Toward Unsupervised Text Content Manipulation

Abstract

Controlled generation of text is of high practical use. Recent efforts have made impressive progress in generating or editing sentences with given textual attributes (e.g., sentiment). This work studies a new practical setting of text content manipulation. Given a structured record, such as `(PLAYER: Lebron, POINTS: 20, ASSISTS: 10)', and a reference sentence, such as `Kobe easily dropped 30 points', we aim to generate a sentence that accurately describes the full content in the record, with the same writing style (e.g., wording, transitions) of the reference. The problem is unsupervised due to lack of parallel data in practice, and is challenging to minimally yet effectively manipulate the text (by rewriting/adding/deleting text portions) to ensure fidelity to the structured content. We derive a dataset from a basketball game report corpus as our testbed, and develop a neural method with unsupervised competing objectives and explicit content coverage constraints. Automatic and human evaluations show superiority of our approach over competitive methods including a strong rule-based baseline and prior approaches designed for style transfer.

Introduction

Generating natural language text to describe structured content, such as a database record or a table, is of ubiquitous use in real-life applications including data report generation BIBREF0 , article writing BIBREF1 , BIBREF2 , dialog systems BIBREF3 , BIBREF4 , and many others. Recent efforts have developed many techniques to improve fidelity to the source content, such as new powerful neural architectures BIBREF5 , BIBREF6 , hybrid generation and retrieval BIBREF7 , BIBREF8 , and so forth, most of which are applied in supervised context.

Language is rich with variation–given a data record, there are diverse possible ways of saying the same content, with different word choices, expressions, transitions, tone, etc. Previous data-to-text work has largely focused only on content fidelity, while ignoring and lacking control over the rich stylistic properties of language. It can be practically useful to generate text that is not only describing the conditioning content, but also following a designated writing style, e.g., as provided in a piece of reference text.

In this work, we study the new yet practical problem in which we aim to express given content with a sentence and mimic the writing style of a reference sentence (Table TABREF1). More specifically, we are given a structured data record containing the content to describe, along with a sentence about a similar but different matter. Our goal is to generate a new sentence that precisely depicts all content in the record, while at the same time using as much of the writing style of reference sentence as possible. As above, the problem differs critically from the supervised data-to-text BIBREF0 or retrieval-and-rewriting work BIBREF7 , BIBREF8 as we have imposed an additional goal of preserving the reference text style. The resulting problem is typically unsupervised due to lack of parallel data.

The problem also differs in important ways from the emerging task of text style transfer BIBREF9 , BIBREF10 which assumes an existing sentence of certain content, and modifies single or multiple textual attributes of the sentence (e.g., transferring negative sentiment to positive) without changing the content. Our task, on the contrary, assumes abstract style is encoded in a reference sentence and attempts to modify its concrete content to express new information from the structured record. The different setting can lead to different application scenarios in practice, and pose unique technical challenges. In particular, though the most recent style transfer research BIBREF11 , BIBREF12 has controlled multiple categorical attributes which are largely independent or loosely correlated to each other, a content record in our task, in comparison, can contain varying number of entries which are of different types (e.g., player, points, defensive/offensive rebounds, etc), having many possible values (e.g., hundreds of players), and are structurally coupled (e.g., 32 points by Lebron). A model must understand the content structure, and minimally yet sufficiently manipulate the reference sentence by rewriting, adding, or deleting text portions, with necessary polishing for grammatical correctness and fluency. We name the problem text content manipulation. Our empirical studies show the most recent models designed for style transfer fail to perform well in the task.

In this paper, we first develop a large unsupervised dataset as a testbed of the new task. The dataset is derived from an NBA game report corpus BIBREF0 . In each data instance, besides a content record and a reference sentence as the problem inputs, we also collect side information useful for unsupervised learning. Specifically, each instance has an auxiliary sentence that was originally written by human reporters to describe the content record without seeing (and thus stylistically irrelevant to) the reference sentence. We also provide the structured record of the reference sentence. The side information can provide valuable clues for models to understand the content structure and text semantics at training time. We do not rely on the side information at test time.

We then propose a neural method to tackle the problem. With a hybrid attention and copy mechanism, the model effectively encodes the reference and faithfully copies content from the record. The model is learned with two competing objectives of reconstructing the auxiliary sentence (for content fidelity) and the reference sentence (for style preservation). We further improve the model with an explicit content coverage constraint which encourages to precisely and fully convey the structured content.

For empirical study, we devise automatic metrics to measure content fidelity and style preservation, respectively. We also perform human evaluations to compare different approaches. Results demonstrate the proposed method significantly improves over others, including a strong rule-based baseline and the recent style transfer models.

Related Work

Generating text conditioning on structured input has been widely studied in recent work, such as BIBREF3 , BIBREF1 , BIBREF4 , BIBREF0 . Those methods are based on neural sequence to sequence models and trained with supervised data. This line of work has focused primarily on generating more accurate description of the given data, while does not study the problem of controlling the writing style of outputs. Our task takes a step forward to simultaneously describing desired content and controlling stylistic properties. Furthermore, our task is challenging due to its unsupervised setting in practice.

Beyond generating text from scratch, there is another line of work that first retrieves a similar sentence and then rewrites it to express desired information BIBREF8 , BIBREF7 , BIBREF13 , BIBREF14 . For example, BIBREF8 used the framework to generate response in dialogues, while BIBREF7 studied programming code generation. The goal of the work is to manifest useful information from neighbors, usually in a supervised context, without aiming at controlling writing characteristics, and thus has fundamentally different assumptions to ours.

Recently, there has been growing interest in text style transfer, in which many techniques for controlled text generation are developed BIBREF9 , BIBREF10 , BIBREF15 , BIBREF16 , BIBREF17 , BIBREF11 , BIBREF12 . The main idea underlying those models is to learn disentangled representations of text so as modify textual attributes or style of interest. Those papers used different objectives to encourage learning disentangled representations. BIBREF9 used pre-trained classifiers as the supervision. BIBREF10 used a GAN-based approach in which binary classifiers were used as discriminators. BIBREF15 proposed to use more structured discriminators such as language models to provide better supervision to the generator. BIBREF16 , BIBREF11 further augmented prior work using back-translation technique to incorporate cycle-consistency loss. Both BIBREF11 and BIBREF12 generalized the task to controlling multiple categorical attributes at the same time. Our work differs from those in that we assume an existing sentence to provide the source of style and a structured record as the source of content. The input content record in our task is also more structured than the style attributes which are typically loosely connected and of a pre-fixed number. The resulting content manipulation setting poses unique challenges in controlling, as discussed more in the empirical study.

Task and Dataset

We first formally define the problem of unsupervised text content manipulation, and establish the notations. We then present a large dataset for the task.

Task Definition

Without loss of generality, consider a content record INLINEFORM0 , where each element INLINEFORM1 is a data tuple which typically includes a data type (e.g., points), a value (e.g., 32), and other information (such as the associated player, e.g., Lebron_James). INLINEFORM2 is the number of tuples in record INLINEFORM3 , which can vary across different records. We are also given a reference sentence INLINEFORM4 which is assumed to describe content that has a similar but not exact the same structure with that of the record INLINEFORM5 . For example, in Table TABREF1 , both the content record and the reference sentence involve two players, respectively, but the number of associated data tuples as well as the types are different (e.g., Lebron_James in the record has 3 box-score entries, while Jrue_Holiday in the reference has only 2).

We may also have access to other side information at training time. For example, in the dataset developed below, each content record INLINEFORM0 is associated with an auxiliary sentence INLINEFORM1 that was originally written to describe INLINEFORM2 without following the reference INLINEFORM3 . Each reference sentence INLINEFORM4 also has its corresponding record INLINEFORM5 containing the content information. The side information can provide valuable clues for models to understand the content structure and text semantics at training time. For example, the auxiliary sentence provides a hint on how the desired content can be presented in natural language, though it is stylistically irrelevant to the reference sentence. Note that, at test time, a solution to the task should only rely on the inputs INLINEFORM6 without using the side information.

The goal of the task is to generate a new realistic sentence INLINEFORM0 that achieves (1) content fidelity by accurately describing the full content in INLINEFORM1 , and at the same time (2) style preservation by retaining as much of the writing style and characteristics of reference INLINEFORM2 as possible. The task is unsupervised as there is no ground-truth sentence for training.

Dataset

We now present a dataset developed for the task. Our dataset is derived from a recent large table-to-document corpus BIBREF0 which consists of box-score tables of NBA basketball games and associated documents as game reports. The corpus is originally used for studying supervised game report generation which has attracted increasing research interest BIBREF18 , BIBREF0 .

To obtain our data, we first split each game report into individual sentences, and, for each sentence, find its corresponding data in the box-score table as the content record. A record can contain a varying number of tuples, with each tuple containing three fields, namely a data type, a value, and an associated player or team, e.g., (team_points, 106, Lakers). As the original corpus is already largely clean, we found some simple rules are sufficient to obtain high-quality results in this step. Please see the supplementary materials for more details. Each of the resulting record-sentence pairs is treated as a pair of INLINEFORM0 , namely (content record, auxiliary sentence). The next step is to find a suitable reference sentence INLINEFORM1 for each content record INLINEFORM2 . As defined above, the reference sentence should cover similar but not the same content as in record INLINEFORM3 . We achieve this by retrieving from the data another record-sentence pair using INLINEFORM4 , where the retrieved record is designated to have a slightly different structure than that of INLINEFORM5 by having less or more tuples and different data types. More details of the retrieval method are deferred to supplements. The retrieved record-sentence pair thus plays the role of INLINEFORM6 and is paired with INLINEFORM7 to form an instance.

Table TABREF6 summarizes the statistics of the final dataset. The vocabulary size is 8.4K. We can see that the training set contains over 31K instances. Each content record contains around 5 tuples, each of which takes one of the 34 data types.

Model

We next develop methods to tackle the problem. As shown in the empirical study (section SECREF5), a simple rule-based method that matches INLINEFORM0 with INLINEFORM1 and performs text replacement would fail in terms of content fidelity due to the different structures between INLINEFORM2 and INLINEFORM3 . Previous approaches for (multi-attribute) style transfer do not apply well either, because of the different underlying task assumptions and the rich content structures of records with varying lengths.

In the following, we present a new neural approach that addresses the challenges of text content manipulation. We first describe the model architecture, then develop unsupervised learning objectives, and finally add a content coverage constraint to improve learning. Figure FIGREF7 provides an illustration of the proposed approach.

Let INLINEFORM0 denote the model that takes in a record INLINEFORM1 and a reference sentence INLINEFORM2 , and generates an output sentence INLINEFORM3 . Here INLINEFORM4 is the model parameter.

Experiments

We conduct both automatic and human evaluations to assess the model performance. For automatic evaluation, we use two metrics to measure content fidelity and style preservation, respectively. Results show our model balances well between the two goals, and outperforms a variety of comparison methods. All code will be released soon.

Experimental Setup

We compare with a diverse set of approaches:

[leftmargin=*]

AttnCopy-S2S. We first evaluate a base sequence-to-sequence BIBREF22 model with the above attention-copy mechanism, which takes in record INLINEFORM0 and generates its descriptive sentence INLINEFORM1 . The evaluation provides a sense of the difficulty in describing desired content.

Rule-based Method. A straightforward way for text content manipulation is to match between INLINEFORM0 , INLINEFORM1 and INLINEFORM2 with certain rules, and replace corresponding portions in INLINEFORM3 with those in INLINEFORM4 . Specifically, we first build a mapping between the tuples of INLINEFORM5 and INLINEFORM6 through their data types, and a mapping between INLINEFORM7 and INLINEFORM8 through data values, types and indicative tokens (e.g., “12 points” in INLINEFORM9 indicates 12 is of type player points or team_points). The two mappings connect INLINEFORM10 and INLINEFORM11 , enabling us to swap appropriate text in INLINEFORM12 to express content INLINEFORM13 .

In theory, rule-based method sets the best possible style preservation performance, as it only replaces content related tokens (particularly numbers) without modifying other parts of the reference sentence. The output, however, tends to miss or contain extra content compared to the content record of interest.

Multi-Attribute Style Transfer (MAST) BIBREF11 . We compare with the most recent style transfer approach that models multiple attributes. To apply to our setting, we treat content record INLINEFORM0 as the attributes. The method is based on back-translation BIBREF23 that first generates a target sentence INLINEFORM1 conditioning on INLINEFORM2 , and then treat it as the reference to reconstruct INLINEFORM3 conditioning on INLINEFORM4 . Auxiliary sentence INLINEFORM5 is used in an extra auto-encoding loss.

Adversarial Style Transfer (AdvST) BIBREF12 . As another latest style transfer approach capable of handling more than one attributes, the model also mixes back-translation with auto-encoding as the above method, and additionally uses adversarial training to disentangle content and style representations.

Ours w/o Coverage. For ablation study, we compare with a model variant that omits the content coverage constraint. That is, the model is trained by maximizing only Eq.(EQREF13).

We use single-layer LSTM RNNs in all encoders and decoders, and use the Luong attention BIBREF19 . Both the embedding dimensions and hidden dimensions are set to 384. During training, we first set INLINEFORM0 and pre-train the model to convergence so that the model captures the full characteristics of the reference sentence. We then set INLINEFORM1 for full training. We apply Adam optimization BIBREF24 with an initial learning rate of 0.001 and gradient norm clipping of 15. For inference we use beam search with beam-width 5. The maximum decoding length is set to 50.

Automatic Evaluation

As no ground truth annotations are available, we first set up automatic metrics for quantitatively measuring the key aspects of model performance.

We use separate metrics to evaluate in terms of the two primary goals of the task, namely content fidelity and style preservation, respectively. A desired solution should balance and excel on both metrics.

[leftmargin=*]

Content fidelity. Following the table-to-document task BIBREF0 where our dataset is derived from, we use an information extraction (IE) approach to measure content fidelity. That is, given a generated sentence INLINEFORM0 and the conditioning content record INLINEFORM1 , we extract data tuples from INLINEFORM2 with an IE tool, and compute the precision and recall against INLINEFORM3 . We use the IE model provided in BIBREF0 and re-train with INLINEFORM4 pairs in our dataset. The IE model achieves around 87% precision and 76% recall on the test set, which is comparable to the one used in BIBREF0 .

Style preservation. A generated sentence is desired to retain stylistic properties, such as word choice and expressions, of the input reference sentence. Inspired by the text style transfer literature BIBREF15 , BIBREF11 , we measure the BLEU score between generated and reference sentences. To reduce the influence of new content, we first mask in both sentences all obvious content tokens, including player/team names and numbers, by replacing them with a special token <M>, and then compute the BLEU score. In this way, the above rule-based method has a maximum BLEU score of 100, which is consistent with our intuition above.

We now compare the performance of different methods in terms of the above metrics. Table TABREF29 shows the results.

The first block shows the two baseline models providing reference performance. The AttnCopy-S2S model only concerns about content fidelity, and achieves a high content precision score (but a low recall). However, its style BLEU is particularly low, which verifies the rich variation in language and that direct supervised learning is incapable of controlling the variation. We can see that the rule-based method achieves reasonably good precision and recall, setting a strong baseline for content fidelity. As discussed above, the rule-based method can reach the maximum BLEU (100) after masking out content tokens. To improve over the strong rule-based baseline, we would expect a method that provides significantly higher precision/recall, while keeping a high BLEU score. The two style transfer methods (MAST and AdvST) fail the expectation, as their content fidelity performance is greatly inferior or merely comparable to the rule-based method. This is partially because these models are built on a different task assumption (i.e., modifying independent textual attributes) and cannot manipulate content well. In comparison, our proposed model achieves better content precision/recall, substantially improving over other methods (e.g., with a 15-point precision boost in comparison with the rule-based baseline) except for AttnCopy-S2S which has failed in style control. Our method also manages to preserve a high BLEU score of over 80. The superior performance of the full model compared to the variant Ours-w/o-Coverage demonstrates the usefulness of the content coverage constraint (Eq. EQREF15). By explicitly encouraging the model to mention each of the data tuples exactly once—a common pattern of human-written descriptions—the model achieves higher content fidelity with less style-preservation ability "sacrificed".

Human Evaluation

We also carried out human evaluation for a more thorough and accurate comparison. Following the experimental settings in prior work BIBREF11 , BIBREF12 , BIBREF10 , we undertook two types of human studies: (1) We asked human turkers to score generated sentences in three aspects, namely content fidelity, style preservation, and sentence fluency. Each score is from 1 (strongly bad) to 5 (strongly good); (2) We present to annotators a pair of generated sentences, one from our model and the other from a comparison method. We then ask the annotators to rank the two sentences by considering all the criteria. Annotators can also choose “no preference” if the sentences are equally good or bad. For each study, we evaluate on 80 test instances, and compare our model with the rule-based method, AdvST style transfer model (which has shown better performance on the task than the other style transfer model MAST), and the model variant without coverage constraint.

Table TABREF31 shows the human evaluation results. From the top block of the table, as expected and discussed above, the rule-based method sets the records of style preservation and fluency scores, as it only conducts lightweight token replacement on reference sentences. However, its content fidelity score is very low. In contrast, our model achieves a reasonably high content score of 3.88, which is much higher than those of other methods. The model is also more balanced across the three criteria, achieving reasonably high scores in both style preservation and language fluency. The fluency of the full model is slightly inferior to the variant without coverage constraint, which is not unexpected since the full model has modified more portions of reference sentence in order to better describe the desired content, which would tend to introduce more language mistakes as well.

The bottom block of Table TABREF31 shows the results of ranking sentence pairs. We can see that our model consistently outperforms the comparison methods with over 50% wins.

Qualitative Study

We take a closer look at the model performance by studying generated sentences from different models.

Table TABREF33 shows example outputs on three test cases given content record INLINEFORM0 and reference sentence INLINEFORM1 . We can see that, in general, the proposed full model can manipulate the reference sentence more accurately to express the new content. For example, in the first case, the rule-based method was confused between the winning and losing teams, due to its incapacity of understanding the semantics of text such as “held off”. The style transfer model AdvST failed to comprehend the content record well and generated irrelevant data “100 - 100”. The simplified variant without explicit coverage constraint copied the content of Bulls twice. In contrast, the full model successfully generates the desired sentence. Similarly, in the second and third cases, other methods tend to keep irrelevant content originally in the reference sentence (e.g., “and 5 rebounds” in the second case), or miss necessary information in the record (e.g., one of the player names was missed in the third case). The proposed model performs better in properly adding or deleting text portions for accurate content descriptions, though sometimes it can yield sentences of lower language quality (e.g., in the third case).

Table TABREF34 shows some failure cases by the proposed model along with the respective desired outputs. Despite the enhanced performance over other methods, the model can still get confused in presence of complicated content records or non-straightforward correspondence between the semantic structures of content record and reference sentence. It is desirable to further improve the modeling of both content and reference to better understand the underlying semantics and achieve better manipulation results.

Conclusions

We have proposed a new and practical task of text content manipulation which aims to generate a sentence that describes desired content from a structured record (content fidelity) and meanwhile follows the writing style of a reference sentence (style preservation). To study the unsupervised problem, we derived a new dataset, and developed a method with competing learning objectives and an explicit coverage constraint. For empirical study, we devised two automatic metrics to measure different aspects of model performance. Both automatic and human evaluations showed superiority of the proposed approach.