall Impression

Mehri and Eskenazi, SIGDIAL 2020

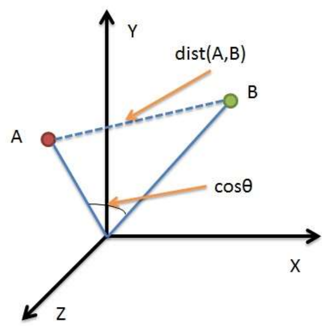
Human Evaluations: Other Aspects

- Evaluators
- Inter-Annotator Aggreement
  - Percent agreement, Cohen's $\varkappa$, Fleiss's $\varkappa$, Krippendorff's α
- Evaluation experiment design
  - Side-by-side or singleton?
  - The amount context (e.g., dialog or summarization)
  - How many models to compare at a given time?

Untrained Automatic Evaluation Metrics

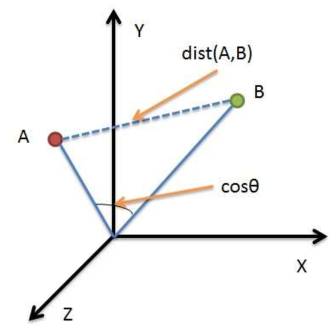


Untrained Automatic Evaluation Metrics



- Measure the effectiveness of the models that generate text
- Compute a score that indicates the similarity between *generated* and *gold-standard* (*human-written*) *text*
- Fast and efficient and widely used

Untrained Automatic Evaluation Metrics



1. *n*-gram overlap metrics
2. distance-based metrics
3. *n*-gram based diversity metrics
4. content overlap metrics

1. N-Gram Overlap Metrics

Metric	Property	MT	IC	SR	SUM	DG	QG	RG
BLEU	*n*-gram precision	✓	✓			✓	✓	✓
NIST	*n*-gram precision	✓						
F-SCORE	precision and recall	✓	✓	✓	✓	✓	✓	✓
WER	% of insert,delete,replace			✓				
ROUGE	*n*-gram recall				✓	✓		
METEOR	*n*-gram w/ synonym matching	✓	✓			✓		
HLEPOR	unigrams harmonic mean	✓						
RIBES	unigrams harmonic mean							
CIDER	*tf-idf* weighted *n*-gram similarity		✓					
EDIT DIST.	cosine similarity	✓	✓	✓	✓	✓	✓	
TER	translation edit rate	✓						
WMD	earth mover distance on words		✓		✓			
SMD	earth mover distance on sentences		✓	✓	✓			
PYRAMID				✓				
SPICE	scene graph similarity		✓					
SPIDER	scene graph similarity		✓					

MT: Machine Translation

DG: Document Generation

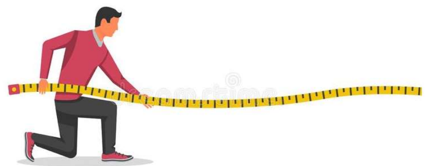
RG: Response Generation

IC: Image Captioning

SUM: Summarization

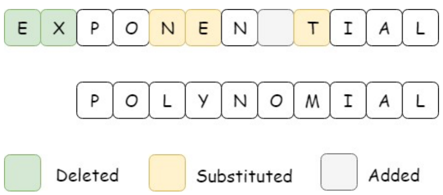
QG: Question Generation

2. Distance Based Metrics



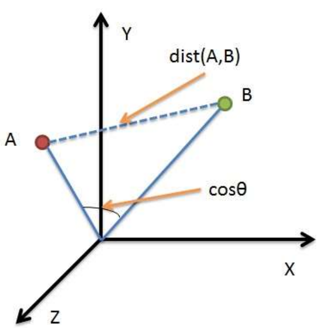
- Distance function to measure similarity between two text units
- Text units are represented as vectors → embeddings!
- Even though embeddings are pretrained, distance metrics used to measure the similarity are not!

2. Distance Based Metrics



Edit Distance:

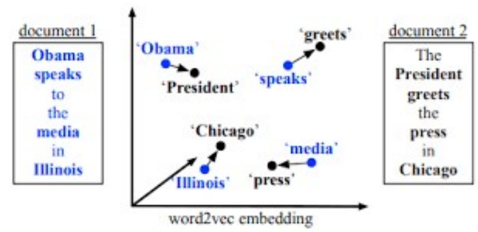
Measures how dissimilar two text units are based on the minimum number of operations required to transform one text into another.



Vector Similarity:

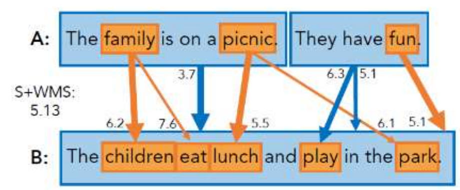
Embedding based similarity for semantic distance between text.

MEANT
YISI
Word Movers Distance
Sentence Movers similarity



Word Mover's Distance:

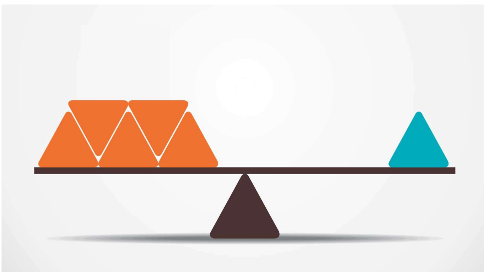
Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), represented with relative word frequencies. It combines item similarity on bag-of-word histogram representations of text with word embedding similarity.



Sentence Movers Similarity :

Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings (Clark, et.al. 2019)

3. *n*-gram Based Diversity Metrics

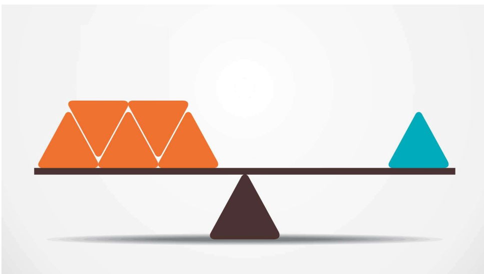


Type-to-Token Ratio (TTR):

- The ratio of types to tokens in a corpus:
  *"**The** cat sat on **the** mat new **the** log fire"*
  TTR = 8 /10

- Used to measure the lexical variety in a text:
  The higher the TTR, the more varied the text vocabulary

3. *n*-gram Based Diversity Metrics



Type-to-Token Ratio (TTR):

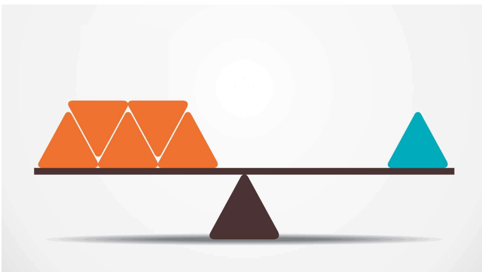
- The ratio of types to tokens in a corpus:
  *“**The** cat sat on **the** mat new **the** log fire”*
  TTR = 8 /10

- Used to measure the lexical variety in a text:
  The higher the TTR, the more varied the text vocabulary

`Self-BLEU`:

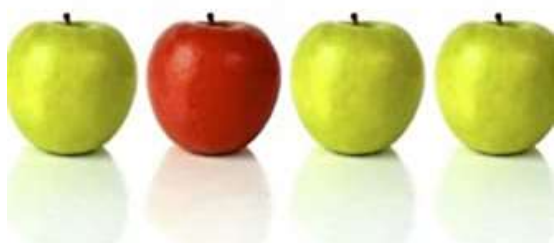
Measures the distance between generated sentence to reference or other generated sentences.
Calculates `BLEU` score for every generated sentence and defines the average of these `BLEU` scores as the `SELF-BLEU` score.
(Zhu et.al. 2018)

3. *n*-gram Based Diversity Metrics



Type-to-Token Ratio (TTR):

- The ratio of types to tokens in a corpus:
  *"**The** cat sat on **the** mat new **the** log fire"*
  TTR = 8 /10
- Used to measure the lexical variety in a text:
  The higher the TTR, the more varied the text vocabulary



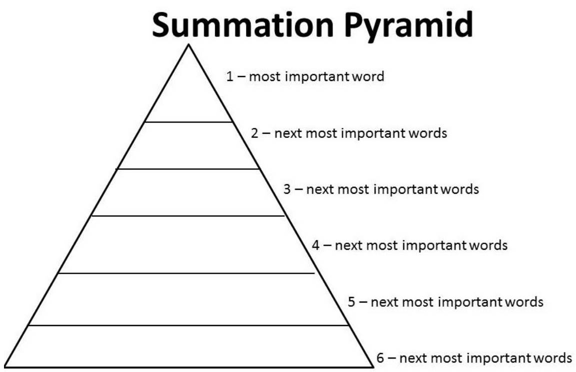
Self-BLEU:

Measures the distance between generated sentence to reference or other generated sentences. Calculates BLEU score for every generated sentence and defines the average of these BLEU scores as the SELF-BLEU score.
(Zhu et.al. 2018)

Textual Lexical Diversity:

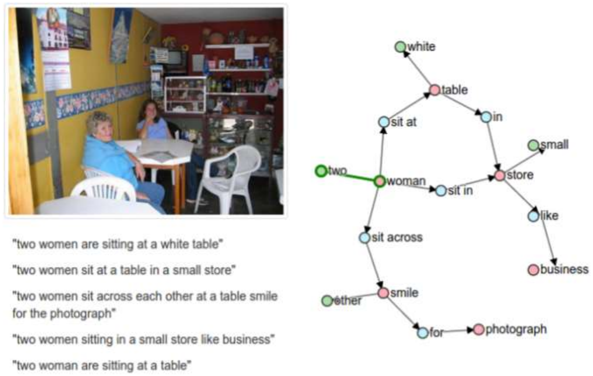
TTR can be sensitive to the length of the text. This metric (HD-D) assumes that if a text sample consists of many tokens of a specific word, then there is a high probability of drawing a text sample that contains at least one token of that word. Used to evaluate story generation and summarization tasks.
(McCarthy and Jarvis, 2010)

4- Content Overlap Metrics



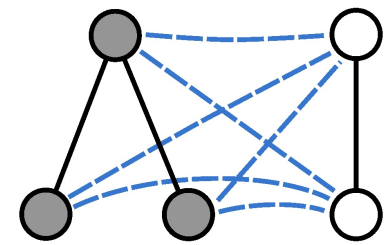
PYRAMID:

- Semi-automatic metric for evaluating document summarization models.
- Requires reference text as well as human annoations for **Summarization Content Units (SCU)**
- **SCUs** are phrases labeled by human judges as, that express the text spans with the same meaning.



SPICE:

Semantic propositional image caption evaluation is an image captioning metric that initially parses the reference text to derive an abstract scene graph representation. The generated caption is also parsed and the parsed graphs are compared against each other using F-score metric.
(Anderson et.al. 2016)



SPIDER:

A combination of semantic graph similarity (SPICE) and *n*-gram similarity measure (CIDER), the SPICE metric yields a more complete quality evaluation metric.
(Liu, et.al., 2017)

Machine Learnt Metrics

	Dialog Response Generation	Image Captioning
Context	**Speaker A**: Hey John, what do you want to do tonight? **Speaker B**: Why don't we go see a movie?	
Ground-Truth	**Response:** Nah, I hate that stuff, let's do something active.	**Caption:** a man wearing a red life jacket is sitting in a canoe on a lake
Model/Distorted Output	**Response:** Oh sure! Heard the film about Turing is out!	**Caption:** a guy wearing a life vest is in a small boat on a lake
BLEU	0.0	0.20
ROUGE	0.0	0.57
WMD	0.0	0.10

Machine Learnt Evaluation Metrics

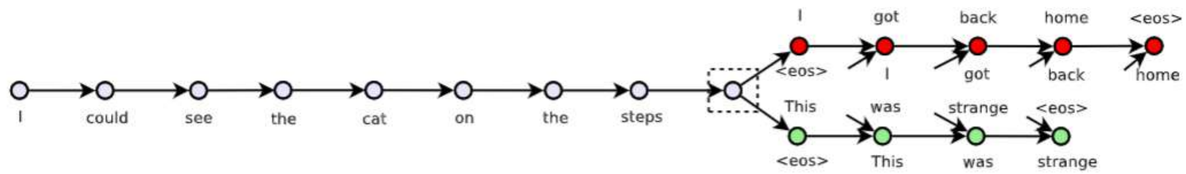
1. Sentence similarity metrics
2. Regression Based Metrics
3. Learning from Human Feedback
4. BERT-Based Evaluation
5. Composite Metrics
6. Factual Correctness metrics

Machine Learnt Evaluation Metrics

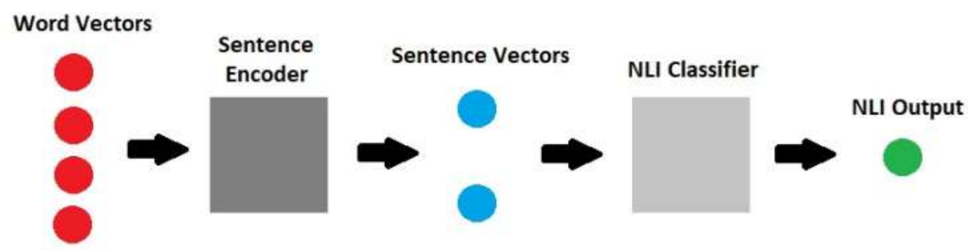
1. Sentence similarity metrics
2. Regression Based Metrics
3. Learning from Human Feedback
4. BERT-Based Evaluation
5. Composite Metrics
6. Factual Correctness metrics

Sentence Similarity Metrics

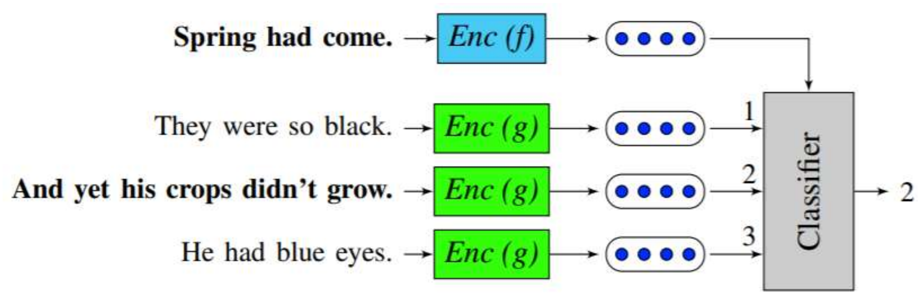
- **Skip Thoughts Vectors:** Unsupervised LSTM based model to encode rich contextual information by considering the surrounding context. (Kiros,et.al. 2015)



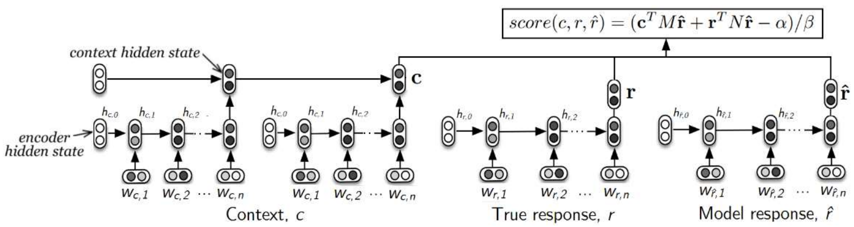
- **INFERSENT:** encode LSTM based Siamese networks to encode word-worder and is trained on high quality sentence inference dataset. (Conneau, et.al. 2017)



- **Quick Thoughts Vectors :** Unsupervised model of universal sentence embeddings trained on consecutive sentences. A classifier is trained to distinguish a context sentence from other contrastive sentences based on their embeddings. (Logeswaran and Lee, 2018)

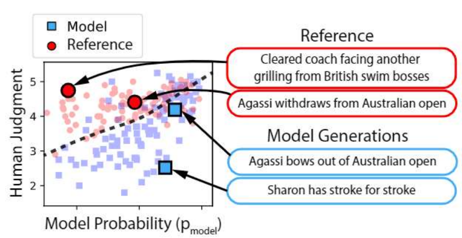


Learning from Human Feedback



ADEM:

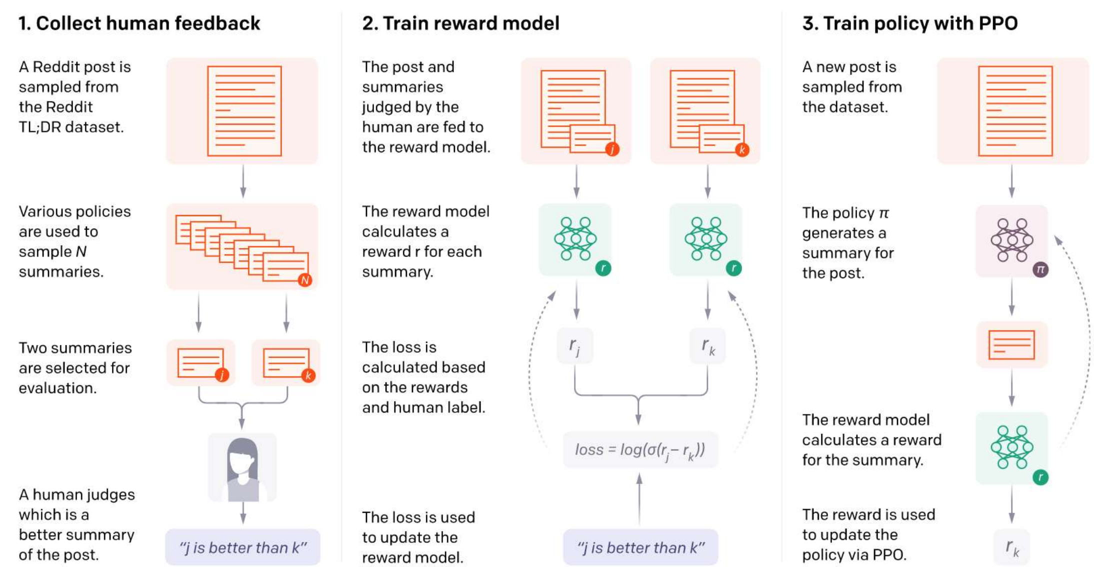
- A learned metric from human judgments for dialog system evaluation in a chatbot setting.
- A latent variational recurrent encoder-decoder model is pretrained on dialog dataset
- The model is trained to evaluate the similarity between the dialog context, reference response and the generated response.



HUSE:

Human Unified with Statistical Evaluation (HUSE), fetermines the similarity of the output distribution and a human generation reference distribution.

(Hashimoto et.al. 2019)



OPENAI – Learning to Summarize with Human Feedback:

A reinforcement learning (RL) based evalation framework with human feedback to train language models that are better at summarization Reward model via supervised learning predicts which summaries humans will prefer. Then a fine-tuned language model with RL produces summaries that score highly according to that reward model.

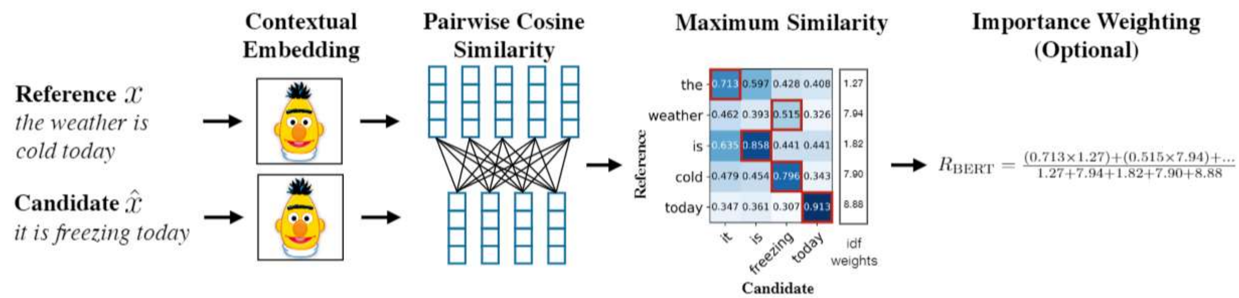
(Lowe, et.al., 2020)

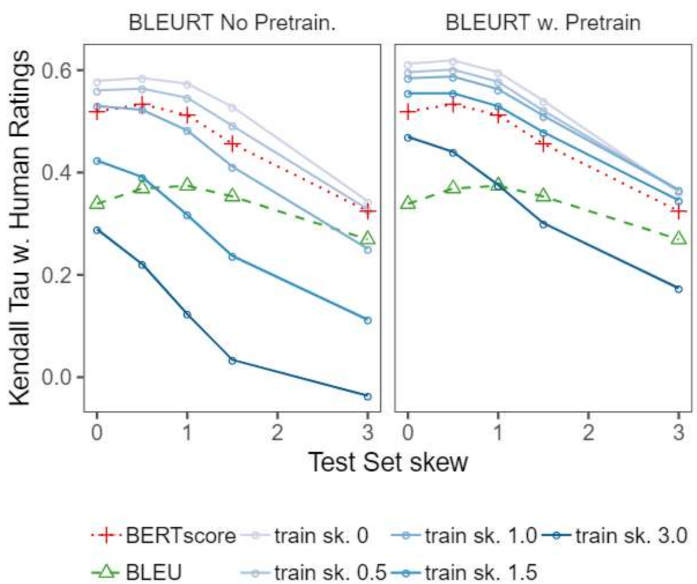
BERT Based Evaluation

BERTSCORE:

- Leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
- Computes precision, recall, and F1 measures, which are useful for evaluating a range of NLG tasks.
- It has been shown to correlate well with human judgments on sentence-level and system-level evaluations.

(Zhang et.al. 2020)



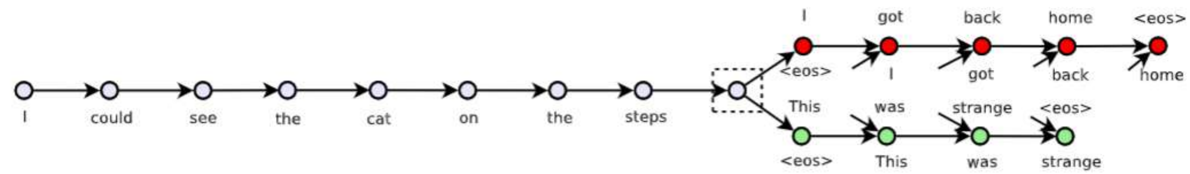


BLEURT:

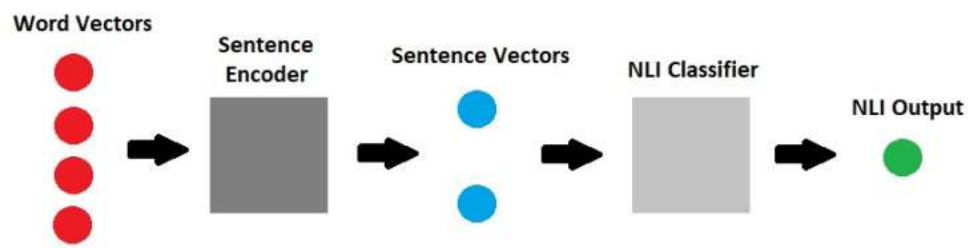
- A checkpoint from BERT is taken and fine-tuned on synthetically generated sentence pairs using automatic evaluation scores such as BLEU or ROUGE, and then further fine-tuned on system-generated outputs and human-written references using human ratings and automatic metrics as labels.
- The fine-tuning of BLEURT on synthetic pairs is an important step because it improves the robustness to quality drifts of generation systems.
- (Sellam et.al. 2020)

Trained Factual Correctness Metrics

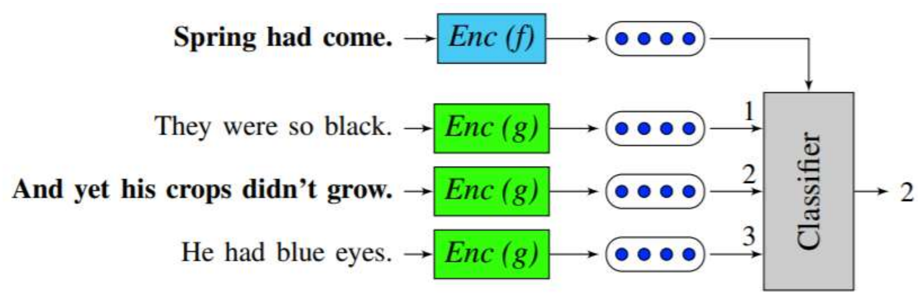
- **Skip Thoughts Vectors:** Unsupervised LSTM based model to encode rich contextual information by considering the surrounding context. (Kiros,et.al. 2015)



- **INFERSENT:** encode LSTM based Siamese networks to encode word-worder and is trained on high quality sentence inference dataset. (Conneau, et.al. 2017)

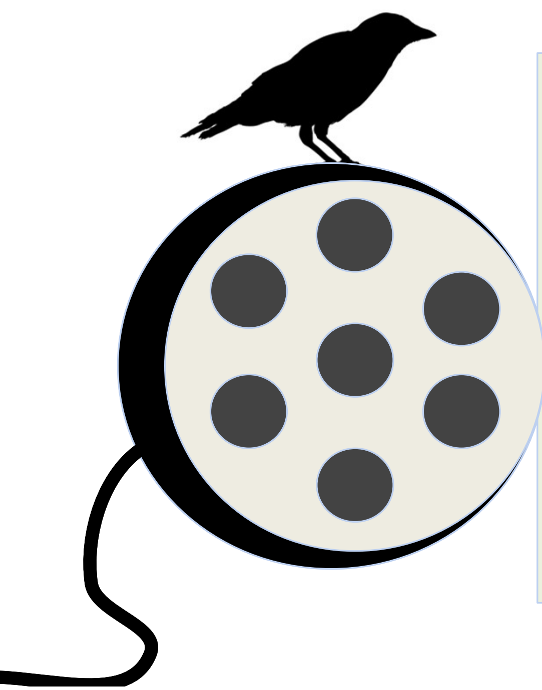


- **Quick Thoughts Vectors :** Unsupervised model of universal sentence embeddings trained on consecutive sentences. A classifier is trained to distinguish a context sentence from other contrastive sentences based on their embeddings. (Logeswaran and Lee, 2018)



Factual Consistency

Models are generating increasingly convincing text...



A device called the crow box could enable bird watchers to make money from their hobby as well As watch birds develop new skills.

The training aid can be used for teaching bullied crows how to collect coins in return of peanuts or simply test wild corvids' intelligence.

CNN\DM news summary generated from T5 language model

Factual Consistency

However this text is often very extractive or factually incorrect

A device called the crow box could enable bird watchers to make money from their hobby as well As watch birds develop new skills.

The training aid can be used for teaching bullied crows how to collect coins in return of peanuts or simply test wild corvids' intelligence.

Snippets from article

The sight of birds pecking at seed or nuts from a garden feeder fills many people with joy . Now , a device called the crow box could enable bird watchers to make money from their hobby.

… the training aid can be used to teach crows to collect coins in return for peanuts , or simply test the intelligence of wild corvids .

Factually Inconsistent Summaries

Generated Summary

A solar system has landed in the US stat of Ohio.

A lorry has been caught on camera overtaking a van at Grasshoppers' Park.

Irish President Leo Varadkar has said he is "very happy" with the way he is treating Canada.

Reference Summary

Solar impulse has landed in the US state of Ohio following the 12th stage of its circumnavigation of the globe.

Factually Inconsistent Summaries

Generated Summary

A solar system has landed in the US stat of Ohio.

Solar systems don't land on states.

A lorry has been caught on camera overtaking a van at Grasshoppers' Park.

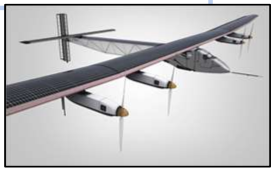
Wrong location, this happened in Lincolnshire.

Irish President Leo Varadkar has said he is "very happy" with the way he is treating Canada.

Varadkar is a prime minister, and he never said this (at least in the article).

Reference Summary

Solar impulse has landed in the US state of Ohio following the 12th stage of its circumnavigation of the globe.



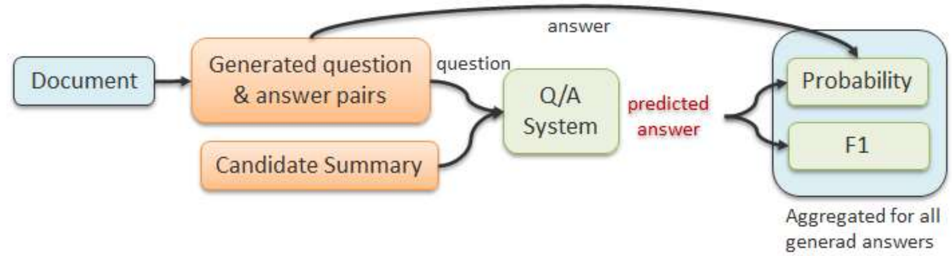
Solar impulse is a plane not a solar system.

Most Factual Correctness Metrics rely on:

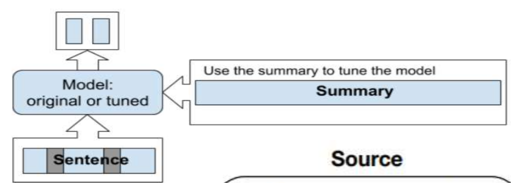
Keyword overlap, ignoring structure	Ngram-based metrics like ROUGE (Lin et al., 2014)
Contextual similarity	Metrics like BertScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020)
Proxy objective for coherence (and factuality?)	NLI metrics, Cloze task metrics and QA metrics like SummaQA (Scialom et al., 2020)

Trained Factual Correctness Metrics

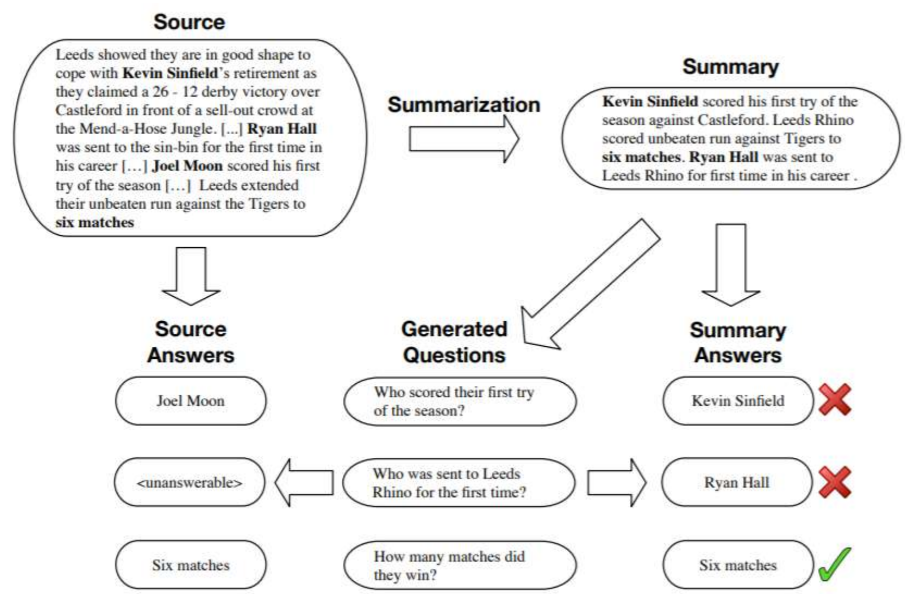
- **SummaQA:** BERT-based question-answering model to answer cloze-style questions using generated summaries. Named entities in source documents are masked to generate questions. (Scialom et.al. 2020)



- **BLANC:** as a measure of how well a summary helps an independent pre-trained language model while it performs its language understanding task on a document. (Vasilyev et.al. 2020)

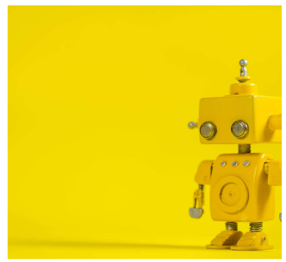


- **QAGS :** a question-answering and generation based automatic evaluation protocol that is designed to identify factual inconsistencies in a generated summary. They use fairseq for generation and BERT for QA model as a backbone (Wang et.al., 2020)

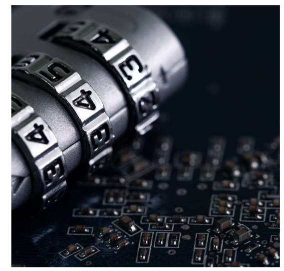


Summary of Challenges of Evaluating Text Generation

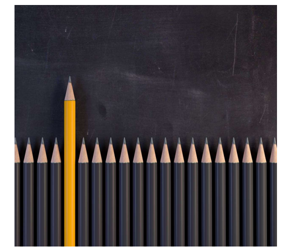
Making evaluation explainable



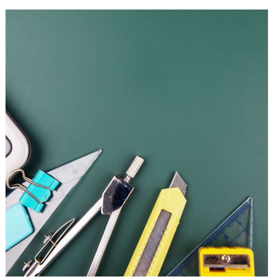
Detecting machine-generated text



Detecting and fake news

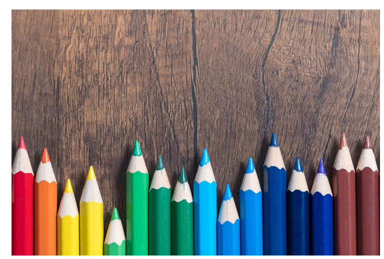


Improve corpus quality



Standardizing evaluation methods

Developing effective human evaluations



Evaluating ethical issues

Benchmarks

- Support research on open-domain text generation models.
- Evaluate the ***diversity***, the ***quality*** and the ***consistency*** of the generated texts on various datasets/domains
- Facilitate ***sharing*** of fine-tuned open-source implementations among researchers

Benchmarks

- Text generation benchmarks:
  – Generic text evaluation tasks
  – Specific text generation tasks
    - Machine Translation, Dialog Modeling, Summarization, etc.

Benchmarks

generic text evaluation tasks

General Text Evaluation Platforms

Features	OpenML	Kaggle	Topcoder	CrowdAI	ParlAI	CodaLab	EvalAI
AI Challenge Hosting		☑	☑	☑		☑	☑
Custom Metrics				☑	☑	☑	☑
Multiple phrases/splits				☑		☑	☑
Open Source	☑			☑	☑	☑	☑
Remote Evaluation					☑	☑	☑
Human Evaluation					☑		☑
Environments				☑			☑

Yadav et.al, 2020

EvalAI

Benchmarks

task specific text evaluation platforms

DialoGLUE

Dialogue Language Understanding Evaluation

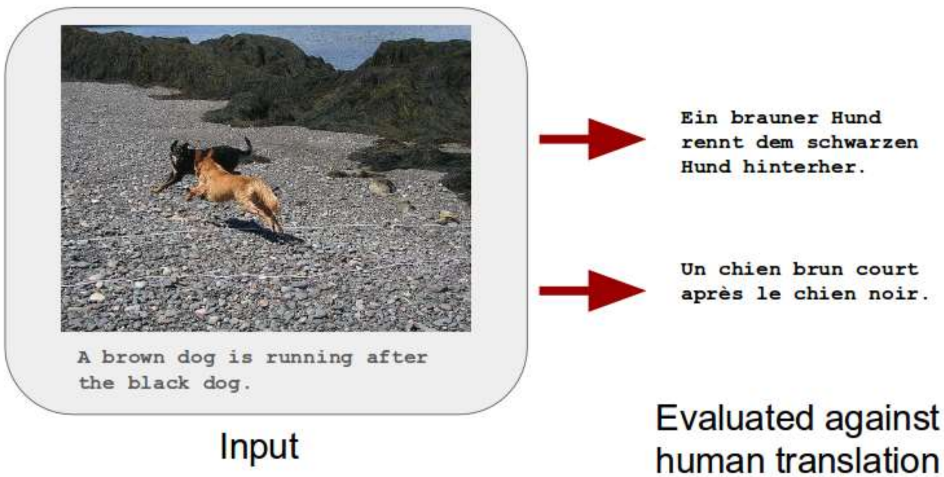
- Banking
- HWU
- Clinc
- Restaurant8k
- DSTC8 SGD
- TOP
- MultiWOZ 2.1

by Geek.AI
Texygen

- SeqGAN - SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient
- MaliGAN - Maximum-Likelihood Augmented Discrete Generative Adversarial Networks
- RankGAN - Adversarial ranking for language generation
- LeakGAN - Long Text Generation via Adversarial Training with Leaked Information
- TextGAN - Adversarial Feature Matching for Text Generation
- GSGAN - GANS for Sequences of Discrete Elements with the Gumbel-softmax Distribution

https://github.com/geek-ai/Texygen

WMT: Workshop on Machine Translation



- Builds on a series of annual workshops and conferences on statistical machine translation, going back to 2006
- It features shared tasks, evaluation metrics and datasets.
- BLUE has been standardized as MT evaluation metric in WMT

http://www.statmt.org/

Sitemap

- SMT Book
- Research Survey Wiki
- Moses MT System
- Europarl Corpus
- News Commentary Corpus
- Online Evaluation
- Online Moses Demo
- Translation Tool
- WMT Workshop 2014
- WMT Workshop 2013
- WMT Workshop 2012
- WMT Workshop 2011
- WMT Workshop 2010
- WMT Workshop 2009
- WMT Workshop 2008
- WMT Workshop 2007
- WMT Workshop 2006
- WPT Workshop 2005
- ACL SIG MT
- Edinburgh SMT Group
- SE Times Corpus

Statistical Machine Translation

This website is dedicated to research in statistical machine translation, i.e. the translation of text from one human language to another by a computer that learned how to translate from vast amounts of translated text.

Introduction to Statistical MT Research

- The Mathematics of Statistical Machine Translation by Brown, Della Petra, Della Pietra, and Mercer
- Statistical MT Handbook by Kevin Knight
- SMT Tutorial (2003) by Kevin Knight and Philipp Koehn
- ESSLLI Summer Course on SMT (2005), day1, 2, 3, 4, 5 by Chris Callison-Burch and Philipp Koehn.
- MT Archive by John Hutchins, electronic repository and bibliography of articles, books and papers on topics in machine translation and computer-based translation tools

Conferences and Workshops

See comprehensive list of NLP meetings.

Software

- Giza++ a training tool for IBM Model 1-5 (version for gcc-4)
- Moses, a complete SMT system
- UCAM-SMT, the Cambridge Statistical Machine Translation system
- Phrasal, a toolkit for phrase-based SMT
- cdec, a decoder for syntax-based SMT
- Joshua, a decoder for syntax-based SMT
- Jane, decoder for syntax-based SMT
- Pharaoh a decoder for phrase-based SMT
- Rewrite a decoder for IBM Model 4
- BLEU scoring tool for machine translation evaluation

Parallel Corpora

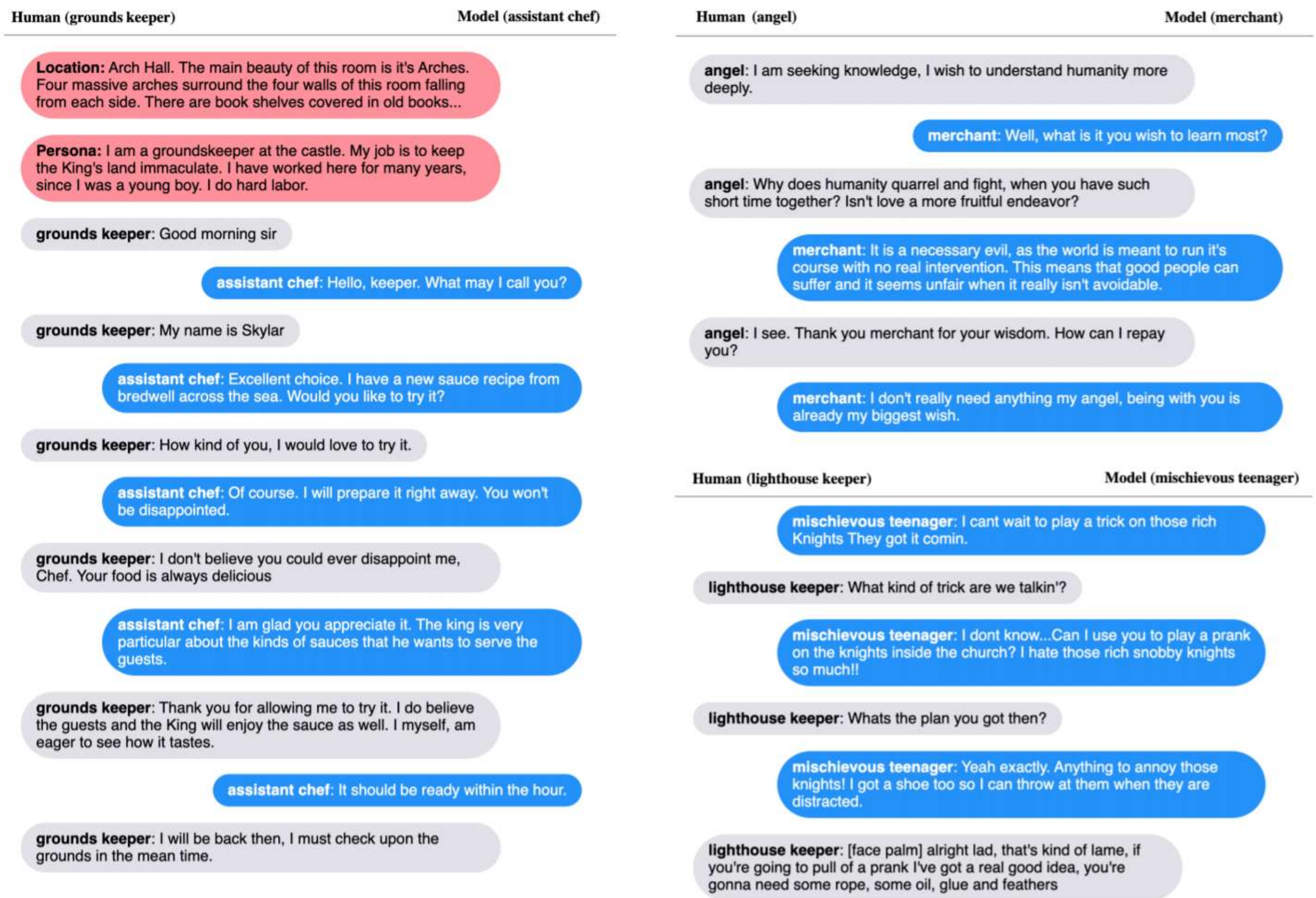
- LDC Linguistic Data Consortium
- Canadian Hansards

http://www.statmt.org/

SummEval

- Provides data and evaluation platform for summarization tasks
- Enables benchmarks for more than 10 different trained and un-trained evaluation metrics

Lifelong Open-Domain Dialog Learning



Shuster, et.al, 2020

References and Additional Reading

[1] Evaluation of Text Generation, Asli Celikyilmaz, Elizabeth Clark, Jianfeng Gao