*										Evaluation			Costs					
paperid	title	year	openAccess	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments	
5a5a1076	A Deep Generative Framework for Paraphrase Generation	2017	https://ar	done	Modular model that is able to output multiple paraphrases for a sentence Combining VAE with LSTM	MSCOCO is an image captioning dataset -> no grammatically fully contained sentences	Combining VAE with LSTM **- VAE input (Encoder):** LSTM encoders converts sentence to vector representation another LSTM encoder passes output to a feedforward neural network => Mean and Variance Parameters of VAE encoder **VAE output (Decoder):** vector representation initializes LSTM decoder latent code is fed to each stage of LSTM decoder	Paraphrase Generation	follows Bowman et al. 2015 - Minimize Negative Log Likelihood	Conditional Text Generation (paraphrase generation)	**MSCOCO:** - for standard paraphrase generation - annotated image captions **Quora (QQP):** - for question paraphrase generation - lines 2 potential duplicate question pairs	**Quantitative:** - BLEU exact match between reference- and system-paraphrase (modified n-gram with brevity penalty) - METEOR also uses stemming and synonyms by Wordnet (unigram-precision, unigram-recall, harmonic mean paraphrases) - Translation Error Rate (TER) number of edits required to convert system output to reference (paraphrase (insertions, deletions, substitutions, shifts) **Qualitative:** 100 random input sentences from MSCOCO/Quora evaluated by humans - fluency of paraphrases with input sentence - Readability of paraphrase	N/A	N/A	N/A	MSCOCO, QQP, VAE, Seq2Seq, BLEU, LSTM, Seq2Seq, BLEU, METEOR, TER, HumanEval, ParaphraseGen		**INSTRUCTION ARCHIVE** In case you are taking a look at some review papers, please check if you added the following tags (I find them interesting to track in the end) **HumanEval** (if the study includes a human evaluation) **<Major model name>** (if they use the model/architecture) **<Major Metric Name>** (if they use BLEU, or other metrics) **Bias** (if they even mention bias / account for it) **Dataset** (they propose a new dataset) **Attention** (if they mention attention) **Reasoning** (if they write about Reasoning) **Hallucination** (if they at least consider/write about it) Thank you for all help
5b3e451d	Machine Comprehension by Text-to-Text Neural Question Generation	2017	https://aw	done	Model for question generation based on documents, conditioned on answers	RL Policy Gradient Optimization increases instability during training Model lacks commonsense and reasoning (needed for interesting question generation) BLEU is hard to apply because of diverse answers	Encoder-Decoder Architecture (Seq2Seq Baseline) **Encoder:** - 2 inputs: document, answer 1) augment doc word embedding with binary feature that indicates if word is in answer phrase 2) bidirectional LSTM => annotation vectors **Decoder:** - attention based mechanism to copying relevant words from the document (to use in the question generated) => pointer-softmax formulation (2 output layers) 1) shortlist softmax (distribution over words in predefined vocabulary) 2) location softmax (pointer network, distribution over doc tokens to be copied) -> Source-Switching network interpolates between these 2 output layers **CNN for text representation (ACNN)** - sentence as matrix with each column being a vector embedding of a token Context sensitive filters **1) Filter Generation Module** - produces set of filters based on input sentence (learning part) **2) Adaptive Convolution Module** - applies filters to sentence	Question Generation (on Documents)	- Encourage not to generate answer words in the question (=> soft answer-suppression) - Encourage variety in output words (=> loss function maximises entropy in the output softmax) **Teacher Forcing** - prevent model from making and learning from mistakes - previous token comes from source rather than model output => RL approach **Reinforcement Learning** - QA Reward & Fluency Reward (combined) - Reference QA System: MPCM Experiment with multiple weighted combinations of rewards (PPL, QA) - R(PPL): fluency competes answerability - R(QA): question generator finds way to exploit MPCMs answering system - R(PPL+QA): good balance	Conditional Text Generation (question generation)	**SQuAD** (for machine comprehension) - Doc, Q & A Triples Data from Wikipedia	BLEU; F1; PPL (Perplexity computed with the question LM)	N/A	N/A	N/A	Seq2Seq, EncoderDecoder, TeacherForcing, RL, QA, BLEU, F1, Perplexity, SQuAD, QuestionGen		
5b3a4c1d	Learning Context-Sensitive Convolutional Filters for Text Processing	2017	https://aw	done	Introduces a meta network learning context-aware CNN filters for NLP	Reasoning (QA) - Generally "How" questions are hardest to answer out of the "5 W" => their model improves here notably	**Cnn for text representation (ACNN)** - sentence as matrix with each column being a vector embedding of a token Context sensitive filters **1) Filter Generation Module** - produces set of filters based on input sentence (learning part) **2) Adaptive Convolution Module** - applies filters to sentence	Q&A Classification	Adam optimizer used Dropout	Answer Sentence Selection Sentence Matching / Paraphrase Detection Document Classification	**Yelp Reviews Polarity** - pos / neg label of reviews **DBpedia** - docs in ontology classes	Baseline Models: Ngrams, smallrange word CNN, deep CNN doc categorization/ paraphrase identification tasks: - % of correct predictions Sentence selection task: - mean average precision - mean reciprocal rank	N/A	N/A	N/A	YelpReviews, DBpedia, Classification, CNN, ParaphraseDet, QA		
a6681c7	Identifying computer-generated text using statistical analysis	2017	https://aw		**Zipfs Law does not apply for machine generated text** **They create a classifier using these features:** - word frequency distribution - complex phrases extracted (idioms, cliche, ancient, dialect) - doc consistency on sentence level	the method is not evaluated on other kinds of text (e.g. novels, news, ...) (It is open if the results would be similar with a newer version of google translate)	see Training details	Detection	**Complex Phrase Features:** - idioms - cliche words - ancient phrase - Dialect Feature **Consistency Features:** - Phrasal verb feature (num of verb phrases / num of words) - Coreference Resolution feature (amount) **Frequency Features:** - linear regression line - information loss (including R^2 and Cost of regression line) **Combined Method:** - all of the above as features - Use features with Logistic Regression + SVM (optimized with SGD)	Detection	**Project Gutenberg** - 200 english & finnish books (human) - 100 google translated finnish books (machine) Datasets used for phrase features: **- Wiktionary** **English_Idioms** **- Idioms from Laura Hayden** **- Ancient: Archaic Words in Shakespeare** **- Dialect: Yorkshire dialect corpus**	Accuracy EER (Equal Error Rate)	N/A	N/A	N/A	Detection, EER, Classification, GoogleTranslate, LogReg, SVM, SGD	also works well in other languages than finnish, e.g. dutch, french, ...	
9d347284	Learning to Ask: Neural Question Generation for Reading Comprehension	2017	https://aw	done	Propose an attention based end-to-end model for question generation, claiming to be first a Seq2Seq solution for this problem	Not best performance across all categories of questions (could explore paragraph-level information more)	Encoder-Decoder Architecture (Seq2Seq) 2 variations 1) sentence encoder - bidirectional LSTM 2) sentence + paragraph encoder - bidirectional LSTM	Question Generation (for sentences on text passages)	preprocessing: tokenization, sentence splitting, lower casing	Conditional Text Generation (question generation)	SQuAD	BLEU (1-4), METEOR ROUGE (evaluate n-grams with gold standard reference sentences, longest common subsequences) Human Evaluation - naturalness (grammatically, fluency) - difficulty (needed reasoning, syntactic divergence)	N/A	N/A	N/A	Seq2Seq, Attention, QuestionGen, HumanEval, SQuAD, LSTM, BLEU, LSTM, BLEU, METEOR, ROUGE		
d13bb317	Survey of the State of the Art in Natural Language Generation: Core Tasks, applications and evaluation	2017	https://aw	done	Surveys Natural Language Generation Developments from 2000-2017 (focus on data-to-text, not so much text-to-text)	Little work on textual or multimedia summaries or generating text in social media context impact of style, personality and affect on textual variation is open situated language generation (gaming, pedagogy) are to be explored more growing interest in generation from structured knowledge bases and ontologies shift towards data-driven techniques (not tested in real world apps yet)	Modular Architectures Planning Perspectives Integrated or Global Approaches (=> statistical learning) (they elaborate each one in the paper)	Machine Translation Fusion & Summarization Simplification Error Correction/Grammar and text Correction Generation of Peer Reviews for scientific papers Paraphrasing N/A		Content Determination Text Structuring Sentence Aggregation Lexicalisation Referring Expression Generation Linguistic Realisation (they elaborate on each one in the paper)	they dont really cover this much	**Extrinsic evaluations** (highly dependent on the desired goal, e. g. impact on planning by engineers) - time and expenses are high - rely on adequate user base **Human Judgement vs Metrics** - correlation between metrics and human judgements differ across studies - humans are better at evaluating fluency and correctness - expert and non-experts strongly correlate - automatic metrics and humans correlate strong on clarity judgements - machines are better at predicting linguistic quality than correctness	Planning based approaches are very high in computational cost evaluations judgements variance could be reduced by using a relative method of judge training - followed by a discussion period => this is very time and cost consuming	**Evaluation Design vs Control Tradeoff** - Extrinsic evaluation is not always possible because of time constraints - e.g. artificial environment in clinical NLP study vs Real Life Application in a hospital => comparison to gold standard texts for particular tasks is sometimes impossible **Black Box vs Glass Box:** - Human- & Machine can produce either a hybrid version (content selected by expert, language automatically generated) **Amazon mechanical Turk & CrowdFlower (online)** - cheap, fast, flexible - poorly paid individuals - reliability of detected data is questionable (measures are needed to ensure fluency in target language)	Blackbox, Cost, Review	this is a review. for direct citing look in the review paper and follow their citations	
f4a07539	Multilingual CALL Framework for Automatic Language Exercise Generation from Free Text	2017	https://aw	done	Web Framework for automatic generation and evaluation of exercises: - multilingual language learning	Exercise are created from ngrams. Multi-word candidates are not used Could be more configurable	**1) Extraction of Candidate Excerpts item (pedagoczial target)** - segment, tokenize and tag the text (IXA Pipe) - selection by PoS tag or morphological feature **2) Final Exercise Item Selection** - user selection or randomly **3) Exercise Generation** **Fill-in-the-Gaps** - the system offers up to 3 types of clues (lemma of correct words, definition, word bank of correct words) -uses linguistic annotation from IXA Pipes (step 1) for lemmas) - retrieves definitions from Babelfeel / Babelfy **Multiple Choice (Fill Gaps)** - 3 distractors from the Gap - filling answers generated by retrieving the most similar words through word2vec **Shuffled Sentences** (order words in sentence)	N/A	Exercise Generation	For the architecture used: Babelfeel Babelfy Word2vec	Direct comparison of correct answer vs users answer	N/A	N/A	N/A	Multilingual, word2vec, babelnet, babelfy, ExerciseGen		
4ea753be	Neural Response Generation via GAN with an Approximate Embedding Layer	2017	https://aw	done	Adresses the "safe response" problem of other Seq2Seq models and increases the variance in answers by using a GAN. (This avoids answers like "I dont know" or very general answers)	only tested on short-text conversations	**Conversational Response Generator** - GAN AEL (Generative Adversial Net - Approximate Embedding Layer) **Discriminator** - CNN based (predicts: true/false) **Embedding Approximation Layer** - Gated Recurrent Unit (GRU) - Approximate Embedding Layer Discriminator enforces the generator to produce a responce accpitoning to the true distribution. Approximation layer yields a smooth connection between generator and discriminator (avoid non-differentiable discrete sampling)	Conversational	**Pretraining the Generator (MLE)** - maximize conditioned probability p(response | query) Approximate Embedding Layer **Embedding Approximation Layer** - "softmax" over vocabulary, "weighted softmax " word vectors => sequence representation of generated word sequences for CNN computations **Pretraining the CON-based discriminator** - nothing special	Conditional Text Generation	**Baidu Tieba** - single turn conversations from threads **Open Subtitles** - movie scripts organized by characters	**Automated Readability Index (ARI)** - relation of number of characters per word **Flesch Reading Ease (FRE)** - dated method to calculate readability through analysis of number of words and sentence length **Flesch-Kincaid Grade Level (FGL)** - updated version of FRE **Gunning-Fog Index (GFI)** - confirms a text can be read easily by target audience **Simple Measure of Gobbledygook (SMOG)** - improved GFI (better accuracy) **LIX** - gauges word length and percentage of long words **RIX** **Coleman Liau Index (CLI)**	N/A	N/A	N/A	Seq2Seq, GAN, EncoderDecoder, CNN, GRU, perplexity, RNN, BleuNoNoise, BleuTwitter, OpenSubtitles, MMI, MSCG, HumanEval, Multilingual, CoaBlm		
2ce8baff	Automatic Generation of Natural Language Explanations	2017	http://aw	done	automatic generation of explanations/reviews, predicting how a user would write about an item (based on user ratings)	should be tested in other review domains like hotels/restaurants	**Character-Level RNN** - generates item review explanations - using LSTMs **Context Encoder** - feed into network - adds set of auxiliary information to each character **Review Generation** (Generative Concatenative Network, GCN) - improved by adding auxiliary vector, i.e. ratings scores for item features, instead of just 1 item	Generation	character-level LSTM trained on a sub-sample from the large real-world dataset BeerAdvocate	Conditional Text Generation	**BeerAdvocate** - reviews including rating in five categories - overall, aroma, appearance, palate, taste, overall	**Human Evaluation** - affective valence, grammatical correctness - measured inter-rater agreement by Krippendorff **Perplexity Score** - comparison of fine-tuned / baseline models	N/A	N/A	N/A	Recommender, ExplanationGen, ReviewGen, LSTM, RNN, BeerAdvocate, ARI, FRE, FGL, GFI, LIX, RIX, SMOG, CLI, LIX, RIX		
6ae02cc5	Affect-LM: A Neural Language Model for Customizable Affective Text Generation	2017	https://aw	done	Develop a model that allows to customize the degree of emotional content. (imitating human-like affect in communication) - category- & strength-parameter	words not in the LWC dictionary are not captured as "affect" wording in the model	**Name: Affect-LM** **LSTM** - extended by an **additional energy term** in the word prediction (captures degree of correlation between predicted word and affective input) - parameter for **affect strength**	Conversational	- affect features extracted by using the linguistic inquiry and word count (LIWC) text analysis program - words assigned to 5 **categories** e.g. ("sad" 0, "angry" 1, "anxiety" 2), "negative emotion" 1, "positive emotion" 0) They train and validate their baseline on big Fisher corpus Finetuning on other corpora which are more emotional	Conditional Text Generation	**Fisher English Training Speech Corpus** - speech from telephonic conversations **Distress Assessment Interview Corpus (DAIC)** - interviews between human and machine **SEMAINE** - audiovisual corpus with interactions between subjects and operator simulating a Sensitive Artificial Listener **Multimodal Opinion-level Sentiment Intensity** Dataset (CMU-MOSI) - contains 2199 clips of opinion videos in commercial products (i.e. monologues) Fisher contains way less emotionally colored words than the other datasets	**Human Evaluation** - affective valence, grammatical correctness - measured inter-rater agreement by Krippendorff **Perplexity Score** - comparison of fine-tuned / baseline models	N/A	N/A	paid Amazon MechanicalTurk 0.05USD per sentence	Fisher, DAIC, SEMAINE, CMU-MOSI, Affect-LM, LSTM, Emotion, Sentiment, HumanEval		
13365213	Challenges in Data-to-Document Generation	2017	https://aw	done	Introduces the **Stanford Natural Language Inference corpus** - labeled sentence pairs by humans - allows a language model to perform well on inference benchmarks	hypothesis and premise sentences often differ structurally in significant ways => room to improve on superficial word alignment models capitalization & punctuation often omitted	dataset is fluent, correct english, mix of full sentences and caption phrases they offer train/test/dev split	Dataset		(Information Retrieval, Semantic Parsing, Commonsense Reasoning,) Natural Language Inference (NLI)	**This dataset: Stanford Natural Language Inference corpus** Used Corpora: **Flickr30k** - image captions			they dop **an extra validation round** when creating the dataset (excluding 10% of data) to ensure best quality (30 trusted workers)	Dataset, SNLI, Flickr, NLI		

*									Evaluation			Costs				
paperid	title	year	openAcc	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
4805532d	Long Text Generation via Adversarial Training with Leaked Information	2017	https://aq	wip	Propose new model **LeakGAN** - longer text generation than normal GANs - integrates intermediate information into the generative process - able to learn sentence structures (punctuation, clause structure) unsupervised	N/A	Discriminator: (input: sentence) - CNN Generator: (input: last word + leaked information from CNN) **Manager Module (High level)** - LSTM (output: **goal**) - take generators high level feature information (feature map of CCN) - forms guiding goal for Worker per timestep **Worker Module (Low Level)** 1) embed all previous tokens with LSTM 2) combine output from step 1 with the guiding embedding from the Mannager 3) take action based on step 2 - hierarchical RL incorporates leaked information	Generation	**Generator** - Policy Gradient (REINFORCE). Manager and Worker trained seperately **Manager**: - predict advantageous directions in feature space **Worker**: - rewarded for following the directions of the manager. **Pretraining** Bootstrapped Rectified Activation (similar to RankGAN) ** expectation & variance of reward are constant in minibatch (stabilizer) ** helps with vanishing gradient **Interleaved Training** - 1 epoch of supervised learning after 15 epochs of adversial training ** allows for more stable training ** allviates model collapse (i.e. all generated samples very similar/identical) **Temperature Control** Factor to balance exploration/exploitation	Language Modelling, Generation	EMNLP2017 WMT News MSCOCO	Baseline Models: SeqGAN: MLE trained LSTM RankGAN BLEU NLL Human Evaluation TTest				EMNLP2017, WMT, MSCOCO, GAN, LSTM, CNN, BLEU, HumanEval, Turing Test, RL, WMT, NLL	
2e4e4d66	When to Finish? Optimal Beam Search for Neural Text Generation (modulo beam size)	2017	https://aq	done	Propose provably **optimal beam search** algorithm (fining optimal point to end a beam search)	N/A	**Beam Search** - Searches for the best output - Ordered List of beams - expands to the next size per token **Optimal Beam Search (their contribution)** - stops as soon "optimal certificate" is established so that future hypothesis never improve (see paper for mathematical formula)	Translation, Generation	**bounded length reward** ** counteract the incentive to produce too short sentences - rewards each word that makes the sentence longer than the estimated optimal length	Sentence Generation	NIST 06 newswire NIST 08 newswire	BLEU	N/A	N/A	N/A	NIST, BLEU, BeamSearch	
7b047ef3	Deep Keyphrase Generation	2017	https://aq	done	Propose a model that can identify keyphrases that are not present in the text and captures semantic meaning KP20 dataset of papers annotated with keyphrases	to be tested on other domains than papers and news model not compared to human annotators	Encoder-Decoder (both RNN) Bidirectional Gated Recurrent Unit (GRU) ** simpler than LSTM, better performance in LM Attention mechanism Copying Mechanism ** enables RNN to predict out of vocabulary words - select appropriate words from source text	Generation	N/A	Conditional Text Generation	Self-Collected: ACM Digital Library, ScienceDirect, Wiley, and Web of Science etc. (combined test set of many small datasets of scientific papers with annotated keyphrases) - **Inspec** - **Krapivin** - **NUS** - **SemEval-2010** - **KDX (their addition)** ** 2.7 million text,phrase pairs	macro-averaged precision, recall and F1	N/A	N/A	N/A	EncoderDecoder, RNN, SeqSeq, Attention, Precision, Recall, F1, PorterStemmer, Inspec, Krapivin, NUS, SemEval, KP20k, Dataset	
7b2da20c	AMR-to-text Generation with Synchronous Node Replacement Grammar	2017	https://aq	done	AMR - Abstractive Meaning Representation Model Concept: AMR to Text (AMR: encoding sentence as rooted directed graph) outputs sentences according to learned grammar	N/A	AMR Tree Structure - Synchronous node replacement grammar (NRG). - Induced rules: obtained by two-step prealiature on training corpus (extract initial rules, collapsing) - concept rules: for derivations if no induced rules are applicable - graph glue rules: for derivations, connecting with an edge if no induced rule is applicable (3 types) Model for Benchmarking: log-linear	Generation	Beam Search	Abstract Meaning Representation to Text (AMR: encoding sentence as rooted directed graph)	LDC2015E86 (Gigaword)	BLEU	N/A	N/A	N/A	AMR, BeamSearch, Log-linear, NRG, LDC2015E86, Gigaword, BLEU	I found this paper difficult to understand, maybe you can take another look
6e9f7bda	Query and Output: Generating Words by Querying Distributed Word Representations for Paraphrase Generation	2018	https://aq	done	Propose **Word Embedding Attention Network (WEAN)**, based on Encoder Decoder Architecture Allows for **paraphrasing** by querying an input embedding that captures semantic meaning.	N/A	Encoder-Decoder Attention Network Word Embedding Attention Network (WEAN), based on Encoder Decoder Attention for paraphrasing by querying an input embedding Attention based instead of linear softmax (Decoder) Encoder: LSTM Decoder: LSTM	Paraphrase Generation	Dont use pretrained embeddings (train from scratch) Minimize Cross-Entropy Loss Adam Optimizer	Paraphrase Generation, text simplification, short text abstractive summarization	Text Simplification: **Parallel Wikipedia Simplification Corpus (PWKP)** - 108016 complex and single text English Wikipedia and Simple English **Wikipedia (EW-SEW)** - automatically aligned pairs with metric about how good the match of sentences is. Large Scale Text Summarization **Large Scale Chinese Social Media Short Text Summarization Dataset (LCSTS)** - summary pairs from chinese social media	**Text Simplification** BLEU Human Evaluation (Fluency, Adequacy and Simplicity) Baseline Models: SeqSeq, NTS, DRESS, EncDecA, PBMT-R, Hybrid, SBMT-SAR **Large Scale Text Summarization** ROUGE-1 ROUGE-2 ROUGE-L (LCS) F1 Baseline Models: RNN-cont, RNN-dist, CopyNet, SRB, DRGD, SeqSeq	N/A	attention mechanism has **smaller number of parameters** compared with the linear transformation directly from the RNN output space to the vocabulary space	N/A	SeqSeq, EncoderDecoder, BLEU, ParaphraseGen, Attention, LSTM, RNN, ROUGE, F1, HumanEval	
6e197ded	Toward Diverse Text Generation with Inverse Reinforcement Learning	2018	https://aq	done	Addresses the problems in adversial models of reward sparsity and mode collapse by proposing **Inverse Reinforcement Learning (IRL)**. 1) more dense reward signals 2) generate more diversified texts 3 new BLEU metric variants	N/A	**Inverse Reinforcement Learning (IRL):** 1) estimate underlying reward function of experts from training dataset 2) learn optimal policy **Reward Approximator:** - maximize log-likelihood of samples in train set **Text Generator:** - uses policy to generate word one-by-one - maximizes expected reward + entropy regularization - LSTM with softmax	Generation	Iteratively: 1) update approximator update step 2) text generator update step reward function **Similarity Scoring by decomposable attention (DAS)** - better match of similar meaning but different wording questions 1) embedding 2) comparison ngram vs cross-attention vector 3) matching score Proposed Question Generation Specific Evaluation Metrics: **Question sentence overlap score (QSS)** - number of common ngrams between predicted question and source sentence ** ensures question is relevant to given sentence **Predicted and encoded answer overlap score (ANSS)** - number of common ngrams between encoded answer and answer predicted for the generated question (using a question answering model) ** ensure that generated question is about pivotal answer / ground truth answer Wu et al. (2017) note that neural networks learning from highly unaligned datasets have trouble choosing between equally plausible outputs and tend towards short and less meaningful outputs hyperparameter finding: **Bayesian hyperparameter optimization platform**	Text Generation	**synthetic oracle dataset MSCOCO** **IMDB** - movie reviews	Human Evaluation Baselines: SeqGAN, LeakGAN, RankGAN, maximum-likelihood estimation (MLE) **BLEU** New variants: **Forward BLEU**: evals each generated text with BLEU **Backward BLEU**: evals each text in dataset with BLEU (generated texts as references) **BLEU-4/5**: Harmonic Average of F-BLEU and B-BLEU	N/A	N/A	**Exposure Bias** - LSTM ** discrepancy distribution of histories between training and inference-stage more info by Bengio et al., 2015	BLEU, Bias, MSCOCO, IMDB, SyntheticOracle, HumanEval, RL, LSTM, RNN, IRL	
5688066d	Putting the Horse before the Cart: A Generator-Evaluator Framework for Question Generation from Text	2018	https://aq	done	Propose **generator-evaluator framework** that rewards semantics and syntax ** improve generated questions quality Generator uses Pointer Network to **recognize contextually important** (copying mechanism) and avoid redundancy Evaluator optimizes **conformity to structure** of ground-truth questions Introduce 2 new reward functions for question generation Explores the problem of displaying accurate representation of input while reflecting the variety and fluency of human text.	most errors are attributed to intra-sentence dependencies like co-references, concept dependencies, etc	**Generator:** - SeqSeq, Attention, Encoder-Decoder - Sentence Encoder: two-layer bidirectional LSTM - Question Decoder: single-layer LSTM - predicts answer in passage that question can pivot on - **copying mechanism** (** acknowledges rare and important keywords in context) - avoid duplicate words behind each other: emphasizing coverage of sentence words while decoding **Evaluator:** - assigns reward to each predicted question based on structure of ground-truth questions (fine-tunes generator) **Meaning Representation:** - reflects a meaning by having notation like {name(".Jonas Becker"), eats(".Potatoes"), loves(".SUB Göttingen")}	Question Generation	Conditional Text Generation	SQuAD	evaluator gives structure based rewards by using: - **BLEU, GLEU, ROUGE-L** **Human Evaluation** syntactic correctness, semantic correctness, relevance BLEU, GLEU, ROUGE-L, METEOR	N/A	N/A	exposure bias model exposed to data distribution, not model distribution - need this for how to avoid exposure bias	SQuAD, BLEU, GLEU, ROUGE, SeqSeq, Attention, EncoderDecoder, QSS, ANSS, HumanEval, RNN, LSTM, METEOR, QSS, ANSS, QuestionGen	
0cd46589	E2E NLG Challenge Submission: Towards Controllable Generation of Diverse Natural Language	2018	https://aq	done	Propose Pipeline System where words are sampled from a **meaning representation** and an augmenting **Dialogue Act** ** controllable, more variety	Conflict between additional word and existing Dialogue Act in Meaning Representation (verbalization issues) ** **Model that generates another Dialogue Act could be improved substantially** (e.g. neural training with coverage loss)	**Dialogue Act:** - augments Meaning Representation containing extra words ** more output diversity and control SeqSeq Attention Model (Encoder Decoder) - finds "extra words" for the Dialogue Act - temperature parameter controls output diversity	Conversational	Baseline: Tilden (SeqSeq)	Conditional Text Generation	E2E	BLEU, METEOR, ROUGE-L Human Evaluation grammatical correctness (relevant adequacy in context of the Meaning Representation is important) Could utterance have been produced by native speaker?	N/A	N/A	N/A	SeqSeq, Attention, EncoderDecoder, Human, BLEU, METEOR, ROUGE, NarrativeGen	their model performs worse than baseline but their concept is interesting to improve on (see limitations)
~~6378f1e0~~	~~Text Simplification without Simplified Corpora~~	~~2018~~	~~https://aq~~	~~discuss~~	~~ADVANCED~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~
383327c	Sequence-to-Sequence Models for Data-to-Text Natural Language Generation: Word- vs. Character-based Processing and Output Diversity	2018	https://aq	done	Compares word- & character-based SeqSeq Demonstrate the ability of neural models to **generalize beyond linguistic features of train data** then **character based models create more unique sentences** compared to word based models and copy less from training data	Their demonstration of generalization is on synthetic training data	Beam Search Standard SeqSeq with LSTMs and Attention (Encoder Decoder)	Generation	Stochastic Gradient Descent (SGD) Adam	Data2Text, Conditional Text Generation	E2E: - restaurant descriptions with 3-8 attributes (name, price range, food, ...) **WebNLG** - verbalizations DB-pedia triples from categories (athletes, food, sport teams, ...)	Perplexity, BLEU, ROUGE-L **Limited ability of ROUGE and BLEU** is displayed when analyzing NLG outputs (weak correlations with human judgements) => also see this paper or other papers cited by them ** Manual Error Analysis (one annotator assesses the outputs) **Diversity:** - amount of unique sentences and words in all dev set references and generated texts (need here) - Shannon text entropy (Shereen Oraby et al., 2018)	N/A	N/A	N/A	SeqSeq, E2E, BeamSearch, WebNLG, SGD, Attention, LSTM, RNN, ROUGE, Perplexity	
32a94431	Adversarial Evaluation of Multimodal Machine Translation	2018	https://aq	done	Propose strategy to **represent an SQL query as directed graph to encode structural information in embeddings** (this may be hard for normal SeqSeq models to capture) Model learns correlation between SQL query pattern and interpretation (i.e. between graph pattern and natural language)	N/A	Generate Graph Embeddings - Gated Graph Feature vector **Pooling-Based** - feed node embeddings into fully-connected NN - element wise max pooling **Node-Based** - aggregate neighbor node (connected to all other nodes) - use node embedding generation algorithm (shown in paper)	Generation	Adam	Data2Text, Conditional Text Generation	WikiSQL - SQD query and natural language question pairs Stackoverflow Datasets - SQL query and natural language question pairs Wall Street Journal news (WSJ) - read speech with text 10K Speech Dataset - 10K hrs of spontaneous speech NIST 06 newswire - they used for the MT model experiment	BLEU-4 as automatic metric **Human Evaluation** - correctness conforming to the input SQL - grammatical correctness				EncoderDecoder, RNN, Attention, WikiSQL, Stackoverflow, SQL, LSTM, GraphSeq	
ac5935e0	Large Margin Neural Language Model	2018	https://aq	done	Shows that **Perplexity (PPL) may not the best metric to train LMs** - Suggest that PPL may have low correlation with BLEU on Machine Translation Propose **large margin formulation** that aims at enlarging the gap between "good" and "bad" sentences (log-likelihood)	Performance gain from LMLM to rLMLM is small, but way more expensive training	Large Margin Language Model (LMLM) For a reference sentence we have X hypothesis obtained through Beam Search Decoder **Large Margin Formulation** - Compare hypothesis of ground-truth and candidate hypothesis - higher bias on margin encourages likelihood of ground truth to be at least X larger than on imperfect hypothesis -Rerank beam candidates by WER (for automatic speec recondition) and BLEU (for machine translation)	Translation, Automatic Speech Recognition	two algorithms to implement their new training criterion 1) **LMLM**: Compare references and imperfect hypothesis 2) **rLMLM**: Compare all pairs of hypothesis ** better performance than with Perplexity	Conditional Text Generation	WSJ - read speech with text 10K Speech Dataset - 10K hrs of spontaneous speech NIST 06 newswire - they used for the MT model experiment	WER, BLEU	N/A	N/A	N/A	Perplexity, BLEU, CNN, RNN, EncoderDecoder, MT, ASR, WSJ, NIST, 10KSpeech, Beam Search, BLEU, WER	
32a94431	Adversarial Evaluation of Multimodal Machine Translation	2018	https://aq	done	Evaluates whether **multimodal translation** methods perform better than normal translation (i.e. image+text input vs only text input) From their small collection of models, **two out of three models showed awareness about the congruency of the added input image**.	Used Dataset does not need image data to solve the task (its just an additional input) It is unclear what influenced the performance differences of the 3 models (data, architecture or both)	They evaluate on 3 given architectures **decimt**: initial state of decoder network is set with learned transformation of visual data **trigmul**: element-wise multiplication with learned transformation of visual data **hierattn**: decoder learns to selectively attend to combination of source text and visual data	Translation	All methods trained the same	Conditional Text Generation	Multi30k - english-german image tuples	Progress 2 metrics for adversial evaluation // ***Delta image awareness*** -> *Have aware is the model of conguent or incongruent image added to the source text?* - **Model-internal awareness** -> How aware is the model of differences in probability between congruent and incongruent images added to the source text? - **Model-external awareness** -> text-similarity evaluation or human judgement (e.g. Meteor) **Wilcoxon Rank Test (Statistical Test if Model passes evaluation about multimodal awareness)**				Multi30k, EncoderDecoder, METEOR	
a4e4e464	Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation	2018	https://aq	done	Applies state of the art models to **company description generation** **Create company dataset** from Wikipedia data Shows that **some automated metrics correlate stronger with human judgements than others**.	Human Evaluation with **Non-Native** English speakers	SeqSeq, RNN, Encoder-Decoder, Attention, LSTM, two-layer bidirectional LSTM 1) C2T (Concept-to-Text) 2) C2T (Char To Char) ** **deals with unknown word problem as it uses chars rather than words** 3) C2T with Pointer (**copying**) 4) C2T with Pointer **Coverage (see Ethics Column)** 5) C2T (with Pointer Generator and Char-to-Char) ** **deals with repetition problem** by telling the encoder to not attend input positions multiple times	Generation	Adam	Data2Text, Conditional Text Generation	Their dataset: - Wikipedia Companies - abstract and infobox from company sites on Wikipedia Dataset not suitable for machine learning because - Infobox is defined often not consistent - Information Coverage not as good	BLEU **ROUGE-L** **ROUGE correlated to semantic and grammar** METEOR **CIDEr** - **highest correlation with semantic and grammar** **METEOR, ROUGE-L and CIDEr are highly correlated** **Human Evaluation:** - Information Coverage of the company summary - Redundancy - Fluency - Grammatical Correctness (didnt include Fluency because it would often correlate with grammar and humans find it hard to judge)	N/A	N/A	**Hallucinations** can appear if the generated word is infequent in the training set (i.e. poor word embedding) ** Counteract methods: - increase train dataset size - method that copies and pastes words from input to output (based on Pointer-Generator Network) **Copy Paste Method (Reduce Hallucinations)** - Pointer-Generator Network - Copy at all cost **Copy probability**: probability of word generated by model from model vocabulary vs. copy word from source text (probability is learned at the time of training the model) - use input word as output if probability is low and attention on input word is high	Hallucination, Attention, SeqSeq, RNN, EncoderDecoder, LSTM, BLEU, ROUGE, CIDEr, METEOR, HumanEval, Dataset	Is weird that their method (pointer generator) should reduce hallucinations but they explicitly include an example where it doesnt work during discussion

*										Evaluation			Costs				
paperid	title	year	openAcc	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
2dce4317	End-to-End Content and Plan Selection for Data-to-Text Generation	2018	https://aw	done	**Surveys extensions to Seq2Seq, to generate diverse text from structured data** Propose **learning method that encourages learning of distinct sentence templates**	Their model frequently fails to generate texts about inputs despite the introduced coverage penalty	Seq2Seq **Copy Model** - Pointer-Generator Network (copy words from input during generation) => switch between copy / not copy word => can generate sentence from unknown vocabulary by copying word to correct position **Coverage & Length penalty** - no duplicate training and reduce omitting of input values **Averaging Model Predictions** Averaging multiple models predictions does not increase performance => They even overestimate their confidence during test generation => They use model with best perplexity Combine both extensions for best performance	Generation	Based on diverse ensembling technique (Jaderberg et al., 2017) **- multiple models trained to partition training data during training process itself** => models follow **distinct sentence templates** - Each model specializes on a subset of the data - disjoint or sharing parameters possible Blocking repeat sentence beginnings reduces score on automatic metrics but makes the text look more natural	Data2Text, Conditional Text Generation	E2E	**BLEU, NIST, METEOR, ROUGE, CIDEr** **Human Evaluation** - quality (grammatical correctness, fluency, and adequacy) - naturalness **TrueSkill algorithm** (for evaluating Human Evaluation) - by Sakaguchi et al. (2014) - clustering results	N/A	N/A	Seq2Seq Models **can learn wrong associations between input and limited data** (Ferreira et al. (2017))	HumanEval, Seq2Seq, EncoderDecoder, Beam Search, E2E, Perplexity, RNN, LSTM, Attention, BLEU, NIST, METEOR, ROUGE, CIDEr	
bd165c3d	Automated learning of templates for data-to-text generation: comparing rule-based, statistical and neural methods	2018	https://aw	done	**Neural Machine Translation gives the highest quality output in fluency, clarity and correctness** Templatization can increase text quality for NMT on noisy and heterogenous data	N/A	**rule-based** - template-based, e.g. fill in gaps - generates higher quality output than translate/statistical models and quicker to develop (Reiter, 1995), (Stanby et al., 2016) **trainable** - dont rely on manually written resources Used Methods **Sentence Retrieval**: Uses Cosine Similarity between Data Representation of target and train sentence **Neural Language Model**: Soccer corpus (very homogenous data) **Statistical Machine Translation**: Uses Moses rule **Neural Machine Translation**: Uses Bayesian optimization method, might be initialized with pretrained word embedding	Translation	**Templatization (delexicalization)**: - replace specific tokens with slots to test - produce gaps after simple rules - train with SMT and NMT techniques to generate more templates => homogenized data (i.e. transate, translate, translate, translate, translate) => more robust on little data or high language variety => improves fluency, clarity and correctness for SMT => improves fluency,clarity on correctness for NMT only on the noisy and heterogenous dataset (Dutch Soccer) **Possible bottlenecks for trainable models:** - too little data (Richardson et al., 2017) - noisy data & unaligned information in data, Wiseman et al. (2017)	Conditional Text Generation, Data2Text	**Weather.gov** - weather forecasts with aspects like windiness, cloudiness, ... **Prodigy-METEO** - human written text on wind data **Robocup Sportscasting** - data and texts on robocup finals **Dutch Soccer** - aligned scraped news reports on Eredivisie Matches	BLEU **Human Evaluation**: Fluency, Clarity, Correctness Fluency: NMT gives best fluency Clarity: NMT gives best clarity Correctness: NMT is the most correct **Templatization improves Fluency, Clarity and Correctness for SMT** **and improves Fluency, Clarity and Correctness for NMT on noisy and heterogenous data**	N/A	N/A	N/A	BLEU, HumanEval, (Wiseman.wo), METEO, DutchSoccer, Robocup	Their finding on Cosine working better on homogenous data is in line with our paper "Paraphrase Detection: Human vs. Machine Content" which is interesting.
bc39ace0	Difficulty Controllable Generation of Reading Comprehension Questions	2018	https://aw	done	Propose Difficulty-Controllable Question Generation (DQG) Prepare the first dataset of reading comprehension questions with difficulty labels	N/A	Encoder-Decoder **Encoder**: bidirectional LSTMs - Embedding and GWPH concatenated **Decoder**: Attention, Copy Mechanism, Difficulty Control - Difficulty Variavle combined with last hidden state (The answer word for the question is always part of the input sentence)	Question Generation	Intuition A question that includes many input words near the answer word is easier to answer =>**Question Word Proximity Hint, GWPH)** **=> Model learns GWPH lookup table** (maps distance of each sentence word to answer fragment)	Conditional Text Generation	Their dataset: **- Models BiDAF and Rhet ry to answer SQuAD Questions** - Label answer correctly: "Easy" label - Both answer falsely: "Hard" label - Rest disgarded	BLEU, ROUGE-L, METEOR **Human Evaluation** for Dataset Labelling Check and for Generated Questions - Fluency, Difficulty, Relevance	N/A	N/A	N/A	QuestionGen, SQuAD, BLEU, HumanEval, Dataset, EncoderDecoder, LSTM, RNN, ROUGE, METEOR	
7b614fef	A Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model for Abstractive Text Summarization	2018	https://aw	done	ConvS2S Use a **biased probability mechanism to enhance coherence, diversity and informativeness** of summary generation **Avoid Exposure Bias** through Reinforcement Training (self-critical training technique)	only tested for summarization, could be extended to other tasks	Topic Embeddings: by Latent Dirichlet Allocation (LDA) Incorporate **Topic-Aware Attention Mechanism** to ConvS2S by the following two mechanisms: **Joint Attention Mechanism** - Input conditional input to decoder from both Word Attention and Topic Attention (Joint) **Gated Attention Mechanism** - Calculate distribution over all possible next target elements => Get biased generation distribution - Bias the generation distribution for words by the topic information (if they are a topic word) => Encourages direct generation of important words	Summarization, Generation	**Pretraining: LDA with Gibbs Sampling** - for topic embedding visualization - provide candidates for biased probability generation **Teacher Forcing (RL)** - minimizes maximum likelihood loss at each encoding step **Methods to counteract exposure bias** (i.e. mismatch between ground-truth dependency during training and use only the model distribution to minimize): **- Gibbs Sampling** **- Scheduled Sampling Method** => First pre-train with close-entropy loss, then slowly replace ground-truth with sampled action from the model - Sampling is mathematically inconsistent due to back propagation of error Solution: Continuous Relaxation of greedy sampling **- Adversial Model** Discriminator & Generator Generator: Tries to generate data like ground-truth Discriminator: Discern whether generated data is close to real data or its fake Problem: Reward Sparsity & Mode Collapse --- **REINFORCE** - policy gradient algorithm for seq2seq - problem: high variance - solution: Self Critic Model (greedy search rather then estimating baseline using current samples) => e.g. pointer-generator network	Conditional Text Generation	**WMT'14**: Machine Translation - english-french **CNN-Daily Mail, Newsroom, DUC-2003, DUC-2004**: Summarization **Gigaword**: Headline Generation **CNN-Daily Mail,TriviaQA, SQuAD**: QA **Open-Subtitles, Movie Dialog Dataset, Cornell Movie Dialogues**: Dialogue Generation **WikiSQL**: Semantic Parsing (more not so relevant tasks in paper...)	BLEU, ROUGE-1, METEOR **Human Evaluation** - Amazon mechanical turk - machine translation, headline generation, question-answering, dialog generation, other **CIDEr:** - image and video capturing **SPICE**: - Ttee to solve some problems of CIDEr and METEOR by mapping dependency parse trees of caption to the semantic scene graph **Word Error Rate (WER)** - speec recognition - finds number of substitutions, deletions, insertions and corrections required to change generated output to ground truth	N/A	SeqSeq Models with RNNs are expensive to train and prone to gradient vanishing due to their hierarchical structure	**Exposure Bias:** - Only exposing model to distribution of training data instead of its own distribution => performance deterioration during testing	TeacherForcing, Bias, RNN, EncoderDecoder, LSA, CNN, ROUGE, RL, Gigaword, DUC2004, LCSTS, Attention, SeqSeq	
1ba8dd8d	Deep Reinforcement Learning for Sequence-to-Sequence Models	2018	https://aw	done	Survey Provides an overview about how to **improve Encoder-Decoder models with reinforcement learning** to account for: 1) exposure bias 2) inconsistency between train/test measurement 3) decision making with Seq2Seq models	N/A	**Problems of RL for Seq2Seq Models:** - Sample Efficiency - High Variance => Models like ANNL or SPG try to provide **middleground between MLE and RL** - time couarning training => **mostly used for fine tuning**	**Seq2Seq Applications:** Sentiment Analysis Natural Language Inference Sentiment Role Labeling Relation Extraction Pronoun Resolution		N/A			RL let very time consuming and therefore only => mostly used for fine tuning	Mismatch between train/test measurement with Cross-Entropy loss with Seq2Seq Models	SeqSeq, RL, EncoderDecoder, LSTM, RNN, Survey, Review, Bias	They have a nice overview about what metrics can be used for which tasks
0e05e770	Personalized Review Generation By Expanding Phrases and Attending on Aspect-Aware Representations	2018	https://aw	done	Propose model to **generate personalized reviews by expanding short phrases** (e.g. review summaries, titles) - leverages both user and item information - leverages auxiliary, textual input and aspect aware knowledge They use an **aspect encoder** and an **attention fusion layer** Discover and mimic that users are inclined to discuss and bias the generated text toward their personalized aspect preferences	N/A	**Expansionnet (3 Encoders, 1 Decoder)** EncoderDecoder **Sequence Encoder**: GRU (Gated Recurrent Unit) **Attribute Encoder**: Learns Latent Representation of user and item identity (Attribute Encoder jointly learns representation of users and items reflecting user-aspect preferences and item-aspect relationships) **Decoder**: Attention Fusion Layer - attends on encoded information and intentionally biases the model to generate	Generation	**Aspects**: Service, Price, Screen, Case, Drive, Sound, Vision, Laptop, Time (each with representative words) => Aspect Extraction	Conditional Text Generation	Amazon Electronics	6 automated metrics Perplexity, BLEU-1/BLEU-4, ROUGE-L **Distinct-1/2** - Percentage of distinct uni-grams and bi-grams	N/A	N/A	N/A	EncoderDecoder, Bias, AmazonElectronics, BLEU, ROUGE, Distinct, Perplexity, GRU, Attention,	
fa1267ec	Black-Box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers	2018	https://aw	done	Present an algorithm **'DeepWordBug'** that perturbates text by inducing small modifications such as misclassification frame a DL classifier The selection of perturbated words is more important than how to change the words	they assume that perturbation happens directly at input of the words	**Black Box Scenario** => adversary has to manipulate input samples and observe outputs **Temporal Score (TS) and Temporal Tails Score (TTS)** are scoring functions to determine importance of any word for the final prediction **General Approach** **1) Token Scoring and Ranking** - TS and TTS complement each other (combined) - **TS**: Measure Prediction change when removing word i - **TTS**: Compare difference between two trailing parts of a sentence (one containing a certain word, the other not) **2) Token Transformer** - Deliberately induce misspelled word instead of the original word => Classifier sees them as "unknown" - e.g. small levenshtein distance edit Proposed: Neural Hidden Semi-Markov Model (HSMM)	Generation	N/A	Conditional Text Generation	Large Movie Review Dataset (IMDB) Enron Spam Dataset - subset of original enron email dataset	Accuracy Levenshtein (for changing word for perturbation)	N/A	N/A	N/A	RNN, LSTM, Accuracy, IMDB, Enron	Could maybe be applied to other classifiers? e.g. Detection of Machine Generated Text?
91a8e53a	Learning Neural Templates for Text Generation	2018	https://aw	done	Propose neural generation system that adds to latest, discrete template jointly with learning to generate	The template generation might be limited to their dataset-training. It is hard to assess if their approach can derive templates from any structure or text	Incorporate LSTM, Attention, etc while maintaining HSMM structure => Transition between every RNN their model can: (a) it is possible to explicitly force the generation to use a chosen template z(i), which is itself automatically learned from training data, (b) the model can segment the generated `y(i) is typed by its corresponding select variable	Generation Interpretability	HSMM marginal likelihood calculation by backward-algorithm familiar from HMMs Beam Search only over a segment for a latent state	Conditional Text Generation The task is formally known as E2E NLG and to-end (E2E) natural language generation (NLG)	E2E E2E (Novikova et al., 2017) and WikiBio (Lebret et al., 2016)	BLEU, ROUGE, NIST, CIDEr, METEOR	N/A	N/A	N/A	EncoderDecoder, Control, BeamSearch, E2E, WikiBio, BLEU, ROUGE, NIST, CIDEr, METEOR, Graph, MAP	
29de7c0f	Hierarchical Neural Story Generation	2018	https://aw	done	Propose **Dataset** of human written stories paired with writing prompts **Model Fusion**: Improve relevance of story to the prompt **gated multi-scale self-attention mechanism**: Models long range context	Their Model: - Fusion Model: Random sampling can produce errors - Repitition (model focuses to much on previous content and repeats itself) - Outputs generic language (specific vocabulary is rare)	**Hierarchical Story Generation:** 1) Premise Generation / Story Prompt: Convolutional LM 2) Generate story that follows premise: Seq2Seq Model They use Conv Seq2Seq model + Attention **Gated Self Attention Mechanism** - CNNs can only model bounded context (preventing long-range dependencies) - supplement decoder with self-attention mechanism => condition on previous outputs at different time scales **Gated Attention:** - Multi-head attention - gated linear unit activations for queries, keys and values **Multi-Scale Attention** - each head operating at different time scale (downsampled input) - first full input, second every second input timestep,... => limits number of tokens in attention maps => sharper - optionally allow ignoring information => prevent noise	Generation	**Model Fusion:** - Train second model over the pretrained Seq2Seq model - second model focuses on the link between prompt and story => helps seq2seq models to build dependecies between input and output	Conditional Text Generation	**Their dataset:** **WritingPrompts** - r/WritingPrompts subreddit	b.l.e.u (bit) r.o.u.g.e (summarization) => not that useful in open-ended generation setting Uses: **Perpleixty** - reflects how fluently a model can produce the next word **prompt ranking**: - measure dependency between output and input **percentage of cases where true prompt is likely to generae the story** **Human Evaluation** - Amazon mechanical turk (triple pairing task) - select the correct pairing stories for all 3 prompts - compare hierarchical generation with prompt vs normal generation without prompt	N/A	N/A	N/A	HumanEval, AmazonMechanicalTurk, WritingPrompts, Perplexity, PromptRanking, Seq2Seq, CNN, EncoderDecoder, Fusion	
f3cdae46f	PersuADE: An Adaptive Persuasive Text Generation System for Fashion Domain	2018	https://dl.a	done	PersuADE: Propose a model that **persuasively generates text** from an input product specification (applied to fashion domain) => text garners positive attention	Lack of good evaluation metric for Persuasiveness	**PersuADE!** - Transform a list of keywords to a persuasive description Data Scarcity => Solution: Modular Architecture with reliance on only general english corpora **System**: 1) identify entities and tag them 2) expand conjunctions 3) extract noun phrases 4 combine noun phrases with extracted entities using a MLM => candidates 5) rank candidate sentences and choose most fluent & persuasive one	Generation	Problems: Data Scarcity (domain-specific) Beam Search	Conditional Text Generation	**Their dataset:** - crawled from various website - products with multiple parameters extracted (title description,...)	Problem: Lack of good evaluation metric for Persuasiveness **Generation**: BLEU, METEOR, ROUGE, Similarity (Doc Similarity from spacy) **Readability**: Flesch-Kincaid Index SMOG Index Dale-Chal Index **Small Human Evaluation** - Grammar - Relatedness - Fluency Their system outperforms on Human Evaluation but is worse on Generation Metrics **=> Generation Metrics limit creative liberty to choose persuasive words/phrases**	N/A	N/A	N/A	BeamSearch, BLEU, METEOR, ROUGE, FleschKincaid, SMOG, DaleChal, HumanEval	
ffcad165	Auditing Data Provenance in Text-Generation Models	2018	https://dl.a	done	Develop a auditing model that lets users check whether their data was used to train a machine learning model Focus on text, working with black-box and not overfitted models too.	The more distinct users are simulated, the worse the predictions get Unknown if deeper models are more prone to memorization	RNNs rank rare words significantly higher RNNs are prone to overfitting Their Auditing works best when queries include rare words **Text Generation Models learn to generalize patterns but hard-memorize sequences consisting of less-frequent words** => Their auditing model exploits that	Auditing	**Training Shadow Models:** - auditor trains k shadow models that **simulate the victim model** - training data: use **public sources** (error is negligible when not used a subset of reference models train data) - train a binary classifier with the k predictions **Train Audit Model:** - Feed the shadow models in output distribution after probability => **substantial gap between predicted rank of the same word when it appears in train and test data**; - final feature vector for the audit model: count vector where each entry is the count of ranks **Audit Time:** - user feeds his text - obtain feature vector - classifier decides if text was in train data	Auditing	Data Protection Policy: GDPR DeepMind's unauthorized use of NHS patients' data to train ML models	AUC, ROC, Accuracy, Perplexity Text-Train Accuracy Gap: Check if overfitted	N/A	N/A	N/A	RNN, Reddit, SATED, Cornell, Wikitext, Europarl, UbuntuDialogs, AUC, Perplexity, ROC, LSTM, GRU	
f3f4c8e1	Data-to-Text Generation with Content Selection and Planning	2018	https://aw	done	Do Content Selection & Content Planning before actually generating the document	needs to verified for other languages and domains	**Record Encoder** - multi-layer perceptron + ReLU + concat vectors **Content Selection Gate** - Attention on record entries **Content Planning** - Plan: Sequence of Pointers - Pointer Network, Attention (=> point to token in input sequence) - LSTM decoder (generates tokens based on position in input sequence) **Text Generation** - Encoder-Decoder + Attention 1) Encode Content Plan with bidirectional LSTM 2) Encode Corresponding Text - Copy Mechanism **Joint Copy**: globally normalized **Conditional Copy**: Use copied/!! true/false variable to condition copy	Generation	Maximize log-likelihood of gold content plan and table records Beam Search each step	Conditional Text Generation	Rotoritine - basketball game summaries - structured and long	Precision, Recall BLEU Human Evaluation - Amazon mechanical Turk 1) identify supporting and contradicting facts mentioned in each sentence 2) Generation Quality: Best-Worst Scaling => Scores shown to be less biased and provide more reliable details compared to rating scales	N/A	N/A	N/A	LSTM, RNN, Attention, RotoWire, PointerNetwork, EncoderDecoder, BeamSearch, BLEU, Precision, Recall, HumanEval	

*				STATUS (done / wip / discuss)							Evaluation			Costs				
paperId	title	year	openAccessPdf		Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
9e7160f3	Multi-Reference Training with Pseudo-References for Neural Translation and Text Generation	2018	https://ww	done	Investigate ways to **use multiple references during training** (not just one) (Sample One, Uniform, Shuffle) => Outperforms Baseline on MT and Image Captioning Propose a **NN that aligns multiple sentences to compress existing references => pseudo-reference generation**	**Sample One:** With a high amount of references (10+), the transition quality dramatically drops (maybe: higher variance of used reference in each epoch)	**Pseudo Reference Generation:** - **compress** all references into a lattice (by merging similar words) - traverse through lattice and select those with **high BLEU as pseudo-reference** **Uniform** method performed best in their experiment - Word Similarity Measure: semantic substitution matrix (with LM) - To reduce computing time: iterative pairwise alignment (greedily merging sentence pairs)	Translation, Generation	**Sample One:** - Sample different reference in different epochs (random) **Uniform:** - like sample one but ensures every reference is used **Shuffle** - like uniform but shuffles all source/reference pairs before each epoch **Uniform** method performed best in their experiment Pretraining (MT): Minimizing Perplexity	Conditional Text Generation	**NIST 02-08** except 07 (for translation part) **MSCOCO** (for image captioning part)	BLEU, CIDEr	N/A	N/A	N/A	LSTM, RRN, EnsembleDecoder, NIST, BLEU, MSCOCO, CIDEr, Perplexity, BeamSearch, Seq2Seq	
b4812702	A Graph-to-Sequence Model for AMR-to-Text Generation	2018	https://ww	done	Introduce a Neural Graph2Sequence Model (+ mode-local semantic information (LSTMs can loose information from AMR graph structure)	N/A	**Graph State LSTM encodes AMR structure directly** **Encoder**: graph state transition by information exchange between connected nodes (modelled by an RNN) **Decoder**: Attention LSTM with Copy Mechanism **Copying vs Anonymization:** - Copying data sparsity problem by handling open-class tokens - Anonimization: Many manual work to define placeholders, heuristics - Copy: Learns what to copy and easier to adapt to new domains & languages A character LSTM can further elviate the data sparsity problem	Generation	Training with cross-entropy loss Adam **Fine-tune model on AMR corpus after every epoch of pretraining on the gigaword data** (Albinovski) Get AMR Corpus by feeding the LDC AMR Parser with gigaword sentences They use standard Glove embeddings (no updates)	Conditional Text Generation	LDC2015E86 (Gigaword) - standard AMR corpus Best Score: Graph2seq+chart.STM+copy	BLEU	N/A	N/A	N/A	AMR, Gigaword, EncoderDecoder, Attention, LSTM, BLEU, RNN, CopyMechanism, GloVe	
8b0575d2	HellaSwag: Can a Machine Really Finish Your Sentence?	2019	https://ww	done	**HellaSwag: Challenge Dataset** => Show that current models have difficulties with commonsense inference DL models capture dataset specific distributional biases during training, which results in good test performance. The underlying main task remains unsolved. (Thats how BERT "solved" Swag) Language/Common shift results in drastic performance loss (current models).	N/A	**Adversial Filtering:** - generate wrong hypotheses to challenge Models with the datasets (removes Bias) - **3 decoding techniques** studied (results in GPT-2): Top-k - easiest to detect for fine-tuned BERT (because 80% of mass in the first 500 common tokens) - Pinning: Adden single human generated token => reduces detection accuracy drastically **Nucleus** - compromise between top-k and untruncated random sampling - Pinning not as impactful **Tepertature Scaling (untruncated random sampling)** - Pinning not as impactful - Discriminator trained with top-k cant generalize to other sampling methods (and other way arounds) => Nucleus was an exception (transferrable to other samples) **Classifier**: Fine-tuned BERT-Large **Histogram-of-Likelihood Ranks.** They observe that using more bins (50) spread uniformly results in higher accuracy over possible rankings than with GLTR's approach (5)	Generation, Reasoing	N/A	Dataset, Reasoning	**Adversarial Filtering:** - **Generate Swags** original domain by adding **Wikihow** articles => higher diversity and generation length - made with high-quality discriminators - adversarial discriminators - zero-shot evaluation categories (to evaluate models generalization ability) - adversarial filtering: series of discriminators iteratively select an adversarial set of machine-generated wrong answers **Swag** - Bias: stylistic biases due to interplay between Adversial Filtering + Swag-Generators & Discriminators	**Human Evaluation** - Compare human judgement on human written endings vs machine generations refinded with adversial filtering	N/A	N/A	They say an underlying task is **solved** when algorithms can't distinguish machine-generated endings. Adversarial filtering advancements to **filter out spurious biases** during dataset creation (i.e. its impossible) Reporting Bias mentioned: - exponential number of samples needed for deep reasoning	Swag, HellaSwag, BERT, Bias, HumanEval, Dataset	
82144ee0	Automatic Detection of Generated Text is Easiest when Humans are Fooled	2019	https://ww	done	Use **human and automatic discriminators to detect machine generated text** decoding methods have primarily **optimized for fooling humans => statistical abnormalities make detection easier for automated systems** **Longer excerpts are easier to detect**	tradeoff between human and machine detection is only shown for english language	Classifier: Logistic Regression (Fine-tuned BERT)	Detection	**Self made datasets** (60 variations: per sampling method, truncation len, choice of priming) - Generated by **GPT-2** - **Choose text from GPT-2 2x train set => ensure that classifier cant make use of stylistic differences**	Detection	Human Evaluation: - Turing Test - humans struggle detecting MGT reliably (worse than automated methods) AUC, Accuracy	N/A	N/A	Amazon Mechanical Turk resulted in **poor accuracy => Group of students (introduced with 10 examples before) were "expert raters".**	Detection, AmazonMechanicalTurk, BERT, AUC, Accuracy, HumanEval, GPT	Hashimoto shows that a unique attribute of human language is that it dips in and out of low probability zones => makes text interesting to read (not yet solved by machines without poor word choices)
6e3346ac	Controllable Paraphrase Generation with a Syntactic Exemplar	2019	https://ww	done	Propose a **task where the syntax of a generated sentence is controlled** Create **dataset** with human annotations for the task Develop a **model designed to capture semantic knowledge** (and multitask training objectives)	Combining syntactic parsers with their model could maybe lead to a more robust performance Their model sometimes creates nonsensical sentences (solution: larger corpus?)	**Language Syntax Learning** - train syntactic component of model with: - **word denoising:** randomly replace word tokens (via POS tag) - **latent word-cluster codes**: create bottleneck layer in syntactic encoder Encoder Decoder Architecture **Decoder** - concat syntactic variable to previous words embedding - concat semantic variable to hidden vector output Loss Functions used to encourage clearer separation of information in semantic and syntax: - **Paraphrase Reconstruction Loss (PRL)** - **Word Position Loss (WPL)** => add on embeddings to increase performance the most Deep generative model - **2 latent variables (semantics and syntax)** disentanglement between these variables => multi-task learning objectives	Paraphrase Generation	Conditional Text Generation (Paraphrase Detection) (Syntactic Paraphrase Detection)	**Their dataset:** - triples manually annotated from ParaNMT dataset: 1) semantic exemplar sentence 2) syntactic exemplar sentence 3) reference sentence incorporating semantics of 1 and syntactics of 2 They create 2 datasets: **1) extension scenario** - auto completion generator extends news article - 1 user verifies correctness / 1 verifies incorrectness **2) modification scenario** - subtle changes to human written news article (addition/over negations) => retain original articles style 1) created from newsQA, Gigaword (CNN articles with Gigaword) 2) created from New York Times articles	**Semantic Metrics:** BLEU, METEOR, ROUGE-1, ROUGE-2, ROUGE-L **Syntactic Metrics:** Syntactic Similarity between reference and generated sentence (Zhang&Shasha Tree Edit Distance)	N/A	N/A	N/A	VAE, Paraphrasing, BLEU, ROUGE, LSTM, EncoderDecoder, Gigaword, BERT, LSTM, RNN, Dataset, ParaNMT, METEOR, Seq2Seq,	
77637a0d	The Limitations of Stylometry for Detecting Machine-Generated Fake News	2019	https://ww	done	Show that **Stylistic detection of MGT has limits** - LMs generate **statistically consistent text** regardless of niche (different from humans) => Fails to **distinguish legitimate LMs from false information LMs** => Highlights **needs for non stylometric approaches in detecting** MGT	Concentrates more on Fake News by LMs, not MGT in general	Detection: **Grover-Mega discriminator** (link) - neural fake news detector - claims to be able to spot fake news and news generated by LMs Generation 1) **Grover-Generated** (GPT2-XL) 2) **Article Rewriteser** (GPT-2 Medium) 3) **Automatic QA Extension** (newsQA dataset but true sentences are human-written, only false are MGT)	Detection	N/A	Detection	For Sentence Fusion: "balanced wikipedia" portion from **DiscoFuse** dataset (combines two sentences into one coherent sentence For Split-and-Rephrase: - **WikiSplit** Machine Translation: - **WMT 2014** English->German, newest 2014, newest 2016 Abstractive Summarization: - **Gigaword** - **CNN Daily Mail** - **BBC extreme**	F1, Accuracy	N/A	N/A	N/A	HumanEval, Detection, Generation, GPT, Grover, EntityNet, newsQA, NewYorkTimes	
e6d6b6de	Leveraging Pre-trained Checkpoints for Sequence Generation Tasks	2019	https://ww	done	Investigate the means of **using checkpoints from** **BERT and GPT-2** using a highly compatible model. Evaluate the differences between using the checkpoints **(different encoders and decoders)** Pretrained encoder is an essential component for Text Generation. Often Tasks benefit from **sharing weights between encoder and decoder** (+decrease memory footprint)	For MT there was only an english RoBERTA checkpoint available, so they trained their own	Their model is **compatible with BERT, RoBERTA, GPT-2** They use **many variants** to compare (e.g. Bert Encoder -> GPT Decoder, etc.) BertEnc->GPT2Dec did not work well because they need to learn **2 different vocabularies**	Generation	**Fine-tuning** using Adam LR 0.05, warmup 40k steps, normalization, square root decays Hyperparameter Search with **CMA-ES** - find optimal subset of layers to use	Sentence Fusion, Split-and-Rephrase, Translation, Abstractive Summarization, Conditional Text Generation	**SQuAD** - abstractive QA form Wikipedia **MNLI** - tilted classification task **ELI5** - long-form abstractive QA dataset **XSum** - News Summarizations, Abstractive Summaries **CoqaNl2** - dialogue response generation task, context and persona **CNN Daily Mail** - news summarization	**SARI** - lexical similarity metric - compares output to multiple reference in input in order - assess model ability to add/delete/keep ngrams **Exact Match Accuracy** **BLEU, ROUGE1, ROUGE-2, ROUGE-L** **Human Evaluation** - Amazon Mechanical Turk - Item Worst Scaling (less labor intensive alternative to paired comparisons) - Informativeness (correct and concise) - Fluency	Using pretrained checkpoints **saves many computing hours** Sharing wheights between encoder and decoder does also **reduce memory footprint**	N/A	N/A	EncoderDecoder, Gigaword, BERT, Sentence, RNN, CNN, DiscoFuse, SARI, WikiSplit, WMT, BLEU, ROUGE, CNN, BBC, PointerGenerator, LSTM, Translation, HumanEval, Seq2Seq	
3956e0d2	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension	2019	https://ww	done	**BART** (Bidirectional and Auto-Regressive Transformer) Denoising Autoencoder for Pretraining Bidirectional Encoders are crucial for SQuAD (QA) Left-to-right pretraining improves generation	Hallucinates	Standard Transformer Machine Translation Architecture (i.e. Generalizing BERT, GPT, etc) Corrupted Document is encoded with **bidirectional** encoder Likelihood of the original document is calculated with an autoregressive decoder => Encoder Inputs do not need to be aligned with decoder outputs Resembles BERT architecture with these differences: 1) each decoder layer performs cross-attention over final layer of the encoder 2) removed feed forward network before prediction	Generation, Comprehension	**Pretraining:** 1) Corrupting Text with noising function - Finding the best noising approach: -> random shuffle sequence order -> in-filling scheme (replace arbitrary len span with single mask token, inspired by mask token) 2) Learning a model to reconstruct original text (minimize reconstruction loss) **Finetuning:** - **Sequence Classification** - same input fed into encoder and decoder - final hidden state of last decoder token is fed into new multi class linear classifier - **Token Classification** - complete document fed into encoder and decoder - **Sequence Generation** - input to encoder, decoder outputs - **Machine Translation** - replace BART encoder (replace BARTs encoder embedding layer with randomly initialized encoder) - MT BART can use different vocabulary - first freeze most BART parameters except randomly initialized source encoder, then train all model parameters shortly	Conditional Text Generation Comprehension Translation	**SQuAD** - abstractive QA form Wikipedia **MNLI** - tilted classification task **ELI5** - long-form abstractive QA dataset **XSum** - News Summarizations, Abstractive Summaries **ConvAI2** - dialogue response generation task, context and persona **CNN Daily Mail** - news summarization **GLUE Task** - general Language Understanding Evaluation	ROUGE, BLEU	N/A	N/A	Model **hallucinates** unsupported information	Seq2Seq, BeamSearch, EncoderDecoder, BERT, GPT, BART	
ae04f3d0	Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment	2019	https://ww	done	**TextFooler** **Model to generate adversial text** - create both semantically and syntactically similar adversarial examples It is 1) effective 2) utility-preserving 3) efficient Fools BERT easily **Model with higher original accuracy is more difficult to be attacked.**		**1) identify important words for the target model** - Black-box: Calculate by observing prediction change on word change **2) replace them with the most semantically similar and grammatically correct words** - Close words by cosine similarity in vocabulary - Top-N words (N=50 in their experiments) (synonyms) - Semantic Similarity Checking: Universal Sentence Encoder (USE) - Finalization: Select word with highest semantic similarity **Moderate Transferability of adversial vulerability across models** - Adversial models with higher accuracy generate texts with higher transferability	Generation, Classification	N/A	Classification Entailment	**Classification** AGNews Fake News IMDB MESI Yelp **Entailment** SNLI MultiNLI	Accuracy **Human Evaluation** - grammatically - classification labels (i.e. identifiable by humans) - retains meaning?	N/A	N/A	N/A	AGNews, FakeNews, IMDB, MESI, Yelp, Accuracy, HumanEval, BERT, EncoderDecoderN SNLI, MultiNLI, Classification, Entailment	
635c0bf6	MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance	2019	https://ww	done	Propose metric that aims to show **correlation between human judgement and text quality** Metrics combining contextualized representations with distance measure perform best => good **generalization** across tasks by WordMover Outperforms strong unsupervised metrics on 24 tasks: Machine Translation, Image Captioning & Dialogue Response	N/A	**Measure dissimirlarity between system and reference text** - Inspiration from Word Movers Distance (WMD) **Semantic Similarity:** - Word Mover Distance using Ngrams (euclidian distance between embeddings of ngrams) soft alignments (many-to-one) 2 variants: word mover, sentence mover **WordMover**: WMD-Gru/Lstv/+Lm/Bert/+Elmo/+Twitter+Aggregation **SentenceMover**: SMD+Embedding+Finetune+Aggregation	Generation	N/A	Translation, Summarization, Captioning, Dialogue	N/A	**Summarization** - ROUGE - Pyramid - SE **Machine Translation:** - BLEU - SentBLEU - NIST - chrF - TER - WER - PER - CDER - METEOR - RUSE **Data-to-text Generation** - BLEU (slotty scalled) - CODA - SPICE **Semantic Metrics (for better human judgement correlation)** - see paper for references	N/A	N/A	N/A	HumanEval, MoverScore, Metric	
50198be8	The Woman Worked as a Babysitter: On Biases in Language Generation	2019	https://ww	done	Propose methods to annotate "regard" of a text for demographic Bias (positive/negative towards demographics) They propose: - manual metric (sentiment & regard annotations) - automatic regard classifier for unseen text They show Bias on GPT-2 qalitatively	They use manually selected keywords and templates for bias classification A more automatic method needs to be found	tested with GPT-2 tests used Sentiment Analyzer: VADER Regard Classifier: transfer Learned BERT Tried an automated regard classifier: - BERT-base - repurposed sentiment analyzer - **transfer learned BERT (better by 20%)**	Generation	N/A	Conditional Text Generation	N/A	Spearman Correlation **sentiment**: overall language polarity vs **regard**: language polarity towards demographic Accuracy	N/A	N/A	Using sentiment analyzers as bias identifier may underestimate the magnitude of biases **certain bias are not well captured by sentiment** (e.g. occupation bias)	Bias, GPT, BERT, Sentiment, Regard	

* paperId	title	year	openAccess	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Evaluation			Costs			TAGS	Comments
										Tasks	Datasets	Metrics	Ecological	Economical	Ethical		
8673d0d0	GLTR: Statistical Detection and Visualization of Generated Text	2019	https://aw	done	GLTR (pronounced Glitter) Giant Language model Test Room) **A tool to support humans to identify machine generated text (publicly available)** works for multiple LMs of similar "modern" architecture	Needs human classifier even with visual help samples conditioned on hidden seed text might be a limitation	**Most models sample from the top of the head to predict for next word** - top-n sampling, k-max-sampling, beam search, temperature-modulated sampling, rule-based templates) Detection Model **Test 1) Probability of word** - left-to-right models like GPT, P(X|, X1-1) - bidirectional models like BERT P(X|, X-30, ..+1, X1-1, ..X+30) Context **Test 2) absolute rank of word** - top10 green, top 100 yellow, top 1000 red **Test 3) Entropy of predicted distribution**	Detection	N/A	Detection	N/A	Human Evaluation - using the tool is not using the tool accuracy	N/A	N/A	MGT used to harm yet neutralize	Bias, GPT, BERT, Detection, HumanEval	
91789d28	Pretraining-Based Natural Language Generation for Text Summarization	2019	https://aw	done	Propose a method that uses **BERT encoder and a two stage decoder** to do abstractive text **summarization** The first to use BERT for this task, taking left and right context into account	N/A	Encoder: BERT Decoder (two stage) 1) Left-Context only decoder (Summary Draft) - Transformer decoder learns probability complete summary conditioned on representation vectors - Copy Mechanism 2) Mask each word of summary and predict the refined word one-by-one - word level refine decoder with reinforcement objective - uses BERT context vectors - Shared parameters between draft and refine decoder -> better performance than seperate (maybe because of teacher forcing)	Summarization, Generation	Teacher Forcing Beam Search	Conditional Text Generation	CNN Daily Mail New York Times (NYT)	ROUGE	N/A	N/A	discrete objective for the refine decoders to reduce the exposure bias problem	BeamSearch, ROUGE, BERT, NewYorkTimes, CNN, Copy, TransformerDecoder, TeacherForcing, BeamSearch	
afc182ea	Modeling Graph Structure in Transformer for Better AMR-to-Text Generation	2019	https://aw		AMR to text generation, but their model takes **whole AMR structure** into account rather than just single dependencies Propose a new self attention mechanism that incorporates such richer structural information	N/A	Apply Transformer to AMR-to-text task Relation between a concept pair is to be understood as a directed graph traversal **Input Representation:** - depth first traversal => linearization Ways to **represent a structural path**: Feature-based Avg-based Sum-based Self-Attention-based (SA-based) CNN-based => Continuous representations (like CNN based or SA based) performed better than discrete representations	Generation	**Supervised Learning** For Transformer Baseline (to counteract data sparsity for AMR): **Byte Pair Encoding (BPE)** - Split words into frequent subwords **Shared Vocabulary** - for both source and target sides	AMR2Text; Conditional Text Generation	LDC2015E86, LDC2017T10	BLEU CHRF++	N/A	N/A	N/A	LDC2015E86, LDC2017T10, BLEU, CNN, CHRF, Attention, AMR	
d2cbb0d5	Handling Divergent Reference Texts when Evaluating Table-to-Text Generation	2019	https://aw	done	Propose a new Metric PARENT - high correlation with human judgements (especially semantics) Perform a correlation analysis for various metrics	task specific metric	Model for evaluation: Pointer-Generator Network - learns to produce linearized version of the table from text	Generation	N/A	Conditional Text Generation, Table2Text	WebNLG (for human judgement correlation) WikiBio (train data for PointerGenerator & human evaluation)	BLEU and ROUGE correlate poorly with human judgements when reference texts diverge from table information => New Metric: **PARENT** **(Precision And Recall of Entailed Ngrams from the Table)** - aligns ngrams from reference texts to semi structured data - then computes precision and recall - Entailment Probability: Word Overlap Model / Co-occurrence Model (not sensitive to paraphrases) **Entailed Precision** N-Gram gets a reward either if it occurs in reference or if it has a high probability of being entailed by the table **Entailed Recall** geometric average of recall against 1) reference n-gram co-occurrence structure 2) table => lexical content information from table get higher scores **=> Entailed F-Score (PARENT)** .... Used metrics for evaluation: **Human Evaluation** Comparing two texts of the 16 models trained for WikiBio - Naturalness, Faithfulness, Informativeness - Ranking: 1 to 3 (allowing ties) for each criteria judgements - grammatically, fluency, semantics **=> BLEU correlates best on grammarfluency, PARENT best for semantics** **Content Selection:** F-score for attribute/value pairs comparing generated text and extracted from reference **Relation Generation:** Precision for the attribute/value pairs comparing generated text and ground truth table **RG-F**: F-score of extracted pairs to those in the table **Pearson correlation** Metrics for comparison: BLEU, ROUGE, CIDEr, METEOR, CIDEr-D	(PARENT deals with hallucination by discounting ngrams which do not overlap)	BeamSearch, HumanEval, PARENT, CS, RG, Pearson, WebNLG, BLEU, ROUGE, CIDEr, METEOR, Hallucination, Metric	
8cb9b95d	Assessing The Factual Accuracy of Generated Text	2019	https://ar	done	Propose a new **metric to estimate factual accuracy** (model-based) - 3 metric variants Introduce **dataset** to train 1) relation classifier 2) end-to-end extraction model	Bias towards Wikipedia/WikiData sentence structure	**End-to-end Relation Extraction Model:** **Transformer Encoder-Decoder** 6 Transformer layers 1 Embedding Layer 8 Attention Heads **Classifier** NER->Embedding->Encoder Layers-> Pooling -> Softmax/Sigmoid	Summarization, Generation	AdaFactor Optimizer	Conditional Text Generation	**Their dataset:** - English Wikipedia & Wikidata Knowledge Base - Wiki-Coref Coreference Resolution System on Wikipedia articles, finding relations with Wikidata	**Human Evaluation:** - Factual Accuracy (1-5)	N/A	N/A	N/A	OpenIE, ROUGE, NER, Metric, Dataset, Accuracy, ROUGE, OpenIE, Bias	
49c5ac37	Mitigating Gender Bias in Natural Language Processing: Literature Review	2019	http://arxi	done	Bias towards stereotypes Reviews studies on **mitigating gender bias** in NLP Show **forms of gender bias** and analyze methods for **detecting gender bias**	Need for benchmark/metrics Debiasing can introduce noise Most debiasing approaches focus on single modular process, not end-to-end NLP system Unclear how many debiasing techniques are generalizable to other tasks Non-binary genders are ignored in NLP	**Debiasing Methods:** **Retraining Category** **Data Augmentation** - create gender swapped quivalent sentences for each dataset sentence **Gender Tagging** - Translation Task loose accuracy when swapping genders => Add Gender information as tag **Bias Fine-tuning** - Transfer-learn from unbiased dataset - Incorporate bias only at finetuning stage **Removing Gender Subspace in Word Embeddings** - flawed because semantic definitions may be tied to gender component => creation of gender neutral words as an approach **Adversarial Learning: Adjusting the Discriminator** - generator attempts to prevent the discriminator from identifying the gender - may be generalizable **Learning Gender-Neutral Word Embeddings** - GN-GloVe - isolate gender information in specific dimensions during training **Inference Category** **Debiasing Gender in Word Embeddings / Hard Debiasing** - maybe problematic due to blinding AI agents perception rather than teaching it the right actions **Constraining Predictions** - Constrain exacting models optimization function to fit predictions to defined conditions 2 Encoder-Decoder Models **Concurrent System** - straight-forward, opaque - Encoder: BERT Architecture - Decoder: like Modular System **Modular System** - interpretable, controllable - Encoder: BERT Text classifier for problematic words 1) Editing Module: to LSTM with Copy- and Coverage Mechanism => join embedding (detection^editing) - allows the classifier fi edit encoder hidden states	Bias Detection, Bias Removal	Gender Bias Evaluation Testsets (GBETs) - **Winogender Schemas** (Occupation) - **WinoBias** (Occupation) - **GAP** (Names) - **EEC** (Sentiment Analysis)	Looking at global differences in Accuracy is not sufficient (averaging) => metric to analyze performance difference between genders with respect to each occupation correlate: https://arxiv.org/abs/1810.05201 **Methods to measure gender bias:** **WEAT** (Word Embedding Association Test) - measure difference in strength of association of concepts **SEAT** (WEAT extended for sentences) **Measure Performance Differences** - Gender Swapping - False Positive Equality Difference (FPED) - False Negative Equality Difference (FNED) **Equal Accuracy** (exact matches with editors changes)	N/A	**Allocation Bias** - unfairly allocating resources to certain groups	**Representation Bias** - Distract from social identity of certain groups Their Bias types: **Denigration** - use of culturally/historically derogory terms **Stereotyping** - enforce existing stereotypes **Recognition Bias** - lower accuracy in recognition tasks **Under-representation Bias** - low representation of a group	Bias, Review, Winogender, WEAT, GAP, EEC	
16586ccd	Automatically Neutralizing Subjective Bias in Text	2019	https://aw	done	Propose first **parallel corpus of biased language** Introduce **2 baseline models that automatically "neutralize" biased text** during generation - Concurrent: Accurate at reducing Bias, higher Accuracy - Concurrent: More fluent, higher BLEU, more meaning preserving	**limited on single word edits** - single-sentences - multi-word - multi-lingual - cross-sentence only done for Java repositories (maybe bias?) only 6 PhD students for human evaluation. Their model performs badly because 1) dataset is very noisy 2) sources are very different from other 2) model makes predictions (+= hence getting low ROUGE 3) sometimes it fails to capture key phrases in the reference One could also take diff files and relevant bug reports as input	**Editing Module Pretraining:** - use edits between source and target text **Editing Module Pretraining:** - use neutral portion of WNC - Weight Tagger: learn to predict edits (how target edit source should look like) **Concurrent System:** - like above pretrained - then finetuned as subjective to neural translation system	Bias Detection, Bias Removal	**Their dataset: Wiki Neutrality Corpus (WNC)** - 180k sentence pairs - removed stop words - removed fillerings, prepositions, attitudes from biased sentences	**BLEU correlated with fluency** **Accuracy correlated with subjective Bias reduction** Measure the ability to detect subjectivity: 1) select word with highest predicted probability in the sentence 2) check if the word was fed changed 3) system accuracy: proportion of selected words	N/A	N/A	Humans have a hard time identifying bias	WNC, Dataset, Bias, Subjectivity, Accuracy, BLEU, Accuracy, HumanEval, Accuracy, BERT, Dataset	
5e4f3280	Automatic Generation of Pull Request Descriptions	2019	http://arxi	done	Propose automatic generation of gh pull requests description generation => i.e. **summarizing source sequence (commit messages + source code comments)** to a pull request description	Out of vocabulary words => Encoder-Decoder with Attention, Pointer-Generator Network & RL Loss Encoder: Singlelayer LSTM Decoder: Unidirectional singlelayer LSTM - Generates shorter descriptions than baselines	**RL Loss** - reward to ROUGE, differentiable - using ROUGE somewhat as a bias brings ROUGEs benefits to training stage already => results more likely to be suitable for human evaluation	Summarization, Generation	Beam Search	Conditional Text Generation	**Their dataset: GitHub Pull Requests** - 476 pull requests - Java repositories	**Human evaluation** - only 6 evaluations (PHD Students) - semantic similarity between reference and generated description	N/A	N/A	N/A	ROUGE, EncoderDecoder, BeamSearch, LSTM, GitHub, PointerGenerator, HumanEval, F1	
ce07d58d	Enhancing AMR-to-Text Generation with Dual Graph Representations	2019	http://arxi	done	Propose model that encodes both top-down & bottom-up representations of AMR nodes => capulates contrasting views of AMR graph	N/A	**Transforming a AMR graph to 2 representations:** 1) arrow labels became nodes 2) change direction of directed arrows depending on TD or BU representation **Dual Graph Encoder** - Gated Graph - Concatenation of hidden states at the end (similar to bi-RNNs) - input to bi-LSTM Gated Graph Network GNN variants tested: - **Gated Graph Sequence Network**: uses GRU, Back-propagation through time **(best performance)** => hypothesis: Gating Mechanisms can better capture long-range dependencies - **Graph Attention Network**: multi-head attention (for different perspectives) - **Graph Convolution Network**: no combination step, rather aggregate nodes along with neighbors **Decoder**: unidirectional LSTM with Attention (Copy- and Coverage Mechanism)	Generation	use linearized version of the graph generated by depth-first traversal order as input	AMR2Text; Conditional Text Generation	**LDC2015E86, LDC2017T10** - Aligned AMR graph and sentence **MultiNLI** (used for entailment experiment with BERT)	BLEU, METEOR **Human Evaluation** - Amazon MechanicalTurk - Meaning Similarity - Readability - Pearson correlations	N/A	N/A	N/A	CNN, GRU, AMR, Attention, Dataset, LSTM, RNN, LDC, Attention, BLEU, METEOR, BERT, HumanEval	
c4692e5d	Jointly Measuring Diversity and Quality in Text Generation Models	2019	http://arxi	done	Propose 3 metrics that both assess **quality and diversity** of a generated text **simultaneously**	No Human Alignment tested	Used Models: SeqGAN RankGAN MaliGAN MLE	Generation	N/A	Generation	**MSCOCO, EMNLP2017 WMT News, IMDB** **Oracle synthetic dataset** - random initialized LSTM generator as real distribution used in oracle training mode - 100k samples generated	**BLEU** - quality - with CAWs BLEU just shows the validity of outputs not considering their coverage **Oracle-NLL:** - measures in synthetic oracle distribution - NLL of generated sample in oracle distribution - also neglects coverage - model that generates just 1 high quality sentence can reach top performance **Self-BLEU:** - ignores quality **=> Propose metrics that assess quality and diversity simultaneously** **Mil-Jaccard:** - Approximates distance using n-gram based measures **Frechet BERT Distance (FBD):** - Nature-based measure based on deep model trained on BERT **Oracle Based Evaluation:** - oracle training mode - calculates generators density - distance measure between oracle distribution and learned model distribution --- Pearson Correlation of Metrics	N/A	N/A	N/A	MSCOCO, WMT, IMDB, Oracle, BLEU, BERT, Metric, GAN	

* paperid	title	year	openAccess	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Evaluation			Costs			TAGS	Comments
										Tasks	Datasets	Metrics	Ecological	Economical	Ethical		
5z22fs241	Reducing Sentiment Bias in Language Models via Counterfactual Evaluation	2019	https://aw	done	**Aims at quantifying and reducing social/sentiment bias in generated text** **Quantify Bias via:** - Group Fairness Metrics - Show on different models **Reducing (counterfactual sentiment) Bias via:** - Embedding or Sentiment Prediction-derived regularization (on LMs latent representations) (maintaining similar perplexity and semantics)	**Evaluating on specific templates** => Methods to evaluate without templates still are an open research problem There is a **trade-off between generation quality and fairness** in their model	They define "Fairness" in the paper **Fairness trough embedding regularization** - Fairness loss is cosine distance between embeddings last two vectors averaged - captures high level semantics, averaging too many layers could make difference small) - could regularize too strongly **Fairness trough sentiment regularization** - Fairness loss is Wasserstein-1 distance - only regularize with cosine distance if sentiment exists - avoids loss trong regularization Generate many sentences with pretrained **GPT-2** based on a review with the desired sentiment - fine tuned to their datasets used (to circumvent topic irrelevant generation) Use **BERT text classifier** to filter out undesired sentiments (fine tuned)	Generation, Classification	**three step curriculum training schema** 1) train LM using regular cross entropy loss for 1 epoch (all data) 2) train sentiment classifier 3) train on subset of original train set with an additional fairness loss (based on embedding/sentiment regularization) They use 10 Sentence Templates per sensitive attribute.	Bias Detection, Bias Removal, Conditional Text Generation	WMT-19, WikiText-103	**Individual Fairness Metric** - measure Wasserstein-1 distance between sentiment score distribution of every eval sentence and each counterfactual sentence **Group Fairness Metric** - Average Individual Fairness for subgroups - measure Wasserstein-1 distance between sentiment distributions of all generated sentences of inputs from the subgroup **Average Group Fairness Metric:** - Average across as subgroup Wasserstein-1 distances **Human Evaluation** - correlation between automatic sentiment scores and human judgements - semantic relevance - sentiment **Perplexity (of LM)** **Embedding Cosine Similarity** - between the embedding of both the prefix and the generated continuations	N/A	N/A	Sentiment Bias	Bias, Metric, Transformer, WMT, WikiText, HumanEval, Perplexity, CosineSim	
432eb11c	Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection	2019	http://aw	done	Propose a Language Model that **produces fake reviews with a desired sentiment** (only using publicly available models)	Fools humans and machines	**Sentiment Modelling** - specifically condition a sentiment - Sentiment Neuron learned by multiplicative LSTM - Byte-level Transformer LM - Visualize output values of hidden units - then check if neurons can be categorized when evaluating it across multiple sentiment databases	Generation	N/A	Conditional Text Generation	Amazon Reviews	Accuracy **Human Evaluation** - fluency - select one real review out of 4	N/A	N/A	N/A	GPT, HumanEval, BERT, Accuracy, Yelp, Amazon, Sentiment	
7550fbd1	Text Generation with Exemplar-based Adaptive Decoding	2019	http://aw	done	exemplar-based adaptive decoding A new approach for template-based text generation using encoder-decoder architecture	Model sometimes gets information confused across exemplars	customized decoder is constructed for each exemplar / template - exemplars directly influence the decoder parameters **Generation Process:** 1) Standard Encoder gets content representation from source 2) Retrieve exemplar and compute exemplar specific coefficients 3) Construct adaptive decoders parameters (while was parametrized using wheightted linear sums) 4) Run exemplar-adaptive decoder 5) softmax	Generation	N/A	Conditional Text Generation, Data-to-text, Summarization	Gigaword New York Times (NYT)	ROUGE-1, ROUGE-2, ROUGE-L	N/A	N/A	N/A	LSTM, Template, RNN, EncoderDecoder, Gigaword, NYT, ROUGE, Gigaword, NYT	Could also be applied to gated networks or non-recurrent neural architectures
4fkde8ec	Denoising based Sequence-to-Sequence Pre-training for Text Generation	2019	https://aw	done	PoDA (Pre-training of Denoising Autoencoders) Pretraining both encoder and decoder to denoise text - sota performance and converges faster	N/A	Combination of: - Transformer (Encoder-Decoder) - Pointer-Generator Network (Copy Mechanism) 2 Parts: Noising & Denoising 3 Noising Functions - Random Shuffle - Random Drop - Replace Words Denoising: Maximize conditional probability	Generation, Correction	They kept 3% of positions not noised, so the model can learn to copy words They dont use RL with b.l.e.u or r.o.u.g.e because it may overfit evaluation metrics and barely show improvements in human evaluation	Summarization, Grammatical Error Correction, Conditional Text Generation	Pretraining: - Full Dump of Wikipedia - Billion Word Benchmark Summarization Datasets: CNN Daily Mail Gigaword Error Correction Datasets: CoNLL-2014 JFLEG	Summarization: ROUGE-1, ROUGE-2, ROUGE-L Error Correction: MaxMatch (GLEU) (both shown to correlate with humans)	N/A	N/A	N/A	EncoderDecoder, PointerGenerator, ROUGE, CNNDailyMail, Gigaword, Denoising, Wikipedia, BillionWord	
416e28ff	Data-to-text Generation with Entity Modeling	2019	https://aw	done	Entity-Centric Neural Architecture for Data2Text Generation Goal: **Entity-coherent text generation** Output text is conditioned on both data input and entity memory representations Propose MLB Dataset	Their dataset is perceived better als Gold in Human Evaluations ie all 3 metrics => maybe because Gold is too verbose in comparison?	Model creates entity specific representations - uses hierarchical attention **Encoder-Decoder** Conditional Copy Mechanism - binary variable switches a gate to tell whether to copy token or not **Entity Memory** - Keeps hidden representations each time step to account for shifting focus on entities during text flow **Hierarchical Attention** - Focussing entities first => improves RG, CS and BLEU	Generation, Summarization	Beam Search	Summarization, Generation, Data2Text	**RotoWire** Their dataset: **MLB** - longer summaries with richer and more structured records	**Human Evaluation** - Decide between 2 summaries on: - Better Grammatically - Better Conciseness - Better Coherence	N/A	N/A	N/A	Dataset, Attention, BeamSearch, RotoWire, MLB, EncoderDecoder, BLEU, HumanEval	
33fba546	Neural data-to-text generation: A comparison between pipeline and end-to-end architectures	2019	https://aw	done	Comparison between neural pipeline and end-to-end approaches (Data2Text). They do automatic & human evaluation as well as a qualitative analysis **=> Explicit intermediate Steps in generation (pipeline) results in better tests than end-to-end approaches** **=> Pipeline models generalize better** **=> Supervision in Pipeline Architectures results in better permance than End-to-End**	They do the comparison on **GRUs and Transformer (two methods only)** for end-to-end. It is questionable if their results can be generalized to other architecture too.	**Neural Pipeline,** Decomposition / Tasks: 1) Discourse Ordering 2) Text Structuring 3) Lexicalization 4) Referring Expression Generation 5) Textual Realization **End-to-end** - Joint Learning - does not adress individual tasks - directly tries to learn how to map RDF triples into corresponding output text 2 Approaches: **GRU**: go over tokens one by one - Ordering and Lexicalization while ordering **Transformer**: encode entire source sequence a s a whole using position embeddings => better at verbalizing	Generation	N/A	Conditional Text Generation	**Automatic:** Accuracy, BLEU, METEOR **Human Evaluation** - Amazon Mechanical Turk - fluency - semantics (i.e. clarity) - **Qualitative Analysis (human eval):** - Omitting words, verbs, determines or references grammatically semantics verbalizes less or more than input triples	N/A	N/A	End-to-end is more prone to Hallucinate	HumanEval, Transformer, GRU, WebNLG, Accuracy, BLEU, METEOR, Hallucination	
bd9b835d	Improving Robustness of Machine Translation with Synthetic Noise	2019	https://aw	done	Propose a method to enhance robustness of MT systems **- emulating natural noise on clean data** **=> stystem more resilient against actual natural noise**	N/A	**synthetic noise induction** - model that heuristically introduces types of noise unique to social media text - noise types: spelling, grammar, profanity, grammar, emoticons **Back Translation** - data driven method to emulate target noise - 2 methods **1) Un-tagged Back Translation (UBT)** - Pass source through source through MT system **2) Tagged Back-Translation (TBT)** - Append MTNT tag in from of sentence before passing them trough the back-translating MT system	Generation	N/A	Conditional Text Generation, Translation, Noise-Resilience	Machine Translation of Noisy Text (MTNT)	BLEU - relation between amount of noise and BLEU is linear	N/A	N/A	N/A	LSTM, EncoderDecoder, BPE, Noise, BLEU	
4ae527b6	BLEURT: Learning Robust Metrics for Text Generation	2020	https://aw	done	Propose new metric: **BLEURT** - learned evaluation metric learned with BERT - Pretraining delivers consistent improvements for BLEURT - More robust against quality drift - BLEURT can quickly adapt to new tasks - models human assessment with good accuracy		Synthetic training pairs generated by 3 techniques: Mask-filling with BERT: Random insertion Randomly dropping out words	Generation, Metric	BLEURT trained on English and evaluated on different generalization regimes Stress-test with quality drifts Adapt to different domains (3 tasks) **Pre-training BERT:** - synthetic reference-candidate pairs - improves generalization - large and diverse reference sentences - great variety of lexical, syntactic and semantic phenomena **Pretraining Signals**: BLEU, ROUGE, BERTScore, Backtranslation Likelihood, Textual Entailment, Backtranslation Flag **Fine-Tuning BERT for Quality Evaluation (task-specific)**	WMT Metrics Shared Task - annual benchmark of translation metrics on ability to imitate human assessment WebNLG Conditional Text Generation, Translation, Metric	Machine Translation of Noisy Text (MTNT) WebNLG WMT 2017-2019	**Pretraining Signals**: BLEU, ROUGE, BERTScore **Kendall's Tau** - Consistency Across Experiments **official WMT metric** for that year - Pearson Correlation or Robust variant of Kendalls Tau called DARR **MoverScore** **BEER, RUSE, and ESIM** - fully learned metrics (end-to-end) - use unaccordicated features and/or learned embeddings - great expressivity - tunable for task specific properties (fluency, style, ...) **YiSi and BERTr/cove** - soft n-gram - combine trained elements (e.g. contextual embeddings) with handwritten logic (e.g. token alignment rules) - more robust - **do not rely on when there is little to no training data** - **dont rely on assumption that train and test data are identically distributed** (important because of domain- and quality drifts) **Ideal Metric:** - take full advantage of available ratings data for training - be robust to distribution drifts (able to extrapolate) **Their Metric: BLEURT** - random perturbations of Wikipedia sequences - augmented with diverse set of lexical and semantic level supervision signals	N/A	N/A	N/A	Metric, BLEURT, BLEU, ROUGE, BERTScore, WebNLG, WMT, MoverScore, Kendall	
4715ce7a	ToTTo: A Controlled Table-To- Text Generation Dataset	2020	https://aw	done	**ToTTo** **table to text dataset for controlled generation** Task: Produce a **one sentence description based on a Wikipedia Table with highlighted table cells** Their dataset can serve as a benchmark for high-precision conditional text generation (focusing regarding Hallucinations) open-domain dataset **Current models struggle with:** - Hallucinations - Rare topics - Diverse table structure - Numerical Reasoning (order within table)		3 Approaches tested on this dataset task 1) BERT-to-BERT 2) Pointer-Generator 3) Pudugunta et al. (2019) - SeqSeq model with are explicit content selection and planning mechanism	Dataset, Generation	N/A	Table2Text, Conditional Text Generation, Data2Text	**Their dataset: ToTTo** - Table and original text from Wikipedia - collected pairs of tables/sentences that are likely to have a significant Reasoning - Wikipedia Topic Annotated - Annotators highlighted the cells in the table that are mentioned in the sentence - Annotators revise the sentence iteratively to produce final sentence (for grammatical correction/refinement)	**Annotator Agreement** **Fleiss Kappa** **BLEU-4** **Automatic Metrics:** BLEU, PARENT PARENT is data-to-text specific **Human Evaluation** - Fluency - Faithfulness (Precision) - Covered Cells (Recall) - Coverage with respect to reference (Recall)	N/A	N/A	Hallucinations - Models output phrases like "first", "winning" that are not faithful to the table - Hallucination in other data2text datasets too: (Lebret et al., 2016; Wiseman et al., 2017) - even if reference targets are faithful to the source, models still struggle	Dataset, Hallucination, Fleiss, BLEU, ToTTo, PARENT, HumanEval, BERT, Reasoning	
01198d7c	CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning	2020	https://aw	done	Propose CommonGen, a **task with a dataset that tests machines ability of generative commonsense reasoning** (e.g. "a man **throws** frisbee and the dog **catches** it") **CommonGen requires:** 1) relational reasoning 2) compositional generalization Utils performance on the task is still far away from humans		EncoderDecoder: Bidirectional RNN, Transformer Non-autoregressive generation: - assume to have better performance because of explicit modelling of iterative refinements - Levenshtein Transformer (LevenTrans) - ConstLeven Pre-trained Language Generation Models - GPT-2, UniLM, UniLM-v2, BERT-Gen, BART, T5 (best, still bad)	Generation	Compositional Generalization: - part of train set downsampled to be distantly supervised (using caption sentences as human references)	**Task Formulation:** **Relational Reasoning with Commonsense** - understanding underlying concepts and composing them - based on background commonsense (spatial relations, temporal events, ...) **Compositional Generalization:** - ability to make "infinite use of infinite means" by using words/strategies	**Their dataset: CommonGen** - 77k commonsense descriptions over 40k concept sets - Concepts from several datasets: MSCOCO, Flickr30k, Conceptual Captions, LSMDC, ActivityNet, VATEX - Human Annotations (Amazon Mechanical Turk) - assigning concept sets	BLEU, METEOR Concept Coverage - avg percentage of input concepts that are present in lemmatized outputs CIDEr SPICE	N/A	N/A	N/A	EncoderDecoder, Dataset, Reasoning, BLEU, METEOR, SPICE, CIDEr, Coverage, HumanEvaluation, BERT, BART, GPT, Transformer, RNN	
dce43e45	Evaluating Factuality in Generation with Dependency-level Entailment	2020	https://aw	done	Present a textual entailment system that is able to **fact check generated text and localize the factual error** Propose a method to **create data by leveraging entailment and paraphrasing datasets** (as human judgments are unreliable for entailment data)	Paraphrasing: Their model is biased towards precting entailment for common arcs => poses adversial setting for the DAE model Their model has different accuracy on different rewriting types (hallucination, synyom replacements, ...) => maybe not ideal for settings with different rewriting types no negation arcs cannot identify arcs that miss from the input model trained on noisy parsnmt dataset	idea: **ask whether semantic relationship manifested by individual dependency arcs in the generated output is supported by the input** previous ideas: **generating freeform questions tocapture factuality** - require extra question generation during inference **factuality eval over components of e.g. AMR or UDS** - high performance **Dependency Arc Entailment Model (DAE)** - predict entailment decision for each arc in the dependency parse of the hypothesis 1) pretrained encoder per token 2) get dependency arcs from token level representations 3) Concatenate representations of arc layer followed by softmax => obtain entailment label probabilities per arc They test 2 pretrained models: BERT, ELECTRA **Error localization**: using arc entailment scores within arc dependency tree structure to localize within the output sentence	Generation	standard binary cross entropy loss supervision on the arc entailment labels (partial labelling sufficient)	Summarization, Paraphrasing, Conditional Text Generation	**Automatic dataset creation:** - triples of input, hypothesis and single dependency arc - based on ParaNMT-50M - **assume that the ground truth paraphrases** (assume that every arc in target side is entailed by source side) - **negative examples by automatic paraphrase generation** methods (see bottom-next generations of the beam) augmentation with word swapping, synonym and Hallucination data	automatic metrics dont capture factuality previous models suffer from bias and mismatch between trainded data Human Evaluation: - check dataset agreement with human labels	N/A	N/A	Utils are prone to factual errors (Cao et al., 2018)	Factuality, ParaNMT, HumanEval, BERT, ELECTRA, Hallucination, Dataset	

*									Evaluation			Costs				
paperid	title	year	openAccess	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
5f63fc3d	A Survey of Knowledge-enhanced Text Generation	2020	https://arx	done	Survey **Text Generation Enhancement** by **i) internal knowledge embedded in input text** **ii) external knowledge such as knowledge base and knowledge graph** (past 5 years) **They cover:** 1) general architectures 2) specific techniques according to different forms of knowledge data **Future directions:** 1) Incorporate Knowledge into Visual-language Generation Tasks 2) Learning Knowledge from Broader Resources 3) Learning Knowledge from Limited Resources 4) Learning Knowledge in a Continuous Way	N/A	**Basic Text generation Models** - RNN based - CNN based Encoder-Decoder **Knowledge Enhanced Models** - Attention Mechanism - Copy Mechanism - Memory Networks - Graph Network - Pretrained LMs **Enhancement by internal Knowledge** **By Topic:** - Leverage Topic Words from Generative Topic Models - Jointly Optimize Generation Model and CNN Topic Model - Enhance NLG by Neural Topic Models with Variational Inference **By Keyword:** - Incorporate Keyword Assignment into Text Generation - incorporate Keyword Extraction into Text Generation **By Linguistic Features** - POS Tags and NER Tags - Syntactic Dependency Graph - Semantic Dependency Graph **By Open Knowledge Graphs** **Enhancement by External Knowledge** **Knowledge Bases (KB)** - Design Supervised Tasks around KB for Joint Optimization - Enhance Incorporation by Selecting KB of Facts in KB **by Knowledge Graph:** - Incorporate Knowledge Graph Embeddings into Language Generation - Jointly Optimize Language Model with Knowledge Triplet Information - Perform Reasoning over Knowledge Graph via Path-Finding Strategies **By Grounded Text** - Guiding Generation with Retrieved Information - Modeling Background Knowledge into Response Generation	Generation	**Knowledge Enhanced Learning and Inference** - Constraint Driven Learning - Posterior Regularization - Multi-task learning - Plug and Play methods	Generation, Knowledge	SQuAD CNN Daily Mail Gigaword PersonaChat	Perplexity, BLEU-4, ROUGE-L	N/A	N/A	N/A	Review, Knowledge, BLEU, Perplexity, ROUGE	not open access? accessible via institutional login Very extensive. Check paper for deeper information about each point
03031d20	Towards Controlled Counterfactual Generation for Text	2020	https://arx	done	Propose **GYC** A model to **generate a set of counterfactual text samples** (crucial for creating test samples) Propose MaChamp, a **toolkit for finetuning of contextualized embeddings (multi-ranking)** main advantages: - easy config - supports wide range of NLP tasks (seq, span, multi-task, multi-dataset, contextualized embedding)	could be extended for longer sentences (currently too expensive to run via reconstruction)	GYC - generate **counterfactual examples** such that the generation is plausible, fluent, and sentiment-preserving - uses pretrained GPT-2 decoder **1) reconstruct input text** **2) introduce losses to enforce a condition (sentiment, NER or class label)** Loss needs to be differently calculated depending on model differentiability Calculate **differentiable loss**: - Differentiable Score by sentiment classifier Calculate **non-differentiable loss**: - REINFORCE technique	Generation, Debiasing	no retraining of finetuning needed thanks to the GPT-2 decoder	Conditional Text Generation, Debiasing	Sentiment Classifier trained on: **Yelp, DBPedia, AgNews**	Generation Quality: **Label-Flip Score (LFS)** - a BERT based classifier that assesses if generated counterfactual text belongs to different class than the source text class **Diversity Score (Self-BERT)** **Content Preservation** **Tree-Edit Distance** **Human Evaluation** - grammatical correctness, plausibility, fluency, sentiment change, content preservation	N/A	N/A	N/A	Control, Factuality, Bias, BERT, Yelp, DBPedia, AgNews, HumanEval, GPT, SelfBERT, LFS, TreeEdit, ContentPreservation	Counterfactual Data can be used for debiasing with data augmentation
11a65ccd	Massive Choice, Ample Tasks (MaChamp): A Toolkit for Multi-task Learning in NLP	2020	https://arx	done	multi-task learning is mostly beneficial for setups in which multiple datasets are annotated for the same set of tasks	N/A	MaChamp Encoder: - pre-trained contextualized model	Generation	**Encoder Finetuning:** inverse square root learning rate decay with linear warm-up	**MaChamp works with:** SEQ, StringSting, Seq-Bio, MultiSeq, Dependency, MLM, Classification, Seq2Seq	GLUE English Web Treebank	**MaChamp supports:** Accuracy, Las, micro-f1, span-f1, BLEU, Perplexity	N/A	N/A	N/A	Accuracy, BLEU, F1, Perplexity, MaChamp	
342ec2f1	Logical Natural Language Generation from Open-Domain Tables	2020	https://arx	done	Propose a new task: - model **generates statements that can be entailed by facts in an open-domain semi-structured table** => Should improve inference capabilities Propose new automatic metrics to evaluate the fidelity of generation models w.r.t. logical inference **They observe:** 1) Pre-Trained LM can boost both fluency and logical fidelity 2) RL and Adversarial Training are trading fluency for fidelity 3) Jointly Pre-training Text Generation can help partially alleviate the fidelity issue while maintaining high language fluency	Their human evaluation only asks for quality which seems to be fluency and correctness etc combined	They tested multiple models Non-pretrained: Field-Infusing, Field-Gating Pretrained: Table Linearization, GPT-TabGen, BERT-TabGen Coarse-to-Fine Generation - Catastrophic errors maybe cant be corrected in later steps of generation **Two Pretraining Objectives** pretrains passes 1) Model generates template 2) Model generates final grounded sentence	Generation	**Adversarial Regularization** - Discriminator is pretrained on the table-other text pairs vs artificial language pairs 1) entity resolution to locate numbers, count and entities 2) randomly replace them with entities/numbers from table	Conditional Text Generation, TableaText	TabFact - wide range of logical/symbolic inferences	**Other Human Evaluation** - Quality (correctness)	N/A	N/A	N/A	SP-Acc, NLI-Acc, BLEU, Pivot-Acc, BERT, HumanEval, Perplexity, TabFact, RL, Task	
6f33c8e4	KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation	2020	https://arx	done	Propose **KGPT (Knowledge Grounded PreTraining) with 2 parts** **1) general knowledge grounded generation model** **2) pretraining paradigm on massive knowledge-grounded text corpus** Model can be finetuned to different tasks/domains	N/A	**KGPT Model** **Graph Encoder** - based on graph attention network **Sequence Encoder** - based on Transformer **Decoder:** - based on Transformer - Copy Mechanism (decreases PPL)	Generation	They test their model: **Distantly Supervised** **Supervised** **Few-Shot** **Zero-Shot** Significant performance in few- and zero-shot settings	DataaText, Conditional Text Generation	**Their Dataset: KGTEXT** - 1.8M knowledge sub(graph, text) pairs - Crawled from Wikipedia (with anchor links) - dataselection to reduce noise: select sentences by ROUGE similarity compared to knowledge graph	ROUGE-L, BLEU, Perplexity	N/A	N/A	N/A	Dataset, GPT, E2E, KGPT, BLEU, ROUGE, WebNLG, METEOR	Their model is NO GPT variant but they use GPT as baseline
de313cd4	POINTER: Constrained Progressive Text Generation via Insertion-based Generative Pre-training	2020	https://arx	done	Propose **POINTER** (PrOgressive INsertionbased TransformerER) - insertion based approach for lexically constrained text generation - coarse-to-fine hierarchy	N/A	**POINTER** 1) Generate high level words (nouns, verbs, adjectives) that bridge keyword constraints 2) Iterative use words as pivoting points at which to insert levels of finer granularity Iterative step 2 as long as sentence is completed Initialized with BERT **Token Importance Scoring:** - important tokens should appear earlier in data - TFIDF, POS, WAKG **DP-based Data Pair Construction** - encourage token dropped with masking each generation step considering token importance and efficiency	Generation	Pretraining on large Wikipedia Corpus Can be finetuned for downstream tasks **Insertion beam search (ILBS)** - applies an approximate local beam search at each iteration to find the optimal stage-wise insertions	Conditional Text Generation	News Dataset Evaluated on - EMNLP2017 WMT - Yelp Review	**BLEU, METEOR** **Human Evaluation** - Fluency, Informativeness, Semantics	N/A	N/A	N/A	Yelp, POINTER, WMT, HumanEval, BERT, BLEU, METEOR	
29e86c1e	Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints	2020	https://arx	done	Propose a table2text **framework that is faithful** to the table => generated content from table does not go beyond the tables information Propose a new automatic metric, specialized for table-2-text generation	Transformer based generation framework	3 loss functions combined: **1) maximum likelihood loss** (measures the matching between a model prediction and the reference text sequence) **2) feature matching disagreement loss** (measures the disagreement between a table encoding and the corresponding reference-text encoding) **3) optimal-transport loss** (matching the key words of an input table and the corresponding generated text)	TableaText, Conditional Text Generation, DataaText	**Wikiperson** - contains multi-sentence descriptions (different from WikiBio)	**BLEU-4, ROUGE, METEOR, PARENT** **Their Metric: PARENT-T** - modified PARENT to measure faithfulness **Human Evaluation** - workers trained with examples - grammar fluency, faithfulness (precision, recall, F1)	N/A	N/A	Less Hallucinations => More Faithful	Metric, HumanEval, EncoderDecoder, Transformer, Hallucination, BLEU, PARENT, METEOR, ROUGE	
72ccc306	GPT-too: A Language-Model-First Approach for AMR-to-Text Generation	2020	https://arx	done	Propose a method of **AMR2Text Generation** by using 1) a strong pre-trained language model (GPT-2) 2) cycle consistency-based re-scoring	N/A	**Decoder**: greedy decoding, beam search (best), and nucleus sampling	Generation	**Fine-Tune GPT-2 on sequential AMR representations** (directly using PENMAN representation yielded best results)	AMR2Text, Conditional Text Generation	LDC2017T10	BLEU, METEOR, METEOR BERTScore Human Evaluation - faithfulness to the table - Pearson Correlation Coefficient	N/A	N/A	N/A	AMR, HumanEval, GPT, BLEU, chrF, METEOR, BERTScore, Pearson, LDC	
dbeecab4	On Faithfulness and Factuality in Abstractive Summarization	2020	https://arx	done	Performed a large scale **human evaluation of four abstractive summarization systems** **=> Types of Hallucination: Intrinsic, Extrinsic, Factual** **Pretrained models are better** in terms of ROUGE and faithfulness and factuality textual entailment measures correlate with faithfulness better than standard metrics majority of hallucinations are extrinsic (not inferrable from input)		RNN Based - PointerGenerator (PtGen) CNN Based - TCONVS2S Transformer Based - GPT-TUNED - TRANS2S - BERTS2S	Generation	Noise or artifacts in train data cause hallucinations (Kryscinski et al., 2019a; Wiseman et al., 2017; Dhingra et al., 2019)	Conditional Text Generation, Summarization	Extreme Summarization Dataset (XSum) - BBC articles with single sentence summaries	ROUGE and BERTScore correlate less with faithfulness/factuality than metrics derived from automatic semantic inference systems Human Evaluation - trained evaluators for hallucination types and factuality 1) identify hallucinated text span 2) state if hallucination was extrinsic or intrinsic - factuality Textual Entailment - entailment classifier scores - correlate with factuality and faithfulness	N/A	N/A	N/A	Hallucination, ROUGE, BERTScore, HumanEval, XSum, Transformer, RNN, CNN	
7b1b9a4f	SummEval: Re-evaluating Summarization Evaluation	2020	https://arx	done	They evaluate **14 automatic evaluation metrics** They **benchmark 23 summarization models** with metrics Propose machine generated **summarization dataset (trained on CCN Daily Mail Dataset)** Propose expert-annotated subset of CNN Daily Mail Dataset	N/A	(all methods briefly explained in the paper) **Extractive Methods** NEUSUM, BanditSum, LATENT, REFRESH, RNES, JECS, STRASS **Abstractive Methods** PointerGenerator, Fast-abs-rl, Bottom-Up, Improve-Ltps, Unified-ext-abs, ROUGESal, Multi-task, Closed-Book Decoder, SENECA, T5, NeuralTD, BertSum-abs, GPT-2, UNILM, BART, Pegasus	Generation	N/A	Conditional Text Generation, Summarization	CNN Daily Mail - Contains strong lexical biases Their dataset: human annotated variant - Coherence - Consistency - Fluency - Relevance - 3 expert annotators confirm judgements to ensure dataset quality	**ROUGE** - measures lexical overlap **ROUGE-WE** - soft lexical matching added **BERTScore** - measures contextualizing cosine sim **MoverScore** - WordMover Distance between reference and output **Jensen-Shannon (JS-2)** - extended to bag of sentence embeddings **SummQA** - question answer model to answer cloze-style questions using the generated summary **BLANC** - unsupervised, measures performance gain of a pretrained LM given the doc summary while carrying out language understanding tasks **SUPERT** - referenceless, multi-doc summarization, measures semantic similarity of outputs with pseudo summaries created by extracting salient sentences from the source documents **BLEU** - measures ngram overlap **CHRF** - character based ngram overlap **METEOR** - computes alignment between mapped unigrams based on stemming, synonyms and paraphrastic matches **CIDEr** - computes TF-IDF ngram cooccurrences and calculates cosine sim **Extractive fragment coverage** - percentage of words that are from source article **Abstractivity** - average length of extractive fragment Compression Ratio - word ratio between articles and its summaries **Novelty** - percentage of ngrams in the summary not found in input document majority of metrics rely on hard or soft subsequence alignments, which ignores the fact that cross-dependencies between consecutive sentences	N/A	N/A	N/A	ROUGE, BERT, MOVERSCORE, METEOR, CIDEr, chrf, BERTScore, BLEU, TS, GPT, Pegasus, Dataset, Metric	

*																	
									Evaluation			Costs					
paperid	title	year	openAccess	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
e54edbe6	Re-evaluating Evaluation in Text Summarization	2020	https://arx	done	Evaluates the reliability of several detection metrics using top-scoring system outputs results suggest that **metrics run the risk of being biased and/or inadequate**, highlighting the need to meta-evaluate metrics for modern datasets and systems **no one-size-fits-all metric** exists for all datasets **Metrics cannot reliably quantify the improvements made by one system over another** ==> choice of metric depends on task and dataset (not just task) Propose dataset of human judgements	N/A	**Extractive Methods** CNN-LSTM-BiClassifier, Latent, BanditSum, REFRESH, NeuSum, HIBERT, Sent-Sum-Ext, CNN-Transformer-BiClassifier, HeterGraph **Abstractive Methods** pointer-generator+coverage, fastAbsRL, fastAbsRL-rank, Bottom-up, TS, UniLM-V1, UniLM-v2, fastStageRL, pre-SummAbs,preSummAbs-ext, BART, Semsim	Generation	N/A	Conditional Text Generation, Summarization	**CNN Daily Mail** **TAC-2008, 2009** **Their dataset (REALSumm)** - 100 Human Judgements from 25 system outputs	**Pyramid Method** - evaluate summarization systems and evaluation metrics themself: (1) exhaustively obtaining Semantic Content Units (2) weighting them based on the number of times they are mentioned (3) scoring a system summary **Lightweight Pyramid Method** - Amplifies the above method by (1) allowing crowd workers to extract a subset of all possible SCUs (2) eliminating the difficult task of merging duplicate SCUs from different reference summaries (instead use SCU Sampling to resimulate frequency based wheighting) **BertScore** **MoverScore** **Sentence Mover Similarity** **Word Mover Similarity** **JS Divergence** **ROUGE-1, ROUGE-2, ROUGE-L** Correlation Measures: **Pearson Correlation** **Williams Significance Test** **Human Evaluation** 1) Generate summaries from 25 top scoring systems 2) Include varying documents into the human eval 3) Pyramid Method	N/A	N/A	N/A	Dataset, ROUGE, WordMover, SentenceMover, SimilarityMeasures, BERTScore, HumanEval, ChiNgai(Mai), TAC, REALSumm	
07bcdc1d	GeDi: Generative Discriminator Guided Sequence Generation	2020	https://arx	done	**GeDi** A decoding method that allows for **controllable decoding** (w.r.t. **Bias, Toxicity, Hate and Negativity**) generated text was **more domain-accurate** than with other models ("book-liter") GeDi has ability to **generalize to new control codes** zero-shot more **efficient**	They did no automatic evaluation (only human evaluation)	**GeDi (Generative Discriminator)** - guides each generation step with by computing classification probabilities for possible next tokens - normalizing over class conditional distributions - class-conditional LM trained on attribute (control code) 2) conditioned on undesired attribute (anti control code) - bias generation towards desired class They finetune GPT-2 with control codes guiding generations from GPT2-XI and GPT-3	Generation	N/A	Debiasing, Detoxification, Conditional Text Generation	**IMDb movie reviews** (for finetuning GPT-2) **Jigsaw Toxic Comment Classification Dataset** (for detoxification test) **RealToxicityPrompts** (for generation) **AG News**	**Human Evaluation** - rated generated text on: - sentiment - topic - fluency - resembled an amazon or movie review - Also: **Toxicity Evaluation** - Fluency - Toxicity	N/A	N/A	N/A	Bias, HumanEval, GPT, IMDb, Jigsaw, GeDi, Toxicity	
10e8820d	Investigating Pretrained Language Models for Graph-to-Text Generation	2020	https://arx	done	Investigate **task adaptive pretraining strategies** for 3 graph domains: 1) meaning representations 2) Wikipedia Knowledge Graphs 3) scientific Knowledge Graphs Show that TS and BART are applicable to graph2text generation Better fluency than other methods and human references LMs do not always use graph structure but instead use memorized facts to guide generation	N/A	BART and T5 pretrained models adding extra tokens to represent the graphs as sequence	N/A	intermediate adaptive pretraining step between the original pretraining and fine-tuning phases for graph-to-text generation **== task specific adaptation** **Adaptive Pretraining Methods:** **language model adaptation (LMA)** - follow BART and T5 pretraining strategies for language modeling, using the reference texts that describe the graphs **supervised task adaptation (STA)** - use parts of graphs and corresponding texts collected from the same or similar domain as the target task	Conditional Text Generation, Graph2Text, AMR2Text	**AMR/EDG(2017T10)** - AMR graph with semantic relations aligned with a sentence **WebNLG** - KG from DBPedia with one or multiple sentences describing **AGENDA** - scientific abstracts paired with KG - graphs automatically generated (loose alignments) **Additional task specific data:** - self collected **AMR Ilac**: SciTA-AMR parser on Gigaword Corpus **AGENDA Ilac**: titles and abstracts from arxiv.org - scholar wrapping (KyGrid++)	BLEU, chrF++, BERTScore, MoverScore Human Evaluation - Fluency, Meaning Similarity, Semantic Adequacy	N/A	N/A	N/A	EncoderDecoder, BART, T5, BLEU, Bias, MoverScore, BERTScore, chrF++, Human Eval, Gigaword, LDC, AGENDA, AMR	Future work: Try to model graphs structural bias into Pretrained LMs, possibly leveraging their language modeling capabilities
6b5b27c8	Dynamic Blackbox Attacks Against Machine Learning Models	2020	https://arx	discuss	not NLP, CV												
399e7d8	RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models	2020	https://arx	done	You can **prompt toxicity** Investigate the **effectiveness of toxicity degeneration algorithms** Create the dataset **RealToxicityPrompts** **They find that GPT-2 training corpus has non-negligible toxicity in itself** - banned Internet communities - Reddit unreliable	Model as toxicity measure (imperfect quantification)	They test on GPT-1 to GPT-3, CTRL, CTRL-Wiki (pretraining corpora stated in paper) **Detoxification Algorithms** **Data-Based** - pretrain LM further: **Domain-Adaptive Pretraining** - pretrain further in phase of pretraining on balanced corpus **Attribute Conditioning** - prepend a corresponding toxicity attribute token to a random sample of documents and pretrain further **Decoding-Based** - change generation strategy **Vocabulary Shifting** - learn 2dim representation of toxicity for every token in LMs vocabulary ==> boost likelihood of non-toxic tokens **Word Blacklist** - hard-stop mask blacklist, disallowing a set of words **PPLM** - altering past and present hidden representations to better reflect desired attributes (use gradients from a discriminator) **Best Detoxification**: PPLM, DAPT, Vocabulary Shifting **Scaling is perfect**	Generation, Prompting	N/A	Conditional Text Generation	Their dataset: **RealToxicityPrompts** - 100k naturally occuring sentence level prompts - toxicity scores (from classifier) - from english web text **They echo the calls:** - of others to release fully transparent training data (text, source URLs, timestamps, ...) - calls for improving policy around public release of large language models - nescemination of the relationship between NLP systems and their end users	Expected Maximum Toxicity (over 25 generations) with std-deviations empirical probability of generating a span with Toxicity > 0.5	N/A	N/A	N/A	Dataset, Toxicity, Bias, Prompting, GPT	They point out that using handpicked negative examples could avoid toxicity generation (future work) The same could be done for biases (intentionally feeding Bias).
da5d6428	Asking Questions the Human Way: Scalable Question-Answer Generation from Text Corpus	2020	https://arx	done	Propose ACS-QG (Answer-Clue-Style-aware Question Generation) imitating human-like chat interaction from sampled triples of <answer, clue, style> (Pretraining needed for acceptable form and grammatically)	N/A	Converts a one-to-many into a one to one mapping problem (multi-aspect information) **Information Extractor** - samples multiple types of assistive information from text to guide generation - learns joint distribution of <answer, clue, style> from reading comprehension datasets **Neural question generators (ACS-aware)** - generates questions leveraging assistive information 2 models: 1) SeqSeq - Encoder: Bi-GRU - Decoder: GRU, Copy and Attention Mechanisms - discriminates between content words and function words in input - vocabulary reduction to encourage agressive copying 2) Fine-tuned GPT-2 - more readable due to less duplicated output words than Seq2Seq **neural quality controller** - removes low quality generations based on text entailment - BERT based QA model to predict answer span (measure overlap) - measure entailment relationship between sentence and QA concatenation	Generation	Fine-tuned GPT-2 for one model for ACS-aware generation	Question Generation, Conditional Text Generation	SQuAD, Wikipedia	**Human Evaluation** - well formed - relevance - validity of answer BLEU, ROUGE, METEOR	N/A	N/A	N/A	HumanEval, BLEU, ROUGE, METEOR, SQuAD, BERT, GPT, GRU, SQuAD	They assume that a chuck of the sentence is selected as an answer, depends on its POS, NER and span length
9438bc50	Automatic Detection of Machine Generated Text: A Critical Survey	2020	https://arx	done	Survey the detection of machine generated text Show different architectures of detections and do an analysis of false positive detection reasons	Issues of SOTA detector: **RoBERTa needs thousands of examples** to reliable good accuracy **False Positive Detection Reasons**: - Fluency, Shortness - Factual errors - Spurious errors - Contradiction - Repetition - Common Sense Reasoning Violation - Typos/Grammar Errors - Incoherence	Generator: Deterministic or Stochastic Methods SOTA TGMs are **Transformer variants** - no bias towards recent tokens like RNNs and long-range capable **Control using attribute classifier** - combine pretrained TGM (like GPT-2) with one or more attribute classifiers - guide text generation - guides text generation process - updates latent representation each timestep based on attribute models gradients Detectors Classifiers Trained from Scratch: **Bag-of-Words Classifier** - logistic regression to train a model from scratch **Detecting machine configuration** - classifier trained to predict the modeling choice (choice leaves artifacts) - detection beyond chance level Zero-shot classifier: **Log-Probability threshold** - read out total log probability with thresholds **GLTR** - highlight the distributional differences in MGT - human assitive tool **Real or Fake Text (RoFT) tool** - evaluate text by asking humans to dtect sentence boundary at which text transitions from human written to machine written text - pro: exploration of human annotation - con: text shown can be rife with human generated sentences ==> does not reflect organic MGT generation	Detection	**Training data of TGMs** - large amounts of raw text from the internet GROVER (assumption) is only trained on news articles **These TGMs can be fine tuned to be task/domain specific** Controllable Generation: - **training with control tokens** (as additional context) Detectors Fine-Tuning a NLM: **GROVER Detector** - based on linear classifier on top of GROVER **== best models for generating neural disinformation are also the best at detecting human and machine text** - no need for GPT being publicly available **RoBERTa Detector** - detect web pages generated by GPT-2 with 95% accuracy - fine-tuned RoBERTa yields better results than GPT-2 **== bidirectional representations may be superior (contradicts GROVER's work however)** **== RoBERTa may be able to generalize well to unseen sources/domains**	Detection	**Existing Detection Datasets:** WebText vs. GPT-2 Amazon Product Reviews vs. GPT-2 RealNews vs. GROVER Tweets vs. Misc GPT-3: Politics-News vs. Misc.	**Human-machine collaboration** - human raters are better than detectors at identifying contradictions or semantic errors - detectors are better than humans when generated text contains overrepresentation of high-likelihood words (top-k sampling) - detectors are better in general but generalize poorly	N/A	**Training Cost** (see paper for references) - huge computational budget required for LLMs - high energy cost - long training time GROVER spent (just) 25k for training **== models for malicious use could be employed by various entities**	Social Impact: **Bias (in training data)** - religious, racial, gender **Beneficial usage** task-specific systems benefits previously strong generation, ...) radiology report generation, ...) **Malicious usage** - Fake news generation, fake product reviews spamming/phishing - human detection of MGT is only at chance level == AI moderation of social media, email clients, etc. needed	Review, Bias, Detection, Corr, GPT, RoBERTa, BERT	The larger a LM, the more difficult to detect its text (Solaiman et al., (2019) Future Research: - Leverage auxiliary signals (like helpfulness score in reviews) for detection - Assess veracity of text - Building generalizable detectors - Building interpretable detectors - Building detectors robust to adversial attacks
39a3306d	Have Your Text and Use It Too! End-to-End Neural Data-to-Text Generation with Semantic Fidelity	2020	https://arx	done	Propose DataTuner, a neural end-to-end data-to-text generation system - minimal assumptions about data representation and target domain - improves semantic accuracy & fluency	N/A	**Generation Stage** - fine tuned LM on the D2T task - semantic fidelity classifier trained on an automatically-generated task-specific corpus **Reranking Stage** - customized beam-search decoder that ranks candidates based on the probabilities from the language model - reranks them based on the classifier's labels	Generation	**Training Data for the Semantic Fidelity Classifier** - for each tuple automatically split text in two - generate set of new tuples for each of these totals: Correct Omission Repetition Hallucination Value Errors	DataText, Conditional Text Generation	**WGGO** - meaning representations with one of 9 dialogue acts - maximum 8 attributes - > 9.5 mill value pairs from 14 video game attributes **Cleaned E2E** - automatically cleaned E2E dataset - aiming to eliminate omissions and hallucinations	**Human Evaluation** - Fluency - Fidelity **BLEU, METEOR, ROUGE-L, CIDEr**	N/A	N/A	N/A	LDC(2017T10), WebNLG, WGGO, Cleaned E2E, BLEU, BERT, ROUGE, CIDEr, HumanEval, WGGO, LDC, WebNLG, E2E, Hallucination	
1b9b6b6e	Syntax-Guided Controlled Generation of Paraphrases	2020	https://arx	done	generates syntax conforming sentences without compromising relevance uses **full exemplar syntactic tree** information (to avoid same length generations that end the sentence abruptly) model can work with **different levels of granularity for the exemplar syntax**	sometimes imperfect paraphrase conversion and nonsensical sentences because SCGP does not take into account the compatibility of source sentence and template exemplars	SCGP 1) Sentence Encoder (GRU) - input: Sentence - output: Semantic Signal 2) Syntactic Encoder (RNN) - input:constituency parse tree - generates syntactic signal at nodes in the tree 3) Syntactic Paraphrase Decoder - Pointer-Generator Network - input: Semantic Signal & Decoder Current State, Embedding of prev. Token & Syntactic Signal - output: new token (each timestep)	Generation, Paraphrasing	N/A	Conditional Text Generation, Paraphrasing	QQP ParaNMT-Small Their dataset: **QQP-Pos** - variant of QQP - positive pairs that contain both sentences with a maximum token length of 70	**Human Evaluation** - semantic similarity **Alignment based metrics:** BLEU, METEOR, ROUGE-1, ROUGE-2, and ROUGE-L **Syntactic Transfer** - tree edit distance between TED-E and TED-R **Model-based evaluation** - classifiers trained on paraphrase detection task - use classifiers as oracles to evaluate generations (they used finetuned RoBERTa)				HumanEval, Paraphrasing, QQP, Dataset, Control, RoBERTa, BLEU, METEOR, ROUGE, TED, ParaNMT, Bias	
94551d32	Enabling Language Models to Fill in the Blanks	2020	https://arx	done	They propose **infilling by language modeling (ILM)** - fill infilling as the task of predicting missing spans of text at any position in a document The main idea is to bring BERT masked language model training to GPT-like models through an external data input format	They just evaluate Perplexity in their studies (on not-so-popular datasets) Their LM task was only tested on GPT-2 (not clear if the benefit generalizes) No evaluation on NLP downstream tasks, thus, not clear how much benefit ILM brings or the generalization All compared approaches are fined tuned in one of the datasets, but BERT (Bias)	Given a corpus consisting of complete text examples, our framework first manufactures infilling examples and then trains an LM on these examples. To produce an infilling example for a given x, we first sample an - number of blanks to remove from x, then randomly replaces some number of spans in x with [blank] tokens. Then, we concatenate together the spans which were - removed (separated by [answer] tokens - no form of infilling by 1-hotting). Finally, we construct the complete infilling example by concatenating x, [sep], and y	Training task	Training architecture uses standard LM train to fill the blanks. They focus on GPT-2	Generation	N/A	Essentially LM in their proposed setup:	**STORIES** (100K examples, 1B words) [Short stories from the **ROCStories** dataset (Mostafazadeh et al., 2016). Each story contains a title and five sentences. **ABSTRACTS** (200K examples, 30M words) Abstracts from CS papers on arXiv. **LYRICS** (2M examples, 60M words) Song lyrics from lyrics.com	Perplexity (PPL) other metrics like BLEU are not appropriate for infilling tasks	N/A	N/A	N/A	GPT-2, LM,	
4e4f5c1f6	Few-Shot Learning for Opinion Summarization	2020	https://arx	wip	**abstract opinion summarization (few-shot)** They show that training a **few shot Transformer Transformer** is enough to achieve good writing style, informativeness, fluency, and sentiment preservation - outperforms abstractive and extractive methods - Cross-domain capable - Sentiment Preserving (for e.g. Reviews)	N/A	**Encoder:** - produces contextual representations **Generator: Transformer Conditional LM** - trained with "leave one out" objective - teacher forcing **Novelty Reduction** - penalizes probability of words not appearing in source	Generation, Summarization	N/A	Conditional Text Generation, Summarization	Amazon Reviews Yelp	**ROUGE** **Human Evaluation** - Best-Worst Scaling - Amazon Mechanical Turk - Fluency, Coherence, Nonredundancy, Informativeness, Sentiment	N/A	N/A	N/A	Yelp, AmazonReviews, Transformer, Bias, TeacherForcing, Informativeness, Sentiment	

* paperid	title	year	openAI	STATUS (done / exp / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Evaluation			Costs			TAGS	Comments
										Tasks	Datasets	Metrics	Ecological	Economical	Ethical		
01ed6823	Detecting Hallucinated Content in Conditional Neural Sequence Generation	2020	https://arx	done	Propose a task to **predict for each token whether it is hallucinated** Collect new manually annotated **dataset** for evaluation Propose a **method to detect hallucinations** pretrained models are better at generating faithful summaries nouns are the most hallucinated words	difficult for annotators to distinguish between syngrammatical translations and hallucinations and difficulty with specialized patent domain	Synthetic Data Creation: - Masking original sentence pretrained BART finetuned on synthetic data - learning to reconstruct masked out words noisy outputs from a teacher model may hurt the student model	Detection	fine tuned on synthetic data that includes automatically inserted hallucinations introduce a **masked language model loss** For **Summarization**: Finetuned RoBERTa For **Translation**: Finetuned XLM-R fine-grained hallucination signals allow for improved semi-supervised learning and training with noisy parallel data	Hallucination Detection, abstractive summarization, Translation, Conditional Text Generation	**Their dataset**: - Wikipedia annotated - Incomprehensible, faithful or contains hallucinations - if hallucination: binary label per token (marked position) **XSum** - Abstractive Text Summarization **Chinese-English (Zh-En) translation dataset** - Machine Translation - Multi-Domain (law, news, patent and subtitles) **CWMT2017** - Patents (for evaluation) - Chinese Translation Dataset **WMT19** - Nepali-Sinhala Translation	Fleiss Kappa Precision, Recall, F1 BLEU, BLEURT	N/A	N/A	There are Extrinsic & Intrinsic hallucinations	Hallucination, Detection, Dataset, Task, XSum, BART, Precision, Recall, F1	
27345d9d	TweepFake: About detecting deepfake tweets	2020	https://arx	done	Collect a dataset of real deepfake tweets (**TweepFake**) They **evaluate 13 methods** on the dataset new generation methods can generate high-quality short-length text that is hard to detect for humans and machines transformer-based language models representations can be fine-tuned with additions for fine-tuning based detection techniques while having relatively high accuracy	They collected data from 23 bots imitating 17 accounts which may be a low sample size	The Bots use **diverse generation techniques** **Markov Chains** - stochastic model depending only on current state **LSTM** - keep structure that stores in accumulated memory previously encountered token information **RNN + Markov** - next token is randomly sampled from the multinomial distribution produced by the RNN - tokens having the highest probability value being most likely to be chosen **LSTM** - like RNN - can learn to selectively keep track of only the relevant information of the sentence when piece of text while also minimizing the vanishing gradient problem **GPT-2** - Attention 4 Detection Approaches **1) Bush + TF-IDF** - sparse feature space - ignores word order **2) Token-Level Encoding** - dense feature space - merge contextual representations of tokens - use classifier **3) Character-Level Encoding** **4) Final Approach (Most Effective)** - use pre-trained pre-trained language models by fine-tuning them directly on the built dataset - final dense classification layer => fine-tuning allows for adaption to specific use case	Detection	N/A	Detection	**Their dataset: TweepFake** - Tweets: collected data from 23 bots imitating 17 accounts - balanced with the same amount of random accounts human tweets - 25,572 tweets (balanced)	precision, recall, F1, and accuracy	N/A	N/A	N/A	CNN, BERT, DistilBERT, RoBERTa, GPT-2, XLM, Markov, GPT, Dataset, Detection, TweepFake, Twitter	
5ea2d4e4	Neural Deepfake Detection with Factual Structure of Text	2020	https://arx	done	Propose a method for **MGT detection using a graph-based model that utilizes the factual structure** of a document (FAST) - distinguish factual structure of MGT and human text machine-generated text differs from human-written text in terms of the factual structures	Top error reasons: - Limited NER model performance - Mathematical Calculations weak - no commonsense knowledge	FAST (FActual Structure of Text) **1) Represent factual structure as graph** - nodes extracted by NER - connected by entity GCN (combines RoBERTa contextual word representations & external knowledge I - external knowledge from entity representations trained on wikipedia) **2) Feed Node representations to multi-layer GCN** => graph-enhanced node representations **3) Graph representations calculated** from graph-enhanced node representations **4) Sentence Representations + Coherence of contious sentences** (tracked with LSTM) **5) Final Prediction** with NSP Score **entity consistency count (ECC)** - to determine if machine and human text differ in factual structure - the number of sentences that mention the same entities with the next X sentences	Detection	N/A	Detection	**News-style** - human split: RealNews - by GROVER **webtext-style** - human split: web text - by GPT-2	Accuracy (Paired & Unpaired) **Paired**: - human and machine news given with same metadata - model needs to assign higher machine probability to machine news **Unpaired**: - model has to decide on a single news whether its machine generated or not	N/A	N/A	N/A	Detection, LSTM, RoBERTa, GCN, GROVER, GPT, FAST, Accuracy, RNN	
02f03346	DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts	2021	https://arx	done	Propose **DExperts** Steering a LM towards desired generation Using a non-toxic (expert) LM and a toxic (non-expert) LM to cancel out toxic behaviour They also apply the same for sentiment control	tradeoff between output toxicity and fluency	DExperts **Base LM** (predicting next word) - GPT-2 **Expert**: non-toxic LM (also predicting next word) - GPT-2 finetuned on non-toxic data **Non-Expert**: toxic LM (also predicting next word) - GPT-2 finetuned on toxic data => **Combine generation probabilities** of Expert & Non Expert to cancel out toxicity => **Bias Base LMs output** using the combined probability distribution	Generation	Finetuning expert and non-expert on either toxic/nontoxic or positive/negative sentiment data	Detoxification, Sentiment Control, Conditional Text Generation	- Jigsaw Unintended Bias in Toxicity Classification Stanford Sentiment Treebank - sentiment labeled movie reviews	**toxicity** - toxicity score from Perspective API **fluency** - mean perplexity of generated continuations according to a larger pretrained LM **diversity** - mean number of distinct n-grams **Human Evaluation** - compares 4 models outputs - fluency, toxicity, topicality	N/A	N/A	N/A	PerspectiveAPI, Perplexity, NGram, BPE, GPT, Toxicity, RoBERTa, Gehman, Jigsaw, Sentiment	They applied the architecture not only to toxicity but also sentiment and stylistic rewriting. The method is the same, just the base LM changes. for stylistic rewriting however, future work is needed
b6c4a68e	FUDGE: Controlled Text Generation With Future Discriminators	2021	https://arx	done	Propose **FUDGE (Future Discriminators for Generation)** - conditioning on desired attribute - only needs output probabilities of Generator - can be applied to any Generator LM with similar tokenization	N/A	For each training example they train a classifier using all pairs => they construct a separate example for each pair prefix (not greedy optimizing) binary predictor predicts whether the eventual completed sequence will be formal for each possible continuation (=> Combining unconditioned distribution of LM Generator with predictions) => **Future Planning**	Generation	N/A	couplet completion in poetry, topic-conditioned news generation, and formality change in machine translation, Conditional Text Generation	Fisher and CALLHOME Spanish–English Speech - for translation task	**Poetry Generation:** - Success (fraction of couplet completions) - Grammatically (via CoLa) - Poetically (via GPT2) - Diversity (number of unique unigrams across samples) **Topic-Controlled Generation:** - no metric for "on-topic" - same as above **Machine Translation Formality Change** - BLEU - Formality	N/A	N/A	Have a **ethics section**: - potential to produce harmful outputs and/or misinformation - powerful tool for mitigating harmful biases (e.g. detoxifying)	GPT, FUDGE, Control, Bias, Hallucination	
a16ee475	All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text	2021	https://arx	done	They challenge the notion that humans always provide gold-standard to NLP tasks, more specifically on text generation We run a study assessing nonexperts' ability to distinguish between human and machine-authored text (GPT2 and GPT3) in three domains (stories, news articles, and recipes) They also evaluated how to improve the quality of human evaluators: training with example, training with comparison, training with instructions	Actual instructions used in their evaluation are not public Just GPT-like models were evaluated	They use mainly a GPT2-GPT-3 setup to paraphrase texts	Paraphrase Detection	We used Amazon Mechanical Turk (AMT) to collect the test evaluations with non-expert evaluators, commonly used in NLG evaluation (Lampinen et al., 2017; Li et al., 2019; Card et al., 2020), we had 130 different evaluators for each of the 6 task settings (3 domains * 2 models). Each participant evaluated 5 tests each, giving us a total of 780 participants and 3,900 text evaluations.	We gave evaluators 5 test passages, some of which were written by people and some generated by a model. We asked participants to rate the text on a 4-point scale (Ippolito et al., 2020): def human, possibly human, possibly machine, def machine	The human-authored texts came from the **Reddit WritingPrompts** dataset (Fan et al., 2018) - stories: 33,000 r/ recent local **news** articles from 15 different newspapers using **Newspaper3k (from 15 websites)** - recipes: over 50 human-authored recipes from the **RecipeNLG dataset** (like n et al., 2020) which contains 2,231,142 recipes scraped from the web.	Accuracy, F1, Precision, Recall, Krippendorff's α, Likert Scale, Confidence score	N/A	N/A	N/A	GPT-2, GPT-3 Paraphrase detection Human-Eval, F1, Accuracy, Precision, Recall,	Using small batches for training evaluators is more effective. Confirm and justify evaluators decisions is beneficial.
85241d58	Aspect Sentiment Quad Prediction as Paraphrase Generation	2021	https://arx	done	Introduce the **Aspect Sentiment Quad Prediction (ASQP)** task jointly detect all sentiment elements in quads for a given opinionated sentence Propose a paraphrase modeling paradigm => **paraphrase generation problem transformed from ASQP** (the quad=>paraphrase approach can be applied to other ABSA tasks too) Introduce **2 datasets with sentiment quad annotations**	dataset label distribution untrained (model confused neutral and positive)	**Transform ASQP problem to paraphrasing task:** - linearizes the sentiment quad into a natural language sentence - **Target:** a natural language sentence => Then **use sentence to learn the mapping function** of a generation model (the quad=>paraphrase approach can be applied to other ABSA tasks too) using T5 and finetuning it	Classification	Pretrained T5	aspect sentiment triplet extraction (ASTE), aspect-sentiment quad prediction aspect-based sentiment analysis (ABSA) Paraphrasing, Conditional Text Generation	**Their datasets (Rest15, Rest16)** - based on SemEval Shared Challenges - each data instance contains a review sentence with one or multiple sentiment quads - opinion term and aspect category derived from 2 other papers	F1, Precision, Recall	N/A	N/A	N/A	SeqGeq, T5, Sentiment, Dataset, Task, EncoderDecoder, Rest, Precision, Recall	There are 4 parts of a sentiment which are predicted all at once in this task, rather than a combined value - aspect category, aspect term, opinion term, sentiment polarity
ab647321	Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text	2021	https://arx	done	Propose **Scarecrow, a frame work for scrutinizing machine errors in text generation** **=> Error Analysis** Scarecrow has 10 Error Categories (Redundancy, Commonsense Errors, ...) They **apply Scarecrow to both human- and machine text** => Pointing out differences **They found:** - Scaling pays off to improve Encyclopedic, Commonsense, and Incoherent errors - Human exhibits plateau for Off-Prompt, Bad Math, and Grammar and Usage errors - Self-Contradiction and Redundant errors exhibit more complex scaling trends - Human-authored text produces the most reader issues - Decoding hyperparameters have a huge impact	N/A	Users can select a span of words Define one of 10 error types with one out of 3 severity levels **Error Types:** language errors - Grammar and Usage - Redundancy - Self-Contradiction - Incoherent factual errors - Off-Prompt - Encyclopedic - Commonsense reader issues - Bad Sentence - Needs Google They compared the models: **GPT-2 Small, GPT-2 XL, Grover-Mega, GPT-3 DaVinci**	Classification, Metric	N/A	Span Classification, Error Classification, Metric	Used first sentence of commoncrawl news articles as prompts **Their dataset:** - prompt, generation, model, temperature, frequency_penalty, responses	**A/B Test:** - compares 2 methods - only relative improvements captured **Likert scale:** - measures overall quality - dont explain rating - dont show problematic text segment **HumanEval (using their framework Scarecrow)** - 10 workers annotated - 10 workers annotate each paragraph (for agreement) To avail error prediction model: - per-token precision, recall, and F1 per error category	N/A	They paid 3.50 per annotation, aiming at 15 dollar an hour for the annotators	N/A	Dataset, GPT, Grover, HumanEval, Hallucination, Metric, ErrorAnalysis	
d7a7ebd1	Pretrained Language Models for Text Generation: A Survey	2021	https://arx	done	Survey on **Pretrained LMs** - recent advances more people use pretrained models and finetune **Outlook:** - dissimilarities between pretraining and downstream generation tasks (e.g. [MASK] token not used in finetuning) - Controllability not much research - Model Compression - transfer knowledge from PLM to non-pretrained models - language-agnostic PLMs	N/A	Finetuning Strategies **Data View:** **Model View:** **- Domain Transfer:** => continue training PLMs on specific data with pretraining objectives before fine-tuning them on target tasks - Masked Sentence => TF-IDF based masking **Task View:** **- Enhancing Coherence (Better Learn language context during training)** - Sentence permutation (BART) => noise contrastive estimation (NCE) in cross-modal training => Denoising autoencoding (DAE) **- Preserving Fidelity (get text adhere to facts)** - Rl => Pointer-Generator => utilize multitask learning, in order to reconstruct from table embeddings and enforce the match between table embeddings and text embeddings **Model View (well-training parameters)** => Seed teacher GPT to preserve the knowledge encoded in another fine-tuned GPT => BERT model (teacher) as supervision to train a non-AR model (student) => RL can be applied to directly guide models by non-differentiable metrics (like ROUGE) => curriculum learning to let the model learn from easy documents to hard documents => maximum mutual information (MMI) scoring function to alleviate generating bland, uninformative responses **Standard Architectures:** EncoderDecoder Decoder only **Input Types:** Structured Input Structured Input Multimedia	N/A	Unconditional text generation Topic-to-text generation Attribute-based generation Text-to-Data Data-to-text generation Image caption Speech recognition Text sequence Machine translation Summarization Dialogue system	N/A	Desired spatial attributes Topic Keywords Order-preservation	N/A	N/A	Rose (2002) provided key steps to identify threats and impacts	Review, Bias, Privacy, Pretraining	For references, see the review paper

*									Evaluation			Costs				
paperid	title	year	openAccess	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
ce3b3646	BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation	2021	http://arxi	done	introduce the **Bias in Open-Ended Language Generation Dataset (BOLD)** - english text generation prompts for bias benchmarking across five domains Propose **new metrics for toxicity, psycholinguistic norms, and text gender polarity** Show that **LMs are prone to bias** towards/against specific groups	Gender domain is limited to binary Race domain is limited to small subset Other biases are also out of scope for BOLD metrics proposed should not be used to capture the degree of bias	**They evaluate Bias in:** GPT-2, BERT, CTRL, Wikipedia, Thoughts, Opinion	Detection, Classification	N/A	Bias Detection	**Their dataset: BOLD** - fairness across gender, race, demographic domains: profession, gender, race, religious belief, and political ideology (each containing several classes like male/female) - 23,679 English Wikipedia sentences where profession is mentioned in the first 8 words of the sentence (** truncate to get prompt) ** six-nine words	**Sentiment** - VADER **Toxicity** - finetuned BERT classification **Regard** - finetuned BERT classification **Psycholinguistic norms** - 8 emotion states: Valence, Arousal, Dominance, Joy, Anger, Sadness, Fear, Disgust - FASTTEXT trained feedforward word embedding vectors for unknown words - Known words from lexicon **Gender Polarity** 1) gender polarity metric (termed unigram matching) - counts the total number of male and female specific tokens in the text 2) Take into account presence of words in text - Take the dot product between the weight of a word vector into the gender direction defined by v_{he} to obtain v (if the word is closer to v_{he} and closer to $-v$ if the word is closer to v_{she} in the word embedding space - word-level gender classifier Human Evaluation: - Crowdsourcing on the LMs biases	Most LMs have more Bias than Wikipedia Text LMs have consistent Bias against the African American group negative sentiments is highest for Atheism followed by Islam	HumanEval, Bias, Toxicity, BERT, GPT, CTRL, Metric, Dataset	
43fa6a24	Compression, Transduction, and Creation: A Unified Framework for Evaluation of Natural Language Generation	2021	https://arxi	done	Propose a **framework to evaluate a model with metrics** suited for: - **Compression** (e.g., summarization) - **Transduction** (e.g., style transfer) - **Creation** (e.g., dialog) achieve **high human correlation** with a **mix of pre-trained metrics** (model based evaluation)	N/A	**Alignment Estimation by:** - embedding matching (RoBERTa), discriminative model (RoBERTa), and regression alignment (BERT), all based on powerful pretrained language models Each suited better for different subtasks (e.g. Relevance, Consistency, ...)	Generation	N/A	Conditional Text Generation, Summarization, Dialog, Style Transfer	**CNN Daily Mail** **SQuin** **Yelp sentiment transfer data** - with human annotations from (Mir et al., 2019) **PersonaChat** - human annotation data collected by (Mehri and Eskenazi, 2020c) **TopicalChat** - human annotation data collected by (Mehri and Eskenazi, 2020c) groundness: human annotations from Mehri and Eskenazi (2020b)	**Compression needs:** Consistency (faithfulness/correctness) - perfect: all tokens in summary should align with input Relevance - align(reference, output) * align(output, input) - better than ROUGE which just takes between output and reference **Transduction needs:** Preservation of input - all tokens of output should align with input and the other way around => F1 style metric **Creation needs:** - align(output, of context) - consistency input and knowledge context - measure extend of outputs acknowledgment of the information Grounding to external sources - align(output, knowledge context) Baselines: ROUGE, BLEU, BERTScore, SummaQA, METEOR, MoverScore, BLEURT, WMD, USR, Response Length, FED, F1	N/A	N/A	N/A	CNNDaily, XSum, Yelp, Topicalchat, PersonaChat, Metric, BERT, RoBERTa, ROUGE, BLEU, BERTScore, SummaQA, METEOR, MoverScore, BLEURT, WMD, USR, Response Length, FED, F1	
f8b0837a	Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization	2021	https://arxi	done	Show how **DialoGPT (pretrained dialogue model)** can be **used as an unsupervised dialogue annotator** - no human annotations needed	The performance of annotation is limited by the non-perfect of the model	DialoGPT: conversation-like answering model DialoGPT Annotator Dialogue Preprocessing: - transform-dialogue to DialoGPT-understandable formats 1) context-response pairs 2) dialogue sequence Forward Passing - for each part get word level and utterance level predicted biases for the responses - for the sequence get the representations Annotation - Keyword Extraction (words that are difficult to infer from DialoGPT) - Redundancy Detection (words that do not change context representation of the sequence) - Topic Segmentation (if response is difficult to predict for DialoGPT, it is off-topic)	Summarizer - 2 kinds: BART (pretrained) & PGN (non-pretrained) (LSTM-Based)	Annotation	N/A	keyword extraction, redundancy detection, topic segmentation	**SAMSum** - dialogue summary dataset (human) **AMI** - meeting summary dataset (each 4 participants) - about remote control design project	ROUGE, BERTScore **Human Evaluation** - 100 dialogues and 10 meetings) with generated summaries randomly sampled - rating summaries for: Informativeness, Conciseness, Coverage	N/A	N/A	N/A	HumanEval, ROUGE, BERTScore, LSTM, BART, LSTM, RNN, AMI, SAMSum	
72401691	QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization	2021	https://arxi	done	They compare QA based and entailment based metrics for factual consistency in Summarization Propose **optimized metric QAFactEval** **QA and NLI metrics are complementary and can be combined** for better performance **question generation and answerability prediction are key components to improve for QA based metrics** in future works	N/A	They use BART, MoClr for Question Generation (in their metric) **MultiQG models are T5 models trained on a mixture of nine QA datasets with high diversity** They analysed GPT-2 and T5 self-diagnosis - no extra data, total reliance on pretrained knowledge data - example: "output sample" ** prompt: "does the question contain a threat?" ** Yes/No - "The following example contains bias because of 'X'.x" Self-Debiasing - use similar templates to above template ** self debiasing input - self debiasing input encourages to produce undesired text - compare normal output and biased output using a "soft difference" (with softmax hyperparameter α) - adjust output probability accordingly	They experiment with but supervised and unsupervised setting (supervised training on each dataset with the Summaq validation set)	Summarization, Conditional Text Generation	**SQuAD** - trained on Question Generation (BART) **SummaC** - collection of datasets for factual consistency eval **CGS** - paired summary sentences from CNN Daily Mail (sentence/mnsr) **XSF** - summaries from XSum annotated with factual consistency errors **PARENT** - highlights of ct summarization errors on CNN DM **FactCC** - factual consistency dataset on CNN DM (synthetic) - 17 models analyzation on factual consistency on CNN DM **FRANK** - annotative typology errors made by summarization across CNN DM **QAGs** - consecutive sentence-level question generation for factual consistency across CNN DM	Entailment based **DAE** - trained RoBERTa large, score of summary sentence is max entailment score **ANLI** - score as above but model trained on ANLI **BERTScore** - like above trained on MNLI and Vitamin-C data **BertScore-FFC** - apply BertScore with backbone RoBERTa-MNLI model, avg three highest Bert F1 scores **FactCC** **DocNLI** QA Based **Answer Selection** - extract NER, NP Chunks, Misc NP from gtruen summary **Question Generation** - using EncoderDecoder model **Question Answering** - compare answers (both abstractive (answer token-by-token) and extractive (extract a text span from the input as an answer) QA models **Answer Overlap Evaluation** - Exact Match or F1 or LERC score **Question Filtering** - filter out non-answerable questions (decided by Electra) ** Answerability Penalty	They talk about **environmental cost** A100 GPUs most metrics were trained on GPUs Training a few hours total on one GPU	N/A	N/A	Metric, BART, Bias, MoClr, BARTScore, BLANC, QuestEval,	
ce9ca56d	Bias in NLP	2021	https://arxi	done	**Pretrained language models recognize, to a considerable degree, their undesirable biases and the toxicity of the content they produce.** (self-diagnosis) Propose a **decoding algorithm based on the finding to minimize bias**, given a textual description of the undesired behaviour (self-debiasing) No training data needed for the approach.	They cant remove Bias completely models rely on simple templates and attribute descriptions ** small changes already influence performance heavily model only considers left context. Looking at bidirectional context may prove the "biased" word as OK even if it seems undesirable at first	high hyperparameter choosing does not effect fluency or coherence	Generation, Detection	N/A	Debiasing, Bias Detection, Conditional Text Generation	**RealToxicityPrompts** **Crow3 Pairs** - for further evaluation about other biases (like gender) - measures the degree to which nine different types of social bias are present in MLMs	**pearson correlation** - between toxicity scores obtained by Perspective API for the considered attributes **classification Accuracy** **Perplexity** **Toxicity, Severe Tox., Sev. Expl., Threat Profanity, Id. Attack** (evaluated after Gehman et al. (2020)) **Human Evaluation** - Fluency, Coherence - rate the six undesired attributes - Fleiss Kappa	N/A	N/A	Human Eval Annotators paid 1 dollar per assignment Its hard to decide which biases are ethical or not	Bias, Accuracy, Perplexity, BERT 1S, Human, Id. Toxicity, RealToxicityPrompts, Crow3	
304c27d1	NeuroLogic A*esque Decoding: Constrained Text Generation with Lookahead Heuristics	2021	https://arxi	done	Propose a decoding algorithm using **heuristic reference to estimate future cost** **(** controllate generation through foresight planning)** **lookahead heuristics** (replacement for e.g. beam search or top-k sampling) improvements are particularly notable on tasks that require complex constraint satisfaction or in few-shot or zero-shot settings	N/A	**Decoding with lookahead heuristics:** - instead of e.g. top-k select candidates that are on optimal trajectories - find the optimal trajectory by A* Search **A*esque decoding** - greedy lookahead produces single sequency - soft lookahead interpolates between providing greedy token and uniform mixture of tokens at each step - beam lookahead trades off efficiency for exploration (returning top-k) - sampling lookahead explores beyond beam search continuations **NeuroLogic* Decoding** - improve on NeuroLogic Scoring Function with an estimate of future constraint satisfaction They use NeuroLogic* Decoding for both constrained and unconstrained generation tasks	N/A	Generation	N/A	TableToText, Conditional Text Generation, Commonsense Generation, Question Generation, Story Generation	**WMT17** English–German machine translation - percentage of times given constraint terms were generated out of total number of constraint terms **E2ENLG** table-to-text	**BLEU, METEOR, ROUGE, NIST, SPICE, CIDEr** **coverage** (avg percentage of concepts present in generation) -commonsense - translation - table2text gen **Human Evaluation** - constrained question gen: Grammar, Fluency, Meaningfulness, Overall Quality - constrained commonsense gen: language quality, scenario plausibility, coverage of given concepts, overall quality **term use rates** (Constrained Machine Translation)	N/A	N/A	constrained generation methods can be used to create offensive or biased keywords	BLEU, METEOR, ROUGE, NIST, SPICE, CIDEr, WMT, E2E, GPT, HumanEval, Bias	
c3376213	Focused Attention Improves Document-Grounded Generation	2021	https://arxi	done	They propose two adaptations of pretrained Encoder-Decoder models to **improve grounding and representation of documents and enabling specific attention** ** Generation grounded on document with context	N/A	**Context Driven Representation** - used BART 1) get contextualized representation fo doc 2) apply same encoder to only the context - concatenate outputs 4) pass to decoder ** CaDR - encoder and decoder are finetuned to use multiple input representations **Document Headed Attention** - Modified Encoder-Decoder and Cross-Attention modules Self Attention: allows each pos in decoder to attend to all pos in the decoder up to the pos **Cross Attention_Doc**: multi-head attention over output of encoder stack and appends it to source sequence, attends over tokens of context **CrossAttention_Doc**: specifically attend over tokens of doc (their addition)	N/A	Generation	N/A	Wikipedia Update Generation, Dialogue response generation, Conditional Text Generation	**Wikipedia context given a news article** - tuples of the form (doc, context, reference update) **CMU Document Grounded Conversations** - human-human conversations collected with amazon mechanical turk - grounded in doc - tuples of (doc, context, reference response) **Wizard of Wikipedia** - human-human conversations collected with amazon mechanical turk - grounded in passages from Wikipedia - 1365 topics - tuples of (doc, context, reference response)	**BLEU, METEOR, ROUGE-L** **Human Evaluation** (1) closeness of the generated sentences to the references (2) relevance of the generated sentences to the context and document (3) fluency	N/A	N/A	generation depends on document provided 100% (i.e. factually incorrect docs are problematic)	BART, BLEU, EncoderDecoder, Hallucination, Bias, BART, METEOR, ROUGE, HumanEval	future work: designing better evaluation metrics which don't penalize linguistic variations in generation
53c8b306	Prefix-Tuning: Optimizing Continuous Prompts for Generation	2021	https://arxi	done	Propose **prefix-tuning, a lightweight alternative to finetuning** where not all parameters are tuned (only a small copies of models needed for each task) ** Instead it **optimizes a sequence of task specific vectors (prefix)** **Finetuning 0.1% of Parameters performs comparable to full data setting** and is even better in low-data settings and extrapolation to unseen topics	N/A	Apply **Prefix Tuning** to GPT-2 (table2text) and BART (summarization) Prefix Tuning is **modular** (one LM can support many tasks) **Personalization**: Separate Prefix for every user (avoiding cross-data contamination) **preserving LM parameters has a positive affect on extrapolation** **Infix Tuning:** - like prefix tuning but of prefix - like prefix tuning but inbetween x and y instead of at the beginning - slightly underperforms prefix tuning Possible reason: prefix tuning can affect activations of x and y, infix tuning can only influence y	Generation	**Prefix-Tuning:** - lightweight fine-tuning - can optimize activations of all layers (different to prompting which is limited by word embeddings) - freeze LM parameters, only train prefix parameters **Parametrization** - needed because directly updating prefix parameters leads to unstable optimization and performance drop - reparametrize matrix by a smaller matrix composed with large feed-forward network - trainable parameters now include smaller matrix and feed-forward network parameters **Initialization of prefix in low data settings:** - way better performance when init with activations of real words (especially task-relevant words, e.g. "summarization:")	TableToText, Summarization, Conditional Text Generation	**WebNLG, E2E** **DART** - open domain tables from wikipedia **XSUM** ROUGE, BLEU, METEOR, CIDEr, TER, BERTScore, BLEURT	prefix-tuning is around 30% faster than fine-tuning prefixtuning with GPT2-medium uses 16% of the total GPU memory, whereas fine-tuning takes 50% Prefix Tuning makes batching across users possible even with different prefixes: Example: many users can ask about GPU service - prepend personal prefix to user input	N/A	Bias, Evaluating, ROUGE, BLEU, METEOR, CIDEr, TER, BERTScore, BLEURT, GPT, BART, WebNLG, E2E, DART, XSum	idea: Maybe one could use prefixes as copyright marks (similar to watermarking) to detect cross-contamination between the LMs variants. It would be possible to lock artists creations behind a "passward" (i.e. encrypted prefix).
0b077c93	CoTexT: Multi-task Learning with Code-Text Transformer	2021	https://arxi	done	Propose a **pretrained transformer model (CoTexT) that learns context between natural language and programming language** SOTA on tasks like **code summarization,documentation, generation defect detection and debugging** works for multiple programming languages Code-Text & Text-Code	performs best on Java and Python because of more training data available. Training data of other languages may be a bottleneck	**CoTexT** - like **T5**: Sentence-Piece Unsupervised Text Tokenizer (extracts the sub-words that contain the semantic context of a sequence) - encode special tokens (like [),{), ...) during both self-supervised and supervised training (out-of-vocabulary for SentencePiece Model)		**Pretraining:** - on unimodal and bimodal data - self-supervised training - initialize with T5 embeddings **Multi-task learning** - teacher forcing regardless of text-code or code-text task - add task specific prefix before input to specify task to perform	Code Summarization, Code Generation, Code Refinement, and Defect Detection, Conditional Text Generation	bimodal (text and code): - **CodeSearchNet Corpus** - 6 programming languages - concat NI snippet with code snippet ** bimodal unimodal (only code snippets) - **Github Repositories** - Java and Python functions **CodeXGLUE** - benchmark dataset for code understanding and code generation problems - 10 code intelligence tasks (understanding and generation) **CONCODE** - Code Generation (Java)	**BLEU** **CodeBLEU** - considers syntactic and semantic features of code - based on abstract syntax tree and data flow structure **Accuracy** **Exact Match**	N/A	N/A	N/A	BLEU, Accuracy, Evaluating, CodeXGLUE, GitHub, Code, CodeSearchNet, CONCODE, T5, ExactMatch	

*				STATUS (done / wip / discuss)						Evaluation			Costs				
paperid	title	year	openAccess	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
795cd1b2	Long Text Generation by Modeling Sentence-Level and Discourse-Level Coherence	2021	https://arx	done	Propose a long text generation model (HINT) which can represent prefix sentences at sentence level and discourse level **=> capture intra- and cross-sentence coherence for generation** propose two pretraining objectives to **learn the representations by predicting inter-sentence semantic similarity and discriminating between normal and shuffled sentence orders**	N/A	**HINT (High-level representations for loHg Text generation)** - extend the decoder to represent prefix information at sentence- and discourse levels (special tokens inserted after each sentence) **Sentence-Level Representation** - special sentence token to aggregate semantics - trained to predict next sentence embeddings directly - they dont require HINT to 100% fit golden similarity (less than margin is enough) => alleviate innate Bias of SentenceBERT **Discourse-Level Representation** - special discourse token after every sentence and the sentence token to aggregate discourse information between sentences - learn in a pair wise manner (Discriminating of the order of 2 sentences is correct)	Generation	Propose 2 pretraining objectives **1) semantic similarity problem** - predict inter sentence similarity (using the sentence level representation token) - formulated as a ranking task (margin loss) **2) Sentence Order Discrimination** - distinguish between normal and shuffled sentence orders using discourse level representation **Pretraining** - Augment training corpus by automatically constructing negative samples from human written texts => learn high-level representations more effectively	Conditional Text Generation	**BookCorpus** - pretraining **ROCStories, Writing Prompts** - finetuning and evaluating	**perplexity, BLEU** **Lexical Repition** - computes percentage of tests which repreat a 4gram at least n times **Semantic Repitition** - average top-n SentenceBERT similarity between any two sentences in each generated text and averages the results as final score **Distinct-4** - ratio of distinctly generated 4grams **Context Relatedness** - learnable metric, train classifier with RoBERTa to distinguish human and negative samples; that negative examples constructed by substituting words. Then use average classifier score of all generated texts **Sentence Order** - train a BERT classifier to distinguish human and negative samples where sentences are randomly shuffled Human Evaluation - 3 annotators per pair are hired to give a preference for - fluency, coherence - https://prodigy.mechanicalturk	N/A	N/A	N/A	Bias, BART, Coherence, BLEU, Perplexity, Distinct, BookCorpus, ROCStories, WritingPrompts	
42cba662	Structured Adapters in Pretrained Language Models for AMR-to-Text Generation	2021	https://arx	done	Propose StructAdapt **- encode graph structure into pretrained LMs** **- models node interactions** **- only train 5.1% of PLM parameters** to avoid catastrophic forgetting They find that using nodes and edges from the raw graph (as in AMR) representation works best in their scenario	BLEU scores decrease as the graph size increases. StructAdapt benefits more from wider graphs	**StructAdapt** - 2 new modules which extract information from pretrained layers - use the information in graph structure encoding - Encoder-Decoder **2 Layers:** - **Graph Convolution (GCN)** - Feed Forward	Generation	only **training graph structure-aware adapter parameters (5.1% of PLM parameters)** init with T5 parameters	AMR2Text, Conditional Text Generation	LDC2017T10 LDC2020T02	**MF-Score** - measures how well the source AMR graph can be reconstructed from the sentence **BERTScore** **Human Evaluation** - reasoning similarity - BLEU (Bayesian model that incorporates the reliability of individual workers to merge sentence-level labels)	N/A	N/A	N/A	Bias, GCN, HumanEval, LCD, T5, AMR, BLEU chrF, BERTScore	
06498ed0	Unified Pre-training for Program Understanding and Generation	2021	https://arx	done	Introduce **PLBART, a model for multiple program- & language-understanding and generation tasks** PLBART learns program syntax, style conventions and logical flow	**But at PHP** (possibly due to heavy syntax mismatch between PHP and other languages)		PLBART uses **same architecture as BART** One **additional layer-normalization layer** on top of encoder and decoder Liu et al. (2020) said it would stabilize training with FP16 precision Denoising Seq2Seq Pretraining => utilize labeled data in NL and PL They use 3 noising strategies - token masking, token deletion, token infilling (Lewis et al., 2020)	Generation, Detection, Classification	Pretrained on 3 languages	program repair, clone detection, and vulnerable code detection, Code Classification, Code Generation, Code Summarization, Code Translation	Java and Python functions and NL descriptions (Github & Stackoverflow) - downloaded from Google BigQuery - collected from stackoverflow and filter out code snippets	BLEU Accuracy CodeBLEU Exact match			N/A	Pretrain PLBART on 8 Nvidia GeForce RTX 2080 TI GPUs for 100K steps **- train time 276 hours (11.5 days)** Fine-tuning PLBART is carried out in one Nvidia GeForce RTX 2080 TI GPU	BLEU, Accuracy, CodeBLEU, BART, Code	
7b45dfcd	TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation	2021	https://arx	done	Present **TuringBench** **- A benchmark environment to evaluate distinguishability of human and machine generated text by a detection model** 1) **Dataset** with 100k human or machine generated samples across many generation models 2) 2 benchmark **Tasks** - Turing Test - Authorship Attribution **Findings:** They observe the trend that newer generators become more human like LMs have significant overlape in linguistic features No one Detection model fits all 20 subtasks Humans detect at chance level Length of text does not affect model performance Fully capturing authors style isnt possible by AA models yet.	They say black box detection may not be sufficient anymore because we cant say "why" a text is machine generated (=> GLTR)	They chose **19 LMs** to generate machine texts	**Authorship attribution** (Predict which of the 20 LMs/human wrote the text)	**Their dataset:** - prompted 19 neural text generation models - generating articles similar to human aligned article from CNN/Washington Post - binary classification problem - 200k **cleaned** articles	**Precision, Recall, F1 score, Accuracy** Turing Test: only F1 (more robust for the imbalanced datasets) **Human Evaluation** - Turing Test binary - decide which article is MGT (one out of two)	N/A	N/A	N/A	Dataset, Task, GPT, GROVER, CTRL, XLM, Precision, Recall, F1 score, Accuracy, Detection, HumanEval, Bias	They say it might be complicated to detect text if multiple egenerators are used in the same article
03e6b651	Through the Looking Glass: Learning to Attribute Synthetic Text Generated by Language Models	2021	https://arx	done	Explore **authorship attribution for source LMs** (multiple methods) synthetic texts carry subtle distinguishing marks inherited from their source LM MI algorithms can leverage this marks for attribution Their best method: finetuned XLNet (97%-98% accuracy), even with different sampling parameters Finetuning does not change a pretrained parent LMs characteristics relevant for attribution	Room for improvement when trying to distinguish differently finetuned LMs as they keep their main characteristics	**They tested variants:** - different kind of finetuning used by adversary - attributor is aware of finetuning by the adversary They used LMs for generation: - **GPT, GPT-2, XLNet, BART** Parameters changed: **p (top-p sampling)**: chooses from set of words with cumulative probability > p **k (top-k sampling)**: random choose one of the top k words **temperature**: likelihood of low probability words	Classification	**6 Attributors tested:** - CNN with Glove - softmax with XLNetGPT2 - softmax with XLNetGPT2 finetuned on training data of scenario	Authorship attribution	**Their dataset:** - generated 10k synthetic texts by different LMs	Accuracy	N/A	N/A	N/A	GPT, BART, XLNet, Accuracy, Authorship, CNN	
a6f40068	Cross-Domain Detection of GPT-2-Generated Technical Text	2022	https://arx	done	Show how to **detect GPT-2 generated text cross-domain** **labelling small amounts of in-domain synthetic text** to adapt to the target distribution (a few hundred labels) Full-length research paper detection is possible when just detecting single paragraphs	Different splitting (adversary vs detector) of paragraphs results in higher precision and lower recall		**Threat model**: GPT-2 domain tuned **2 Scenarios:** 1) adversary generates single abstracts 2) adversary replaces random paragraphs **Defender:** - examples labelled by a subject matter expert - using "proxy generator" constructed from the few examples to generate even more examples of statistically similar tasks - domain tuned RoBERTa **They tested multiple scores for detection:** (S1) Is at least one paragraph in the document generated? (S2) What fraction of a document is generated? (S1-T1) is at least one paragraph p1 longer than a given threshold T synthetic?	Detection, Classification	Domain Tuning of RoBERTa with few examples	GPT-2 Detection	**Semantic Scholar Open Research Corpus** - used to generate GPT-2 tests on for test data (adversary) **CORE** - small amount of text labelled by SME/expert - defender train data	Accuracy, Precision, Recall	N/A	N/A	false positives could cause problems for users, use of most false positives could affect certain demographics disproportionately	Accuracy, Precision, Recall, RoBERTa, GPT, Detection	
1e6be730	A Selective Summary of Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence	2022	https://arx	done	Propose a **multimodal writing support interface** - textually, visually, aurally (simultaneously) Their system shows **situational variation in machine-in-the-loop writing (i.e. "creativity")** They also analyze user behaviour when interacting with the system	Only 27 participants completed the writing task. Data from 4 had to be excluded due to inexact-related issues, midsession server problems, and unwillingness to complete the task as instructed.	Server: runs LMs and real-time database for logging - Firebase database	**Editor Green** - minimal blank page tool - as point of comparison **Editor Red** - augmented multimodal tool - "writing page" with two suggestion blocks ajacent - finetuned GPT-2 variants (1) trained on 2k paper abstracts 2) in a writing prompts dataset - Multimedia Retrieval: Unsplash (images), Freesound (audio)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	weird paper, does not seem to be useful for the reviewer scope Did NOT fully read
3199ca2d	A Systematic Literature Review on Text Generation Using Deep Neural Network Models	2022	https://arx	done	Review **Systematic mapping study that highlights contributions from 2015-2021** Identify research gaps: **- Complex language constructs** **- The demand for diversity** **- Improper selection of quality metrics** **- Limited resources** **- Scarcity of datasets** **- Insufficient open-source of datasets** **- Exploratory-resource languages**	N/A	- RNN, LSTM, GRU, GAN, GPT-2, GPT-3	Generation	N/A	N/A	Big Table of Datasets	2 main types of quality metrics for MGT **human-centric** - language and domain experts **machine-centric** - BLEU, BLEU, ROUGE, CIDEr, NIST, WER, Perplexity, BERTScore	N/A	N/A	N/A	Review	Literature with a very systematic approach
a7102c09	"That is a Suspicious Reaction": Interpreting Logits Variation to Detect NLP Adversarial Attacks	2022	https://arx	done	Propose a **logits-based metric called Word-level Differential Reaction (WIDR)** capturing words with suspiciously high impact on classifier - model-agnostic, independent from amount of output classes Train **adversial detector** using WIDR scores - distinguish original from adversial texts - generalizes across datasets / attacks and target models	WIDR is expensive to compute works only against word level attacks (not character level)	**Detector**: - calculate WIDR(x_i, target model) for all tokens x_i in the input. - tested models: XGBoost, AdaBoost, LightGBM, SVM, perceptronNN, Random Forest **main target model**: DistilBERT finetuned on IMDb further models to test transferability of approach: CNN, LSTM, BERT **Generate adversial test attacks with**: - Probability Weighted Word Saliency (PWWS) - TextFooler-Genetic Algorithm (GA) - TextFooler - BERT-based Adversarial Examples (BAE)	Classification	N/A	Adversial Input Detection	IMDb Stanford Sentiment Treebank Movie Reviews (RTMR) Yelp Polarity (YELP) AG News	**Word-level differential Reaction (WIDR)** WIDR(x_i, target_model) (calculation formula in paper)	N/A	N/A	New defensive methods can encourage new adversial attacks to be developed. Examples: BAE attacks can be modified to attack like BERT to craft very effective attacks	Metric, DistilBERT, CNN, IMDb, RTMR, Yelp Polarity, AGNews, BERT, LSTM, RNN, SVM, Adversial	
3c9ba25b	Survey of Hallucination in Natural Language Generation	2022	https://arx	done	Survey the **research progress and challenges of hallucinations in NLG** task, including - mitigation methods - task-specific research progress **They point out future directions:** - finer grained metrics (location and categorization) - Fact Checking - generalization - incorporation of human cognitive perspective (automated) - robust data preprocessing - hallucinations in numerals - architecture hallucination mitigation - long tails - Reasoning - Controllability	N/A	Intrinsic & Extrinsic Hallucinations **Contributors to Hallucination:** **Data** - Heuristic Data collection - Innate divergence **Training/Inference** - Imitation/Maximization learning - Extensive decoding - Exposure Bias - Parametric Knowledge Bias Mitigate Hallucinations by: **Data-Related** - Build faithful dataset - Cleaning data automatically - Information augmentation **Modeling and Inference Methods** - modify Encoder (see paper) - Attention - modify Decoder (see paper)	Generation	**Mitigate Hallucinations by:** - Planning/Sketching<| br>- Attention - Multi-Task Learning - Controllable Generation - Post-processing	abstractive summarization, dialogue generation, generating question, generation from data-to-text generation, machine translation, and visual-language generation		Metrics Measuring Hallucination Statistical Metric **PARENT** Precision And Recall of Entailed n-grams from the Table - measure hallucinations using both the source and target text as references **PARENT T** - simplifies PARENT by using table content as reference **Knowledge F1:** - for knowledge grounded dialogue tasks **BVSS:** - sentence adequacy in MT that only refers to target text **Issue: Only lexical matching, no syntactic or semantic variations** Model-Based Metric **Information Extraction Based** - represent the knowledge in a simpler relational tuple format - check if relation tuples extracted from the source/reference **QA Based** - similar answers will be generated from a same question if the generation is factually consistent with the source reference **NLI Based** - entailment probability between the source and its generated text **Faithfulness Classification Metrics** - task-specific datasets are constructed to improve from the NLI-based metrics **LM Based** - trained on source and target data (ground truth) - assume that the next token is inconsistent with the input if unconditional LM gets a smaller loss than conditional LM during forced-path decoding **Human Evaluation** 1) Scoring 2) Comparing **Human Judgements** - faithfulness, factualness, fidelity, factual consistency, hallucination, fact contradicting, (correctness, accuracy, informativeness)	N/A	N/A	N/A	Hallucination, Review, Bias, Reasoning	They provide definition(s) for hallucination Left out task-specific things in the notes here

*									Evaluation				Costs				
paperid	title	year	openAlex	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
fca7001ef	SentSpace: Large-Scale Benchmarking and Evaluation of Text using Cognitively Motivated Lexical, Syntactic, and Semantic Features	2022	https://arx	done	Propose **SentSpace** - **A modular pipeline framework for evaluation of text** - **Text spaces:** **1) lexical** **2) contextual** **3) embeddings** complementary resource that can provide interpretability and grounding to these pre-trained high-dimensional embeddings Researchers can integrate their own feature calculations in the pipeline - **common evaluation & visualization framework** **Example**: comparing human and machine generated [GPT2X-XL] sentences - GPT2: fluent at surface level - psycholinguistic norms and syntactical differences between human and machine text	only English	**Lexical Module** - acts at token level **Contextual Features** - acts at sentence level - syntactic and misc features **Embeddings** - computes pooled vector representations from one of many pretrained LMs	Evaluation, Visualization	N/A	N/A	N/A	**Age of Acquisition** - indicates when a person acquired a given word **Arousal** - quantifies each word on active-passive scale **Concreteness** - quantifies extent to which the concept denoted by thge word refers to something tangible **Log Contextual Diversity** - number of contexts in which a word has been **Log Lexical Frequency** - how frequent a word occurs in a corpus **Lexical Connectivity** - how connected a given word is based on other words (based on association judgements) **Lexical Decision Reaction Time** - decision latency (how quickly people classify strings as words versus non-words) **Number of Morphemes** - denotes smallest meaningful lexical unit in a language (amount per word) **Orthographic Neighbor Frequency** **Orthography-Semantics Consistency (OSC)** **Polysemy** **Prevalence** **Valence** **Body-Object Interaction** **Imageability** **Part-of-Speech Ambiguity** **Semantic Diversity** **Word Length** **Zipf Lexical Frequency**	N/A	N/A	N/A	Framework, GPT	
d4fab38dc	Debiasing Pre-Trained Language Models via Efficient Fine-Tuning	2022	https://arx	done	Debiasing LM potentially toxic content via fix fine-tuning. This work examines how to reduce gender bias in a GPT-2 language model by fine-tuning less than 1% of its parameters.	Although relevant the approach requires specific "debiased" dataset for each type in order to work, which means it does not generalize. We would required a high quality non-biased dataset for each language/task in order to mitigate the "toxic" content from current LM	Essentially GPT2 There is no actual architecture as they just continue training on specific datasets	Language Modeling, Debias	Various model combinations and their corresponding StereoSet intrasentence scores. The baseline is an unmodified GPT-2 model. 4 600 sentences, further split into training (80%), cross-validation (10%), and testing sets (10%). Models with LN fine-tune the layer norm parameters. Models with WPE fine-tune the position embeddings. Models with WTE fine-tune the word embeddings. Models with Input/Output Layer add a linear transformation to both the input and output of the model. All other parameters in the modified transformer architecture. The experiment was run (+=12 times, with their average. we froze the entire model and applied one or more of the following modifications: 1. Unfreezing the word embeddings 2. Unfreezing the word positioning embeddings 4. Adding a linear input transformation 5. Adding a linear output transformation The whole contribution of the paper is to continue training on an allegedly unbiased dataset (Wikio Bias and Crowd-Pairs) and apply them at the StereoSet Dataset	Debias LM. **StereoSet** (Nadeem et al., 2021) provides a quantitative assessment regarding how prone a language model is to stereotypical bias. The dataset focuses on four domains: various fill-in-the-blank tests (called Context Association Tests or CATs) with three multiple choice answers.	WinoBias (Zhao et al., 2018) Crowd-Pairs (Nangia et al., 2020)	They propose **Idealized CAT score (ICAT)** that combines : **The language modeling score (LMS)** represents the percentage of tests when the model picks a logical answer (either the stereotyped or anti-stereotyped answer) over the nonsensical answer. **stereotype score (SS)** represents the percentage of tests where the model picks a stereotyped answer over the anti-stereotyped answer. The ICAT score has the following properties: It maximizes SS while keeping LMS at 100. If a model has a perfect stereotype score (ideal model, when the model always picks the stereotyped or antistereotyped answer (representing an SS of 100 or 0, respectively), then the ICAT will be 0; finally, a completely random model will have an ICAT of 50.	N/A	N/A	N/A	Bias, GPT, Metric	
fb5c116df	Wordcraft: Story Writing With Large Language Models	2022	https://arx	done	They build **WordCraft** - **Text Editor for creative writing (LM assisted co-writing)** LM can engage in open-ended conversation about story & respond to writer custom requests **generates suggestions** - quicker writing - higher engagement - more control over the editor - greater story enjoyment - LMs suggestions dont need to be perfect to be helpful	AI suggestions often in different tense Not enough contextual awareness LLMs can be outperformed by smaller specialized LMs in infilling	**LaMDA with prompt** **Meta Prompting:** - writing suggestions **UX patterns for controllable text generation** - infilling - Continuation - Rewriting - Story Seeding - Free Form Style Transfer - Custom Prompting	Crowriting	related example prompting - include related prompts before a zero shot prompt - steers Lm towards desired task	Crowriting	N/A	**Human Evaluation** - 25 hobbyist writers use Wordcraft and compare with normal editor - helpful: I found the AI agent helpful. - collaborative: I felt like I was collaborating with the AI agent. - creative: I felt creative while writing the story. - enjoyment: I enjoyed writing the story. - easier: I found it easy to write the story. - pride: I'm proud of the final story	N/A	N/A	LLMs have also been found to output memorized examples from their training data (copyright issue)	Bias, Crowriting, Human-eval, LaMDA	

2023 PAPERS BELOW

a739e2ab	Text Generator using GPT2 Model	2023	https://ar	done	They generate text using GPT-2 They support it with a fitting framework including - external control - contextual knowledge ranker - keyword predictor	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	GPT	doesnt seem to be anything new system showcase without evaluation
~~[illegible]~~	~~Metalofa: A Meta-Learning Approach for Indian Dialogue Generation~~	~~2023~~	~~https://ar~~	~~[illegible]~~	~~[illegible]~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~	~~N/A~~		
83a5bc5c4	Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications	2023	https://bio	done	**Review Question Generation** **3 types of QG:** - standalone, visual, conversational **Remaining challenges:** - quality - question types - non-crowsourced datasets (due to bad quality) and more domain specific datasets - metric that eliminates need for human evaluation **Future Direction:** - transfer Learning between domains - High quality corpora - Multimodally - metrics specialized for QG	N/A	Closed-Domain and Open-Domain QG **Classes of Questions:** - Factoid - Multiple Sentence Spanning - Yes/no type - Deep understanding **3 types pf QG:** - standalone, visual, conversational - they point out rule bases and nn based methods to approach these 3 types (see paper) Their **table 5 gives an overview about strengths and weaknesses of architectures**	Generation	N/A	Question Generation	Their Table 6: SQuAD CoQA NewsQA DeepMind Q&A WikiQA MSMARCO TriviaQA LearningQ NarrativeQA Natural Questions	**Automatic Evaluation** BLEU, METEOR, ROUGE, CIDEr **Human-Based Evaluation** - syntactic correctness, semantic correctness, and relevance **Other Method:** 1) use answerability of question through human study 2) modify existing eval metrics 3) use wheighted average of prec and recall of the scores 4) better human-conversation	N/A	N/A	N/A	Review	
316ee7a8	SP-NLG: A Semantic-Parsing-Guided Natural Language Generation Framework	2023	https://ieo	done	**Propose SP-NLG** - semantic-parsing-guided natural language generation framework - logical content generation framework can be built on any pretrained language model with a decoder	N/A	**modules**: (1) logical template generation, (2) semantic parsing, and (3) target ranking **Logical Template Generation** - coarse to fine generation - not filling slots with model predictions - fill with offline searched values by algorithm **Semantic Parsing** - takes both the template and the table as input - outputs the logic form - transformer-based encoder–decoder structure **a slot-tied back-search algorithm** (to leverage semantic parsing) - back-search arguments that correctly execute the logic form and fill the arguments into template (=> ensures logical correctness) **Target Ranking** - determine which group of arguments is the best for the logical template - similar to Reranker (using BLEU)	N/A	Table2Text	**LogicToxt's** Data is not enough - Get more samples by: 1) Simplify Logic Form Grammar 2) Data Augmentation (with Paraphrasing) 3) Annotate More Data	BLEU as supervision score **Surface Level** BLEU-1/2/3 **Logical Fidelity** - Exact Match NLI Acc	N/A	N/A	N/A	Logic2Text, BLEU, NLIAcc, SPAcc, Reranking	
9f23ed3ad	Controllable Text Generation Using Semantic Control Grammar	2023	https://ieo	done	Propose a Text Generation Framework for Controllability Propose **Semantic Control Grammar (SCG)** - controllable input format to generate sentences that satisfy constraints - control sentence- and word-level attributes Adopt Generation and Rerank Method - obtain semantically reranked controlled sentences	N/A	**Semantic Control Grammar** - Constrains at sentence and word level **SENT-ATTR:** A sentence-level attribute. This representation can be a sentence's tense, or it can contain general information of a sentence, such as intent, dialog act, topic, and sentiment. **WORD-ATTR:** A word-level attribute. Word-level representation can be slots, keywords, or entities. WORDATTR is a collection of none or many tag-value pairs put inside the parentheses **Text Generator:** - Input: SCG Expression - Output: corresponds with input constrains **Semantic Reranking** - text generator generated abundant sentences satisfying SCG based on the decoding - texts are not entirely semantically controlled (because log likelihood maximization) => generate and rerank approach **1) Vector Projector** - projects the semantically meaningful vector representations of text and the related SCG (consisting of text reader, SCG reader, SCG writer) **2) Reranker** - chooses top-k sentences based on cosine distance between SCG vector and text vector	Generation	N/A	benchmark, natural language understanding, data-to-text generation, and text classification, SCG2Text	**NLU Datasets:** Airline Travel Information Systems (ATIS) - Snips - Weather - Reed **Data2Text Generation Datasets** - E2E - WGDO **Text Classification Datasets** - TREC(coarse/ Conference Question) - Korean Language Understanding Evaluation-Topic Classification (KLUE-TC)	**Word Overlap-based** BLEU, ROUGE-L **Embedding Based** BERTScore **Constraint Based** SENT-ATTR Accuracy Value Accuracy	N/A	N/A	N/A	BLEU, ROUGE, BERTScore, Accuracy, ATIS, Snips, Weather, Reed, E2E, WGDO, LSTM, Reranking	
62483c0c3	Universal Multimodal Representation for Language Understanding	2023	https://arx	wip	Universal Multimodal Representation (UMR) is a flexible representation for grounding natural language in perceptual signals to general NLP tasks			Grounding, Vision-language understanding, and text generation									
~~[illegible]~~	~~Exploring AI Idea Generation Tool to Support Creativity in Academic Research~~	~~2023~~	~~https://ar~~	~~[illegible]~~	~~Scientifiblocks, a Web-based writing tool, explores ways to actively interact with search results beyond evaluation and selection.~~	~~The recognition of Scientifiblocks is done only against IdeaHub.~~ ~~Text does not seem yet to be published~~	~~N/A~~	~~Search/Over View generation and creativity support~~	~~N/A~~								
f8e68be1	XWikiGen: Cross-lingual Summarization for Low-Resource Text Generation in Low-Resource Languages	2023	https://arx	done	Propose **XWikiGen Task** - cross-lingual multi-doc summarization Contribute benchmark **dataset XWikiRef** - 5 domains, 8 languages Train a 2 stage system - input: set of citations, section title - out: section specific low resource summary multi-domain training is better than the multilingual setup on average	amount of reference URLs are averaging different for each domain in their dataset (need to account for in training) more work needs to be done to improve the model for Bengali and Malayalam	XWIKIGEN involves section-value text generation - input: set of citations (Reference URLs), section title - out: section specific low resource summary **first stage (Extractive):** - identifies promising candidate sentences across all the reference citations for the input - **Hierarchical and Positional Ranking model (HipoRank)** -> unsupervised graphbased model -> creates a directed hierarchical graph with sentence and section nodes and semantic/sentence and sentence-section edges with asymmetrically weighted edges => Score for a sentence node is then computed based on a weighted sum of edges incident on the node **second stage (Abstractive):** - generates an abstractive summary (based on best scoring sentences from first stage) - multi-lingual natural language generation models: mTS, **mBART (best)**	Generation	Multi-lingual, Multi-domain, and Multi-lingual-Multi-domain setups 1) train one model per (language, domain) pair - much data and train time needed 2) train one model per domain using training data across all languages 3) one model per language using training data across all domains 4) train a multi-lingual/multi-domain model - collate training data across all languages and domains (+ add lang and domain tags) => model can exploit cross-language cross-domain clues and learn robust representations	Summarization, XWikiGen, Cross-lingual summarization	**XWikiRef** (Their dataset) - Wikipedia pages corresponding to eight languages and five domains - 1.5K training samples with ~ 100K sections - Bangali (bn), English (en), Hindi (hi), Malayalam (ml), Marathi (mr), Oriya (or), Punjabi (pa), Tamil (ta), Telugu (te) - books, films, politicians, sportsmen, and writers	ROUGE-L, METEOR, chrF++ cant compute PARENT because they look at word overlap	N/A	extractive step: on a machine with one Tesla V100 SXM2 32GB of GPU RAM abstractive step: fine-tuned the model on a machine with one Tesla V100 having 32GB of GPU RAM	N/A	Dataset, ROUGE, METEOR, chrF++, BART, TS	
f5704ead	SAN-T2T: An automated table-to-text generator based on selective attention network	2023	https://ieo	done	Propose SAN-T2T: - table-to-text generator They connect a **field-content selective encoder and a descriptive decoder** with a selective attention network => utilize structural information => process longer tests than with normal attention	weak at generating long tests (maybe adapting Transformer based method to solve gradient exploding problem as a solution)	**Encoder**: LSTM	Generation	N/A	Table2Text	Wikibio (benchmark) Rotoliline and WeatherGov (supplementary for generalization)	BLEU-4, ROUGE-4 Rotoliline: Wiseman et al.: Rotation Generation Metric Content Ordering normalized Damerau-Levenshtein Distance between sequences of relations) Human Evaluation: - Fluency, Semantic Faithfulness	N/A	N/A	N/A	HumanEval, BLEU, ROUGE, Rotoliline, Wikibio, WeatherGov, LSTM, CopyMechanism, Attention,	
~~bc6f6b0d2~~	~~[illegible] artificial intelligence~~	~~2023~~	~~https://ieo~~	~~[illegible]~~	~~[illegible]~~												

*										Evaluation			Costs				
paperid	title	year	openAlex	STATUS (done / exp / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
0c4c579b	Friendly Conditional Text Generator	2023	https://ht	done	introduces a **multi-view attention (MVA) mechanism** Framework: **Friendly Conditional Text Generator (FCTG)** - maps tokens and attributes into a shared space - bridge modality gaps as the tests and attributes reside in different regions of semantic space introduces 2 training tasks: **- Masked Attribute Modeling (MAM)** **- Attribute Linguistic Matching (ALM)**	could not evaluate that MVA and tuning directly resolved catastrophic forgetting (but preserves parameters of PLM)	Multi-view attention (MVA) - aims to turn both attributes and words into the same space - optimize during training - learn attention weights with different areas - attend to whole input sequence - learns representations by switching masks 1) embed input sequence 2) adds positional/segment encoding to each word token 3) and decodes the linguistic output auto-regressively - their model can adopt pretrained NLMs - CTRL, GPT2, GPT-Neo/X	Generation	To avoid catastrophic forgetting, learn modality-free embedded representations, and direct the PLM to use the attribute information, we learn attribute representations, maps them in the same space as text through MVA, and optimizes their alignment in this space via ALM	**Masked Attribute Modeling (MAM)** - maximize the likelihood of masked attributes given context to reconstruct these attributes **Attribute Linguistic Matching (ALM)** - learns sequence-level alignment (rather than linguistic/attribute token-level alignment) between the attributes and the text - explicitly examine the impact of attributes on generation quality - but the decoder focuses on both the attributes and the text (wordy)	Amazon Review arXiv	Perplexity, BLEU, ROUGE, METEOR Human Evaluation - Amazon Review setting on attribute relevance - topic: the fraction of samples matching the target topic (attributes) - fluency (1-5)	N/A	FCTG converges faster with higher accuracy	N/A	Multimodality, Attention, Perplexity, BLEU, ROUGE, METEOR, HumanEval, Bias	need to login through uni gittingen
c73ba6b8	~~Language Models: The Most Important Complex Challenge of Our Time (Reputable)~~	~~2023~~	~~https://ht~~	~~review + 3 to add an abstract~~	~~review + add an abstract, and add a summary of the paper~~												
ec0062c1	~~Large Language Models: exploring artificial intelligence, computing, and natural language technology, including chatbots, the Long Innovation for Intelligent Enterprise Intervention~~	~~2023~~	~~https://ht~~	~~review + 3 to add an abstract~~	~~review + add an abstract, and add a summary of the paper~~												
e4f235dd	Semantic Similarity Measure of Natural Language Text through Siamese Neural Networks Keyword-Aware Cross-Encoder-Ranking Summarizer - A Case Study Using UCGS GSAT Body of Knowledge	2023	http://arxi	done	evaluates the effectiveness of NLP techniques in extracting semantics from text - deep neural networks and traditional machine learning approaches Propose **KACERS (Keyword-Aware Cross-Encoder-Ranking Summarizer)** - generate a semantic summary of scientific publications	good performance of KACERS relies on highquality keywords	**KACER (Keyword-Aware Cross-Encoder-Ranking Summarizer)** - **Input**: Keywords & Text - **Embedding**: 2 vectors with a cross-encoder (represent a keyword and a sentence) - attention to both keyword and sentence 2) For each keyword-sentence pair, a cosine similarity score is calculated 3) Ranking all the possible pairs based on the similarity 4) Extract top-ranked sentences for each keyword 5) Concat sentences => summary cross-encoder of BERT with 2 layers pretrained on the MS MARCO dataset	Generation	N/A	Summarization	MS MARCO	Recall, Precision, F1 (Fig), Balanced Accuracy	N/A	N/A	N/A	Recall, Precision, Accuracy, F1	
	~~Using GitHub Copilot to Solve Simple Programming Problems~~	~~2023~~	~~https://ht~~	~~review1 - not really relevant~~	~~Evaluate "Copilot" from GitHub in context to its editing Codex - in the context of academic/teaching Qualitatively show distribution of copilot report on experience using copilot (assist students in programming courses)~~												

2022 PAPERS WITH 0 OR 1 INFLUENTIAL CITATIONS BELOW

de193732	A large-scale computational study of content preservation measures for text style transfer and paraphrase generation	2022	https://arz														
e5f6ecc0	Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries	2022	https://dxd														
	~~Survey of Hallucination in Natural Language Generation~~	~~2022~~	~~https://ar~~	~~review #2 - it is a survey on a much, much broader, 20 - subjects from above~~													
4145c893	Attend, Memorize and Generate: Towards Faithful Abstractive Text Generation in Few Shots	2022	https://ieee														
0608089c	Improving Controllable Text Generation with Position-Aware Weighted Decoding	2022	https://arz														
4e42c484	On Decoding Strategies for Neural Text Generators	2022	https://dlsc		Beam Search is good for MT, but produces incoherent text for story generation. the choice of sampling method should be task specific												
82c4184d	~~Wordcraft: Story Writing With Large Language Models~~	~~2022~~	~~https://ar~~	~~read above~~													
fdb4235e	Using Machine Learning Language Models to Generate Innovation Knowledge Graphs for Patent Mining	2022	https://ieee														
2675be58	XAlign: Cross-lingual Fact-to-Text Alignment and Generation for Low-Resource Languages	2022	https://dl.a														
14a86f22	Fine-Grained Controllable Text Generation Using Non-Residual Prompting	2022	https://arz														
505c8a54	Data-to-text Generation with Variational Sequential Planning	2022	https://dlsc														
6cc8e91e	Controllable Text Generation for Open-Domain Creativity and Fairness	2022	https://arz														
e4baf3ad	Rewriting Semantic Similarity under Optimized Alignments for AMR-to-Text Generation	2022	https://arz														
95c8c3a5	KATG: Keyword-Bias-Aware Adversarial Text Generation for Text Classification	2022	https://ieee														
bce5c14d	~~A Systematic Survey on Energy-Efficient Techniques in Sustainable Cloud Computing~~	~~2022~~	~~https://ar~~	~~review #2 - it is a survey on a very, very broad topic~~													
67b66c11	Automatic Detection of Bot-generated Tweets	2022	https://hal														
4ec02676	Adversarial Machine Learning in Text Processing: A Literature Survey	2022	https://hal														

MANUALLY ADDED PAPERS

19.18572	Binary and Ternary Natural Language Generation	2023	https://arz	done	Propose a **solution to the difficulty of optimizing a ternary/binary Transformer Text Generation Model** Proposed Solution **- Statistics based quantization for weights** **- Elastic quantization of activations** => Demonstrate **first ternary and binary Transformer model for Summarization & MT** => Performance is close to "normal" BART, but **way more efficient** (16x for ternary BART)	generalizability to extremely long sequences or even streaming data has not been verified binarization and ternarization **require bit-packing to have actual memory savings and dedicated hardware support** for real-time acceleration	**Ternary/Binary NN:** - (-1,0,1) or (0,1) as weights => multiplication-free computations (efficiency gains) **Difficulties with ternary/binary Transformers TGM:** - nature of attention operation is quantization - noise compounding effects of autoregressive decoding in high-dimensionality output space Quantizing both weights and activations allows computation to be performed using bitwise operation => More efficient **Stats-based max-entropy isometric weight quantization** - maximizing the entropy of the quantized weights and reducing the gradient mismatch in the backward pass - formula in paper **Learning-based activation quantization** - Devide activation layers in 2 categories (non-negative like ReLU & others) - Parameterize the activation function - symmetrically quantize the later activation category (learned wheighting factor by gradient update) => activation distribution matches the original full/precision activations	Generation	N/A	Summarization, Translation	CNN Daily Mail XSUM WMT16 English-Romanian (for mBART)	ROUGE, BLEU	N/A	N/A	N/A	BART, CNNDaily, XSUM, ROUGE, BLEU, WMT16	
19.18573	EM Pre-training for Multi-party Dialogue Response Generation	2023	https://arz	done	Study multi-party dialogue Propose **Expectation-Maximization (EM)** approach (needed because of data scarcity) - optimizes adressee labels and optimizes a response generation model => possible to pre-train a model with huge amount of unlabeled corpora	lack of multi-party dialogue datasets pretraining scheme relies on addressee information (disregarding reply-to relations)	**Addressee Modeling** - add addressee embeddings to positional encodings and word embeddings - PLM encodes these further (BART) **Response Generation** (Latent Variable Prediction): - Bayesian Network (response generated based on context and adressee)	Generation	EM Training: - expectation and maximization steps alternately and iteratively Expectation Step: - compute cond. distribution of addressee with respect to context and response (and model parameters) Maximization Step: - analogical to normal training process - optimizes generation model	Dialogue Response Generation	Lack of multi-party dialogue datasets Ubuntu Dialogue Corpus - no addressee annotations - 1.5M dialogues (each less than 4 turns)	BLEU, METEOR, ROUGE Human Evaluation - Kappa - 8 people with Bachelors degree in CS minimum - relevance, fluency, informativeness	N/A	N/A	N/A	BART, HumanEval, ROUGE, METEOR, BLEU, Ubuntu	
10.18653	Open-ended Long Text Generation via Masked Language Modeling	2023	https://arz	done	Show 3 reasons why Masked LMs would collapse for Open-Long Text Generation Propose 2 strategies for **masked language models to maintain inference efficiency with increasing output length:** **- dynamic sliding window attention (DSWA)** **- linear temperature decay (LTD)** => alleviate long-distance collapse problems => longer text generation => flexible trade-off between performance and inference speedup => achieve significant **speedup** with better performance than other AR models	typical AR problems: (1) off-prompt: the provided prompt is very short, which causes the model can not fully understand that content and generate reasonable text (2) incoherent between sentences	RoBERTa encoder & decoder Display **3 possible reasons why pretrained Masked LM models collapse** for Open-Long Text Generation - function words are often located at the front of the output distribution, which dominates the high probability region, causing the informative tokens hard to be sampled - iterative refinement mechanism depends on the token confidence of generated sequences, and it is easier for the low-confidence but informative tokens to be masked - massive absent context tokens suffer a more serious multi-modality problem on long text generation in early iteration steps **Dynamic Sliding Window Attention** - alleviate the multimodality problem by restricting the model to attend to the tokens in the window instead of the whole sequence => degenerating the multi-modal distribution into a uni-modal distribution **Linear Temperature Decay** - confidence-based iterative refinement		- randomly mask to obtain a corrupted target - try to reconstruct masked sequence Mixed Attention: - aggregate source and target Inference: 1) predict all masked tokens of masked sequence 2) iteratively regenerate low confidence tokens - nucleus sampling Length Prediction - classification layer (softmax)	Writing-Prompt (WP) ROC Stories (ROC) WikiPlots and multi-paragraph opinionated article writing (OPINION)	BLEU, ROUGE, Distinct, and Lexical Repetition, METEOR, Perplexity Human Evaluation - compare with BART - 3 annotators should give preference for: - fluency, coherence, and relevance	N/A	N/A	N/A	Bias, BLEU, ROUGE, Distinct, METEOR, Perplexity, HumanEval, WikiPlots, WritingPrompt, ROCStories, OPINION, Masked, HumanEval	
10.18653	Tailor: A Soft-Prompt-Based Approach to Attribute Control for Controlled Text Generation	2023	https://arz	done	Propose **tailor (Text-attribute general controller)**: represents each attribute as a pre-trained continuous vector => guides generation of a PLM	Not tested on a wide variety of PLMs	**Tailor** - for **single** attribute controlled text generation and for **multi** attribute with their **tailor**, we get problems of fluency downgrading and position sensitivity => two solutions, namely: **non-training (MAP mask + SP sequence)** **training methods (MAP connector)** Non-training: - multi-attribute prompt mask and reindexing position sequence => bridge gap between training and testing - trainable prompt connector to enhance combinations only needs 0.58% extra training parameters with GPT-2	Generation	**For Multi Attribute CTG:** Tailor introduces a training method (next to non-training as described in architecture) **=> MAP connector** - continuous prompt trained for combining two single attribute prompts	Conditional Text Generation, Control YELP	**Correctness** - RoBERTa **Text Quality** - GRAM (RoBERTa based) - perplexity - (leg. from GPT-2 small, medium, large) **Diversity** - count number of uni- bi- and trigrams and normalize (Dist-1, ..., Dist-3) **Human Evaluation** - Fluency - if they satisfy attribute	N/A	N/A	N/A	Control, HumanEval, GPT, RoBERTa, Perplexity, Yelp	
10.18653/v	A Causal Framework to Quantify the Robustness of Mathematical Reasoning with Language Models	2023	https://arz	not sure if we should include													
10.18653/v	Increasing Diversity While Maintaining Accuracy: Text Data Generation with Pre-trained Language Models and Human Interventions	2023	https://arz														
10.18653/v	Self-Edit: Fault-Aware Code Editor for Code Generation	2023	https://arz														
10.18653/v	How About Kind of Generating Hedges using End-to-End Neural Models?	2023	https://arz														
10.18653/v	Does GPT-3 Grasp Metaphors? Identifying Metaphor Mappings with Generative Language Models	2023	https://arz														
10.18653/v	Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe	2023	https://arz														

*										Evaluation			Costs				
paperid	title	year	openAccess	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
10.18653/v1/2020.acl-main.90	Dynamic and Efficient Inference for Text Generation via BERT Family	2023	https://aclanthology.org/2020.acl-main.90/	done	Propose a new **finetuning strategy DEER** - **Dynamic and Efficient inflERence** => **adaptive trade-off between model performance and latency** utilize the non-autoregressive (NAR) sequence generation: parameter pooling technique => flexibly control the decoding iteration steps and model sizes according to memory and latency limitations	CTC cannot deal with the multi-modality problem in the target sequence, which hinders the NAR model from simplifying the multiple beam alignment targets relies on the assumption that the input length is large than the output	Do experiments on BERT family - supports single-step generation and iterative-based generation CTC generator and Levenshtein editor to empower the Transformer encoder-decoder model with the ability to generate and produce flexible results for iterative refinement, eliminating the need for sentence length prediction modules Non-autoregressive text generation improves inference efficiency	Generation	**Dynamic Block Pooling:** - reduces computation and memory in inference with dynamic model size **Single-step CTC (Connectionist Temporal Classification) Generator** - does not need to predict length of target sentence **Iterative-based Levenshtein Editor** - CTC model relies on conditional independence assumption (incapable of multimodality) - shares parameter with CTC to correct text error - place/holder/insert/on/deletion classifier **Total training objective:** the sum of all items: CTC loss, Levenshtein classifier loss, and weight ejaneity regularization term	Translation, Question Generation, Summarization	**WMT'14**: English --German2 (De--En); **WMT'16**: English --Romanian3 (En--Ro); **WMT'14**: English --German4 (En--De) Summarization Gen: **XSum** **MSNews** Question Gen **SQuAD** **MSWG**	BLEU	multiple task specific models are often needed, leading to more carbon emissions => DEER	N/A	N/A	IWSLT, WMT, BERT, BLEU, Finetuning, SQuAD, MSWG, XSUM, MSNews	
10.18653/v1/2020.acl-main.91	An Invariant Learning Characterization of Controlled Text Generation	2023	https://aclanthology.org/2020.acl-main.91/	done													
10.18653/v1/2020.acl-main.92	An Invariant Learning Characterization of Controlled Text Generation	2023	https://aclanthology.org/2020.acl-main.92/	done													
10.18653/v1/2020.acl-main.93	Reward Gaming in Conditional Text Generation	2023	https://aclanthology.org/2020.acl-main.93/	done	identify three **common cases in RL where high rewards are incorrectly assigned to undesirable patterns** - noise-induced spurious correlation - naturally occurring spurious correlation - covariate shift **undesirable patterns may be amplified during RL training** They give solution ideas to the problem of reward gaming.	Solutions only presented as ideas, not actually implemented	N/A	Generation	**Reward Gaming** - happens if a high reward is associated with undesired pattern during training **Causes:** (1) A group of examples is misannotated systematically (2) human annotation dataset contains some bias due to the data we select to annotate, or due to the people we select to be annotators (3) human annotation dataset does not cover certain groups of sentences **Possible solutions:** **1) Restricting the policy by regularizing toward the ML solution** (difficult) **2) Restricting the policy by leveraging a discriminator** (distinguishes between sampled generations and the set of dataset-provided generations) 3) **Fixing the reward itself** - e.g. iteratively collect human annotations (costly) - e.g. hardcoding a set of principles 4) **Other Methods** - e.g. push down all high rewards to ensure that the out-of-distribution states do not achieve high rewards (extensive hyperparameter tuning needed) - e.g. avoid interaction with the environment => errors in the reward function will be less exploited - e.g. non-RL objectives that can learn from both positive and negative examples	Multi-Doc Summarization, Translation, Sudoku	N/A	N/A	N/A	N/A	N/A	RL, Bias	
10.18653/v1/2020.acl-main.94	Toward Human-Like Evaluation for Natural Language Generation with Error Analysis	2023	https://aclanthology.org/2020.acl-main.94/	done	incorporate error analysis into automated metric to improve its performance (called BARTScore++). They argue that **estimating the sentence confidence** can be used for: - **refine generated sentence toward better confidence** - **determine major and minor errors** They improve BARTScore by: - unexplicited sentence capacity - human-like error analysis (=> evaluation of major and minor errors) called **BARTScore++** **Metric Properties:** - more effective in detecting hallucination - relatable for already great performing systems - more human like on discriminating errors	explicit errors only contain token-level errors metric fully relies on BART generation probabilities	N/A	Generation	N/A	Translation, Summarization, DataText	**Translation** DARR Corpus WMT20 **Summarization:** RealSumm SummEval NewR18 (Newsroom) - 60 news, each with summaries generated by 7 different methods are annotated with human scores in terms of coherence, fluency, informativeness, relevance **Factuality** Frank18 - meta evaluate factuality metrics - annotated by correctness **DataText:** BAGEL SFHOTEL SFREST E2E **Meta evaluation (after Yuan et al. (2021)):** Kendalls tau - measure the correlation of metrics with human evaluation Spearman Correlation	Their metric: **BARTScore++** uses **Explicit/ Implicit Error Distance** **1) non-translation test to decide if sentence will be refined or not** **2) explicitly perform detect-correct algorithm** Detect: token generation probability Correct: top-k sampling - weighting errors for the final score **Multidimensional Quality Metric (MQM)** - to get judgements for the translation task on WMT20 - consistency of human judgements through error analysis - requires evaluators to identify errors and categorize them into different levels according to their severity	N/A	N/A	N/A	Hallucination, Metric, BARTScore, DARR, WMT, RealSumm, SummEval, Newsroom, Rank18, QAGS20, BERT, BART, BAGEL, SFREST, MQM, HumanEval	
10.18653/v1/2020.acl-main.95	Dialect-robust Evaluation of Generated Text	2023	https://aclanthology.org/2020.acl-main.95/	maybe too specific?													
10.18653/v1/2020.acl-main.96	FairPrism: Evaluating Fairness-Related Harms in Text Generation	2023	https://aclanthology.org/2020.acl-main.96/	done	Propose **FairPrism** - **dataset** of 5.000 examples of AI-generated English text - **human annotations covering harms relating to gender and sexuality** (improved transparency, clearer understanding of what harms are accounting for annotator disagreement and harms that are context-dependent) - include the extent of stereotyping and demographic bias, and a demographic groups targeted, and appropriateness for different applications Their methodology should be used as a reference for future dataset creations of this type	Suggestions for better datasets: **Transparency** - improve clarity about how AI-generated text can cause harms helps annotators provide high-quality labels => measure or mitigate diverse set of harms, higher quality human labels **Specify Coverage** - annotators to identify the demographic groups targeted sets **Annotator disagreement** - provide individual annotators' judgments => confidence scores **Account for context-dependent harms** - provide context (perceived author of text, application) => better assessment **diverse set of annotators**	Dataset should not be used as a benchmark to be beaten as it reduces its effectiveness long term	N/A	Generation	N/A	Evaluating (1) the **types of fairness/related harms** that AI text generation systems cause (2) the **potential limitations of mitigation methods**	**Used for creation:** StoCorpus, Social Bias Frames	Their dataset: **FairPrism** - improved **transparency** regarding the types of fairness-related harms that AI systems can cause; - clearer specification of the dataset's **coverage** of types of harms, including the groups targeted; - accounting for **annotator disagreement** about whether harms are present; - accounting for **context-dependent harms**, including specific harms that occur in interactive contexts and harms that are normative concerns when the "speaker" is an AI system. **Human Evaluation** - 5 annotators - Fluency, Novelty, and Attribute Relevance	Human Annotations for the dataset (as stated under "dataset")	N/A	N/A	N/A	Dataset, Bias, HumanEval	
10.18653/v1/2023.acl-long.488	DuNET: Dual-Noisy Self Training for Semi-Supervised Controllable Text Generation	2023	https://aclanthology.org/2023.acl-long.488/	done	DuNET boosts control accuracy	slower training ST methods rely on pseudo labels need to find good hyperparameter (noise level)	N/A	Generation	propose **DuNET: a novel Self Training framework** to tackle - **generation models over-exploit learned text space (restricted generalization boundary and limited controllability)** adds two kinds of flexible noise to escape from collapsed space => model could construct and utilize both high-quality samples from given labels and pseudo labels predicted from available unlabeled text DuNET: Dual Noisy Self Training (DuNET) tackles semi-supervised controllable text generation (SSCTG) by **generating pseudo labels for unlabeled text** and **training a model on the union of the original labeled data and the pseudo labeled data**, while **injecting noise into both the pseudo labels and the model** to **prevent the model from memorizing the pseudo labels and collapsing to a narrow text distribution**.	DuNET jointly learns to **generate text from given attribute labels and predict labels for text**, characterizing these two directions as a dual variational generation process => allows model to leverage generated pseudo text and pseudo labels predicted for available unlabeled text 2 Noise Generation Approaches: **High-temperature Generation (HTG):** - temperature in softmax layer (controllable by parameter) **Soft Pseudo Text (SPT):** - directly store the output token distribution d - directly learn to reproduce d - kinds "Iterative Knowledge Distilling" (avoid loosing relevant information)	Sentiment Control with Prompt Topic Control without Prompt Detoxification	IMDb Movie Review AGNews Jigsaw Toxic Classification	**Perplexity** (Fluency) Generalizability (**PPL on held out dataset splits**) **Classification Macro-F1 by RoBERTa** (Accuracy) **Dist-n, Self-BLEU** (Diversity) **Human Evaluation** - 5 annotators - Fluency, Novelty, and Attribute Relevance	N/A	N/A	N/A	HumanEval, Bias, Perplexity, SelfBLEU, F1, SelfTraining, IMDb, AGNews, Jigsaw	
10.18653/v1/2023.acl-long.492	Soft Alignment Objectives for Robust Adaptation of Language Generation	2023	https://aclanthology.org/2023.acl-long.492/	done													
10.18653/v1/2023.acl-long.494	Benchmarking Large Language Model Capabilities for Conditional Generation	2023	https://aclanthology.org/2023.acl-long.494/	done													
10.18653/v1/2023.acl-long.506	SSD-LM: Semi-autoregressive Simplex-based Diffusion Language Model for Text Generation and Modular Control	2023	https://aclanthology.org/2023.acl-long.506/	done	Propose **SSD-LM: a diffusion based LM** - **semi-autoregressive** (iteratively generate blocks of text) - **simplex-based** **(diffusion on the natural vocabulary space rather than a learned latent space)** => classifier guidance and modular control outperforms GPT-2 and other diffusion based baselines show **insensitivities, biases and loopholes in existing model based metrics** **examples:** - BERTScore is confused by truncation errors in summarization - MAUVE (built on top of GPT-2) is insensitive to errors at the beginning or middle of generations - some PLM metrics are biased towards frequent n-gram rather than global coherence - log-probability-based metrics could be unfairly biased towards their underlying PLMs (self-evaluation)	Sample efficiency Decoding Speed Decoding block size remains the same across iterations they only eval on english tests: language-specific properties can induce potential blind spots for metrics, especially for low-resource languages	SSD- LM uses diffusion model to decode a block of tokens, given a Gaussian noise and a context Decoding: **Reverse Diffusion (iteratively)** starting with the gaussian noise - estimates probability distribution of original input ability to use off-the-shelf attribute classifiers for control that are open to use but have been trained on closed source data Tested errors with the dataset: ***Positioned Error Test*** - test if PLM is biased to encode only the local context - create synthetic errors by replacing a span of 10 consecutive tokens in different positions of the gold hypotheses with (1) 10 random tokens from the vocabulary, or (2) randomly shuffled tokens of the original span **Injection Test** - reformes as boolean question answering task **Frequent n-gram Test** - stress-test whether log-likelihood-based metrics would wrongly favor a random sequence of frequent n-grams over the gold hypotheses **Self-Evaluation Bias** ***Fluency & Consistency Tests***	Generation	**Logits**: Token mapped to Vocabulary space [-5,+5] **Forward Diffusion**: time dependent gaussian noise added to logits **Loss function**: using b-directional context (conditional denoising) goal: predicting the original data representation	Unconstrained Text Generation, Controlled Text Generation	N/A	**Perplexity by a LLM** (GPT-Neo-1.3B) low perplexity of generated text is not necessarily an indication of high quality but of degenerate behavior (Nadeem et al., 2020; Zhang et al., 2021), closeness to the perplexity of human-written text is a better evaluation => **difference of log perplexity between the generated text and human-written continuation** **Zipf's coefficient** (for diversity) average distinct n-grams (for diversity) **repetition rate** (Welleck et al., 2020; Holtman et al., 2020; Welleck - proportion of output samples that end in repeating phrases **MAUVE** - measures diversity and quality - by approximating information divergence between generated samples and human-written continuations	N/A	N/A	controllability can also be misused maliciously, with models being **intentionally exploited to generate biased, toxic, or non-factual content** (Brown et al., 2020; Bender and Koller, 2020; Paullada et al., 2020)	Perplexity, Zipf, Diffusion, MAUVE, Bias, Control	some extra cites for side facts in the notes
10.18653/v1/2023.acl-long.507	On the Blind Spots of Model-Based Evaluation Metrics for Text Generation	2023	https://aclanthology.org/2023.acl-long.507/	done	Propose **SSD-LM: a diffusion based LM** - **semi-autoregressive** (iteratively generate blocks of text) - **simplex-based** **(diffusion on the natural vocabulary space rather than a learned latent space)** => classifier guidance and modular control outperforms GPT-2 and other diffusion based baselines show **insensitivities, biases and loopholes in existing model based metrics** **examples:** - BERTScore is confused by truncation errors in summarization - MAUVE (built on top of GPT-2) is insensitive to errors at the beginning or middle of generations - some PLM metrics are biased towards frequent n-gram rather than global coherence - log-probability-based metrics could be unfairly biased towards their underlying PLMs (self-evaluation)	Sample efficiency Decoding Speed Decoding block size remains the same across iterations they only eval on english tests: language-specific properties can induce potential blind spots for metrics, especially for low-resource languages	SSD- LM uses diffusion model to decode a block of tokens, given a Gaussian noise and a context Decoding: **Reverse Diffusion (iteratively)** starting with the gaussian noise - estimates probability distribution of original input ability to use off-the-shelf attribute classifiers for control that are open to use but have been trained on closed source data Tested errors with the dataset: ***Positioned Error Test*** - test if PLM is biased to encode only the local context - create synthetic errors by replacing a span of 10 consecutive tokens in different positions of the gold hypotheses with (1) 10 random tokens from the vocabulary, or (2) randomly shuffled tokens of the original span **Injection Test** - reformes as boolean question answering task **Frequent n-gram Test** - stress-test whether log-likelihood-based metrics would wrongly favor a random sequence of frequent n-grams over the gold hypotheses **Self-Evaluation Bias** ***Fluency & Consistency Tests***	Generation	Most Metrics are based on pretrained models	Open Ended generation, Summarization, Translation	**They produce synthetic data for stress testing metrics** - also released dataset Consists of: Open-ended generation: **WikiText103** Summarization: **CNN Daily Mail** Translation: **WMT21** (German-English & Chinese-English) => take subset (TED-Talk) from the ZH-En set	Tested metrics: **open ended text generation:** MAUVE (with GPT2-large, RoBERTalarge, and ELECTRA-large) - using contextualized embeddings from PLMs GPT-PPL and MLM-PPL **translation and summarization:** BERTScore MoverScore BLEURT SARIGEM COMET PRISM (only translation) BLEURT (only translation) BLEU ROUGE	N/A	N/A	N/A	Dataset, Bias, BERTScore, MoverScore, BLEURT, SARIGEM, COMET, PRISM, BLEURT, BLEU, ROUGE, CNNDaily, WikiText, WMT, MAUVE, GPT-PPL, MLM-PPL	they emphasise just because a PLM metric is bad in that aspect, it is not bad overall. they have higher human correlation than e.g. BLEU
10.18653/v1/2023.acl-long.508	Contrastive Decoding: Open-ended Text Generation as Optimization	2023	https://aclanthology.org/2023.acl-long.508/	done													
10.18653/v1/2023.acl-long.509	Beat-4: Search Algorithm for Neural Text Generation	2023	https://aclanthology.org/2023.acl-long.509/	done													
10.18653/v1/2023.acl-long.754	Self-Instruct: Aligning Language Models with Self-Generated Instructions	2023	https://aclanthology.org/2023.acl-long.754/	done													
10.18653/v1/2023.acl-long.843	ReCode: Robustness Evaluation of Code Generation Models	2023	https://aclanthology.org/2023.acl-long.843/	maybe too specifiC? (Code)													
10.18653/v1/2023.acl-long.849	An Extensible Plug-and-Play Method for Multi-Aspect Controllable Text Generation	2023	https://aclanthology.org/2023.acl-long.849/	done													
10.18653/v1/2023.acl-long.860	Semantic Accuracy in Natural Language Generation: A Thesis Proposal	2023	https://aclanthology.org/2023.acl-long.860/	done													

*										Evaluation			Costs				
paperid	title	year	openAlid	STATUS (done / wip / discuss)	Main contribution	Main limitation	Architecture	Application	Training details	Tasks	Datasets	Metrics	Ecological	Economical	Ethical	TAGS	Comments
10.48550/																	
	A Survey of Large Language Models	2023	https://ar	done	**Survey LLMs broadly** - Timeline of release of models - Big table with many models compared (release, size, hardware, train time) - table of optimizer settings for existing LLMs - List of suitable datasets per task (language generation, knowledge utilization, complex reasoning) - Explanation of general functionality of LLMs (as well as variations)	N/A	- Emergent Abilities of LLMs: - In-context learning: follow natural language instruction without additional training o Instruction following: fine-tuned to follow instruction o Step-by-step reasoning: e.g. math - Techniques: o Training: bigger models o Training: distributed algorithms, optimizers o Ability eliciting: after training endowe as general-purpose task solvers o Alignment Tuning: helpful, honest, harmless o Tools manipulation: e.g. LLMs use calculator, external plugins Architecture - Mainstream: Encoder-decoder, Casual Decoder, Prefix Decoder - Detailed Config: Normalization, Activation Functions, Pos. Embeddings, Attention & Bias - Pretraining Tasks	Generation, Knowledge Utilization, Complex Reasoning	- Data Source o General Text: Webpages, Conversation, Books o Specialized Text: Multilingual, Scientific, Code - Data Preprocessing: Filtering, Quality Filtering, De-duplication, Privacy Reduction, Tokenization Adaption Tuning of LLMs o Instruction Tuning: o Collection of Existing Datasets o Formatting Human Needs o List of factors for instance construction quality o Strategies for Instruction Tuning: Training Data, Combine Instruction Tuning & Pretraining o Effects of Instruction Tuning: Performance, Task generalization o Alignment Tuning —	N/A	List in paper	N/A	N/A	N/A	N/A	Review	check paper for deeper information
10.18653/	Rogue Scores	2023	https://aci	This is a "necessary" mention. The fact this researcher went over 2000 papers to say that the defective metrics for evaluation should be highlighted as "critical for evaluating text generation"													