What Would Nabokov Say?

Daniel Kharitonov
Department of Computer Science
Stanford University
Stanford, CA 94305
dkh@cs.stanford.edu

Abstract

In this paper we consider the problem of unsupervised style transfer in the natural language. We develop a novel RPS framework for constructing the parallel representation of input content using pretrained sequence-to-sequence generators, and propose a three-dimensional metric for testing the quality of style transfer.

1 Introduction to style transfer

Problem of general style transfer have been extensively researched in the last ten years and led to impressive results in applying stylistic cues to still and moving images where content (contours and shapes) is easily separable from style (colors, strokes and spatial transformations). Natural language does not exhibit such separability, and therefore progress in text style transfer was hitherto limited to domains where the parallel text representations already existed.

One example of style transfer with available parallel representation is translation to foreign languages [1], where extensive corpus of references was made possible by work of human interpreters. Another example is artificial languages (like program source code), where different inputs can be considered equivalent if they lead to identical (or comparable) execution results on a target machine. A third example is the direct-supervision environment, where were the styled output can be evaluated by human experts.

Unsupervised style transfer in natural language without a parallel representation, however, remains an unsolved problem. While neural nets were shown to be effective in the "no-task specific" training on large amounts of human-produced text and proven to be capable of generating credible-looking “news snippets” or “fake tweets” conditioned on seed phrases [2], style and content of such output remained largely random, and deviated from the seed prompt in stochastic directions. This raises a research question on whether directed-content style transfer is possible with no parallel representation, or whether such representation can also be constructed by the machine in unsupervised way. The many practical applications for such algorithms include (but are not limited to) automated correction of writing styles, adaptation of texts to specific audiences, and production of accessible content. In this paper we intend to provide some answers to these research questions.

1.1 Style

Further in this paper we will be focusing on the natural language examples, so we need to define "style" more formally. For the purposes of this study we will consider "style" to be the same as "writing style", with a colloquial definition of being "the essential elements of spelling, grammar, and punctuation, including the choice of words, sentence structure, and paragraph structure". It is important to note that while this definition is rather subjective, it has features that can be easily extracted – most importantly, dictionary ("choice of words"), and sentence features (length, paragraphs, punctuation).

CS230 class project (Fall 2019), Stanford, CA

Embedded in this definition of style is the implicit expectation that a writing style must be coherent – that is, conform to word conjugation and sentence formation rules, while conveying ideas in form easily understandable by human readers.

1.2 Content

Style transfer involves grafting the writing style from one input source onto the content (story arc) from another. Therefore, a style transfer application features two inputs and one output, and is expected to produce the text that remains clear and approachable for an average human. It is important to notice that the same general content can be expressed with a large variety of linguistic devices (modifiers, metaphors, sentence formations and so on), so the total number of admissible style modifications can be very large. We will refer to texts following the same story arc as "parallel representations", expecting such representations to share the content but differ in their expressive ways.

1.3 Evaluation

Since our intention is to provide unsupervised style transfer, we need to concern ourselves with subject of output evaluation. It follows from our discussion of styling that such score needs to include features characteristic of style source, narrative fluency, and content equivalence. Since the ultimate goal is a perfect imitation of source style conditioned on story from content source, missing either of the aforementioned factors will not yield a satisfactory result.

For one example, if the output text does not employ the vocabulary and sentence structure of style donor, it will result in the stylistic miss. For another example, if the output employs the style but departs from the content, it will fail to form a parallel representation. For a third example, if the output text successfully fuses the content with style of input sources but violates general language and writing norms, it will result in a poor reading experience. Therefore, to evaluate the quality of style transfer, we need to take all those considerations into account.

2 The scoring metric

There are several approaches to scoring machine generated texts. In one of the best-known examples, BLEU [3] offers a metric to judge the quality of automatic translations. BLEU assumes the parallel text representations in different languages are close in semantics, and uses n-gram frequencies to approximate preservation of content by preservation of word combinations. Unfortunately, BLEU-style score does not perform well for stylistic shifts, which can be seen from a simple counterexample: *"A cow was slaughtered in Fall."* If a human writer would try to shift the style of this phrase to Shakespeare, it could sound like this: *"In green and warm grass is where I once layed, My life was consumed, 'twas eaten by sun, Basking in the affects of heaven's rays, Til start of Fall when my end had begun. "* (author: Josh Bell). This is a perfectly admissible style transfer, yet it does not perform well according to BLEU because it features some n-grams not present in the original, and lacks others (like "cow" and "slaughter") that were present. Therefore, an issue with BLEU is that style transfer often departs from the content source by use of speech figures, which throws off any metric expecting close match in parallel representations.

These limitations of BLEU gave rise to a different family of evaluation scores that focus on theme, sentiment, sentence length, and other properties of output not linked to n-grams directly. Such metrics encode some analytical properties of text and try to find a match between the style source and the output [4]. Ultimately, an even more sophisticated style score can be constructed via "brute force" sequence classifier training which can pick the aspects of semantics, sentence structure, and vocabulary of the style source without the need to encode those features explicitly. It is also common for researchers to combine different metrics, trying to capture the elements of style and content separately.

In this paper we take the composite metric approach and build upon the evaluation paradigm proposed in [5], where the authors employ LSTM authorship classifier for style evaluation, and word-level embeddings for checking the content equivalence. In addition to adopting authorship classification for style evaluation, we extend the idea of word embedding to phrase embedding implemented by Google in their Universal Sentence Encoder [6]. Additionally, since none of the aforementioned scores directly consider the output quality in terms of the general readability, we chose to add a third metric dimension based on the Harvard GLTR tool originally designed to detect machine-generated text sequences [7]. GLTR employs a classifier trained on the massive amount of text to evaluate the percentage of "unexpected words" in the input. The premise of GLTR is that human-generated content tends to have a certain percentage of "unexpected words" – that is, tokens with relatively small probability of occurrence. Such "unexpected words" are usually a result of the story arc changes, or the reflections of creativity on behalf of the original author. By way of contrast, machine-generated texts display much smaller percentage of "unexpected" words because most of generational models are optimized for the "most probable" expansion of sequences. If we flip this idea around and use GLTR to evaluate texts coming from sequence generating networks constrained to specific tasks (like style transfer), we can employ GLTR to gauge the "smoothness" of the final result by using percentage of "unexpected" words as a proxy for conformance to language norms.

Consequently, our scoring metric has components producing a number in the three-dimensional space: style score (as detected by classifier), content (as determined by phrase embedding), and readability (as produced by GLTR). It is worth noting that these dimensions are not strictly orthogonal: for example, using uncommon words that affect GLTR could also be a part of the writer's style. Our choice to employ a three-dimensional metric is therefore motivated by the desire to have evaluation criteria that is easy to explain and understand, while our ultimate ambition is to design a style transfer algorithm that performs well in all three dimensions.

3 The RPS style transfer algorithm

In this paper, we propose to implement style transfer with a form of rejection sampling, where the sample comes from a sequence-to-sequence generational model pre-trained on the style source input, with a seed phrase coming from content input. We suggest to build a parallel representation with repeated phrase sampling (RPS), which passes through an acceptance/rejection filter (see Fig. 1). The main idea is to apply the seed content recurrently to generate a large number of samples, which are admitted or rejected based on how closely they match the embedding of the content source in a sliding window. Note that this process may take many samples to generate a match, and in some cases (when the seed is far off the style source, or the content is highly unexpected) an fitting sample may not be found despite many attempts. In such scenario, the algorithm takes content fragment verbatim and moves on in hope to find style change opportunities later in the document.

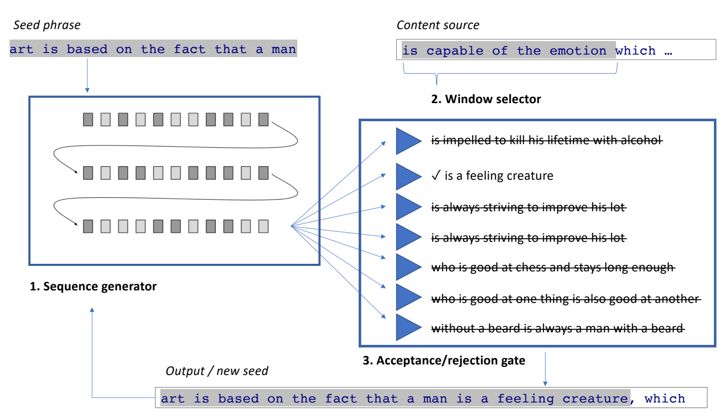

Figure 1: Reconstruction of parallel representation with phrase sampling.

Regardless of success of failure in sample acceptance, the new source becomes a part of the seed phrase and thus moves the sliding window forward. Repeated application of this algorithm allows building sequences of arbitrary length that are guaranteed to include phrase embeddings similar to the content source.

3.1 RPS component: sequence-to-sequence generator

Our algorithm relies on using a sequence-to-sequence text generator pre-trained on the style source. Recent advances in sequence modeling using RNN/LSTM/GRU networks allow for variety of models to select from. The only model-specific requirement is the ability to set the conditioning seed phrase, and adjust probability (temperature) of output samples. Since an acceptable continuation may require combing through many samples, batch output generation is also highly desirable.

3.2 RPS component: sliding window selector

Once the sequence generator is up and running, the next RPS task is to choose the length of the content window to match the samples against. This task involves considering trade-offs between stylistic richness, logical completeness and matching probability. A larger window (longer content quote to match) offers more opportunities to apply stylistic transformations (such as lengthier metaphors), but reduces the probability of generating a sample that matches the content well. At the same time, a window too short diminishes algorithm to an n-gram replacement, thus shutting the opportunity for deeper style alterations.

3.3 RPS component: acceptance/rejection gate

In the RPS architecture, acceptance/rejection gate is responsible for choosing the candidate continuation matching the content source. If no candidates are deemed acceptable, the gate can choose to accept the identity transformation (verbatim source) in order to move forward. This behavior guarantees the algorithm is not stuck over some content for which a sequence generator cannot get out of local optima and produce a globally matching sample.

4 Worked Example

For a worked example of RPS in action let us deviate from the common route of applying a style transfer from an Old English or some other archaic writing and try a more ambitious task of the style graft between texts of the two modern authors, Ayn Rand and Vladimir Nabokov.

Here is an excerpt from Ayn Rand's "Atlas Shrugged" (1957):

```
Man, at his highest potentiality, is realized and fulfilled within each creator
himself. Whether the creator is alone, or finds only a handful of others like him,
or is among the majority of mankind, is of no importance or consequence whatever;
numbers have nothing to do with it. He alone or he and a few others like him are
mankind, in the proper sense of being the proof of what man actually is, man at his
best, the essential man, man at his highest possibility.
```

In the breakdown of Rand's text by GLTR tool, her text appears relatively low on rarely expected words, while having moderate prediction results on the rest of the distribution (see Fig. 2)

For the style input we will be consider a corpus of Nabokov's works published between 1928 and 1975, with a total size of 9.8MB. To get a sense of his writing style, we may consider this excerpt from "The Luzhin Defense":

```
The fact that his life was illumined first of all from this side eased his return.
For a short while longer those harsh eminences, the gods of his being, remained in
shadow. A tender optical illusion took place: he returned to life from a direction
other than the one he had left it in, and the work of redistributing his
recollections was assumed by the wondrous happiness that welcomed him first.
```

As we observe, the difference in writing styles is rather noticeable: where Rand employs straightforward adjectives and concise sentence structures, Nabokov commands much richer and more intricate verbiage; his sentences tend to run long and his metaphors are complicated to the point they become difficult to understand. GLTR analysis confirms this intuition, detecting an unusually high count of hard-to-predict words (shown in red and violet in Fig. 3).

Let us see if we can replicate this unique style using Ayn Rand's content.

4.1 The baseline

Before we proceed with evaluating the RPS performance, let us do a sanity check and see how naive model would fare. For baseline assessment of style change, we implemented a unigram vocabulary shifter trained on Nabokov's corpus that aims to replace words in the content source one-by-one given they come from Nabokov's dictionary and have similar embeddings to content source. If we apply this transformer to Ayn Rand's excerpt, we observe the following:

```
Reading thesaurus... please be patient.
==============================>
Read 30243 words from thesaurus.
Added 70800 definitions

Original story file name? input.txt
Essay file name? rand.txt

d)etailed f)low  > d
Man{Him,His}, at his highest{over,top} potentiality{hope,might}, is
realized{achieved,attained} and fulfilled within each{all} creator{mother,father}
himself. Whether the creator{master,mother} is alone{just,once}, or finds
only{however} a handful{little} of others like{ask,as} him,
or is among the majority{thrust,most} of mankind, is of no{not}
importance{position} or consequence{control,line} whatever;
numbers{whole,matter} have{allow} nothing{wind,trifle} to do{fetch} with it. He
alone{but,once} or he and a few{short,small} others like{mate,even} him are
mankind, in the proper{only,befitting} sense{have,point} of being{new,as} the
proof{why,light} of what man{him} actually is, man{arm,him} at his
best{lick,first}, the essential{need,must} man{guy,personality}, man{body,he} at
his highest{good,over} possibility.

d)etailed f)low s)ave q)uit > f
Man, at his head might, is known and fulfilled within aside maker
himself. Whether the creator is only, or finds but a little of others like him,
or is among the favor of mankind, is of not enchantment or say whatever;
many have nothing to range with it. He alone or he and a small others thus him are
mankind, in the even scope of head the strong of what one actually is, heat at his
utmost, the occasion his, his at his good possibility.
```

In this output, detailed view shows top possible replacements based on Nabokov's dictionary, while the flow mode samples from those replacements to generate a final text. Let us proceed with analyzing this output using our evaluation metric.

First, we will measure the content preservation. For this, we employ Google Universal Semantic encoder, which produces the following identity matrix (inner product of sentence encoding vectors) of the three excerpts shown above (in the order Rand, Nabokov, Synthetic).

```
[[0.99999976 0.46428353 0.8164073 ]
 [0.46428353 1.0000005  0.47638744]
 [0.8164073  0.47638744 1.0000004 ]]
```

As we might have expected, the synthetic output appears much closer in the semantic meaning to the original Rand's paragraph (similarity 0.81) compared to semantically unrelated piece by Nabokov, so some content was definitely preserved.

Second, let us inspect how good unigram shifting performed in matching the writer's style. Here is a result of using a simple 2-means algorithm implementation acting on nltk-style bag-of-words, lexical richness and punctuation features of the three excerpts above:

```
BoW [0 1 0] / Lexical [0 1 1] / Punct: [0 1 0]
```

As expected, synthetic excerpt is clustered together with Nabokov's piece on lexical richness and use of words; it is nonetheless assigned together with Rand on punctuation, which is normal given that unigram replacement does not alter the sentence structure.

A more sophisticated authorship analysis across several authors, however, has no problems detecting that unigram replacement is not sufficient for style shifting. The sentence-level confusion matrix shown below breaks down attribution of pieces between a list of different authors ['Alcott', 'Austen', 'Bronte', 'Collins', 'Doyle', 'Montgomery', 'Stoker', 'Twain', 'Rand', 'Nabokov', 'Unigram Shift'] produced by the CNN trained on bigram and trigram frequency in the full-size texts that we took the above excerpts from. As we can see, this CNN has no problem attributing our shifted-unigram document back to Ayn Rand's authorship despite changes in unigram vocabulary:

```
Confusion Matrix:
 [[ 878   57   74   70   56  138   84  107   58   75   74]
 [   56  985  109  108  111   72   83   40   59   52   70]
 [   93   99  639  177  107   73  147   86   79  118  104]
 [   52   87  134  788  146   44  117   58   83   68   79]
 [   61   93  131  128  685   67  164  105  105  103  112]
 [  110   58   47   41   58 1025   55  129   46   53   76]
 [   71   61  104  103  176   55  824   78   49   75   78]
 [   90   34   64   49   63   94   76  979   74   68   93]
 [   60   27   59   68   82   72   44   95  526   75  599]
 [  108   61  111   94  149  103  113  103   99  658  113]
 [   57   36   58   63   71   69   65  100  [583]   98  477]]
```

Finally, let us consider the GLTR score (Fig. 4) for our unigram shifting sample. As we could have expected by grafting a richer Nabokov's dictionary onto Ayn Rand's text, percentage of "unexpected" words goes up. However, the abnormally high score also indicates that unigram replacement produces more unexpected inputs than the Nabokov's original input we analyzed in Fig. 3. This is a sign of the low text fluency: since our unigram transformer does not consider the context before word replacement, it frequently ends up with sequences that appear erratic, fragmented or otherwise out of place, and GLTR shows that.

Altogether, we can say that a simple unigram-based style shifter produces the output that shows some evidence of the new style and content preservation, but does not yield high-quality representation as replacements appear out of place with surrounding words.

4.2 RPS transform

Now let us see how the RPS model would deal with the same task. In this paper, we will be using a variant of GPT-2 network pre-trained on Nabokov's text corpus as a sequence-to-sequence generator.

We will also employ the following sliding window selector algorithm: the block for replacement must be no less than three and no more that ten words, preferably bookended by separators (period, question mark, exclamation, comma, or semicolon). For the candidate replacement, we will consider the output of generational network truncated at the sentence separator positioned within 1.5x the original word count we are trying to replace. If there are multiple candidates, we prefer end-of-sentence marker to a semicolon, and semicolon to a colon. This sliding window formation rule will allow us to generate expansions both shorter and longer than the original while not breaking the ideas expressed in the sentence blocks.

For acceptance/rejection criteria we will employ a simple semantic embedding score cutoff of 65%.

Priming the generator with a first sentence block from Ayn Rand *"Man, at his highest potentiality,"* yields us the following list of possible continuations along with embedding scores (first score is for the original Ayn Rand's completion):

```
could reach the moon's cloud tops; could see the circlet of rising water glide
across
====================
```

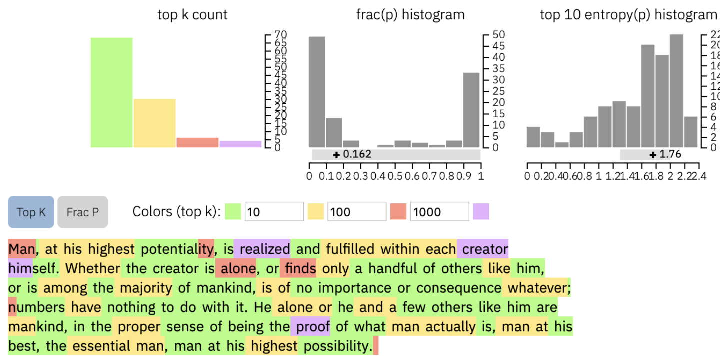

Figure 2: GLTR breakdown of Ayn Rand's excerpt.

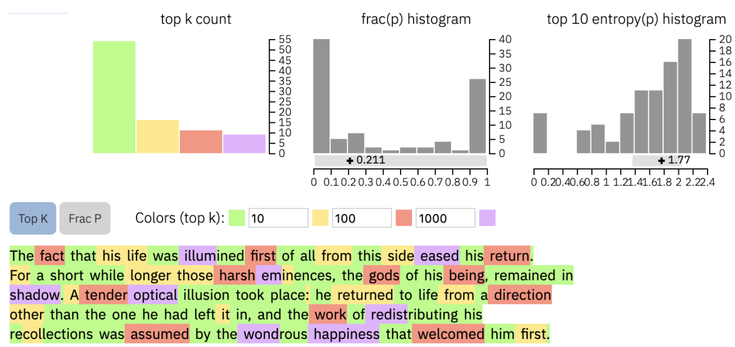

Figure 3: GLTR breakdown of Nabokov's excerpt.

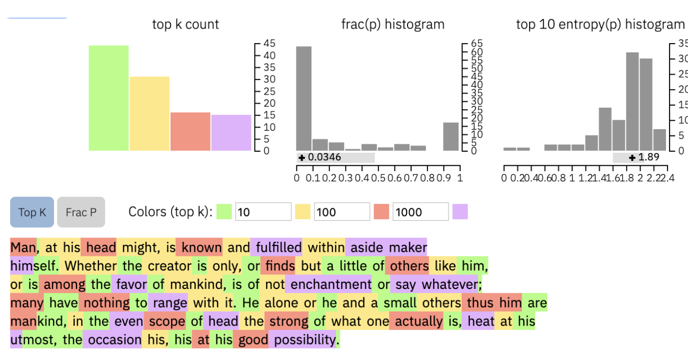

Figure 4: GLTR breakdown of unigram transform of Ayn Rand's source.

```
could in no circumstances ever have been brought as low as he was on Monday,
because on Monday the
====================
an army surgeon, who might have saved me.... No, I am not even interested in his
====================
with the poetic gesture of an airraid, had taken aim and was aiming it at the
heart of
====================
could still convey his feelings to others, but the expressiveness of his gift
was limited by his being
====================
at his impossible and irredeemably fateful apex.
====================
in the course of a struggle for existence or for the survival of a given notion
of self,
====================
could be thunderously rude and enigmatic;
====================
had made a movie out of "The God Conspiracy," for which he had received a
====================
will never resign.

[[1.0000001  0.10747568 0.3770424  0.32227987 0.16586092 0.53050757
  0.31584007 0.2512065  0.19759285 0.37974247 0.15550733]
```

As we see, output 5 has a score above 65%, so we can accept it and move on.

Now we have the following seed expansion: *"Man, at his highest potentiality, could still convey his feelings to others, but the expressiveness of his gift was limited by his being."*

This new seed gives us the next set of candidates:

```
It is not for nothing that his greatest fable, the "Contract with the Devil,"
====================
I believe that, handicapped by his being, he preferred to the art of seduction
to the science of warfare
====================
It is not for nothing that a man who is fully conscious of his mortality, yet
perfectly calm and in touch with his inner self
====================
Emotionally, he was on a par with other men of his caliber
====================
There was something about his face, about the brilliancy of his smile, about
the majestic eyes, about the slenderness of his neck
====================
And as he sits in the shade today, thinking of tomorrow's tennis match, feeling
of the utmost happiness, sure of being able to forgive everything
====================
In order to express his love he required secret expressions to be carefully
managed, and since his love was then in a very good state of mind
====================
Even if someone did have the knack of transforming 'himself' into a commercial
success, how would they want to sell it to the public?
====================
Moreover, he was afflicted with a general repulsion to all things human, and his
fear of offending a people
====================
In other words, his happiness could only be fully appreciated when it had been
fully realized
```

```
[[1.         0.35251617 0.37522185 0.34436636 0.25628358 0.20970061
  0.26284617 0.42134833 0.2947085  0.36970093 0.34587758]]
```

The original Ayn Rand's expansion was *"Whether the creator is alone,"*. This is the beginning of a new sentence and the new idea, so our generational network had trouble matching it, and no suitable expansion was found. We therefore accept the identity transformation and move on to the next prompt, *"Man, at his highest potentiality, could still convey his feelings to others, but the expressiveness of his gift was limited by his being. Whether the creator is alone,"*. Proceeding in this manner and several iterations later we arrive to the following final result:

```
Man, at his highest potentiality, could still convey his feelings to others,
but the expressiveness of his gift was limited by his being. Whether the creator
is alone, or is made up of several creators, or is a reflection of our dozen or so
minds, there is one thing that is clear: our fear of loneliness is a lethal secret;
numbers have nothing to do with it. And if we start counting, there wither we roam
and shrivel numbers, vanish and fade altogether. In the proper sense of being the
proof of what man actually is, in the proper sense of the word, the real dawning of
the Idea of Man, I think that there is no reason for our troubles.
```

Let us evaluate the quality of this output.

First and foremost, the content preservation by Google Semantic Encoder shows a score of 0.8 which is comparable to our baseline:

```
[[0.99999976 0.8096925 ]
```

Second, GLTR score shows a vast improvement over the baseline score: percentage of rare words went down and matches the original Nabokov's score nicely; moreover, it still remains higher that Ayn Rand's, which is exactly what we wanted for style transfer (Fig. 5). This effect is consistent with RPS design that alternates between the two sources of input: high-probability sequence generator and low-probability content constraint. The right mix between the two is expected to be similar to the style source.

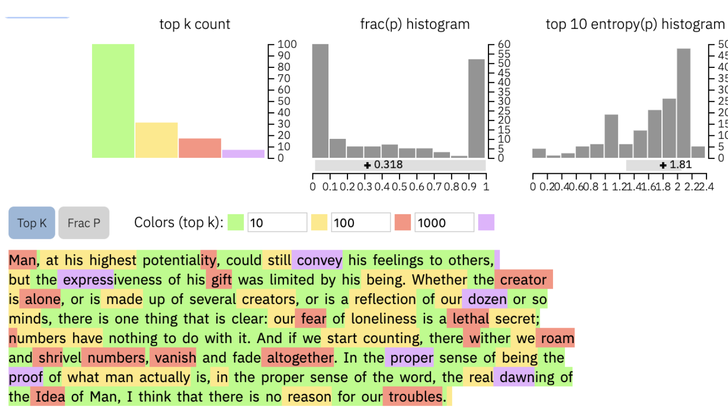

Figure 5: GLTR breakdown of RPS-generated excerpt.

Finally, authorship attribution results also look rather promising:

```
Confusion Matrix:
 [[ 878   57   74   70   56  138   84  107   58   75   74]
 [   56  985  109  108  111   72   83   40   59   52   70]
 [   93   99  639  177  107   73  147   86   79  118  104]
 [   52   87  134  788  146   44  117   58   83   68   79]
 [   61   93  131  128  685   67  164  105  105  103  112]
 [  110   58   47   41   58 1025   55  129   46   53   76]
```

```
[  71   61  104  103  176   55  824   78   49   75   78]
[  90   34   64   49   63   94   76  979   74   68   93]
[  60   27   59   68   82   72   44   95  526   75  599]
[ 108   61  111   94  149  103  113  103   99  658  113]
[  77   46   98   72  171   89  105  60    181  298  377]]
```

The newly generated text was not assigned to Nabokov, but appears quite far from Ayn Rand as well, and displays more similarity to Nabokov's corpus than any other text in the matrix. This is a very reasonable result because RPS resorts to the content source for low-probability expansions, so some degree of the style crossover is inevitable.

4.3 Discussion and open problems

The small example we worked through demonstrates the RPS potential in creating consistent and meaningful text styling. Our algorithm, however, is not perfect, and some limitations can be easily seen. For one example, the final output does not exactly match the original Ayn Rand's piece in the meaning; it reads more like a companion text rather than a faithful attempt to recreate Rand's line of thinking.

This artifact is, of course, related to the fact that our acceptance cutoff for expansion was rather low and allowed the generational network to babble on topics not covered by content source. Our experimentation with this cutoff parameter, however, have shown that acceptance rates for sentence blocks fall rapidly above the cutoff of around 70%, which means the generational network had trouble producing matches within a reasonable number of tries given more stringent acceptance criteria. This can be partially alleviated by reducing the sliding window length, but shorter replacement sequences that do not fit the natural sentence structures also limit the opportunities for stylistic changes.

For another example of RPS algorithm artifact, we can consider the loss of the keywords from the original content source. In the example above, Ayn Rand made an argument appealing to presence of the "creator" within every man. While RPS picked on the large theme of "man in the universe", it dropped the "creator" part, which made the subsequent expansion less meaningful. In other cases, RPS operation may cause keywords to vanish and reappear several times (e.g. after accepting an identity transformation), which adds to the reader's confusion due to in-out phasing of conversational subjects.

As a third difficulty we observed while experimenting with RPS, we should mention the treatment of proper names and proper nouns. It is rather common in content sources to cite and reuse proper names, such as "Peter", "Africa", "Paris" and so on. This creates a problem for RPS because proper names have no embeddings and no semantically coherent meanings. Furthermore, style source may have its own system of proper name objects partially overlapping with content source, which has an effect of triggering the unintended responses when spurious proper names are introduced, or when correct proper names become misused.

The combination of the three problems listed above limits the immediate usefulness of RPS to texts that either use proper names sparingly or match style donor in naming (an example of such content would be an essay written on subject of style source). Nonetheless, RPS is still frequently able to produce surprisingly good results in unsupervised way, and in its current form is definitely viable as a style guide or a writing aid.

5 Conclusions and future work

This paper outlines an original style transfer algorithm built alongside a new evaluation metric, which improves upon style transfers available from comparable projects known to authors. The idea of RPS is straightforward, modular and easy to implement, with results that can improve once better sequence generation and embedding models become available.

One particularly promising direction in upping the quality of RPS output could be shifting attention of generational network to keywords found in content input. Such keywords can be extracted by tools like GLTR, and appear to be worth preserving in text moving forward, which can be an interesting target of future research.

We made the software used to generate the examples available as Google Colab notebook, and it can be loaded

Acknowledgments

The authors owe a lot of debt to many researchers contributing to computational natural language field in the last decade. This project would not be possible without Google Colab and Google Cloud facilities, and OpenAI GPT-2 models.

References

[1] Orch, Franz (April 28, 2006). "Statistical machine translation live". Google Research Blog. Google.

[2] Radford, A., WuDario J., Amodei D., Amodei J., Clark M.B., Sutskever I. (2019) "Better Language Models and Their Implications". OpenAI.org

[3] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation".

[4] Ficler J, Goldberg Y., "Controlling Linguistic Style Aspects in Neural Language Generation"

[5] Fu Z.,Tan X., Peng N., Zhao D., Yan, R., "Style Transfer in Text: Exploration and Evaluation"

[6] Cera D., Yinfei Y., Sneg0yi K., Nan H., Nicole L., Rhomni J., Noah C., Marco G., Steve Y., Chrs T., Yun-Hsuan S., Brian S., Ray K., "Universal Sentence Encoder", Google Research 2018

[7] Strobelt H., Gehrmann S., "GLTR: Giant Language model Test Room" Harvad-MIT-IBM Watcon Lab