

Research Statement

Varun Gangal
Carnegie Mellon University

1 Summary

My research has a **two-pronged focus**, nestled on the *primary fulcrum of natural language generation* (NLG) and the *secondary pivot of data augmentation* (DA).

NLG is the field of study which aims to endow agents with ability to generate language to satisfy any stated *communicative goal*. From the earliest days of AI, NLG has been a part of its holy grail, with ability to converse in human-indistinguishable fashion being the crux of Turing’s “*Imitation Game*” to tell if an agent was truly “intelligent”. In the past decade, NLG models have made persistent strides in terms of their efficacy of learning¹, as well as the basic properties of generated language outputs such as grammaticality, fluency and intra-sentence coherence. However, this newfound success has also exposed the disparity between model generated and human language on *higher-order aspects* such as discourse coherence, commonsense plausibility, faithfulness, pragmatics, creativity and novelty. Enhancing performance of NLG models on tasks attaching importance to these aspects is one motivation driving my research.

As formalized by the sociolinguist Halliday in his magnum opus *Theory of Systemic Functional Linguistics* [15] (SFL), communication is driven not just by *textual* goals (lexical choice, referring expressions, content order inter alia) but also by *interpersonal* (speaker persona, speaker-addressee relationship) and *ideational* goals (pragmatic intent). However, the current mix of NLG tasks and benchmarks e.g summarization, table-to-text generation etc seldom test ability to fulfill such goals.² Adapting extant architectures for existing tasks requiring these goals as well as devising new NLG tasks and benchmarks involving them has been another motivation underlining my NLG research. I together refer to these *higher-order aspects* and *extra-textual goals* as *complex facets*.

In summary, within NLG, the goal of my research has been to *assess and bridge* the gap between NLG models and *human faculties of language generation* by *evaluating and enhancing* their abilities at *tasks involving complex facets such as style, creativity and commonsense*. In pursuit of this goal, I have made contributions along four fronts:

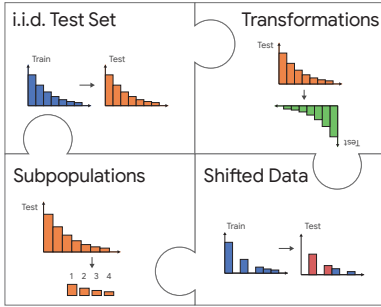
1. **Handling Creative Constructs** : I investigated abilities of NLG models to understand and respond to figurative language [16] as well as generate creative devices like portmanteaus [37] and personification [27]. (§2)^[16, 37, 27]
2. **Commonsense Plausibility Through Input Augmentation** : I devised two distinct approaches [9, 10] to augment per-example inputs with additional information to enhance the plausibility of NLG model outputs for concept-to-text generation tasks such as CommonGen [24]. (§3)^[9, 10]
3. **Incorporating Commonsense Knowledge** : I devised approaches to incorporate a multitude of commonsense knowledge sources such as frame-based [3], game-specific and general-purpose [2], thus building viable NLG models for generating causal explanations [20], game commentary [17] and pseudo-references for dialog evaluation [36]. (§4)^[20, 17, 36]
4. **Transferring & Annotating Style** : I formulated ways to adapt SOTA sequence-to-sequence architectures for diachronic (Modern→Shakespearean English) [18] and narrative style transfer [34]. Further, I proposed novel task settings and annotation studies to curate new datasets for persona [19] and narrative style transfer. (§5)^[18, 19, 34]

Data augmentation (DA) refers to methods for increasing the pool of dataset examples without explicitly collecting new data. DA has recently seen increased interest in NLP due to more work in low resource domains, new tasks, and the popularity of large-scale neural networks that require large amounts of training data. Despite this upsurge, the area is still relatively underexplored, perhaps due to the challenges posed by the discrete nature of language data, which rules out continuous noising and makes it more difficult to maintain invariance. For many nonclassification NLP tasks such as span-based tasks and generation, DA research is relatively sparse . My own foray into DA research aims to close some of the aforementioned gaps:

1. **DA for Learning and Evaluating NLG Better** : In [11], I explore a suite of DA strategies (Fig. 9) for finetuning GPT-2 on low-resource domains. DA is not just useful for training set expansion. In [36], I show how DA methods using commonsense and instance-based knowledge (Fig. 2) can expand reference sets to better automatic evaluation of dialog. Further, in [29], I leverage DA methods like backtranslation to generate NLG evaluation suites (Fig. 1).
2. **Large Scale DA Through Community-Scale Participation** : To sensitize the NLP community to the bespoke lacunae in DA research, I penned a survey [12] of DA work in NLP. Furthermore, I co-organized the **NL-Augmenter participative repository and benchmark** [6, 7], which provides a structure for NLPers to contribute and evaluate task-specific DA methods. NL-Augmenter has curated a large suite of 100+ peer-reviewed and tested methods by using wisdom-of-the-crowd — opening doors to more comprehensive evaluation of robustness. (Fig.4)(Under Prep Submission [7])

¹both in terms of number of training examples and across different tasks/domains

²barring some notable exceptions e.g text simplification.



Subpopulations
 Input Types [Dialog Acts, Topics]
 Frequency [Complexity, Overlap]
 Named Entity Features [Demographics]
 Shape [Syntactic Structure, Properties]
 Size [Length]
 ...

Transformations
 Back-Translation [great -> toll -> fantastic]
 Typos [English -> English]
 Punctuation [English, -> English]
 Numerical Values [66 -> 79]
 Scrambling [Cheap English -> English Cheap]
 ...

Shifted Data
 Time-Shifted [COVID]
 Data Samples [Train/Validation Examples]
 ...

Figure 1: Illustration of the types of evaluation suites [29] that can be constructed from a given NLG dataset.

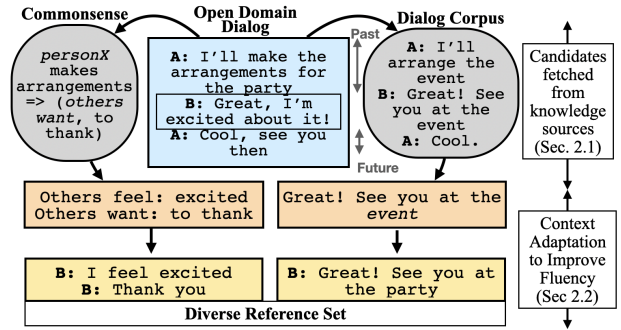


Figure 2: We propose automatic ways to collect references sans any crowd-sourcing [36], through two types of knowledge sources: commonsense and retrieved instance knowledge, followed by automated adaptation to make them more fluent in the target contexts.

Method	Text
Original Review	got sick from the food . overpriced and the only decent thing was the bread pudding . wouldn't go back even if i was paid a million dollars to do so .
Synthetic Noise (10%)	got sick from the food . overpriced and the only decent thing was the bread pudding . wouldn't go back even if i was paid a million dollars to do so .
Synonym Replacement (3 keywords)	got sick from the food . overpriced and the only decent thing was the scratch pud . wouldn't go back even if i was paid a one thousand thousand dollars to do so .
Hyponym Replacement (3 keywords)	got sick from the food . overpriced and the only decent thing was the creoscent roll corn pudding . wouldn't go back even if i was paid a million kiribati dollar to do so .
Hypernym Replacement (3 keywords)	got sick from the food . overpriced and the only decent thing was the baked goods dish . wouldn't go back even if i was paid a large integer dollars to do so .
Random Insertion (10%)	got sick from the food nauseous . overpriced and the only decent thing was the bread pudding . wouldn't go back even if i was paid a million dollars hoodle to do so .
Semantic Text Exchange (60% MRT)	got sick from the coffee . overpriced and the food was good . wouldn't come back if i was in a long hand washing machine .

Figure 3: An example Yelp review and its variations using our augmentation methods for generator fine-tuning [11]. Changes are bolded.

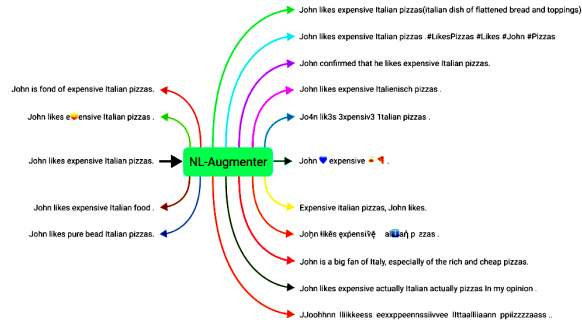


Figure 4: A few randomly chosen augmentations of NL-Augmenter for the original sentence *John likes expensive pizzas*. While the meaning (almost) always remains the same and identifiable by humans, models can have a much harder time representing the augmented sentences. [6, 7]

2 Understanding and Generating Creative Constructs

Can Dialog Models Handle Figurative Language ?: Figures of speech such as idioms (e.g It's hard *filling her shoes*.) and similes (e.g He *fought like a lion*.) are key constructs used by speakers to creatively express themselves. Though easily understood by humans, figurative constructs in the input can be hard to understand and generate outputs in response to for NLG models. This is a result of their long-tailedness (e.g <3% at utterance level in DailyDialog, as we find), non-decompositional semantics (e.g *filling her shoes* = replacing her *job role* and has nothing to do with *shoes*), and commonsense dependent interpretation (lions fight fiercely). In [16], we investigated the ensuing research questions — Can SOTA dialog systems respond properly to figurative contexts? Does "literalizing" improve response quality? Consider the example in Fig. 5 — its seen that taking the context's metaphor literally derails the model responses, and also that literalizing the metaphor makes model response adequate. We see this behaviour even at the aggregate, corpus level - across 5 distinct SOTA models and for both human and automatic measures, drastic drops of >20% (even going to 100% for BLEU-4 on 3/5 models) in response quality are observed for figurative contexts, with effects on both immediate and later-in-dialog responses. Further, encouragingly, literalization can salvage response quality in part, improving metrics by ≈5-10%.

Generating Creative Constructs: Besides responding robustly to creative language, NLG models can also benefit from having the ability to generate it - this could improve the novelty and diversity of their outputs, making them sound more human-like. Furthermore, certain NLG applications such as writing assistants for advertising and news-reporting may require these abilities as part of their communicative goals e.g to make content memorable, interesting, or appealing. A common challenge encountered in learning to generate creative constructs is deficient training data in terms of both size and diversity, the first being since they are used sparingly, the second being since entirely novel creative artifacts are seldom created e.g there are almost about < 8K unique idioms enlisted in Wiktionary [39].

In our first foray on this theme, we explored portmanteau generation [37]. Portmanteaus are a creative construct where two root words blend to form a new word, with meaning derived from but distinct to their original meanings e.g *wiki + etiquette* →

Input	FORWARD	BACKWARD	BASELINE	G.TRUTH
shopping;marathon	shopparathon	shoathon	shon	shopathon
fashion;fascism	fashism	fashism	fashism	fashism
clown;president	clowident	clownsident	clownt	clownsident
car;hijack	carjack	carjack	cack	carjack
tinder;disaster	tinter	tindersaster	tindisaster	tindisaster
chopstick;fork	chopstork	chopfork	chork	chork
happy;harmonius	happonius	happonius	harmonius	happymonius
flexible;vegetarian	flexarian	flexetarian	flegetarian	flexitarian
laughter;orgasm	lauggasm	laughtergasm	lasm	laughgasm

Table 1: Example outputs from different portmanteau generation models. Outputs are from best performing model configurations [37]. G.TRUTH denotes the ground truth portmanteau.

Type of ATTRIBUTE	Description	Example
Part-of	The ATTRIBUTE is a human-like possession of the TOPIC.	The tongue of the engine reached out for the catalyst fumes.
Verb	The ATTRIBUTE is an action that the TOPIC is performing.	My alarm clock yells at me to get of bed every morning.
Adverb	The ATTRIBUTE is a modifier describing the action that the topic is performing.	The tornado ran through the whole town without a care .
Adjective	The ATTRIBUTE is a word or phrase describing the topic.	Justice is blind and, at times, deaf .

Table 2: Examples of different types of personification ATTRIBUTES [27]. (TOPICS in blue, ATTRIBUTES in red)

wikiquette, *fashion* + *fascism* \rightarrow *fashism*. Learning a portmanteau generator requires surmounting a key hurdle – How do you learn a stable model which can generate character sequences \hat{y} which are novel, while also being “English word-like” and faithful to root words $x^{(1)}, x^{(2)}$, given only ≈ 500 training instances? We propose a noisy-channel-style, BACKWARD model $\hat{y} = \text{argmax}_y P(x|y)P(y)$, which allows incorporation of vocabulary-scale unsupervised word lists to pre-learn $P(y)$, improving performance over a standard, source-to-target, FORWARD model $\hat{y} = \text{argmax}_y P(y|x)$. Thus, by learning $P(y)$ from the English vocabulary, one can overcome paucity of training data as well as ensure the outputs are “English word-like”, as seen in Table 1.

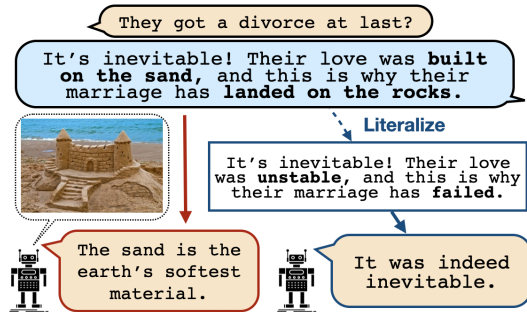


Figure 5: Example illustrating how DialogGPT responses are affected by figurative constructs in dialog context [16]. Here, the model conflates the metaphorical use of *built on the sand* with its literal meaning, leading to an inappropriate, atypical response.

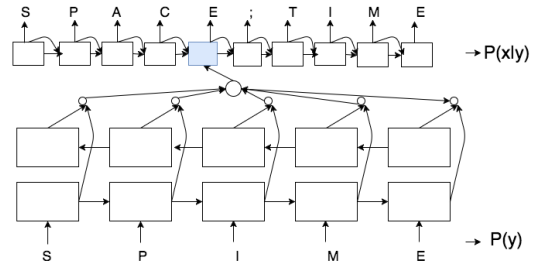


Figure 6: A sketch of our BACKWARD, SMT-style, noisy-channel model for portmanteau generation [37]. The attentional S2S model with bidirectional encoder gives $P(x|y)$ and next-character model gives $P(y)$, where y (*spime*) is the portmanteau and $x = \text{concat}(x^{(1)}, “;”, x^{(2)})$ are concatenated root words (*space,time*).

Next, in a recently presented abstract [27] and its extension, we examine the task of characterizing, identifying and generating personification — the figure of speech where inanimate things are endowed with animate abilities, e.g *My alarm clock yelled at me*. Devoid of any gold parallel training data, we use a silver dataset collection strategy — i) Build a personification identifier ii) Use the identifier to identify instances of personification y_i^{fig} from a large corpus e.g Reddit iii) Use heuristics based on underlying theories to approximately literalize each such instance $y_i^{fig} \rightarrow y_i^{lit}$. The (y_i^{lit}, y_i^{fig}) examples so generated can be used to train a generator. First addressing Steps i) and ii) in [27], we note that, unlike simpler or simpler figures of speech, personification can be hard to identify. As seen in Table 2, we develop a few common characterizations of personification based on the syntactic relation between its underlying TOPIC (the inanimate “thing”) and ATTRIBUTE (the animate property). We are currently exploring heuristics leveraging BERT-based infilling and COMET to accomplish Step iii).

3 Inducing Commonsense Plausibility Through Input Augmentation

Concept-to-text generation tasks such as WebNLG [13], E2ENLG [8] and CommonGen require generating from loosely structured inputs which are collections of keywords, ngrams, triples etc, and maybe thought of as *concept sets*. The NLG model has to then generate a plausible, faithful description featuring all the concepts with appropriate roles and relations. In this thread of work, we particularly focus on the task of generative commonsense reasoning or CommonGen, where the communicative goal is to construct a plausible situation given a set of non-abstract concepts. CommonGen was specifically designed to test the plausibility of NLG models, particularly large pretrained ones like T5. we first conduct a large-scale qualitative analysis of outputs from T5 and BART, revealing significant lacunae that hurt their plausibility, as seen in Table 3.

Concept Set	Baseline Generation	Human Reference	Issues
{horse, carriage, draw }	horse drawn in a carriage	The carriage is drawn by the horse.	Implausible Roles , Incomplete Sentence
{fish, catch, pole }	fish caught on a pole	The man used a fishing pole to catch fish.	Implausible Roles , Incomplete Sentence
{listen, talk, sit }	Someone sits and listens to someone talk.	The man told the boy to sit down and listen to him talk.	Dull Response Problem
{bathtub, bath, dog, give }	A dog giving a bath in a bathtub.	The teenager made a big mess in the bathtub giving her dog a bath.	Implausible Roles Incomplete Sentences Missing Arguments for Roles

Table 3: Example generations from baseline models versus human references [9, 10]

Self-Introspection: As first step to address the low plausibility, in [9] we posit that insufficient conditioning provided by input concept set could be a reason for low plausibility, proposing the SAPPHIRE approach based off this intuition. At training time, we can use keywords extracted off-the-shelf from references themselves to expand the input concept set; this also lessens divergence between reference and input, dampening the tendency to hallucinate. At test time, in the absence of references, we use BASELINE model outputs themselves. Evaluation reveals that SAPPHIRE enhances output plausibility as well as fluency atop either T5 or BART as the BASELINE model. This work won the Best Paper Award 🏆 at INLG 2021.



<p>{cat, bed, pet, lay}</p>  <p>BASELINE: A cat is laying on a bed and petting capt: a cat laying on a bed with a stuffed animal VisCTG: A cat laying on a bed being petted.</p>	<p>{food, eat, hand, bird}</p>  <p>BASELINE: hand of a bird eating food capt: a person holding a small bird in their hand VisCTG: A bird eats food from a hand.</p>
--	--

Table 4: Examples of retrieved images, associated captions, BASELINE and VisCTG (our visually grounded model’s) generations [10] for 2 example concept sets. Note that the images and captions are used as an intermediary to guide the final generation and thus the final generation need not be faithful to them. E.g. there is nobody petting the cat in the image or caption, but since the VisCTG output is conditioned on both the concept set and the caption, it includes *being petted*.

Grounding Through Vision: In [10], we posit that another potential reason for the low plausibility could be the reporting bias [14] and implicit understatement due to Gricean maxims prevalent in text, and it maybe beneficial to ground through the visual modality. we propose VisCTG, a method to accomplish this through retrieving images and captioning them, finally using the captions to augment example input. Thus, we implicitly ground on the visual modality by conditioning on captions of related images. Consider the right example in Table 4. The BASELINE output *hand of a bird eating food* is clearly problematic, birds do not have hands, even if they did, the semantic role of hands should be as INSTRUMENT, not AGENT (*A bird eats food using its hands*). VisCTG augments the input with the caption shown, which discusses how a bird can fit in a person’s hand, resulting in a plausible model output *A bird eats food from a hand*.

4 Incorporating Commonsense Knowledge To Fill The Gaps

For many NLG tasks, it is hard for models to learn *tabula rasa* i.e just from the training examples, to interpret, disentangle relationships in, and represent the input at a granularity sufficient to then literalize it to the output text distribution. This challenge often surfaces through poor quality of model outputs, which exhibit defects such as *hallucinations*, *repetitions* and the *dull response problem*. It is only via incorporation of knowledge commonsensical to the task that a viable NLG model can be learnt for such scenarios.

Explaining Time Series Causes: In [20], we are tasked with generating post-facto explanations of why a *feature* corpus ngram’s time series e.g *senate republicans* causally affects a given *target* time series e.g *facebook’s stock* i.e a chain of ngrams from *feature*→*target* (See Fig. 7). We posit that this requires a commonsense causative knowledge base (KB) over which an efficient abductive reasoning strategy could be devised to mine desired *feature*→*target* paths. We construct this KB by parsing corpus sentences with a Framenet-based parser [3] and aggregating identified triples, further augmenting for completeness with triples from Freebase. We first try symbolic reasoning over the KB graph, using a BFS-like backward chained search from the target. Next, to ensure richer interpretability, lexical choice and reduced dependence on heuristic matching of node surface

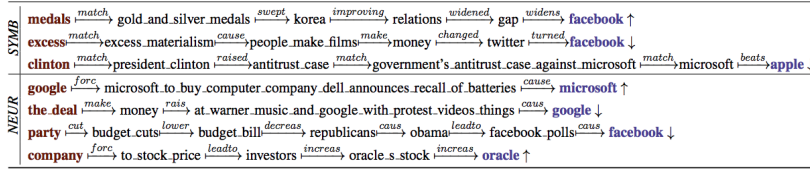


Figure 7: Causal chains for explaining [20] the rise (\uparrow) and fall (\downarrow) of companies’ stock price. The temporally causal *feature* and *target* are linked via a sequence of predicted cause-effect tuples by different reasoning algorithms: a graph reasoner SYMB and a neural reasoner NEUR.

forms, we combine symbolic and neural representations. Specifically, instead of the raw KB, we reason using a neural one-step model trained on the KB tuples to predict the next cause step. Together, our approach shows how frame-based i.e. Framenet and factual i.e. Freebase knowledge can fill in the gaps and further synergize with a neural model to solve our task.³

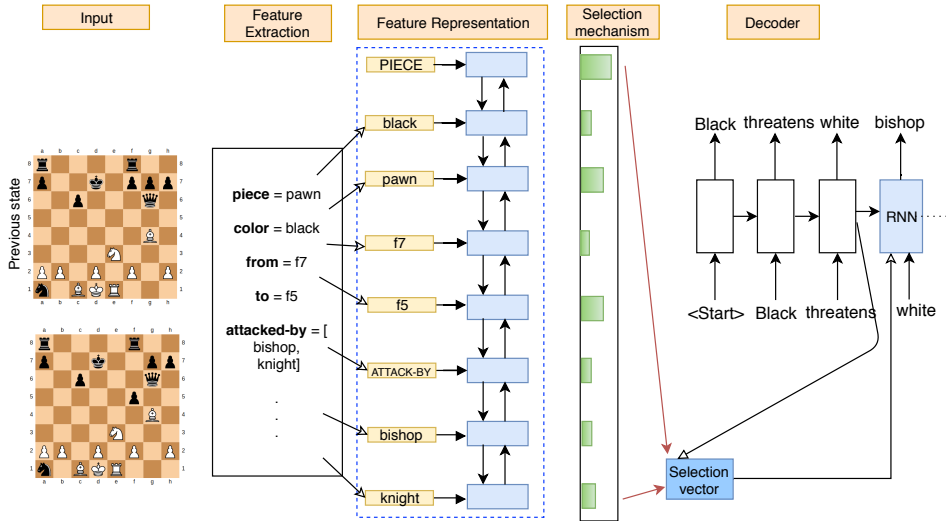


Figure 8: An overview of our chess commentary model [17]. We first extract various semantic and pragmatic features from the previous and current chess board states using a Python chess library. We represent features through embedding in a space shared with the output language embeddings. We observe that feeding in feature conjunctions helps a lot. We consider a selection mechanism for the model to choose salient attributes from the input at every decoder step.

Generating Game Commentary: We again encounter such a scenario for the task of chess commentary generation which we propose and make a first attempt to solve in [17]. This task requires generating a commentary sentence given a chess move and its previous and later board states, as shown in Fig. 8. We find that attentional encoder-decoder models are severely affected by dull responses and hallucination, often producing outputs like “The white knight moved.” and “I take the pawn” agnostic of the input move and board states. These models end up doing even worse than template and retrieval-based baselines. By simply introducing an additional layer which uses the game-based commonsense implicit in a python chess library to extract semantically and pragmatically “interesting” features such as piece positions and pairwise configurations from the board states, we get large gains in output diversity and fluency, outdoing aforementioned baselines.

5 Transferring and Annotating Style

As users get ever more habituated to using, interacting and even co-authoring with NLG systems, there is increasing expectation on them to exhibit *consistent personality* [33] and also be *accommodative* [32] towards *user preferences* and *situation of use*. Together, one can think of these as aspects of target style. Hence, NLG systems should be able to transfer their content to match aspect values for each aspect of target style. From the perspective of Halliday’s SFL, style transfer can be seen as changing *extra-textual aspects* of the communicative goal, while keeping the textual aspect constant - these aspects could be either *interpersonal* or *ideational* in nature.

Shakespearean Style Transfer: Diachronic register of language is one example of an *interpersonal* aspect. In an early 2017 work, [18], we investigated diachronic style transfer from Modern English \rightarrow Early Modern, Shakespearean English, given only $\approx 10K$ parallel pairs for training. We enumerate simple strategies to adapt the attentional sequence-to-sequence framework

³This paradigm of learning a neural reasoner from a symbolic KB has also in later literature been used to create the COMET reasoner from ConceptNet.[2]

for style transfer, exploiting distinctive task properties. First, near-similar languages on input and output sides allows sharing their embedding spaces, which can also be jointly pretrained on larger corpora . Second, since style transfer should preserve content, the model could benefit from simply learning to copy over strongly content-based words e.g topical keywords and entities without having to learn to generate them afresh — we capture this by making our output distribution a learnt mixture of “copy” (distribution over input words) and “generate” (distribution over vocabulary) components, as shown in Fig. 10. Our strategies are effective both individually and in unison, causing multifold increase in test BLEU (12 → 31), surpassing previous SOTA of ≈25. This work, being one of the earliest exploring neural style transfer, was positively received by the community, being used as a baseline by over 30+ papers and cited 120 times.

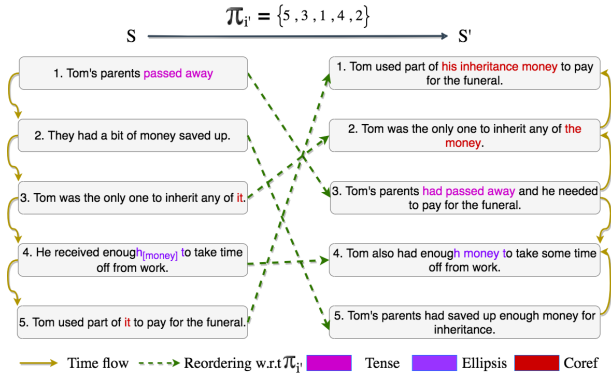


Figure 9: Example of our NAREOR task and dataset [34], with original input story S on the left, target narrative order π_i on the top, and human rewritten story S' on the right. π_i is specified in terms of sentence indices of S - e.g plot events making up sentence 5 of S should be used to make up the 1st sentence of S'

The Narrative Reordering Problem: Narrative refers to the manner in which a plot is presented in text. Two narratives may present the same plot through very disparate texts – the facets on which they make choices and differ are called narratological variables, and are, in general, an understudied topic in NLP [31]. One key variable is *narrative order* i.e the order in which the narrative presents plot events, which can differ from in-plot *chronological order*. Narrative order is one example of an *ideational aspect*.

In [34], we propose and investigate the task of *Narrative Reordering* (NAREOR) which involves rewriting a given story in a different, target narrative order while preserving its plot. Reordering narrative can impact the temporal, causal and other inferences readers draw as they read through it, which in turn affect both its interpretation and interestingness. First, we curate a dataset, NAREORC, with human rewritings of stories in non-linear orders, and analyze it in detail. Next, we propose novel task-specific training methods with suitable evaluation metrics. We perform experiments on NAREORC using state-of-the-art models such as BART and T5 and conduct extensive automatic and human evaluations. We show that though our models perform decently, NAREOR is a challenging task with potential for further exploration.

Lastly, we illustrate two applications where NAREOR is useful: for *generating more interesting* variations of stories and serving as *adversarial sets for temporal/event-related* tasks, besides discussing other prospective ones, such as for *pedagogical* setups related to language skills like *essay writing* and applications to *medicine* involving *clinical narratives*.

Annotating Persona Style Transfer: Persona i.e the values of demographic attributes for a speaker e.g gender, age and ethnicity can be seen as an *interpersonal aspect*. In [19], we collect a novel, parallel dataset with multiple rephrased variants (connotations) of the same underlying *plot* (denotation), each written by different annotators having their respective personae. We find that choice of *denotation setting* - exact plot information shown to annotators before asking them to rephrase, critically influences dataset quality e.g complete plot *text* leads to high content consistency but trivial variation in style whereas just providing *images* leads to the reverse. Comparing a wide range of settings, we find *images+keywords* reasonably trades off between consistency and diversity. Our work underscores centrality of annotation study design to the study of stylistic variation.

6 Research Agenda

Through my *primary thrust* into NLG research, I have evaluated and enhanced abilities of NLG models at tasks involving *complex facets*. NLG is *unique* amongst areas of AI in that its output can be *instantly assessed* even by a *layman*. Just like *peeling onions* reveals *new layers* each step, reaching *sufficient human-likeness* on one facet reveals *gaps on more complex facets* e.g good fluency and local coherence leads to exposure of *poor discourse-level coherence* . Further, as user get accustomed to NLG systems, their expectations on such facets increase. Evaluating and honing NLG systems towards *more*

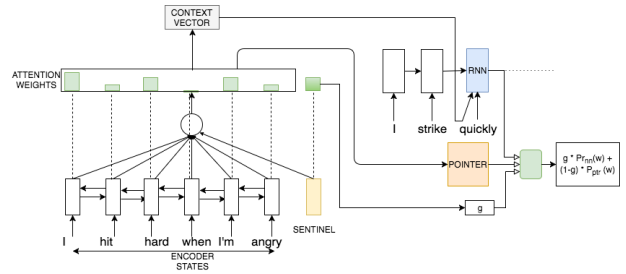


Figure 10: Depiction of our overall architecture for Shakespearean style transfer [18]. Attention weights are computed using previous decoder hidden state, encoder representations, and sentinel vector. Attention weights are shared by decoder RNN and pointer models. The final probability distribution over vocabulary comes from both the decoder RNN and the pointer network.

human-like outputs and to conform to *ever-rising user expectations* will hence be an endeavour in-progress in the near long term, providing *fertile ground* for new and interesting research directions.

Through my *secondary thrust* into DA research, I sought to complement my primary one, devising DA methods to both learn and evaluate NLG in specific and NLP in general, including *identifying research gaps* for the community and designing *mechanisms* to drive *large-scale, participative collaboration*, thereby creating standardized solutions to address them.

In the future, I am excited to explore the following directions, building upon my work along the two aforementioned thrusts.

1. **Controlling Narratives and Studying Their Downstream Application** : Narratives are an *understudied* topic in NLP [31]. As detailed in §5, I already explored controlling *narrative order* [34]. There are several other narratological variables like narrator *person* (1st vs 3rd), narrator *omniscience* (omnipresent vs limited) and character focus etc studying and controlling which would constitute equally worthy directions. My eventual goal is to build a *fully controllable narrator* that can *retell* any story as per a target configuration of all such variables.
Controllable narrators have *downstream relevance* for many education and health applications. For instance, they can *automate pedagogical setups* for fine-grained language skills like *essay argumentation* [38]. I intend to seek out collaborations with AI for Education and HCI researchers to actually *deploy narrators in practical settings* of this nature.
2. **DEI Issues in NLG tasks** : When we transfer style for some particular aspect (e.g politeness), do our models *preserve content* reasonably well for *different subgroups* along others? (e.g gender)? Do dialog models with reasonable overall response quality respond fluently and adequately to *minority and underrepresented* registers like AAVE⁴ and *Singlish*? Which strategy works well to finetune an English dialog model to a new, *code-mixed* target domain such as *Hinglish chit-chat*? These *seemingly disparate* research questions can be seen in an *unified* lens as addressing *Diversity, Equity and Inclusion* (DEI) issues in NLG tasks. Investigating and solving these questions is, in my opinion, critical to building *ethically compliant* NLG systems deployable in the real world. Hence, I hope to explore this direction in my future work. I also plan to explore collaborations with Linguistics (for challenging target domains/tasks/phenomena/applications) and Philosophy/Economics researchers (for posing new ethical norms and fairness notions to address) on this direction.
3. **Unifying Theory, Definition and Concepts For Controllable Generation** : Controllable generation refers to tasks where, the communicative goal G includes, in addition to the task description T , and an input I , a set of *control variables* or controls, whose values can be *dynamically varied* by the user at test time. Each control variable can correspond to some condition or set of values which some computable property or aspect of the output must satisfy e.g its output length, or its probability as per some classifier. An example controllable generation task is text simplification where simplicity level is controllable (high school, undergraduate, graduate). There has been a flurry of interest in controllable generation in recent years, with a rich variety of task settings [30, 28, 1] as well as novel architectures [5, 21] being explored. However, several fundamental questions remain unaddressed.
 - (a) **Hardness**: How do we *measure hardness* of a controllable generation task?
 - (b) **Cross-control comparison**: Are certain aspects of harder to control compared to others? Can we characterize categories of aspects based on aggregate hardness? Are interpersonal aspects harder to control than ideational?
 - (c) **Automatic Evaluation**: What is the best practice to *collect test references* for controllable generation evaluation? How many values of controlled aspect per input test example should reference output be annotated for to ensure adequate evaluation?
 - (d) **Input Fidelity vs Controllability**: Is a model that is less faithful to the input but more controllable better than one that's sufficiently faithful but less controllable? How do we *evaluate and balance* this tradeoff?

As a *first step* towards answering some of these questions, we are co-organizing the *Controllable Generative Modeling in Language and Vision* (CTRLGEN) [workshop](#) at NEURIPS'21. We aim to bring together researchers from the NLP, Vision, and ML communities to discuss the *current challenges* and *explore potential directions* for controllable generation and improve its quality, correctness, and diversity. Furthermore, we are also currently working towards a *survey paper* categorizing and deriving shared concepts from existing work, besides identifying future challenges.

4. **NLG for Law** : As society becomes more urbanized and institutionally complex, it has also become *more litigious* - resulting in judges, lawyers and juries ever being overburdened, with *burgeoning case backlogs*. NLP systems streamlining information access and guiding decisions can ease this backlog and help provide *timely delivery* of justice. Though there has been much growth in the field of AI for Law in general, NLG's potential remains grossly underexplored. Most research studies as well as commercial applications using NLP for Law are restricted to information extraction, retrieval, classification and other NLU problems. At the 2021 edition of [ICAIL](#), which is the key conference of the AI for Law community, *only* 12% of the papers discussed or involved NLG.

I hope to *tap this potential*, both via research studies as well as building real-world legal applications using NLG (e.g *style transfer to simplify* case briefs, laws policies and precedents for law students or clients), particularly as part of

⁴African American Vernacular English

funded projects/collaborations or grants on this theme. I am currently curating a AI for law reading list to better plan my foray.

5. **Discourse Level DA** : Most DA methods for NLP e.g random token shuffle, mask-and-infill and synonym replacement are defined at the “*sentence level*”. Even when they operate on multi-sentence paragraphs, stories or documents, they either are applied sentence by sentence (e.g for backtranslation) or treat the whole input *as if it were one sentence*. Hence, there is a dearth of methods which exploit *invariants* based on *discourse level phenomena* such as ellipsis, coreference, RST structure etc to perform DA . Devising such methods could drive up performance on low-resource, multi-sentence tasks such as document classification, and hence represent a promising avenue for future investigation.
6. **NLP for Finance, Accounting, Auditing and Economics** : In one of my earliest research works at EMNLP’17, we explored the problem of *identifying* and *explaining* text features predictive of *stock market prices* based on a parallelly collected Twitter corpus (See §4). This led to interesting conversations with NLP researchers working at *financial firms* during the poster session, though we couldn’t fructify these into further collaborations. Through 2018, I worked on a funded collaborative project with the *core auditing arm* of PricewaterhouseCoopers (PwC) and an interdisciplinary team of CMU computer science and management faculty and graduate students. Our objective was to use AI to automate the *cash confirmation process* - where client spreadsheet data with arbitrary schemas need to be aggregated into a canonical quasi-tabular format which is vetted by auditors. We successfully met this objective with a *few-shot learnable NLU pipeline*. Moving ahead, I aim to actively pursue and foster such modalities of collaboration with financial institutions such as auditing & accounting firms, investment banks, consulting firms and hedge funds. Besides providing a rich source for novel NLP tasks and learning scenarios, this can also open up a steady stream of funding.
7. **Understanding Extra-Sentential Abilities of Contextual Embeddings** : Designing probes to evaluate the *intrinsic abilities* of contextual embeddings has been an active research area, driven by the annual BlackBoxNLP workshops [25]. A wealth of probing work investigates *intra-sentential* abilities such as word and number agreement [26], function words [22], constituency and dependency structures [4]. However, their *extra-sentential* abilities have been relatively underexplored. Understanding these abilities has great relevance for *multi-sentence tasks* like document classification and *story completion*, and could also *motivate novel pretraining objectives*. Hence, in my view, it is a worthy direction for future research. My two recent contributions on *locating event arguments* using attention heads [35] and a benchmark for *infilling whole sentences* a.k.a “sentence cloze” tests [23], represent an initial foray towards the same.

6.1 Collaboration

Through the course of my research journey, I have had the good fortune of collaborating with **over 60 researchers** spanning over **28 institutions**, of which **18 were universities** and **1 a government agency (DARPA)**, with the rest being corporations. This has primed me to *learn* and *adapt* to work with collaborators following a diverse *variation of work cultures, career stages, skill sets and research motivations*.

Going forward, as a potential faculty member, I will continue to *actively seek out* and *foster* such collaboration — for instance, just recently I’ve begun collaborating with a U.Belfast NLP researcher on *fairness issues in textual style transfer*. An *added advantage* of such collaborations which I hope to leverage is the *funding opportunities* these can open up — e.g internal sources in the *collaborator’s institution*, or tied to the institution’s geography e.g *country-specific grants*.

6.2 Funding

Although funding is an oft *understated and less glamorized* aspect of *research dynamics*, I have come to realize that its in fact the *lifeblood that sustains* academic research positions and *steers* long term research agenda and *priorities*. I aim to slowly *shape my thought process* so as to put *funding potential* at the centre of each candidate research direction I consider, before *actually expending resources* to explore it.

I also realize that every funding agency and source have their *own idiosyncrasies* and strategies of pitching to them (DARPA, NSF, NIH) , and if given a chance would be grateful to *collaborate with more experienced peers*, trying my best to grasp the *art of grant authoring and subsequent execution*. I have also worked under *tech* (Facebook ConvAI) and *non-tech industrial* (PwC) and agency funding (DARPA) and understand the *difference in processes, goals and expectations* between these three.

References (*=equal contribution)

- [1] AMPLAYO, R. K., AND LAPATA, M. Informative and controllable opinion summarization. *arXiv preprint arXiv:1909.02322* (2019).

- [2] BOSSELUT, A., RASHKIN, H., SAP, M., MALAVIYA, C., CELIKYILMAZ, A., AND CHOI, Y. Comet: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 4762–4779.
- [3] CHEN, D., SCHNEIDER, N., DAS, D., AND SMITH, N. A. Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th international workshop on semantic evaluation* (2010), pp. 264–267.
- [4] CLARK, K., KHANDELWAL, U., LEVY, O., AND MANNING, C. D. What does bert look at? an analysis of bert’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (2019), pp. 276–286.
- [5] DATHATHRI, S., MADOTTO, A., LAN, J., HUNG, J., FRANK, E., MOLINO, P., YOSINSKI, J., AND LIU, R. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations* (2019).
- [6] DHOLE, K., GANGAL, VARUN, GEHRMANN, S., GUPTA, A., LI, Z., MAHAMOOD, S., AND OTHERS. NI-Augmenter. <https://github.com/GEM-benchmark/NL-Augmenter>, 2021.
- [7] DHOLE, K. D., GANGAL, VARUN, GEHRMANN, S., GUPTA, A., LI, Z., MAHAMOOD, S., MAHENDIRAN, A., MILLE, S., SRIVASTAVA, A., TAN, S., ET AL. NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation. *arXiv preprint arXiv:2112.02721* (2021).
- [8] DUŠEK, O., NOVIKOVA, J., AND RIESER, V. Findings of the E2E NLG Challenge. In *Proceedings of the 11th International Conference on Natural Language Generation* (2018), pp. 322–328.
- [9] FENG, S., HUYNH, J., NARISSETY, C. P., HOVY, E., AND GANGAL, VARUN. SAPPHERE: Approaches for Enhanced Concept-to-Text Generation. In *Proceedings of the 14th International Conference on Natural Language Generation* (2021), pp. 212–225.
- [10] FENG, S. Y., LU, K., TAO, Z., ALIKHANI, M., MITAMURA, T., HOVY, E., AND GANGAL, VARUN. Retrieve, Caption, Generate: Visual Grounding for Enhancing Commonsense in Text Generation Models. In *Proceedings of the AAAI Conference on Artificial Intelligence (To appear)* (2022), vol. 34, pp. 7764–7771.
- [11] FENG*, S. Y., GANGAL*, VARUN, KANG, D., MITAMURA, T., AND HOVY, E. GenAug: Data Augmentation for Finetuning Text Generators. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures* (2020), pp. 29–42.
- [12] FENG*, S. Y., GANGAL*, VARUN, WEI, J., CHANDAR, S., VOSOUGH, S., MITAMURA, T., AND HOVY, E. A Survey of Data Augmentation Approaches for NLP. *arXiv preprint arXiv:2105.03075* (2021).
- [13] GARDENT, C., SHIMORINA, A., NARAYAN, S., AND PEREZ-BELTRACHINI, L. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation* (2017), pp. 124–133.
- [14] GORDON, J., AND VAN DURME, B. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction* (2013), pp. 25–30.
- [15] HALLIDAY, M. A. K. Language as social semiotic. *The Discourse Studies Reader*. Amsterdam: John Benjamins (1978), 263–272.
- [16] JHAMTANI*, H., GANGAL*, VARUN, HOVY, E., AND BERG-KIRKPATRICK, T. Investigating Robustness of Dialog Models to Popular Figurative Language Constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 7476–7485.
- [17] JHAMTANI*, H., GANGAL*, VARUN, HOVY, E., NEUBIG, G., AND BERG-KIRKPATRICK, T. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 1661–1671.
- [18] JHAMTANI*, H., GANGAL*, VARUN, HOVY, E., AND NYBERG, E. Shakespearizing Modern Language Using Copy-Enriched Sequence to Sequence Models. In *Proceedings of the Workshop on Stylistic Variation* (2017), pp. 10–19.
- [19] KANG*, D., GANGAL*, VARUN, AND HOVY, E. (Male, Bachelor) and (Female, Ph. D) have different connotations: Parallely Annotated Stylistic Language Dataset with Multiple Personas. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019* (2020), Association for Computational Linguistics, pp. 1696–1706.

- [20] KANG, D., GANGAL, VARUN, LU, A., CHEN, Z., AND HOVY, E. Detecting and Explaining Causes From Text For a Time Series Event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017), pp. 2758–2767.
- [21] KESKAR, N. S., McCANN, B., VARSHNEY, L. R., XIONG, C., AND SOCHER, R. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).
- [22] KIM, N., PATEL, R., POLIAK, A., XIA, P., WANG, A., MCCOY, T., TENNEY, I., ROSS, A., LINZEN, T., VAN DURME, B., ET AL. Probing what different nlp tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)* (2019), pp. 235–249.
- [23] KONG*, X., GANGAL*, VARUN, AND HOVY, E. SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 5668–5683.
- [24] LIN, B. Y., ZHOU, W., SHEN, M., ZHOU, P., BHAGAVATULA, C., CHOI, Y., AND REN, X. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (2020), pp. 1823–1840.
- [25] LINZEN, T., CHRUPALA, G., AND ALISHAHI, A. Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (2018).
- [26] LINZEN, T., DUPOUX, E., AND GOLDBERG, Y. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics 4* (2016), 521–535.
- [27] LU, K., FENG*, S. Y., GANGAL*, VARUN, JHAMTANI, H., AND HOVY, E. Personifications are Cunning: Exploring Approaches For Personification Identification. *New Directions in Analyzing Text as Data (TADA)* (2021).
- [28] MARTIN, L., DE LA CLERGERIE, É. V., SAGOT, B., AND BORDES, A. Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference* (2020), pp. 4689–4698.
- [29] MILLE, S., DHOLE, K. D., MAHAMOOD, S., PEREZ-BELTRACHINI, L., GANGAL, VARUN, KALE, M., VAN MILTENBURG, E., AND GEHRMANN, S. Automatic Construction of Evaluation Suites for Natural Language Generation Datasets. *arXiv preprint arXiv:2106.09069* (2021).
- [30] PENG, N., GHAZVININEJAD, M., MAY, J., AND KNIGHT, K. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling* (2018), pp. 43–49.
- [31] PIPER, A., SO, R. J., AND BAMMAN, D. Narrative Theory for Computational Narrative Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 298–311.
- [32] SINGHAL, A. Would You Interact with a Chatbot that’s Unfriendly? 67% users say no! <https://www.entrepreneur.com/article/339248>, 2021.
- [33] STAFF, V. People, not tech companies should pick their AI Assistant’s personality. <https://venturebeat.com/2017/10/>, 2021.
- [34] GANGAL*, VARUN, FENG*, S. Y., ALIKHANI, M., MITAMURA, T., AND HOVY, E. Nareor: The Narrative Reordering Problem. In *Proceedings of the AAAI Conference on Artificial Intelligence (To appear)* (2022), vol. 34, pp. 7764–7771.
- [35] GANGAL*, VARUN, AND HOVY, E. BERTering RAMS: What and How Much does BERT already Know About Event Arguments?-A Study on the RAMS dataset. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (2020), pp. 1–10.
- [36] GANGAL*, VARUN, JHAMTANI*, H., HOVY, E., AND BERG-KIRKPATRICK, T. Improving Automated Evaluation of Open Domain Dialog via Diverse Reference Augmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Online, Aug. 2021), Association for Computational Linguistics, pp. 4079–4090.
- [37] GANGAL*, VARUN, JHAMTANI*, H., NEUBIG, G., HOVY, E., AND NYBERG, E. Charmanteau: Character Embedding Models For Portmanteau Creation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017), pp. 2917–2922.

- [38] WINGATE, U. ‘argument!’helping students understand what essay writing is about. *Journal of English for academic purposes 11*, 2 (2012), 145–154.
- [39] ZESCH, T., MÜLLER, C., AND GUREVYCH, I. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *LREC* (2008), vol. 8, pp. 1646–1652.