Tanya Goyal

Research Statement

The Internet is a valuable but vast source of knowledge, providing more information than we can consume for any topic. AI systems that can engage with end user needs using natural language and convey key points in the form of smaller digestible pieces are going to be invaluable. We have already made great strides in this direction in recent years, owing to breakthroughs like pre-trained language models such as GPT-3. At the same time, successes in this area are greatly overestimated due to outdated evaluation tools that ignore critical failure modes.

When we ask a large language model for information (e.g. *What are herbivores?*), it has a lot of flexibility in how that information is presented. For example, it may output *"animals that only eat plants,"* or *"animals whose primary food source is plant-based."* This makes evaluation here significantly more challenging compared to classification problems with a pre-defined label space. Not only do evaluation tools need to recognize that both outputs are semantically equivalent and correct, they also need to penalize seemingly innocuous changes like *"animals that **can** eat plants,"* as well as account for numerous other phenomena. Standard metrics rely on superficial surface cues and struggle with this. One direction of my research (Figure 1b) builds evaluation models that can operate over such tricky cases and render judgment for important aspects like factuality and coherence of generated text [1, 2, 3]. We design our model outputs to be understandable by humans; errors are localized to sub-parts of the generated text instead of a single uninformative judgment for the whole text.

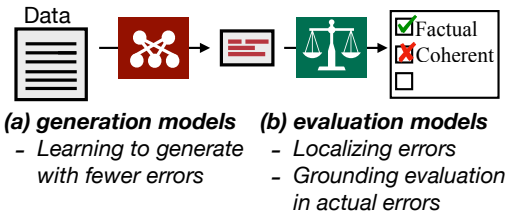

Figure 1: Overall research vision

The other challenge is ensuring generation models themselves output higher quality outputs. A basic requirement here is training data that exactly reflects our goals for models to emulate. But human annotation for generation is expensive and automatically collected data can be noisy and encourage the very behaviors we want to avoid, e.g. bias. The second direction of my work develops training techniques that can disambiguate between *good* and *noisy* signals [2, 4, 5] in existing data and only learn to emphasize the former (Figure 1a). Overall, my work provides different options of easily adaptable modifications to standard training pipelines that help learn better generation distributions.

Beyond explicit generation tasks, there has been a recent shift towards using natural language generation as an integral component of other research pipelines such as reasoning. Despite using stronger models like GPT-3, these retain flavors of the same errors we see in typical generation models. I am excited to continue my research on detecting and fixing generation errors in these other domains.

Localizing errors for better evaluation

Generation models can exhibit different types of failures, even for relatively constrained problems like summarization, with varying degrees of consequences. The most critical ones, however, are those where the model generates text that is non-factual, i.e. unsupported or contradictory to the given input. Foremost, we need evaluation tools to detect these errors before we can safely deploy models.

First, how is model-generated text usually evaluated? The traditional way is to measure overall "quality" using word overlap against a gold standard. Although clearly limited (e.g. it ignores

all surrounding context), such metrics are still commonly preferred as their outputs are intuitively understandable. On the other extreme are neural approaches that use distributed representations to render an overall judgment for the whole text, usually targeting specific dimensions like factuality. These can, principally, handle a more complex set of linguistic phenomena but are uninterpretable. Ideally, we want an evaluation system that strikes a balance between these two designs.

Our factual error detection model for conditional generation, called ***dependency arc entailment*** (DAE), follows this principle [1, 2]. DAE decomposes the overall error detection task into smaller entailment tasks that predict whether individual word relationships are entailed by the input (Figure 2). This set of word relationships we evaluate over is informed by the syntactic structure, specifically dependency parse, of the generated text. We fine-tune a strong pre-trained model to predict the arc-level factuality. Our modeling combines the strengths of the two approaches; we can localize factual errors to smaller interpretable units (e.g. to the *woman* → *Chicago* arc in Figure 2) while also utilizing state-of-the-art encoder models to extract the best performance. Our DAE model achieved state-of-the-art results on text summarization and paraphrasing when it was first introduced.

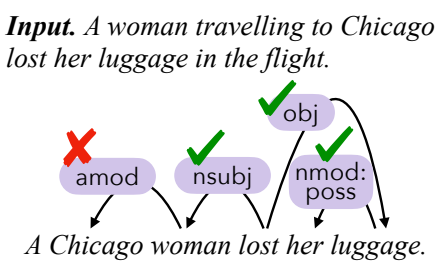

Figure 2: Error localization w/ DAE

In addition to performance, DAE also stands out as one of the only approaches that explicitly models and evaluates localization. Our closest competitors are the family of QA-based models that are specifically designed to localize errors but never tested on this. We addressed this gap in a recent collaboration [6]; we found that DAE localizes errors significantly better than the strongest QA-based models across all settings. DAE's localization has been used in downstream applications for denoising training data [7], improving training itself [8], and extended to other structural representations of word relationships [9].

Grounding evaluation in actual model errors

One of the most challenging aspects of our evaluation problem is the constantly evolving error space. The reason we saw traditional metrics fail is because the error distributions they were designed to target no longer matched those of more recent generation models. Within factuality, my work [2] was the first to show that this mismatch is responsible for poor model performance and grounding in observed errors is essential. Unless we target the right errors, we risk spending valuable resources on meaningless benchmark improvements. My research contributes annotated datasets [3, 10] and tools [11] to this direction.

My work on Summary Narrative Coherence (SNaC) [3] explicitly demonstrates this shift in error space when we move to a more powerful generation model (GPT-3) and a more challenging domain, i.e., long narrative summarization. Most strikingly, coherence errors, which did not surface for earlier summaries, have now emerged as a first-order issue when summarizing complex narratives (Figure 3). But coherence is a broad term that encompasses a number of different criterion. To actually understand model limitations and fix them, we require

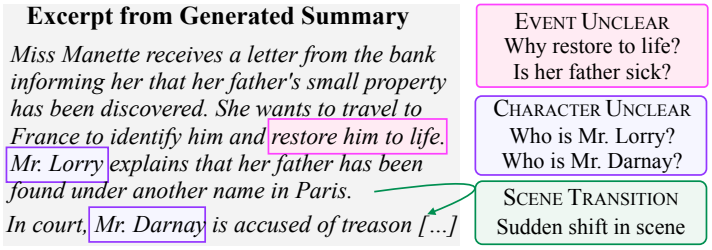

Figure 3: Coherence issues in long narratives

a finer-grained understanding of what *types* of coherence errors models make and how this is impacted by factors like model sizes and specific sub-domains (e.g. screenplay vs book narratives). With SNaC, we developed a span-level error schema that was grounded in actual narrative summarization errors and released a large annotated dataset that contrasted the error distributions across different models and domains. As with DAE, we found this fine-grained treatment of errors to be beneficial during automatic error detection. We show that models trained on SNaC can detect if a sentence has coherence errors with respect to its context and also pinpoint the spans where the errors occur. This opens up exciting opportunities for post-hoc error correction that future work can build on.

Throughout my research, I have worked to steer evaluation towards real system errors. In more recent collaborative work [10], we re-analyzed available factuality datasets and stratified them based on error types and models. Our work revealed that much of the improvements in factuality work in recent years applied to older, obsolete models instead of current ones. In a similar vein, my research also showed that GPT-3 style models bring yet another paradigm shift in generation and that we need updated evaluation tools to audit these models going forward [12].

Learning to generate with fewer errors

With access to reliable evaluation, we are now better poised to measure model quality along meaningful axes like factuality and pursue ways to improve it. First, why do models generate poor quality outputs? This can be attributed to many different reasons (e.g. imperfect training objectives, autoregressive nature of the models, etc.), but the most prominent one is the lack of quality training data. Because we still need decently sized training datasets and manually curating them for generation is expensive, they are usually built by automatically scraping any publicly available task-adjacent data. For example, most summarization datasets are built by combining inputs with any accompanying summary-like text. But this does not guarantee *good* summaries that present the most salient information succinctly and factually. To build useful systems on top of such flawed datasets, we need ways to separate out the *good* training signals from the rest and only train on the former.

One important consideration here is the dataset size after filtering, which can significantly impact model performance. We faced this issue while training summarization models to be more factual; eliminating non-factual summaries meant reducing the dataset size by more than 50% for some datasets. In my work, we show that we can sidestep such issues through error localization [2, 4]. We already know from our evaluation research that most noisy instances are only *locally noisy*, i.e., errors are usually restricted to a fraction of words or word relationships. Separately, we know generation models operate at the token-level. Our core insight is that these two factors can be effectively combined to remove the training signal from noisy tokens only instead of removing whole sequences. This works very well in practice. On text summarization, we show that using DAE to localize errors followed by this loss truncation strategy can improve factuality by more than 20% [2]. In follow-up work, we showed that we can use token-level loss changes during training (*"training dynamics"*) as factuality indicators [4], instead of DAE, allowing us to extend our technique to domains that do not have high performing error localizers.

I have also done work that equips generation models with controllable knobs to vary output properties. Our framework for paraphrasing provides syntactic control [13] that we use to improve downstream data augmentation [1]. With HydraSum [5], we introduced modular generation architectures where multiple experts learn mutually-distinct generation strategies. At inference, Hy-

drasum provides users the flexibility to sample from individual experts or combine them to tailor outputs to their personal preferences.

Future directions

Building sustainable evaluation frameworks As models continue to evolve, either in terms of scale, architectures, or the underlying training distributions, their behaviors also change in unpredictable ways. While their performance on classification can be benchmarked using existing static test sets, generation evaluation is much trickier as metrics might themselves become obsolete [12, 14]. Not only do we need ways to dynamically update metrics to reflect current systems and errors, but also techniques to detect *when* updates are required. I believe that one way to achieve both these goals is by incorporating cues from the target models themselves into our evaluation pipelines. In my previous work [4], I showed that a generation model's comparisons with its previous training iterations can provide useful signals about failures modes and ambiguous examples under its current distribution. Recently, similar ideas were successfully used to improve generation decoding [15]. Can we design methods that effectively integrate such additional signals into existing evaluation models? If we ensure that our integration is modular, i.e. can be easily replaced by updated models, we can develop adaptive evaluation systems that require minimal, if any, manual interventions.

Aligning generation models to users' needs The ultimate goal of generation models is to cater to end users. Despite this, the actual task formulations have been influenced more by easy availability of data than actual user needs. In preliminary work, I showed that humans do not actually prefer outputs that emulate the gold standard in text summarization [12]. As our models progress from research prototypes to user-facing systems, we need to align research goals with actual application scenarios ("purpose factors") [16]. This involves actively engaging with end users during all stages, e.g. during task design to understand how they would use, say, an automatic summarizer (are these intended to be writing assistants or end products, what kind of summaries support each of these goals), data curation to understand their preference functions (can users provide training signal instead of automatically procured datasets), and system evaluation. I look forward to collaborations with HCI researchers and domain-experts to chart a more user-driver research agenda for generation models.

Beyond unstructured text The strength of large language models can be attributed to the pre-training on massive amounts of unstructured text. While these handle natural language prompts quite well, they are not well-equipped to handle other modalities of data, e.g. tables. This has resulted in a clear difference in the model performances on text-to-text tasks and data-to-text tasks, in stark contrast with humans who are good at interpreting both. The ability to reason over both textual and structured knowledge is an important requirement for AI evaluators (or writers), e.g. consider use cases where data records need to be automatically verified (or updated) based on new information represented in either format. A common way to use tabular data as inputs to language models is to linearize it, but that gets rid of the very structure that makes tables excellent choices for storing knowledge. Instead, can we use the superior generation capabilities of language models to decompose textual data into atomic facts that can be more readily compared against data tuples in the table? I am interested in pursuing ways to bridge the gap between these modalities, especially within the framework of large pre-trained models.

References

[1] **Tanya Goyal** and Greg Durrett. Evaluating factuality in generation with dependency-level entailment. *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

[2] **Tanya Goyal** and Greg Durrett. Annotating and modeling fine-grained factuality in summarization. *NAACL*, 2021.

[3] **Tanya Goyal**, Junyi Jessy Li, and Greg Durrett. SNaC: Coherence error detection for narrative summarization. *EMNLP*, 2022.

[4] **Tanya Goyal**, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. Training dynamics for text summarization models. *Findings of the Association for Computational Linguistics: ACL 2022*.

[5] **Tanya Goyal**, Nazneen Fatema Rajani, Wenhao Liu, and Wojciech Kryściński. HydraSum: Disentangling stylistic features in text summarization using multi-decoder models. *EMNLP*, 2022.

[6] Ryo Kamoi, **Tanya Goyal**, and Greg Durrett. Shortcomings of question answering based factuality frameworks for error localization. *arXiv*, 2022 (In submission).

[7] Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. Questioning the validity of summarization datasets and improving their factual consistency. *EMNLP*, 2022.

[8] Prafulla Kumar Choubey, Jesse Vig, Wenhao Liu, and Nazneen Fatema Rajani. CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization. *arXiv*, 2021.

[9] Leonardo FR Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. FactGraph: Evaluating factuality in summarization with semantic graph representations. *NAACL*, 2022.

[10] Liyan Tang, **Tanya Goyal**, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv*, 2022 (In submission).

[11] **Tanya Goyal**, Junyi Jessy Li, and Greg Durrett. FALTE: A toolkit for fine-grained annotation for long text evaluation. *EMNLP: System Demonstrations*, 2022.

[12] **Tanya Goyal**, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of GPT-3. *arXiv*, 2022.

[13] **Tanya Goyal** and Greg Durrett. Neural syntactic preordering for controlled paraphrase generation. *ACL*, 2020.

[14] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv*, 2022.

[15] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv*, 2022.

[16] Karen Spärck Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 2007.