Towards Content Transfer through Grounded Text Generation

Shrimai Prabhumoye
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15219
sprabhum@andrew.cmu.edu

Chris Quirk, Michel Galley
Microsoft Research
One Microsoft Way
Redmond, WA 98052
{chrisq,mgalley}@microsoft.com

Abstract

Recent work in neural generation has attracted significant interest in controlling the *form* of text, such as style, persona, and politeness. However, there has been less work on controlling neural text generation for content. This paper introduces the notion of Content Transfer for long-form text generation, where the task is to generate a next sentence in a document that both fits its context and is grounded in a content-rich external textual source such as a news story. Our experiments on Wikipedia data show significant improvements against competitive baselines. As another contribution of this paper, we release a benchmark dataset of 640k Wikipedia referenced sentences paired with the source articles to encourage exploration of this new task.

1 Introduction

Recent work in neural natural language generation (NLG) has witnessed a growing interest in controlling text for various form-related and linguistic properties, such as style (Ficler and Goldberg, 2017), affect (Ghosh et al., 2017), politeness (Sennrich et al., 2016), persona (Li et al., 2016b) voice (Yamagishi et al., 2016), grammatical correctness (Ji et al., 2017), and length (Kikuchi et al., 2016). This trend offers the promise of empowering existing authoring tools such as Grammarly, Google Smart Compose, and Microsoft Word with the ability to control a much greater variety of textual properties, which are currently mostly limited to grammar, spelling, word choice, and wordiness. What has been relatively less explored in neural NLG research is the ability to control the generation of a current sentence not only in its *form*, but also its *content*.[1] Consider for example Fig. 1, which illustrates a situation where an author edits a document (here a Wikipedia article),

Figure 1: Example of content transfer: Given existing curated text (yellow) and a document with additional relevant information (green), the task is to update the curated text (orange) to reflect the most salient updates.

and the goal is to generate or suggest a next sentence (shown in orange) to the author. This type of unconstrained, long-form text generation task (Mostafazadeh et al., 2016; Fan et al., 2018) is of course extremely difficult. Free-form generation can easily go astray due to two opposing factors. On one hand, ensuring that the generated output is of relatively good quality often comes at the cost of making it bland and devoid of factual content (Li et al., 2016a). On the other hand, existing techniques can help steer neural models away

[1] Historically, NLG has focused on generation from structured content such as a database or semantic representation, but this paper is interested in generation from free-form text.

from blandness in order to produce more contentful outputs (using temperature sampling (Fan et al., 2018), GAN (Goodfellow et al., 2014), etc.), but often at the cost of "hallucinating" (Wiseman et al., 2017) words or concepts that are totally irrelevant. Neither situation provides a compelling experience to the user.

What is clearly missing from the aforementioned authoring scenario is the notion of *grounding*: there is often a profusion of online resources that bear at least some relevance to any given document currently being written. Much of the general-purpose world knowledge is available in the form of encyclopedias (e.g., Wikipedia), books (e.g., Project Gutenberg, Google Books), and news articles. While the generation of good quality texts without any conditioning on "external" sources (Fan et al., 2018) might be an interesting research endeavor on its own, we argue that grounding can make the generation task much easier, e.g., as shown in Fig. 1 where a passage of a news article (green) can be reformulated considering the current context of the document (yellow) in order to produce a natural next sentence (orange). In light of this desideratum, this paper addresses the problem of grounded text generation, where the goal is to infuse the content or knowledge from an external source (e.g., a news article as in Fig. 1) in order to generate a follow-up sentence of an existing document. We see this as a form of *Content Transfer*, as other characteristics of the external source—such as style and linguistic form—are not controlled.

In addition to formulating this new task, our work makes the following contributions: We provide a large dataset of 640k instances that contain parallel data of a source document (news articles), a context, and sentence to be produced. The latter two are extracted from Wikipedia, which is an attractive dataset for grounded generation as many of the statements in Wikipedia cite external sources (i.e., grounded in an external article). Finally, we also provide simple yet efficient models that condition both on the external article and the context of the current document. We compare our models against extractive and abstractive baselines, including summarization methods that simply try to condense the external article without considering the context of the document. Our experiments show that our models which incorporate the context gain 7.0 ROUGE-L F1 points —in other words, treating our task as a summarization problem is not enough. Our human evaluations also show that models that are aware of the context generate relevant and fluent sentences that are coherent to the context.

2 Task

This research is concerned with the general problem of grounded authorship assistance, i.e., the task of suggesting text to insert or append in an existing document draft, in such a way that all the added *content* reflects information from external sources, such as news articles and books. This type of grounded generation task could take many forms, so we decided to formalize the task as follows, while still keeping the task both challenging and practically interesting. Given an external **document** (green in Fig. 1), and some existing **curated text** (yellow), the task is to generate a single **update sentence** (orange). This update sentence should be both relevant to the context and reflective of the information contained in the document.

This task bears some similarity with automatic summarization (Nenkova and McKeown, 2011), as a naïve approach to the above problem is to append a one-sentence summary of the document to the curated text. While indeed related, the two tasks differ in two key points. First, the one-sentence summary must be contextually appropriate given the previous context of the curated text. Second, summarization is mostly concerned with finding *salient* information, but—in the case of our task—information relevant to the context might actually only be auxiliary within the external document. Section 6 (Related Work) further contrasts our task with summarization.

Formally we define our task as follows: given an existing curated text $\boldsymbol{s}$ and a document $\boldsymbol{d}$ describing novel information relevant to that text, the system must produce a revised text $\boldsymbol{s'}$ that incorporates the most salient information from $\boldsymbol{d}$. We restrict our focus to the cases where the revised text $\boldsymbol{s'}$ can be obtained by appending the new information from $\boldsymbol{d}$ to the original curated text $\boldsymbol{s}$.[2] In particular, we assume we can transform the old curated text $\boldsymbol{s}$ into the new text $\boldsymbol{s}'$ by appending one additional update sentence $\boldsymbol{x}$ to $\boldsymbol{s}$.

[2] In general, updated information from $\boldsymbol{d}$ might demand substantial changes to $\boldsymbol{s}$: perhaps core assumptions of $\boldsymbol{s}$ were contradicted, necessitating many removed and rewritten sentences. We postpone this complex setting to future work.

3 Models

This paper operates in a conventional supervised learning setting. For training data, we rely on a large dataset of existing curated text $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, corresponding documents with novel information $\boldsymbol{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n\}$, and the update sentences $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Our task is to generate the update sentence $\boldsymbol{x}_i$ that could be appended to the curated text $\boldsymbol{s}_i$ in order to incorporate the additional information from document $\boldsymbol{d}_i$. The goal would be to identify new information (in particular, $\boldsymbol{d}_i \setminus \boldsymbol{s}_i$) that is most salient to the topic or focus of the text, then generate a single sentence that represents this information.

3.1 Generative models

A natural though difficult means of generating this additional update sentence $\boldsymbol{x}$ is to use a generative model conditioned on the information in the curated text $\boldsymbol{s}$ and the new document $\boldsymbol{d}$. Recent methods inspired by successful neural machine translation systems have produced impressive results in abstractive summarization (Nallapati et al., 2016). Hence, our first step is to use the sequence-to-sequence encoder-decoder model (Bahdanau et al., 2015) with attention (Luong et al., 2015) for our task. This kind of model assumes that the output sentence can be generated word-by-word. Each output word $\boldsymbol{x}_i^t$ generated is conditioned on all prior words $\boldsymbol{x}_i^{<t}$ and an encoded representation of the context $\boldsymbol{z}$:

$$\prod_t p(\hat{\boldsymbol{x}}_i^t | \hat{\boldsymbol{x}}_i^{<t}, \boldsymbol{z}) \tag{1}$$

Context Agnostic Generative (CAG) Model: One simple baseline is to train a sequence-to-sequence model for the document $\boldsymbol{d}$ alone that does not directly incorporate information from the curated text $\boldsymbol{s}$. Here, the algorithm is trained to generate the most likely update sentence $\hat{\boldsymbol{x}} = \arg\max p(\boldsymbol{x}|\boldsymbol{d})$. In this setting, we consider the reference document $\boldsymbol{d}_i$ as the source and the update sentence to be generated $\boldsymbol{x}_i$ as the target.

$$\boldsymbol{z} = \text{Encoder}(\boldsymbol{d}_i, \boldsymbol{\theta}) \tag{2}$$

The encoder and decoder do not directly see the information from the curated text $\boldsymbol{s}$, but the update $\boldsymbol{x}$ inherently carries some information about it. The parameters of the model are learned from updates that were authored given the knowledge of the curated text. Hence, the model may capture some generalizations about the kinds of information and locations in $\boldsymbol{d}$ that are most likely to contribute novel information to $\boldsymbol{s}$.

Context Only Generative (COG) Model: This algorithm is trained to generate the most likely update sentence $\hat{\boldsymbol{x}} = \arg\max p(\boldsymbol{x}|\boldsymbol{s})$. This model is similar to CAG except that we consider the curated $\boldsymbol{s}_i$ as the source. In this setting, there is no grounding of the content to be generated.

Context Informed Generative (CIG) Model: An obvious next step is to incorporate information from the curated text $\boldsymbol{s}$ as well. We can concatenate the document and the curated text, and produce an encoded representation of this sequence.

$$\boldsymbol{z} = \text{Encoder}([\boldsymbol{d}_i; \boldsymbol{s}_i], \boldsymbol{\theta}) \tag{3}$$

This approach incorporates information from both sources, though it does not differentiate them clearly. Thus, the model may struggle to identify which pieces of information are novel with respect to the curated text.

To clearly identify the information that is already present in the curated text $\boldsymbol{s}$, a model could encode $\boldsymbol{s}$ and $\boldsymbol{d}$ separately, then incorporate both signals into the generative procedure.

Context Receptive Generative (CRG) Model: Our next step was to condition our generative process more concretely on the curated text $\boldsymbol{s}$. We condition the generative process on the representation of $\boldsymbol{s}$ at each time step. Formally:

$$\boldsymbol{z_d} = \text{Encoder}_{\boldsymbol{d}}(\boldsymbol{d}_i, \boldsymbol{\theta_d}) \tag{4}$$

$$\boldsymbol{z_s} = \text{Encoder}_{\boldsymbol{s}}(\boldsymbol{s}_i, \boldsymbol{\theta_s}) \tag{5}$$

$$\hat{\boldsymbol{x}}_{\boldsymbol{i}} \sim \prod_t p(\hat{x}_i^t | [\hat{\boldsymbol{x}}_i^{<t}; \boldsymbol{z_s}], \boldsymbol{z_d}) \tag{6}$$

where, $\boldsymbol{\theta_d}$ and $\boldsymbol{\theta_s}$ are the parameters of the encoder for the document $\boldsymbol{d}$ and encoder for the curated text $\boldsymbol{s}$ respectively, $\boldsymbol{z_d}$ and $\boldsymbol{z_s}$ are the encoded representations of the document $\boldsymbol{d_i}$ and curated text $\boldsymbol{s_i}$ respectively. At each time step of generation, the output is conditioned on the tokens generated up to the time step t concatenated with $\boldsymbol{z_s}$. Hence, the generative process is receptive of the context at each time step.

3.2 Extractive models

Generative models that construct new sentences conditioned on the relevant context are compelling

but have a number of modeling challenges. Such a model must both select the most relevant content *and* generate a fluent linguistic realization of this information.

We also consider extractive models: approaches that select the most relevant sentence from the document $\boldsymbol{d}$ to append to the curated text $\boldsymbol{s}$. These approaches can focus solely on the content selection problem and ignore the difficulties of generation. This simplification does come at a cost: the most effective sentence to add might require only a subset of information from some sentence in the document, or incorporate information from more than one sentence.

Sum-Basic (SB): One common baseline is Sum-Basic, an extractive summarization technique that relies on word frequency statistics to select salient sentences (Nenkova and Vanderwende, 2005). As an initial step, unigram probabilities are computed from the set of input documents using relative frequency estimation. Then, sentences are selected one-by-one in greedy rounds until the summary budget is saturated. At each round, this model selects the most likely sentence according to the current unigram distribution. The selected sentence is added to the summary and removed from the pool of available sentences. The unigram probabilities of all words in the selected sentence are heuristically discounted (replaced by square root). Select-then-discount operations continue until the summary is written. Discounting is crucial to prevent repetition: once a word (or ideally a concept) has been selected for the summary, it is much less likely to be picked in a subsequent round.

We use Sum-Basic as a Context Agnostic extractive model: we provide the document $\boldsymbol{d}$ as an input to the model and run Sum-Basic for exactly one round. The selected sentence is considered to be the update sentence $\boldsymbol{x}$.

Context Informed Sum-Basic (CISB): We developed a simple modification of the Sum-basic technique to incorporate information from the curated text $\boldsymbol{s}$ as context. Initial unigram probabilities are computed using word counts from *both* the curated text *and* the document. Next, for each sentence in the curated text, we apply just the discount procedure, updating the probability distribution as if those sentences were selected. Finally, we select the single sentence from the document that is most likely according to the resulting discounted unigram probabilities. This simple modification of Sum-Basic helps select a sentence that is novel with respect to the curated text by lowering the probability of all words already present.

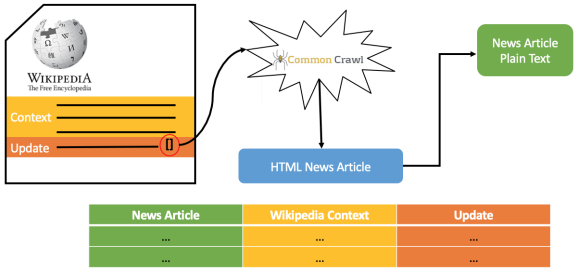

Figure 2: Dataset creation process

Extractive CAG, CIG, CRG Models: Any generative model of $\boldsymbol{x}$ can also be used as an extractive model: we simply estimate the likelihood of each sentence in the document according to the model, and select the most likely one. Generative models may fail because either they are unable to select the most relevant information, or because the resulting sentence is ill-formed. Extractive ranking circumvents all errors due to generation and can help isolate model issues.

Hybrid CAG, CIG, CRG Models: Since the document $\boldsymbol{d}$ can be quite large, a generative model may struggle to pick the most salient information based on the context. To simplify the generative modeling task, we can pre-filter the document toward only the most salient parts. We use the Context Informed Sum-Basic technique to first select the top five sentences from the document. We supply only these five sentences in place of the source document $\boldsymbol{d}$, then apply the CAG, CIG, and CRG techniques described above.

4 Dataset

Our ideal dataset would capture the edits made to some curated reference text in light of a stream of new articles describing changes. For instance, one might maintain reference software documentation about a system, making additions or changes in light of incoming emails describing updates or additions. This type of data is unfortunately difficult to obtain due to privacy considerations.

However, Wikipedia can provide a naturally-occurring body of text with references to primary sources. A substantial fraction of Wikipedia sentences include citations to supporting documentation, a ripe source of data for content transfer. That

Corpus	Input	Output	#Examples	Rouge-1 R
Gigaword (Graff and Cieri, 2003)	10^1	10^1	10^6	78.7
CNN/DailyMail (Nallapati et al., 2016)	10^2–10^3	10^1	10^5	76.1
WikiSum (Liu et al., 2018)	10^2–10^6	10^1–10^3	10^6	59.2
Content Transfer (this paper)	10^1–10^3	10^1–10^2	10^5	66.9

Table 1: Key characteristics of the dataset: approximate size of input and output instances, approximate dataset size, and recall of reference output against the source material, as a measure of dataset difficulty.

said, some of the citations are quite difficult to follow or trust: broken URLs might lead to lost information; citations to books are difficult to consume given the large scope of information; etc. Therefore, we only consider cases where the reference links to some well-known news sources.

Based on citation frequency, we selected a list of 86 domains,[3] primarily news outlets. During the data creation process we only considered citations belonging to one of these eighty six domains. We make this simplifying assumption for several reasons. First, our English Wikipedia dump contained approximately 23.7 million citation URLS belonging to 1.6 million domains; fine-grained filtering would be a daunting task. Our hand-vetted list of domains is a high-precision (albeit low-recall) means of selecting clean data. Second, we wanted to ground the generated text on credible, consistent, and well-written sources of information. Furthermore, well-known domains are readily available on Common Crawl,[4] leading to an easily-reproducible dataset.

Fig. 2 illustrates the procedure used to create a dataset for the task described in Section 2 from Wikipedia. For each Wikipedia article, we extracted the plain text without markdown. When encountering a citation belonging to a selected domain, we considered the sentence just before the citation to be generated based on the content of the citation. This sentence became our reference update sentence: the additional update sentence $\boldsymbol{x}$ added to the curated text $\boldsymbol{s}$ to produce the new text $\boldsymbol{s}'$. The k sentences prior to the target sentence in the Wikipedia article were considered to be the curated text $\boldsymbol{s}$. In our case, we used a window of $k = 3$ sentences to select our context. The cited article acted as the document $\boldsymbol{d}$, from which the appropriate update $\boldsymbol{x}$ can be generated.

The HTML source of the citation was downloaded from Common Crawl for reproducibility and consistency. The HTML derived from Common Crawl is then processed to get the plain text of the news article. The resulting dataset $\boldsymbol{C}$ consists of aligned tuples $\boldsymbol{C} = \left(\boldsymbol{d_i}, \boldsymbol{s_i}, \boldsymbol{x_i}\right)_{i\in[1,n]}$, where n is the total number of samples in the dataset.

Alternatively, one might rely on Wikipedia edit history to create a dataset. In this setting, edits which include a new citation would act as the update $\boldsymbol{x}$. Although this has the upside of identifying potentially complex, multi-sentence updates, preliminary analysis suggested that these edits are noisy. Editors may first generate the content in one edit, then add the citation in a subsequent edit, they may only rephrase a part of the text while adding the citation, or they may check in a range of changes across the document in a single edit. Our simpler sentence-based approach leads to an interesting dataset with fewer complications.

Dataset Statistics and Analysis Table 1 describes some key statistics of our dataset and how it compares with other datasets used for similar tasks. The ROUGE-1 recall scores of reference output $\boldsymbol{x}$ against document $\boldsymbol{d}$ suggest this task will be difficult for conventional extractive summarization techniques.[5] We hypothesize that during content transfer, the language in document $\boldsymbol{d}$ often undergoes substantial transformations to fit the curated text $\boldsymbol{s}$. The average unigram overlap (after stopword removal) between the document $\boldsymbol{d}$ and the reference update sentence $\boldsymbol{x}$ is 55.79%; overlap of the curated text $\boldsymbol{s}$ and the reference update sentence $\boldsymbol{x}$ is 30.12%. This suggests the reference update sentence $\boldsymbol{x}$ can be derived from the document $\boldsymbol{d}$, though not extracted directly. Furthermore, the content of $\boldsymbol{x}$ is very different from the content of $\boldsymbol{s}$ but appears topically related.

Our dataset consists of approximately 290k unique Wikipedia articles. Some heavily-cited

[3] This list is provided in the data release of this paper.

[4] http://commoncrawl.org/

[5] ROUGE-1 recall was computed on a sample of 50k instances from the entire dataset.

articles include 'Timeline of investigations into Trump and Russia (2017)', 'List of England Test cricketers', and '2013 in science'. We randomly split the dataset into 580k training instances, 6049 validation instances, and 50k test instances, ensuring that any Wikipedia article appearing in the train set must not appear in validation or test.

5 Experimental results

We evaluate our models using both automated metrics and, for a subset of promising systems, human assessment. One key evaluation is the similarity between the model generated update sentence and reference update sentence. We also ask human judges to assess grammaticality and coherence.

Hyper-parameter settings: For all our experiments with generative models, we have used bidirectional encoder, 2 layers in encoder and decoder, RNN size of 128, word vector size of 100. We have used sentencepiece toolkit[6] to use byte-pair-encoding (BPE) with a vocabulary size of 32k. We used stochastic gradient descent optimizer and the stopping criterion was perplexity on the validation set. We filtered our dataset to contain instances which have length of the document between 50 and 2000 tokens, length of the curated text between 20 and 500 tokens and the length of the update sentence between 5 and 200 tokens.

5.1 Automated Evaluation

Our primary automated evaluation metric for system-generated update sentences is ROUGE-L F1 against reference update sentence,[7] though we also include BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011) as additional indicators. ROUGE is a standard family of metrics for summarization tasks; ROUGE-L measures the longest common subsequence between the system and the reference, capturing both lexical selection and word order.

Table 2 illustrates that this task is quite difficult for extractive techniques. Furthermore, the results emphasize the importance of having curated text as context when generating the update. In all experimental conditions, models aware of context perform much better than models agnostic of it. In contrast to Liu et al. (2018), generative approaches outperformed hybrid, likely because we only had a single input document. Extractive CAG, CIG, and CRG all outperformed both Sum-Basic and the context informed variant. Extractive CAG was on-par with generative CAG, suggesting the generated sentences were of reasonable quality. However, generative CIG and CRG were substantially better: rewriting to match context was beneficial.

Model	ROUGE-L	BLEU	METEOR
SB	5.6 (5.6–5.7)	0.6	2.0
CISB	7.0 (7.0–7.1)	1.0	2.8
CAG	9.1 (9.0–9.2)	1.2	4.6
COG	13.5 (13.4–13.6)	1.7	3.5
CIG	**16.0** (15.9-16.1)	**3.5**	**5.3**
CRG	14.7 (14.6–14.8)	2.6	4.5
Hybrid CAG	8.0 (7.9–8.0)	1.0	3.8
Hybrid CIG	15.0 (14.9–15.1)	2.7	4.7
Hybrid CRG	13.5 (13.4–13.6)	2.3	4.1
Extractive CAG	9.3 (9.2–9.3)	1.1	3.2
Extractive CIG	9.3 (9.2–9.3)	1.1	3.2
Extractive CRG	9.2 (9.1–9.3)	1.1	3.2
Oracle	*28.8 (28.7–29.0)*	*11.0*	*10.9*

Table 2: Automated metrics; 95% confidence interval in parentheses.

The *Oracle* system of Table 2 aims to establish an upper limit attainable by extractive methods, using the following oracle experiment: For each test instance $(\boldsymbol{d_i}, \boldsymbol{s_i}, \boldsymbol{x_i})$, we enumerate each extracted sentence $\boldsymbol{e}$ of document $\boldsymbol{d_i}$ and select the one with highest ROUGE-L score as *Oracle*'s update sentence $\hat{\boldsymbol{x}}_{\boldsymbol{i}}$ (i.e., $\hat{\boldsymbol{x}}_{\boldsymbol{i}} = \arg\max_{\boldsymbol{e} \in \boldsymbol{d_i}} \text{ROUGE-L}(\boldsymbol{x_i}, \boldsymbol{e})$). Note this yields a very optimistic upper bound, as the same ground truth $\boldsymbol{x_i}$ is used both to select an extractive sentence from a large pool of candidates and for final automatic metric scoring.[8] Nevertheless, these oracle results let us draw two conclusions: (1) They give us better perspective to assess the non-oracle systems, and we believe that their seemingly low

[6] https://github.com/google/sentencepiece

[7] We use the pyrouge toolkit along with ROUGE-1.5.5: https://github.com/bheinzerling/pyrouge

[8] Previous work has shown that this type of oracle can yield upper bounds that are unrealistically high, and they tend to be above human performance (Och et al., 2004, Table 1). One remedy suggested by Och et al. is a round-robin oracle ensuring that the reference (ground truth) used by the argmax is distinct from that of the final automatic evaluation, but that scheme is only possible with a multi-reference test set.

automatic evaluation scores are quite reasonable relative to the optimistic upper bound (e.g., CIGs ROUGE-Ls score is 55% of the oracle). (2) The oracle results suggest that humans are substantially changing the surface realization as they summarize for Wikipedia, as otherwise the oracle results would be much closer to maximum metric scores (i.e., 100%). This shows that extractive methods are not enough for this task, justifying our use of generation techniques.

5.2 Human Evaluations

For careful evaluation of the performance of the most promising configurations (CAG and CIG models) we also asked human judges for quality assessments. We solicited several types of evaluation, including two relative comparisons between pairs of system outputs and an absolute quality evaluation of individual system outputs.

Close to reference (Relative): The first relative comparison measured how accurately the generated update reflected information in the reference update. Here, the annotators saw only the reference update sentence and the outputs of two systems labeled A and B in a randomized order. We asked the annotators "Which system output is closest in meaning to the reference update?" The annotators could pick system A, system B, or indicate that neither was preferred. This is a simple evaluation task though potentially biased toward the sole reference update.

Coherent to context (Relative): The second relative comparison measured whether the generated output contained salient information from the document written in a manner appropriate to the curated text. The annotators saw the document $\boldsymbol{d}$, the curated text $\boldsymbol{s}$, and the outputs of the two systems A and B, again in a random order. They were asked, "Which system output is more accurate relative to the background information given in the snippet of the article?" Each judge had to consider whether the information fits with the curated text and also whether system-generated content could be supported by the document.

Four human judges each annotated 30 unique output pairs for these two relative comparison settings, a total of 240 relative judgments. Table 3 shows the results: the context-aware CIG system was substantially better in both settings.

	prefer		
Evaluation task	CAG	neither	CIG
Close to reference	15.8%	53.3%	30.8%
Coherent to context	7.5%	53.3%	39.2%

Table 3: Human preferences of CAG vs. CIG.

Model	Grammar	Non-redund.	Ref. clarity	Focus	Structure
CAG	2.6	1.8	2.7	2.6	2.4
CIG	**4.3**	**3.9**	**3.6**	**3.5**	**3.2**

Table 4: Human absolute quality assessments.

DUC Guidelines (Absolute): In addition, we performed an absolute quality evaluation following the guidelines from DUC 2007.[9] Each judge was presented with a single system output, then they were asked to evaluate five aspects of system output: grammaticality, non-redundancy, referential clarity, focus, and structure/coherence. For each aspect, the judge provided an assessment on a five-point scale: (1) Very Poor, (2) Poor, (3) Barely Acceptable, (4) Good, (5) Very Good. We gathered 120 additional judgments in this setting (4 judges, 30 outputs). Again, context-aware CIG substantially outperforms CAG across the board, as seen in Table 4.

Observations: Systems unaware of the curated text $\boldsymbol{s}$ tend to generate long updates with repeated frequent words or phrases. Consider the ratio of unique tokens over the total number of tokens in the generated output, which we denote by $\boldsymbol{R}$. A small $\boldsymbol{R}$ indicates many repeated tokens. We find that 88% of the time this ratio $\boldsymbol{R}$ falls below 0.5 for the CAG model, i.e. for 88% instances, more than 50% of the words in the generated output are repeats. This number is relatively small – 14% for CIG and 20% for CRG – in context aware models. In the reference updates only 0.21% instances repeat more than 50% of words.

Figs. 3 and 4 show good and bad examples generated by the CIG model along with the document, curated text and the reference update. Table 5 has a set of updates generated by the CIG model as

[9] http://duc.nist.gov/duc2007/quality-questions.txt

Document (News Article)

sequels are fairly new to bollywood, but director sanjay gadhvi realised there was cash to be made from resurrecting his hit action thriller dhoom, by casting sexy young stars like hrithik rosha, aishwarya rai and abhishek bachchan in an even bigger game of cops and robbes...that the twist in dhoom 2's tail is not explained is yet another shortcoming. it's only roshan's charismatic performance as the criminal mastermind, and the sizzling chemistry he shares with rai's sassy cohort, that rescues this adventure from becoming an elongated tourism commercial.

Curated Text (Wikipedia Context)

it makes no lasting contributions to world cinema, but if two-and-a-half hours of disposable entertainment are all you're after, you could do far worse. "l.a. weekly's david chute stated the film was, "a movie meal as satisfying as this one can make you feel that nothing else matters." jaspreet pandohar of the bbc gave it a two-star rating, writing "by roping in acclaimed action director alan amin to take care of the thrills and spills, you'd expect gadhvi to have spent time crafting out a sophisticated storyline instead of simply sending his cast on a cat-and-mouse chase around the globe.

Reference Update

it's only roshan's charismatic performance as the criminal mastermind, and the sizzling chemistry he shares with rai's sassy cohort, that rescues this adventure from becoming an elongated tourism commercial."

Generated Update

it's only roshan's finest performance as the criminal terrorist, and the sizzling chemistry he shares with rai's sassy anatomy, that attues this adventure from becoming an elongated tourism commercial."

Figure 3: Example of good quality generation, where the system-generated update is close to the reference.

well as the reference update. As we can see in examples 3 and 4, the CIG model misplaces the date but correctly generates the remaining content. In examples 1 and 2, the CIG model appears to successfully select the correct pronouns for coreference resolution, though it gets confused as to when to use the pronoun or the named entity. Examples 5 and 6 represent failure cases due to missing words.

6 Related Work

The proposed content transfer task is clearly related to a long series of papers in summarization, including recent work with neural techniques (Rush et al., 2015; Nallapati et al., 2016). In particular, one recent paper casts the the task of generating an entire Wikipedia article as a multi-document summarization problem (Liu et al., 2018). Their best-performing configuration was a two-stage extractive-abstractive framework; a multi-stage approach helped circumvent the difficulties of purely abstractive methods given quite large input token sequences.

Document (News Article)

anne kirkbride, who portrayed bespectacled, gravelly-voiced deirdre barlow in coronation street for more that four decades, has died. the 60-year-old, whose first appearance in the soap opera was in 1972, died in a manchester hospital after a short illness.... kirkbride had left the soap opera after she was diagnosed with non-hodgkin's lymphoma in 1993 but returned some months later after treatment and spoke candidly about how she had struggled with depression following the diagnosis...

Curated Text (Wikipedia Context)

in 1993, kirkbride was diagnosis with non-hodgkin's lymphoma. she spoke to the british press about her bout of depression following the diagnosis. she was cured within a year of being diagnosed.

Reference Update

anne kirkbride died of breast cancer in a manchester hospital on 19 january 2015, aged 60.

Generated Update

she was diagnosed with non-hodgkin's lymphoma.

Figure 4: Example of lower-quality output: the generated update unnecessarily restates information yet misses the most salient detail from the document.

Looking beyond the clear task similarity of authoring Wikipedia style content, there are several crucial differences in our approach. First, the goal of that paper is to author the whole page, starting from nothing more than a set of primary sources, such as news articles. In practice, however, Wikipedia articles often contain information outside these primary sources, including common sense knowledge, framing statements to set the article in context, and inferences made from those primary sources. Our task restricts the focus to content where a human editor explicitly decided to cite some external source. Hence, it is much more likely that the resulting summary can be derived from the external source content. Furthermore, we focus on the act of adding information to existing articles, rather than writing a complete article without any context. These two scenarios are clearly useful yet complementary: sometimes people want to produce a new reference text where nothing existed before; in other cases the goal is to maintain and update an existing reference.

Another closely related task is update summarization (Dang and Owczarzak, 2008), where systems attempt to provide a brief summary of the novel information in a new article assuming the user has read a known set of prior documents. Our focus on curating an authoritative resource

	Reference Update	Generated Update
1.	rob brydon, the comedian was born in baglan.	he was born in baglan.
2.	in may 2014 he was diagnosed with prostate cancer.	st. clair was diagnosed with prostate cancer.
3.	on april 3, 2014, manning signed a one-year deal with the cincinnati bengals.	on march 9, 2014, manning signed a one-year contract with the cincinnati bengals.
4.	on oct 10, 2013, barrett signed with the memphis grizzlies.	on feb 9, 2013, barrett signed with the memphis grizzlies.
5.	some people think elvis is still alive, but most of us think he's dead and gone."	some people think elvis, but most of us think he's dead and gone."
6.	it's always the goal of the foreign-language film award executive committee to be as inclusive as possible."	it's always the goal of the foreign- entry film award executive to be as possible."

Table 5: Example generations from the CIG system, paired with the human generated updates.

is a substantial difference. Also our datasets are substantially larger, enabling generative models to be used in this space, where prior update summarization techniques have been primarily extractive (Fisher and Roark, 2008; Li et al., 2015).

For any generation task, it is important to address both the content ('what' is being said) as well its style ('how' it is being said). Recently, a great deal of research has focused on the 'how' (Li et al., 2018; Shen et al., 2017), including efforts to collect a parallel dataset that differs in formality (Rao and Tetreault, 2018), to control author characteristics in the generated sentences (Prabhumoye et al., 2018), to control the perceived personality traits of dialog responses (Zhang et al., 2018). We believe this research thread is complementary to our efforts on generating the 'what'.

Another form of content transfer bridges across modalities: text generation given schematized or semi-structured information. Recent research has addressed neural natural language generation techniques given a range of structured sources: selecting relevant database records and generating natural language descriptions of them (Mei et al., 2016), selecting and describing slot-value pairs for task-specific dialog response generation (Wen et al., 2015), and even generating Wikipedia biography abstracts given Infobox information (Lebret et al., 2016). Our task, while grounded in external content, is different in that it leverages *linguistic* grounding as well as prior text context when generating text. This challenging setting enables a huge range of grounded generation tasks: there are vast amounts of unstructured textual data.

7 Conclusions

This article highlights the importance of the task of *content transfer*: generation guided by an existing curated text to set context and tone, and grounded in a new source providing useful information. We demonstrate how multiple models can address this challenging problem on a novel dataset derived from Wikipedia and Common Crawl. This dataset is released to the community along with scripts and models.[10] We find this setting particularly promising given the opportunity for human interaction: in contrast to approaches that do not rely on human-generated context, we establish a collaboration between user and computer. Each newly suggested sentence can be rejected, accepted, or edited before inclusion, and the edits can provide more training data.

We believe there are many natural extensions to this work. The models described here are mostly extensions of existing approaches; approaches targeting novelty detection, focus, and document structure could lead to substantial improvements. We could apply models in series to incorporate changes for a set of documents. Future work could also explore changes that modify existing content rather than simply appending.

Acknowledgments

We are grateful to the anonymous reviewers, as well as Alan W Black, Chris Brockett, Bill Dolan, Sujay Jauhar, Michael Gamon, Jianfeng Gao, Dheeraj Rajagopal, and Xuchao Zhang for their helpful comments and suggestions on this work. We also thank Emily Ahn, Khyati Chandu, Ankush Das, Priyank Lathwal, and Dheeraj Rajagopal for their help with the human evaluation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

[10] https://www.microsoft.com/en-us/research/project/content-transfer/

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 update summarization task. In *In TAC 2008 Workshop - Notebook papers and results*, pages 10–23.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proc. of EMNLP*, page 94.

Seeger Fisher and Brian Roark. 2008. Query-focused supervised sentence ranking for update summaries. In *TAC*.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *ACL*, volume 1, pages 634–642.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680.

David Graff and Christopher Cieri. 2003. English Gigaword LDC2003T05. In *Philadelphia: Linguistic Data Consortium.*

Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. 2017. A nested attention neural hybrid model for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–762, Vancouver, Canada.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213. Association for Computational Linguistics.

Chen Li, Yang Liu, and Lin Zhao. 2015. Improving update summarization via supervised ILP and sentence reranking. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1317–1322.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1865–1874.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.

Hongyuan Mei, TTI UChicago, Mohit Bansal, and Matthew R Walter. 2016. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of NAACL-HLT*, pages 720–730.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural*

Language Learning, pages 280–290. Association for Computational Linguistics.

Ani Nenkova and Kathleen R. McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. Technical report, Microsoft Research.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proc. of HLT-NAACL*, pages 161–168.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 129–140.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6830–6841.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.