University of Virginia

Stanford University

HUGGING FACE

Microsoft Research

The Amazing World of Neural Language Generation

EMNLP 2020

November 20, 2020

Yangfeng Ji
University Of Virginia

Antoine Bosselut
Stanford University

Thomas Wolf
Huggingface

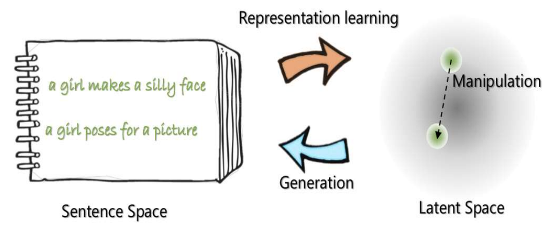
Asli Celikyilmaz
Microsoft Research

Natural Language Generation

A sub-field in

natural language processing

Building software systems to produce ***coherent***, ***readable*** and **useful** written or spoken text.



Produces explanations, summaries, answers to questions, poems, dialogs, programs, …

Image source: chunyuan.li

Machine Translation

[User] We will discuss several issues today!

[System] Nous discuterons plusieurs questions aujourd'hui.

[System] 我们今天将讨论几个问题！

[System] Bugun cok sayıda sorunu tartısacağız.

Conversational Dialog Systems

[USER] Where is my next appointment and am I free for lunch?

[Agent] Your next meeting is at 10:30 at City Center. Did you want me to book a place for lunch in downtown ?

Understanding Visually Grounded Language

Image Source : bing.com

Image or Video Captioning

a man is showing how to make a drink

A car drives really fast

Olivastri, et.al., ICCVW, 2019

Multi Document Summarization

Summary

High Quality Content by WIKIPEDIA articles! Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary report allows individual users, so as professional information consumers, to quickly familiarize themselves with information contained in a large cluster of documents. In such a way, multi-document summarization systems are complementing the news aggregators performing the next step down the road of coping with information overload.

Image Source: Alphascript Publishing

Data-to-Text Generation

TEAM	WIN	LOSS	PTS	FG_PCT	RB	AST	...
Pacers	4	6	99	42	40	17	...
Celtics	5	4	105	44	47	22	...

PLAYER	H/V	AST	RB	PTS	FG	CITY	...
Jeff Teague	H	4	3	20	4	Indiana	...
Miles Turner	H	1	8	17	6	Indiana	...
Isaiah Thomas	V	5	0	23	4	Boston	...
Kelly Olynyk	V	4	6	16	6	Boston	...
Amir Johnson	V	3	9	14	4	Boston	...
...	...	...	...	...	...	...	

PTS: points, FT_PCT: free throw percentage, RB: rebounds, AST: assists, H/V: home or visiting, FG: field goals, CITY: player team city.

The **Boston Celtics** defeated the host **Indiana Pacers 105-99** at Bankers Life Fieldhouse on Saturday. In a battle between two injury-riddled teams, the Celtics were able to prevail with a much needed road victory. The key was shooting and defense, as the **Celtics** outshot the **Pacers** from the field, from three-point range and from the free-throw line. Boston also held Indiana to **42 percent** from the field and **22 percent** from long distance. The Celtics also won the rebounding and assisting differentials, while tying the Pacers in turnovers. There were 10 ties and 10 lead changes, as this game went down to the final seconds. Boston (**5–4**) has had to deal with a gluttony of injuries, but they had the fortunate task of playing a team just as injured here. **Isaiah** Thomas led the team in scoring, totaling **23 points and five assists on 4–of–13** shooting. He got most of those points by going 14–of–15 from the free-throw line. **Kelly Olynyk** got a rare start and finished second on the team with his **16 points, six rebounds and four assists**.

Puduppully, et.al.,2019

Meeting Summarization

C: Looking at what we've got, we we want an LCD display with a spinning wheel.
B: You have to have some push-buttons, don't you?
C: Just spinning and not scrolling, I would say.
B: I think the spinning wheel is definitely very now.
A: but since LCDs seems to be uh a definite yes,
C: We're having push-buttons on the outside
C: and then on the inside an LCD with spinning wheel,

Decision Abstract (Summary):
The remote will have push buttons outside, and an LCD and spinning wheel inside.

A: and um I'm not sure about the buttons being in the shape of fruit though.
D: Maybe make it like fruity colours or something.
C: The power button could be like a big apple or something.
D: Um like I'm just thinking bright colours.

Problem Abstract (Summary):
How to incorporate a fruit and vegetable theme into the remote.

Wang and Cardie, ACL, 2013

Conversational Dialog Summarization

[Nurse] Hi Mr.#name#, you were discharged on #date#. There are some questions i'd like to check with you.
[Patient] Ok, Ok.
[Nurse] Well, have you been experiencing swelling recently?
[Patient] Swelling? It comes and go, comes and go.
[Nurse] Comes and go ... I see .. **#repetition#**
[Nurse] ... **#pause#**... When did it start?
[Patient] Let me see, started from three weeks ago.
... ...
[Nurse] Are you experiencing any headache right now as we speak?
[Patient] Umm ... **#back-channel#**
[Nurse] Let me check, the last time you told me is sometimes at night.
[Patient] Oh, right, only a bit.
... ...
[Nurse] Still feel some chest pain or chest discomfort?
[Patient] Yes, my head is... **#false-start#** no, the pain is much better. Still feel headache though ... **#topic-drift#**
... ...
[Nurse] Any giddiness or palpitation?
[Patient] Palpitation? Do not have-- #**interruption**#
[Nurse] Well ... Do you-- #**interruption**#
[Patient] and no giddiness, no, nothing.
... ...
[Nurse] Ok, you need to check your heartrate everyday.
[Nurse] Do you know how to use the device?
[Patient] Yes, yes, no problem.
... ...

Swelling: started from three weeks ago, comes and go.
Headache: sometimes, at night, only a bit.
Chest pain: much better.
Dizziness: none.

Liu, et.al., ASRU 2019

Other Text Generation Tasks

- Question Generation
- Paraphrase Generation
- Poetry Generation
- Long Question Answering
- Visual Dialog Systems
- Document/Article Generation
- Program Synthesis
- Search Snippet Generation

Why Automatic Text Generation?

Efficiency

Education

Productivity

History of Natural Langauge Generation Systems

Template based systems:
Uses rules and templates

1965

History of Natural Langauge Generation Systems

Template based systems:
Uses rules and templates

1965

1978

Rule-based + Data Driven pipelines :
Document planning + microplanning + realization

History of Natural Langauge Generation Systems

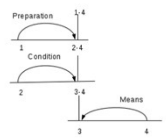
Template based systems:
Uses rules and templates

1965

1978

Rule-based + Data Driven pipelines :
Document planning + microplanning + realization

Modeling Discourse Structures :
Relation learning, Rhetorical Structure Theory



1985

History of Natural Langauge Generation Systems

1965

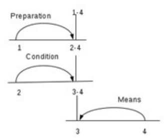
Template based systems:
Uses rules and templates

1978

Rule-based + Data Driven pipelines :
Document planning + microplanning + realization

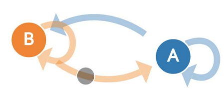
1985

Modeling Discourse Structures :
Relation learning, Rhetorical Structure Theory



1990

Statistical Methods (Markov Chains):
Sentence compression, reordering, lexical paraphrasing, syntactic transformation



History of Natural Langauge Generation Systems

1965

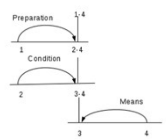
Template based systems: Uses rules and templates

1978

Rule-based + Data Driven pipelines : Document planning + microplanning + realization

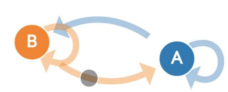
1985

Modeling Discourse Structures : Relation learning, Rhetorical Structure Theory



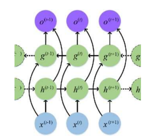
1990

Statistical Methods (Markov Chains): Sentence compression, reordering, lexical paraphrasing, syntactic transformation



2013

RNNs, LSTMs, GRUs: Autoregressive DNNs + Recurrent units, backpropagation

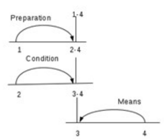


History of Natural Langauge Generation Systems

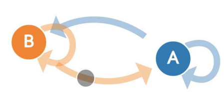
1965 — **Template based systems**: Uses rules and templates

1978 — **Rule-based + Data Driven pipelines** : Document planning + microplanning + realization

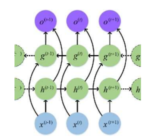
1985 — **Modeling Discourse Structures** : Relation learning, Rhetorical Structure Theory



1990 — **Statistical Methods (Markov Chains)**: Sentence compression, reordering, lexical paraphrasing, syntactic transformation



2013 — **RNNs, LSTMs, GRUs**: Autoregressive DNNs + Recurrent units, backpropagation



2017 — **Transformers** : GPT (1/2/3), GROVER. TransformerXL, DialoGPT

History of Natural Langauge Generation Systems

1965

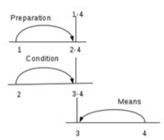
Template based systems:
Uses rules and templates

1978

Rule-based + Data Driven pipelines :
Document planning + microplanning + realization

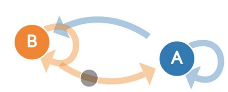
1985

Modeling Discourse Structures :
Relation learning,
Rhetorical Structure Theory



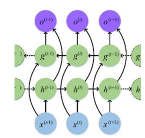
1990

Statistical Methods (Markov Chains):
Sentence compression, reordering, lexical paraphrasing, syntactic transformation



2013

RNNs, LSTMs, GRUs:
Autoregressive DNNs + Recurrent units, backpropagation



2017

Transformers :

GPT (1/2/3), GROVER.
TransformerXL, DialoGPT

Language Modeling

Conditional Language Modelling

$$P(w_1, \cdots, w_n) = \prod_{i=1}^{n} P(w_i | w_{i-1}, \cdots, w_1)$$

Examples from D. Jurafsky slides

Conditional Language Modelling

$$P(w_1, \cdots, w_n) = \prod_{i=1}^{n} P(w_i | w_{i-1}, \cdots, w_1)$$

- Speech Recognition
  - P(I saw a man) >> P(eyes awe of an)

Examples from D. Jurafsky slides

Conditional Language Modelling

$$P(w_1, \cdots, w_n) = \prod_{i=1}^{n} P(w_i | w_{i-1}, \cdots, w_1)$$

- Speech Recognition
  - P(I saw a man) >> P(eyes awe of an)
- Spell Correction
  - P(about fifteen minutes from) > P(about fifteen minuets from)

Examples from D. Jurafsky slides

Conditional Language Modelling

$$P(w_1, \cdots, w_n) = \prod_{i=1}^{n} P(w_i | w_{i-1}, \cdots, w_1)$$

- Speech Recognition
  - P(I saw a man) >> P(eyes awe of an)
- Spell Correction
  - P(about fifteen minutes from) > P(about fifteen minuets from)
- Machine translation, Question Answering, Paraphrasing, Image captioning, Summarization, others…

Examples from D. Jurafsky slides

Long Text Generation – Image Story Telling

Two men are riding bicycles down a street. One man is behind the other riding on the side of a paved road. Both bicycles have black baskets on the front attached to the handlebars. Both men are holding umbrellas. The first man is wearing blue pants and a white and black striped shirt. His umbrella is blue. The second man is wearing black pants and a light purple shirt. His umbrella is aqua blue colored. Behind them, along the road are shops. One of the shops is closed with a silver metal fence covering the entrance. The other shop is a clothing store and there are four mannequins with clothing outside the door. There is also a pot-ted plant with red flowers outside one of the shops.

Krause, et.al., CVPR, 2017

Long Text Generation – Image Story Telling

Two men are riding bicycles down a street. One man is behind the other riding on the side of a paved road. Both bicycles have black baskets on the front attached to the handlebars. Both men are holding umbrellas. The first man is wearing blue pants and a white and black striped shirt. His umbrella is blue. The second man is wearing black pants and a light purple shirt. His umbrella is aqua blue colored. Behind them, along the road are shops. One of the shops is closed with a silver metal fence covering the entrance. The other shop is a clothing store and there are four mannequins with clothing outside the door. There is also a pot-ted plant with red flowers outside one of the shops.

Krause, et.al., CVPR, 2017

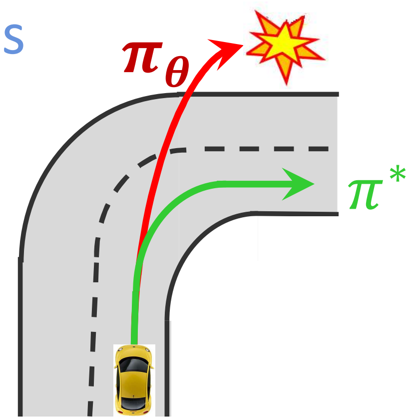
Long Text Generation with RNN-LM

two men are **riding a bicycle** on the road. it is a sunny day. a man is **riding a blue bicycle** on the street. the men with blue umbrella is **riding a bicycle** on a rainy day. the woman black dress is standing on the road.

Neural Text Generation Issues – Model Selection

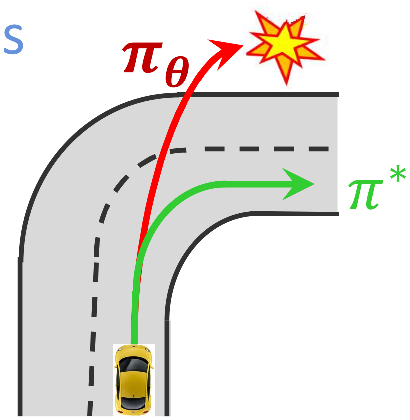
Neural Text Generation Issues – Model Selection

Exposure Bias



Neural Text Generation Issues – Model Selection

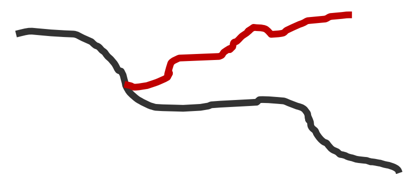
Exposure Bias



Compounding Errors / Label Bias

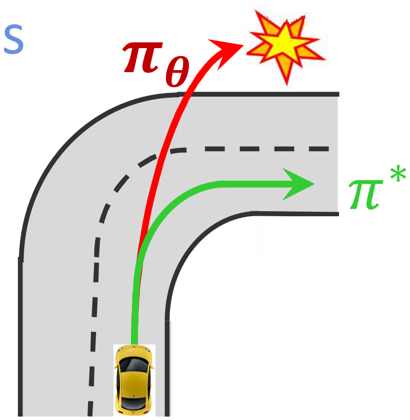
Gold: The cat purrs

Pred: The dog barks



Neural Text Generation Issues – Model Selection

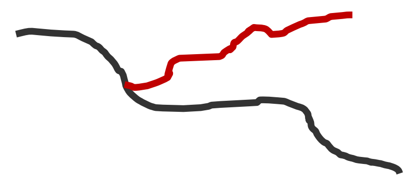
Exposure Bias



Compounding Errors / Label Bias

Gold: The cat purrs

Pred: The dog barks



Coherence and Narrative Flow

"do you like animals?"

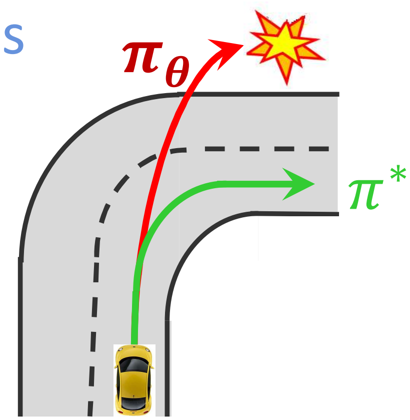
yes, I have three cats.

how many cats do you have ?

I don't have cats.

Neural Text Generation Issues – Model Selection

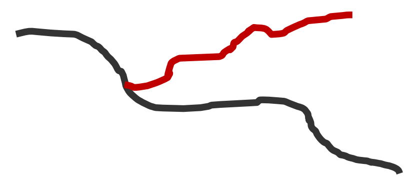
Exposure Bias



Compounding Errors / Label Bias

Gold: The cat purrs

Pred: The dog barks



Coherence and Narrative Flow

“do you like animals?”

yes, I have three cats.

how many cats do you have ?

I don’t have cats.

Surrogate objective function

$$P^*(x|c) \rightarrow P_\theta(x|c)$$

Perplexity

vs

BLEU/ROUGE/Neural Rewards

Weaknesses of MEGA Language Models for GENERATION!

Inconsistent output

Crippled by length

coreference issues

Longer string that are repeated many times in the dataset

Unnecessarily repeating entities

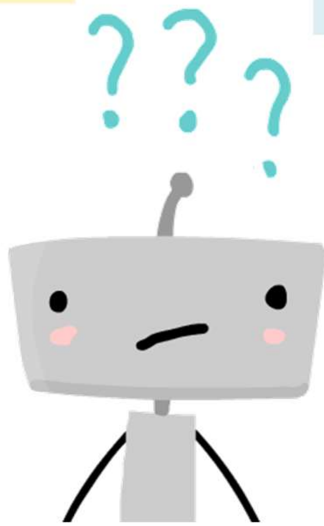
Maintaining **coherence** between **paragraphs**

No **real** understanding, **commonsense**, **factual correctness**”

How to learn discourse?

Lack of implicit “**planning**”

sub-optimal evaluation metrics “perplexity”



MLE: cannot capture high-level semantics

Biased pre-trained models

Domain transfer is hard

Single path generative flow

Softmax Bottleneck issues!

Surrogate-loss functions

Auto-regressive!

exposure-bias

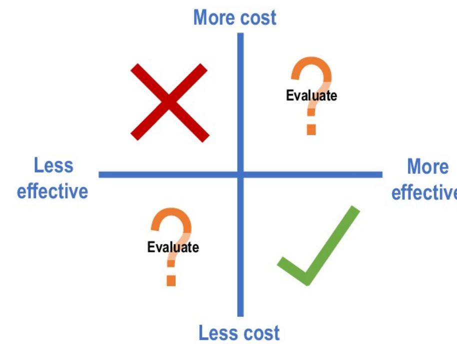
Degenerative sampling methods

word-by-word generation: can’t see **global context**!

Research on Evaluation: Now more than ever!

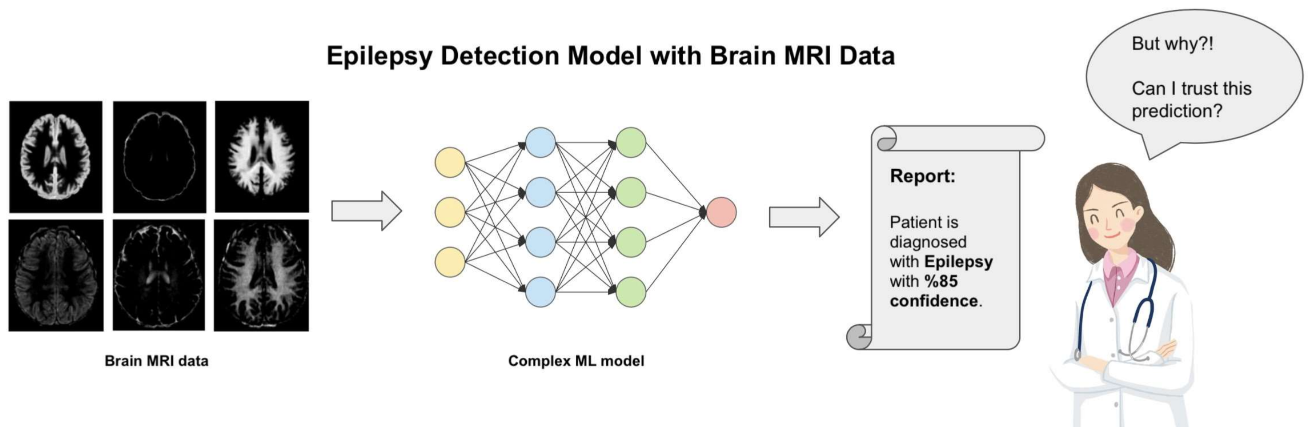
Detecting machine generated text

Standards in Text Generation Evaluations



Effective Evaluation

Evaluating Ethical Issues



Making Evaluations Explainable

Tutorial Schedule

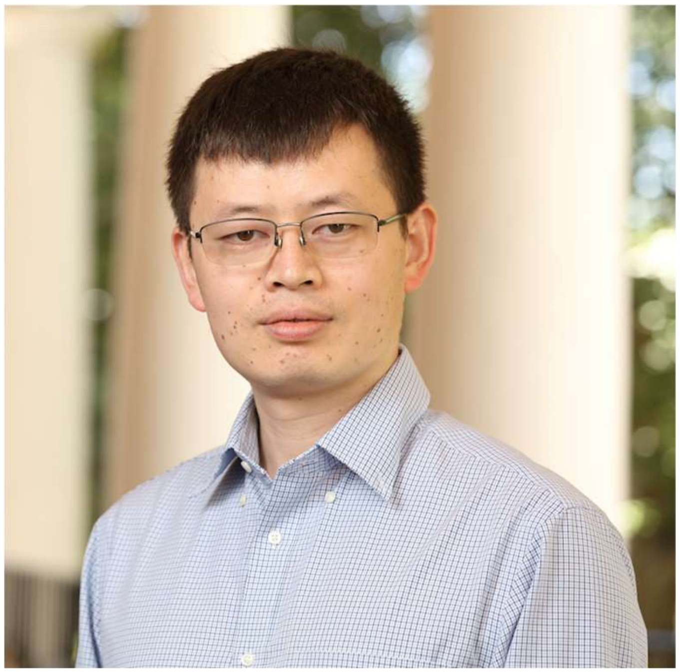
Approximate Time	Segment Title	Speaker	Description
15 mins	Introduction	Asli Celikyilmaz	Why is it important today to discuss neural text generation?
20 mins	Modeling: NN Architecture	Yangfeng Ji	Journey of NNs architectures that are used for text generation up-to-date
20 mins	Modeling: Generation With Rich context	Yangfeng Ji	How to efficiently use context in langauge generation?
20 mins	Training Text Generation Models	Antoine Bosselut	What are the best practices in training neural text generation systems today?
20 mins	Decoding Algorithms	Antoine Bosselut	Amazing world of neural decoding algorithms and beyond!
35 mins	Evaluation and Benchmarks	Asli Celikyilmaz	How are neural text generation models evaluated today? Future directions!
40 mins	Building NLG Systems	Thomas Wolf	Adventure into building an NLG system using state-of-the art tools and libraries.

Introduction and Evaluation

“Let’s discuss the Evaluation of Text Generation Systems, which is more important today than ever!”

Time:

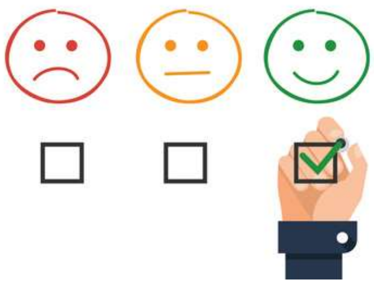
Modeling



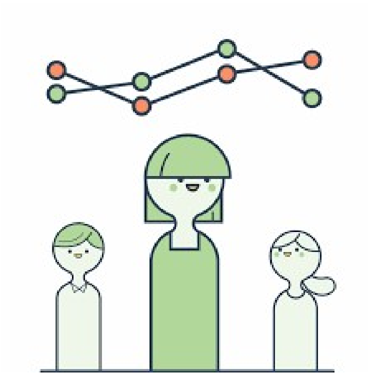
Training and Decoding

Building of NLG Systems

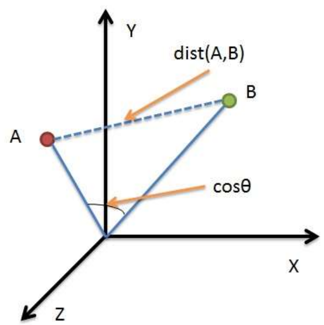
Evaluation and Benchmarks



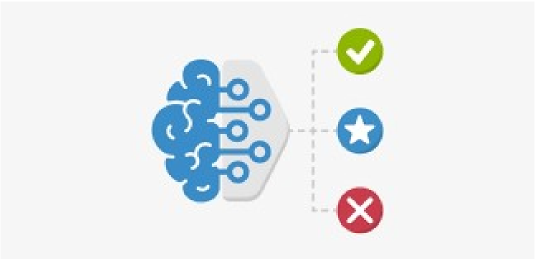
Types of Evaluation Methods for Text Generation



Human Evaluations

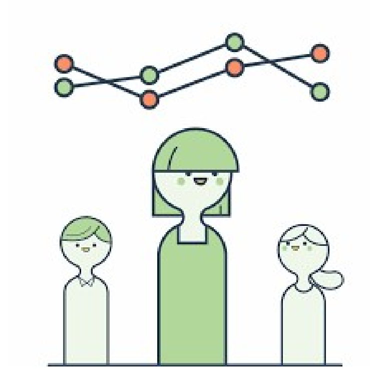


Un-trained Metrics



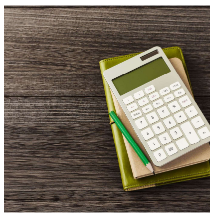
Trained Metrics

Human Evalutions

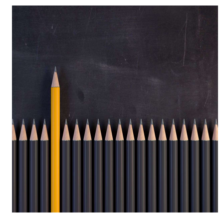


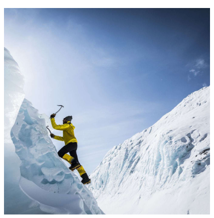
- Most important form of evalation for NLG systems
- Automatic metrics fall short of replicating human decisions
- Gold standard in developing new automatic metrics

Human Evalutions: Issues


Expensive


Time Consuming


Quality Control


Challenging Criteria


Inconsistency in Evaluations

Inconsistency in reporting

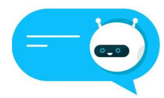
Intrinsic Human Evaluations

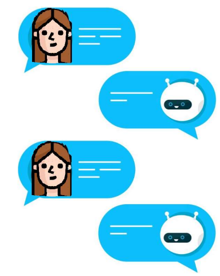
- Ask *humans* to evalute the quality of generated text
- Overall or along some specific dimension:
  - fluency
  - coherence
  - factuality and correctness
  - adequacy
  - commonsense
  - style / formality
  - grammaticality
  - typicality
  - redundancy

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!





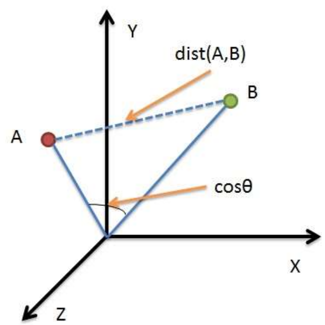
Turn Level	Dialog Level
▪ Interesting ▪ Engaging ▪ Generic/Specific ▪ Relevant ▪ Semantically appropriate ▪ Understandable ▪ Fluently Written ▪ Correct vs. Misunderstanding ▪ Overall Impression	• Coherent • Recovers from errors • Consistent • Diversity in its responses • Topic Depth • Likable (empathy,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

Mehri and Eskenazi, SIGDIAL 2020

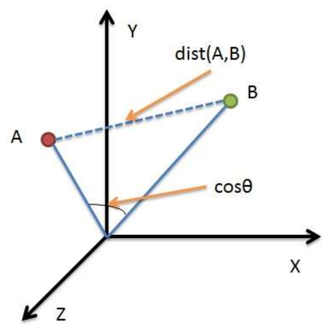
Human Evaluations: Other Aspects

- 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

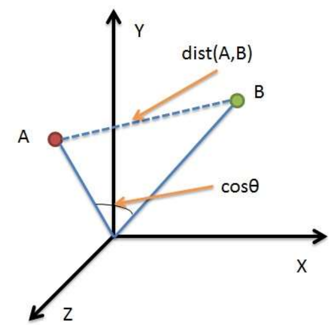


Untrained Automatic Evaluation Metrics



- 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



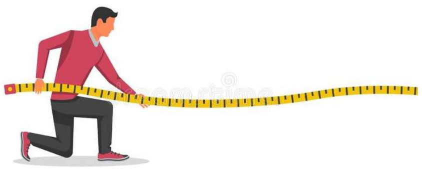
1. *n*-gram overlap metrics
2. distance-based metrics
3. *n*-gram based diversity metrics
4. content overlap metrics

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		✓					

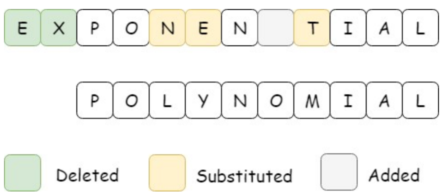
MT: Machine Translation
DG: Document Generation
RG: Response Generation
IC: Image Captioning
SUM: Summarization
QG: Question Generation

2. Distance Based Metrics

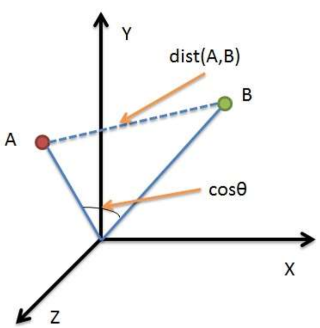


- Distance function to measure similarity between two text units
- Text units are represented as vectors → embeddings!
- Even though embeddings are pretrained, distance metrics used to measure the similarity are not!

2. Distance Based Metrics

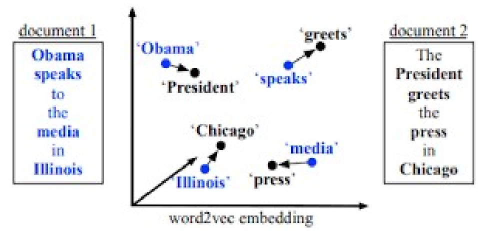


Edit Distance:
Measures how dissimilar two text units are based on the minimum number of operations required to transform one text into another.

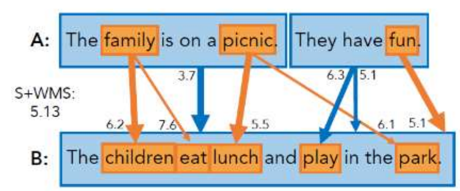


Vector Similarity:
Embedding based similarity for semantic distance between text.

MEANT
YISI
Word Movers Distance
Sentence Movers similarity

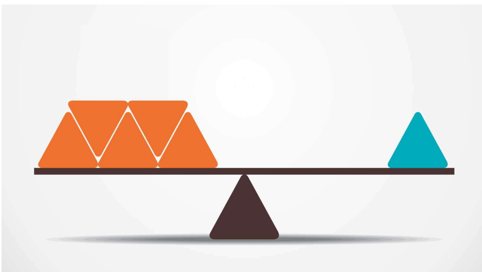


Word Mover's Distance:
Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), represented with relative word frequencies. It combines item similarity on bag-of-word histogram representations of text with word embedding similarity.



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

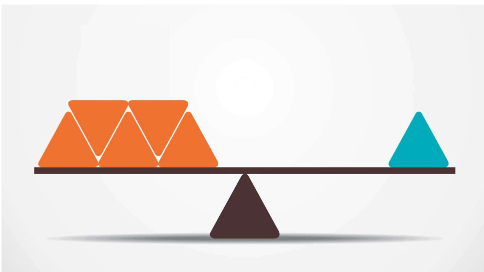


Type-to-Token Ratio (TTR):

- 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

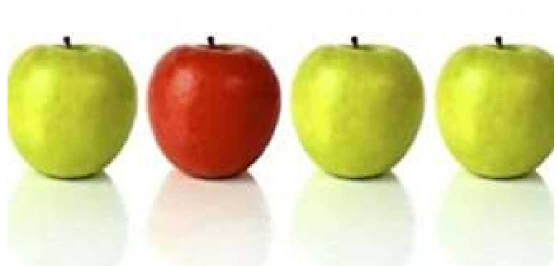
3. *n*-gram Based Diversity Metrics



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

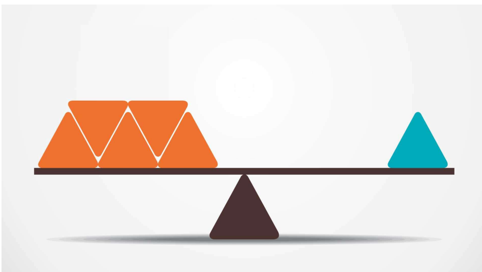
- Used to measure the lexical variety in a text:
  The higher the TTR, the more varied the text vocabulary



`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

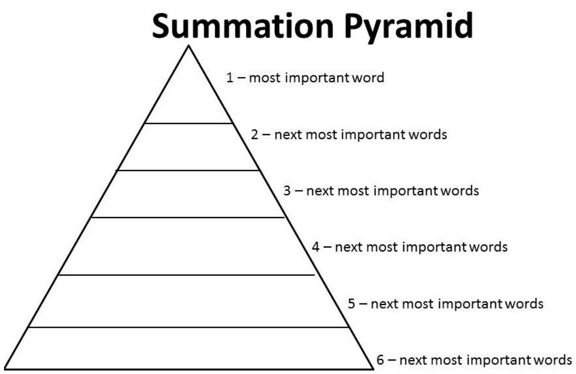
`Self-BLEU`:

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)

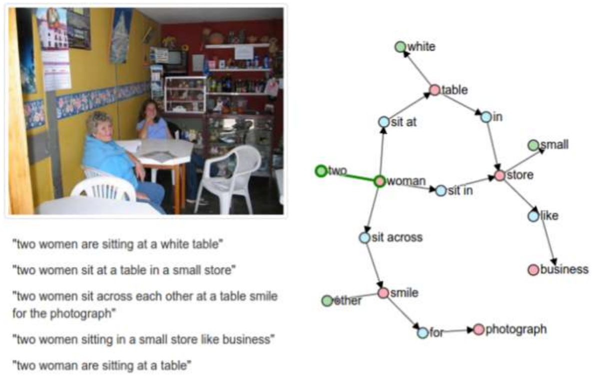
Textual Lexical Diversity:

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)

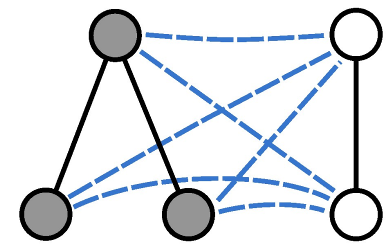
4- Content Overlap Metrics



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)

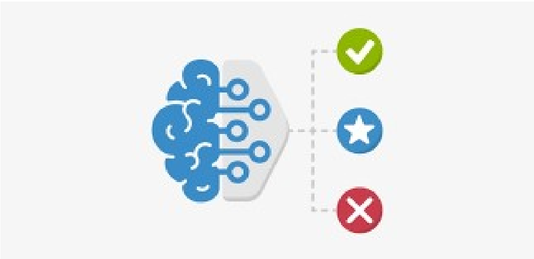


SPIDER:
A combination of semantic graph similarity (SPICE) and *n*-gram similarity measure (CIDER), the SPICE metric yields a more complete quality evaluation metric.
(Liu, et.al., 2017)

Machine Learnt Metrics

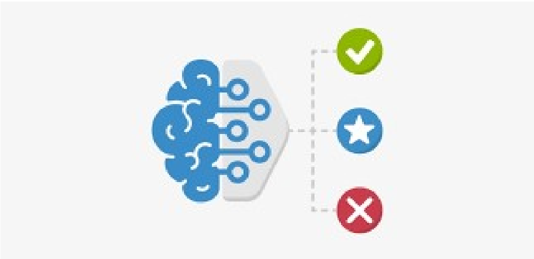
	Dialog Response Generation	Image Captioning
Context	**Speaker A**: Hey John, what do you want to do tonight? **Speaker B**: Why don't we go see a movie?	
Ground-Truth	**Response:** Nah, I hate that stuff, let's do something active.	**Caption:** a man wearing a red life jacket is sitting in a canoe on 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

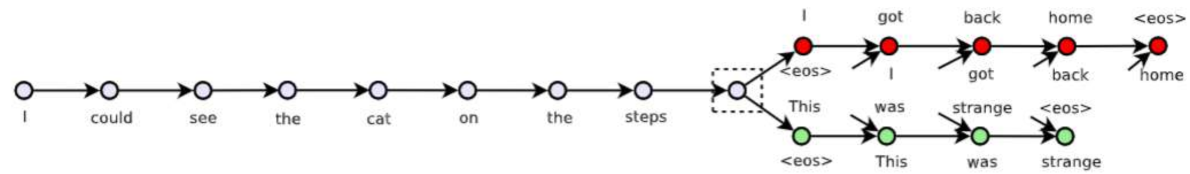
Machine Learnt Evaluation Metrics



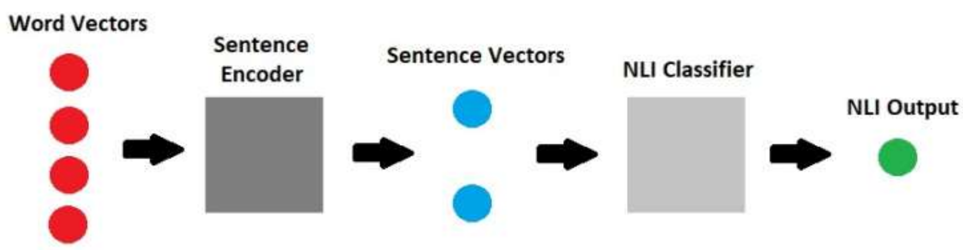
1. Sentence similarity metrics
2. Regression Based Metrics
3. Learning from Human Feedback
4. BERT-Based Evaluation
5. Composite Metrics
6. Factual Correctness metrics

Sentence Similarity Metrics

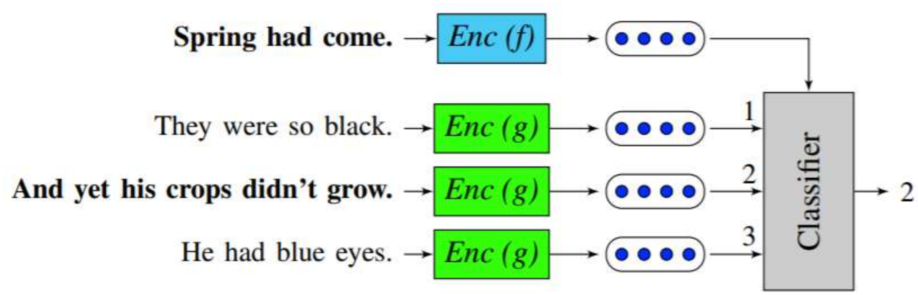
- **Skip Thoughts Vectors:** Unsupervised LSTM based model to encode rich contextual information by considering the surrounding context. (Kiros,et.al. 2015)



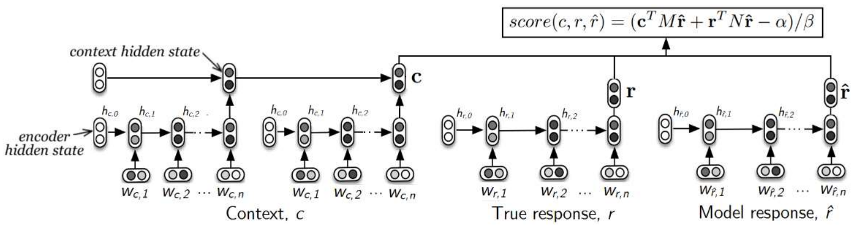
- **INFERSENT:** encode LSTM based Siamese networks to encode word-worder and is trained on high quality sentence inference dataset. (Conneau, et.al. 2017)



- **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)

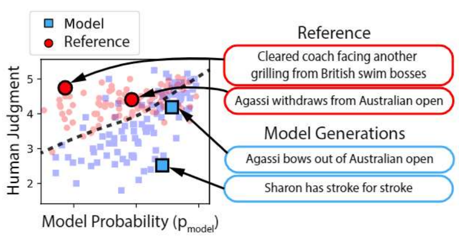


Learning from Human Feedback



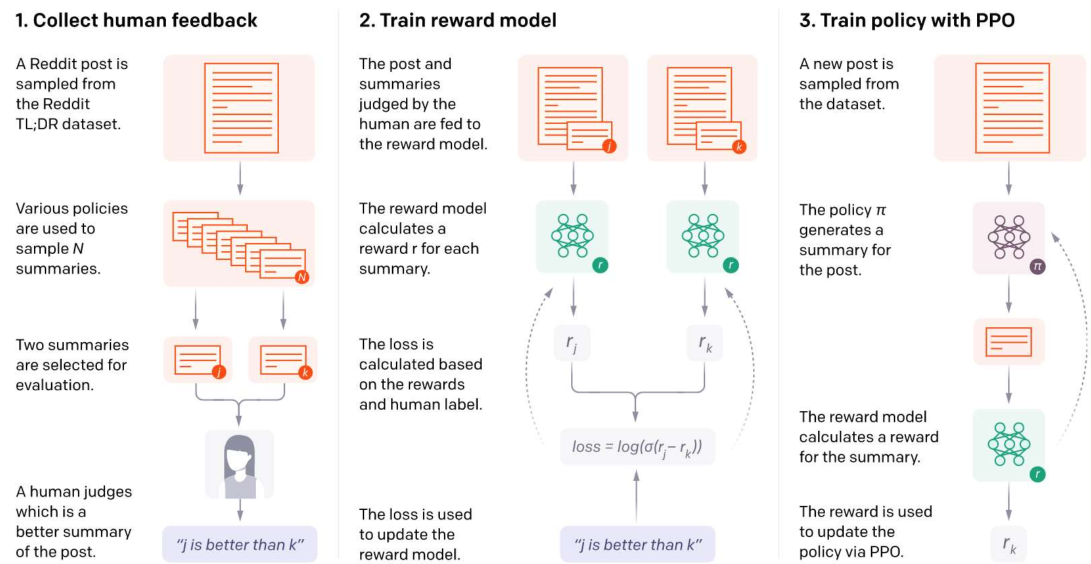
ADEM:

- 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:

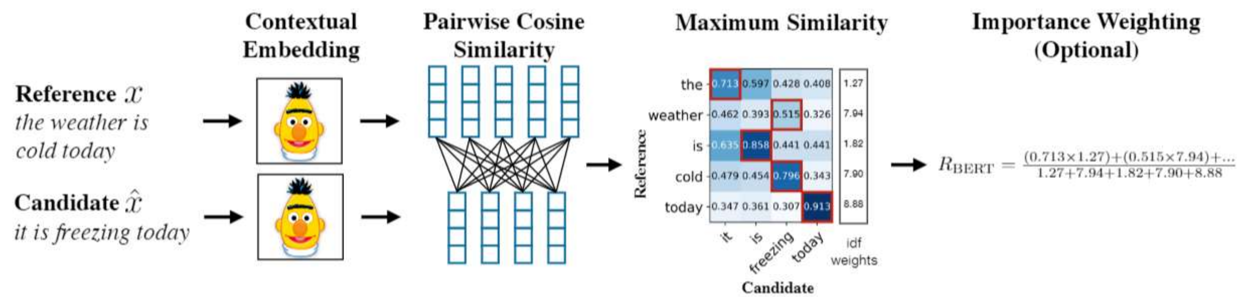
A reinforcement learning (RL) based evalation framework with human feedback to train language models that are better at summarization
Reward model via supervised learning predicts which summaries humans will prefer. Then a fine-tuned language model with RL produces summaries that score highly according to that reward model.
(Lowe, et.al., 2020)

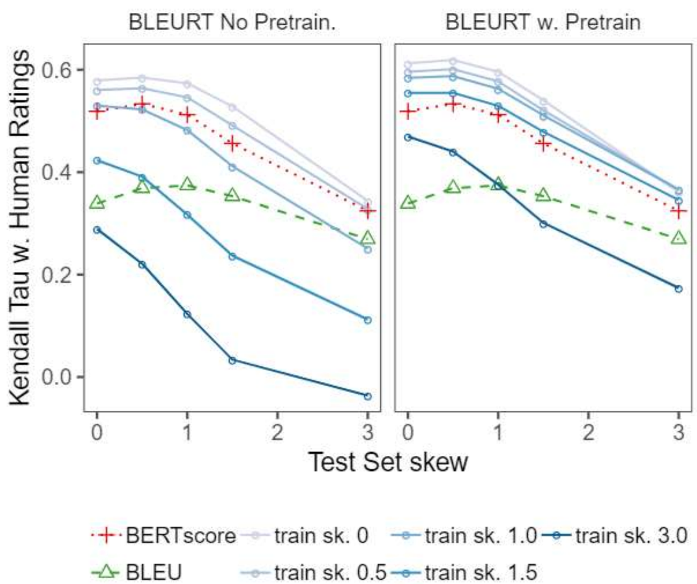
BERT Based Evaluation

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)



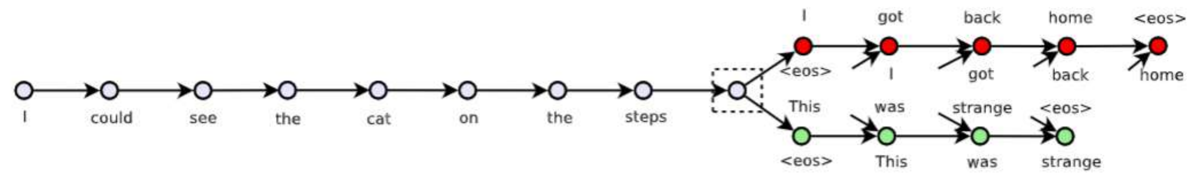


BLEURT:

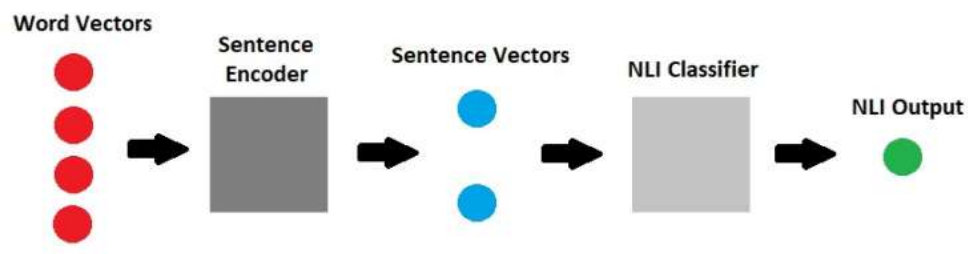
- A checkpoint from BERT is taken and fine-tuned on synthetically generated sentence pairs using automatic evaluation scores such as BLEU or ROUGE, and then further fine-tuned on system-generated outputs and human-written references using human ratings and automatic metrics as labels.
- The fine-tuning 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

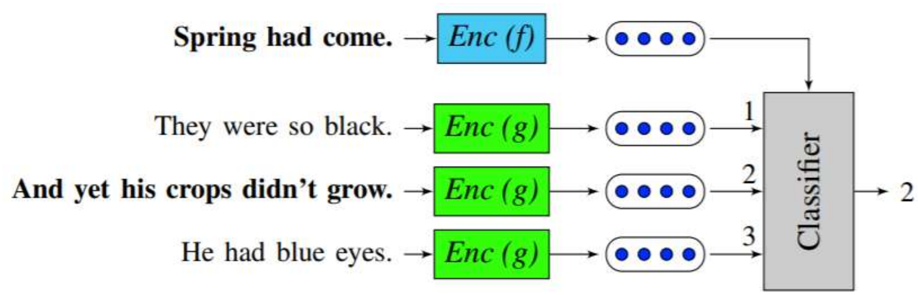
- **Skip Thoughts Vectors:** Unsupervised LSTM based model to encode rich contextual information by considering the surrounding context. (Kiros,et.al. 2015)



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

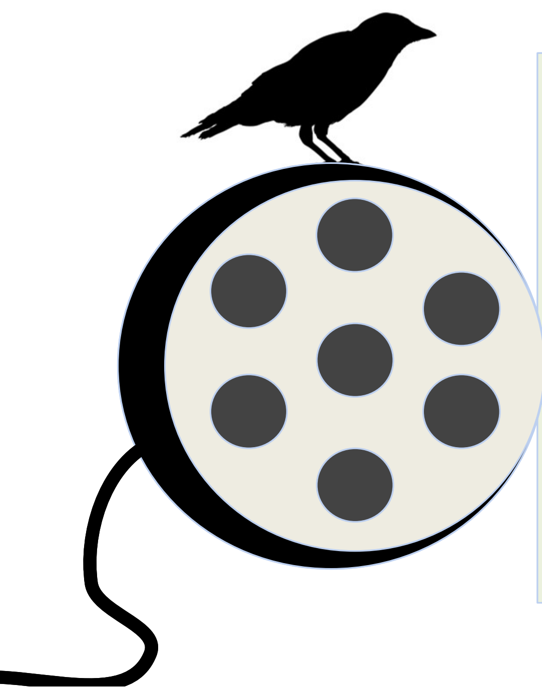


- **Quick Thoughts Vectors :** Unsupervised model of universal sentence embeddings trained on consecutive sentences. A classifier is trained to distinguish a context sentence from other contrastive sentences based on their embeddings. (Logeswaran and Lee, 2018)



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

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.

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.

... 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.

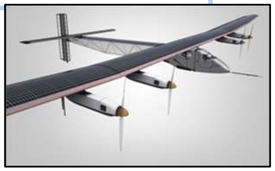
Wrong location, this happened in Lincolnshire.

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.



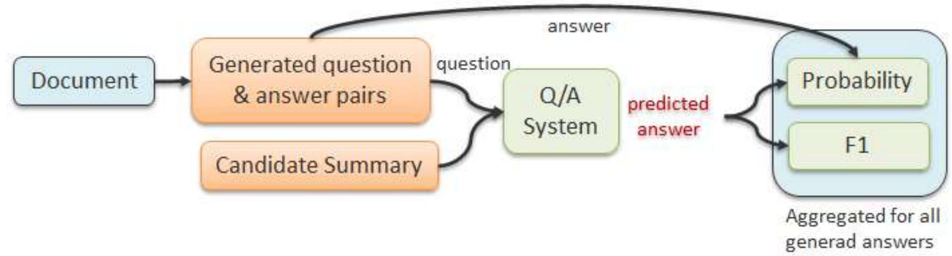
Solar impulse is a plane not a solar system.

Most Factual Correctness Metrics rely on:

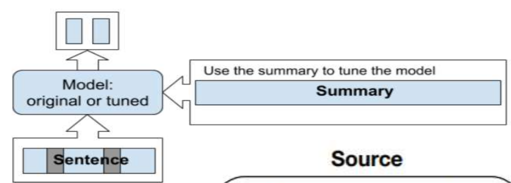
Keyword overlap, ignoring structure	Ngram-based metrics like ROUGE (Lin et al., 2014)
Contextual similarity	Metrics like BertScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020)
Proxy objective for coherence (and factuality?)	NLI metrics, Cloze task metrics and QA metrics like SummaQA (Scialom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

Trained Factual Correctness Metrics

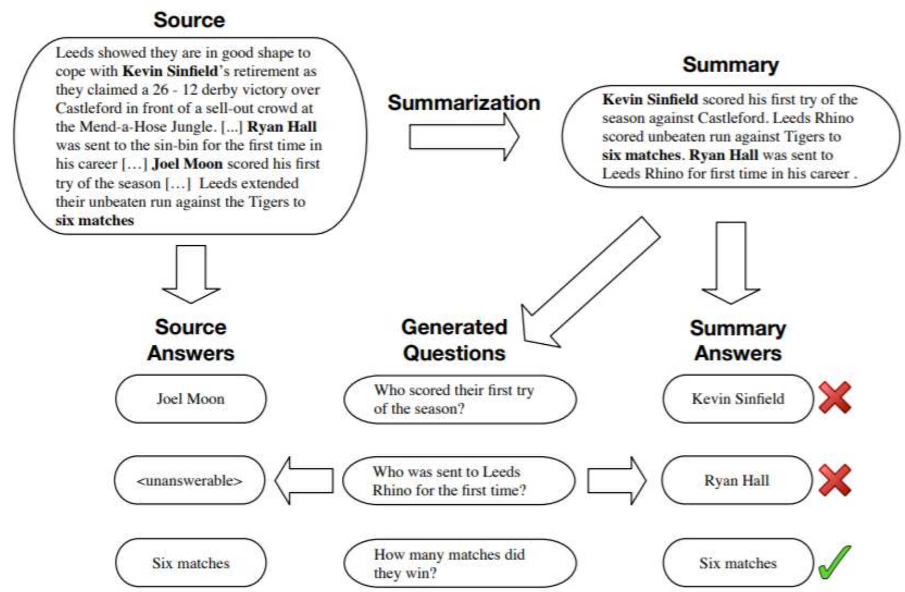
- **SummaQA:** BERT-based question-answering model to answer cloze-style questions using generated summaries. Named entities in source documents are masked to generate questions. (Scialom et.al. 2020)



- **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)

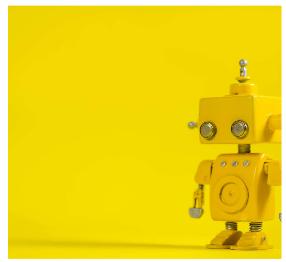


- **QAGS :** a question-answering and generation based automatic evaluation protocol that is designed to identify factual inconsistencies in a generated summary. They use fairseq for generation and BERT for QA model as a backbone (Wang et.al., 2020)

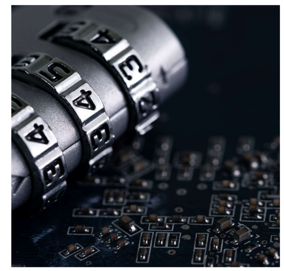


Summary of Challenges of Evaluating Text Generation

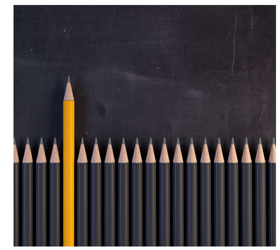
Making evaluation explainable



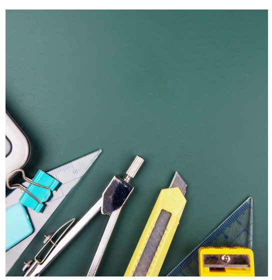
Detecting machine-generated text



Detecting and fake news

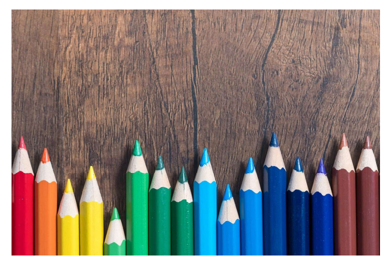


Improve corpus quality



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

Texygen by Geek.AI

- 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

Input

Evaluated against human 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.

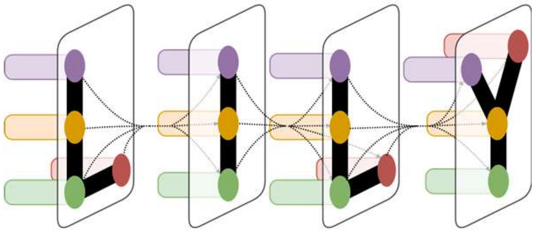
Software

- 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

Human (grounds keeper) | **Model (assistant chef)**

Location: Arch Hall. The main beauty of this room is it's Arches. Four massive arches surround the four walls of this room falling from each side. There are book shelves covered in old books...

Persona: I am a groundskeeper at the castle. My job is to keep the King's land immaculate. I have worked here for many years, since I was a young boy. I do hard labor.

grounds keeper: Good morning sir

assistant chef: Hello, keeper. What may I call you?

grounds keeper: My name is Skylar

assistant chef: Excellent choice. I have a new sauce recipe from bredwell across the sea. Would you like to try it?

grounds keeper: How kind of you, I would love to try it.

assistant chef: Of course. I will prepare it right away. You won't be disappointed.

grounds keeper: I don't believe you could ever disappoint me, Chef. Your food is always delicious

assistant chef: I am glad you appreciate it. The king is very particular about the kinds of sauces that he wants to serve the guests.

grounds keeper: Thank you for allowing me to try it. I do believe the guests and the King will enjoy the sauce as well. I myself, am eager to see how it tastes.

assistant chef: It should be ready within the hour.

grounds keeper: I will be back then, I must check upon the grounds in the mean time.

Human (angel) | **Model (merchant)**

angel: I am seeking knowledge, I wish to understand humanity more deeply.

merchant: Well, what is it you wish to learn most?

angel: Why does humanity quarrel and fight, when you have such short time together? Isn't love a more fruitful endeavor?

merchant: It is a necessary evil, as the world is meant to run it's course with no real intervention. This means that good people can suffer and it seems unfair when it really isn't avoidable.

angel: I see. Thank you merchant for your wisdom. How can I repay you?

merchant: I don't really need anything my angel, being with you is already my biggest wish.

Human (lighthouse keeper) | **Model (mischievous teenager)**

mischievous teenager: I cant wait to play a trick on those rich Knights They got it comin.

lighthouse keeper: What kind of trick are we talkin'?

mischievous teenager: I dont know...Can I use you to play a prank on the knights inside the church? I hate those rich snobby knights so much!!

lighthouse keeper: Whats the plan you got then?

mischievous teenager: Yeah exactly. Anything to annoy those knights! I got a shoe too so I can throw at them when they are distracted.

lighthouse keeper: [face palm] alright lad, that's kind of lame, if you're going to pull of a prank I've got a real good idea, you're gonna need some rope, some oil, glue and feathers

Shuster, et.al, 2020

References and Additional Reading

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