Gröndahl, Tommi; Asokan, N.

Effective writing style transfer via combinatorial paraphrasing

Published in:
Proceedings on Privacy Enhancing Technologies

DOI:
10.2478/popets-2020-0068

Published: 17/08/2020

Document Version
Publisher's PDF, also known as Version of record

Tommi Gröndahl* and N. Asokan*

Effective writing style transfer via combinatorial paraphrasing

Abstract: Stylometry can be used to profile or deanonymize authors against their will based on writing style. Style transfer provides a defence. Current techniques typically use either encoder-decoder architectures or rule-based algorithms. Crucially, style transfer must reliably retain original semantic content to be actually deployable. We conduct a multifaceted evaluation of three state-of-the-art encoder-decoder style transfer techniques, and show that all fail at semantic retainment. In particular, they do not produce appropriate paraphrases, but only retain original content in the trivial case of exactly reproducing the text. To mitigate this problem we propose ParChoice: a technique based on the *combinatorial application of multiple paraphrasing algorithms*. ParChoice strongly outperforms the encoder-decoder baselines in semantic retainment. Additionally, compared to baselines that achieve non-negligible semantic retainment, ParChoice has superior style transfer performance. We also apply ParChoice to multi-author style imitation (not considered by prior work), where we achieve up to 75% imitation success among five authors. Furthermore, when compared to two state-of-the-art rule-based style transfer techniques, ParChoice has markedly better semantic retainment. Combining ParChoice with the best performing rule-based baseline (Mutant-X [34]) also reaches the highest style transfer success on the Brennan-Greenstadt and Extended-Brennan-Greenstadt corpora, with much less impact on original meaning than when using the rule-based baseline techniques alone. Finally, we highlight a critical problem that afflicts *all* current style transfer techniques: the adversary can use the same technique for thwarting style transfer via *adversarial training*. We show that adding randomness to style transfer helps to mitigate the effectiveness of adversarial training.

Keywords: style transfer, style imitation, stylometry, adversarial stylometry, author profiling, profiling, deanonymization, model evasion

DOI 10.2478/popets-2020-0068

Received 2020-02-29; revised 2020-06-15; accepted 2020-06-16.

1 Introduction

Freedom of speech and privacy are threatened by advances in artificial intelligence, including *natural language processing* (NLP). In particular, *stylometry* can be used to identify or profile anonymous authors based on writing style [65, 68]. Institutions or individuals can use stylometry to deanonymize whistle-blowers and dissidents [7, 8, 60]. Deanonymization can put authors in danger of harassment [3] or even legal repercussions [34]. Accordingly, author deanonymization or profiling constitutes an attack on privacy [7, 24, 47, 51].

As a defence, the author can use *style transfer*. This process can consist of several *transformations*, i.e. changes applied to the input text. Prevalent approaches are based on *encoder-decoder networks* [17, 43, 58, 63, 64, 70, 71, 73], but more traditional *rule-based* techniques also continue to be used [36, 45, 61]. Importantly, style transfer is distinguished from mere style-specific generation [32, 35] by the requirement of *semantic retainment*: the transformed text should express equivalent content to the original.

Using both automatic and manual metrics, we conduct a detailed performance evaluation of three state-of-the-art style transfer techniques based on encoder-decoder networks [58, 63, 64] (Sections 5–6). The aim of these techniques is to produce a *style-neutral encoding* of the original sentence's content, and then generate the same content in the target style. However, *they all fail* at producing acceptable *paraphrases* (Section 6.1.1). Semantic retainment only succeeds in the trivial case of *reproducing* the input.

***Corresponding Author: Tommi Gröndahl:** Aalto University, Konemiehentie 2, 02150 Espoo, Finland, E-mail: tommi.grondahl@aalto.fi

***Corresponding Author: N. Asokan:** University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada, E-mail: asokan@acm.org

Such results motivate a reconsideration of alternative approaches, in particular *automatic paraphrasing*. We propose a novel style transfer technique based on *combinatorial paraphrase generation*, and style specific *paraphrase selection*. The technique, which we call *ParChoice*, is inspired by prior work in rule-based style transfer (Section 2) [36–40, 45, 46, 48, 61] but involves substantial additions to existing techniques. In Section 4 we discuss the paraphrasing algorithms in detail.

We compare ParChoice with the three encoder-decoder baselines on four author profiling datasets derived from original work presenting the baselines (Section 6). ParChoice outperforms them all in semantic retainment, especially clearly in human evaluation (Section 6.1.1). In style transfer performance, ParChoice surpasses those baselines that achieve non-negligible semantic retainment (Section 6.1.2). Additionally applying ParChoice to five-author style imitation, we achieve up to 75% imitation success (Section 6.1.2).

We also compare ParChoice to two rule-based techniques that have demonstrated strong performance in prior research [36, 45] (Section 6.2). Experimenting on the Brennan-Greenstadt corpus and the Extended-Brennan-Greenstadt corpus [7], we demonstrate that ParChoice exerts less semantic changes than either baseline. Furthermore, even though one baseline (Mutant-X [45]) achieves a higher style transfer performance than ParChoice alone, applying them both in succession significantly improves the performance of both. This combined application of ParChoice and Mutant-X also retains semantics better than Mutant-X alone.

Finally, affecting *all* state-of-the-art style transfer techniques (including ParChoice) we highlight a serious general problem. A strong adversary who is aware of the style transfer technique can employ *adversarial training*: using the style transfer technique for adding transformed examples to the training data.

We demonstrate that adversarial training thwarts all three encoder-decoder baseline techniques [58, 63, 64] as well as ParChoice, typically with only a minor negative impact on original profiling accuracy (Section 6.3). However, adversarial training fails if paraphrase selection in ParChoice is *random*, which indicates that the problem can be partly mitigated by conducting the transformations randomly instead of using a specific target style. We discuss the relation between style transfer and adversarial training, and suggest directions for future research on this problem (Sections 6.3–7).

We summarize our contributions below.

- We present ParChoice: a style transfer technique based on *combinatorial paraphrasing* (Section 4).
- By comparing ParChoice with three encoder-decoder baselines [58, 63, 64] and two rule-based baselines [36, 45], we demonstrate that:
  - ParChoice retains semantic information better than any baseline (Sections 6.1.1, 6.2.1).
  - ParChoice's style transfer performance exceeds those encoder-decoder baselines that achieve non-negligible semantic retainment (Section 6.1.2).
  - ParChoice significantly outperforms both rule-based techniques in semantic retainment, while ParChoice combined with Mutant-X [45] performs the best in style transfer (Section 6.2).
- We demonstrate that the adversary can counter style transfer by *adversarial training*, except if paraphrases are selected *randomly* (Section 6.3). We discuss possible reasons for this finding, and propose ways in which it can be taken into account in future work on style transfer.
- We make the code for implementing our original contributions available.[1]

2 Background

Author attribution via stylometry has traditionally focused on standard machine learning (ML) algorithms and feature engineering [1, 22, 33, 56, 68, 74], but deep learning methods have become more prominent in recent years [4, 9, 19, 67]. While there is no unanimous agreement on the most effective features [22, 24, 33], the *Writeprints* feature set has been widely applied with success [1–3, 15, 47, 53, 74]. Properties beyond personal identity have also been detected from writing style, including gender and age [61, 62]. We denote the detection of any author attribute as *(author) profiling*, deanonymization being a special case. We use the term *(author) profiler* for ML classifiers used for profiling.

Mitigating author profiling requires *style transfer*, i.e. transforming writing style but not semantic content. Back-and-forth machine translation provides a simple but highly limited technique [3, 7, 10, 14, 44], as it does not allow targeting or avoiding any particular style. Another classical alternative is rule-based paraphrase replacement from knowledge bases [12, 36–40, 45, 61],

1 https://gitlab.com/ssg-research/mlsec/parchoice/

Techniques	Transformations applied
[40, 46]	synonym replacement from WordNet
[45, 61]	word embedding neighbour replacement
[38]	word replacement from GNU Diction [26], hand-crafted rules
[12]	synonym replacement from FreeLing [54], hand-crafted rules
[36]	synonym/hypernym/definition replacement from WordNet or PPDB, hand-crafted rules

Table 1. Prior rule-based style transfer techniques.

Architecture	Training data	
	Same	Different
Same	Query access	Architecture access
Different	Data access	Surrogate access

Table 2. Author types based on how the author's surrogate profiler relates to the adversary's author profiler.

which we expand on with ParChoice. Table 1 summarizes prior rule-based style transfer techniques.

As opposed to rule-based paraphrasing, recent style transfer research has heavily concentrated on sequence-to-sequence mapping via *encoder-decoder networks* [17, 43, 58, 63, 64, 70, 71, 73]. Such techniques aim at producing a *style-neutral encoding* of the original sentence, which serves as the input to a *style-specific decoder*. Their main differences concern the training algorithms used to enforce (i) the style-neutrality of the latent encoding, (ii) the style-specificity of the decoding, (iii) and semantic retainment. In Section 6, we evaluate the performance of three state-of-the-art techniques that aim at reaching (i)–(iii) by different means [58, 63, 64]. As our results illustrate, none attain all three simultaneously.

3 Problem statement

The entities involved in style transfer are the *author* and the *adversary*. The author belongs to a *class* C_1, which is a set of authors. A special case of such a class is *author identity*, which is a singleton set containing only the author. The adversary has an *author profiler* P, which is a ML classifier used to profile texts by author class. We denote the predicted class of a text T as $P(T)$. If the author has written T, profiling succeeds when $P(T) = C_1$ and fails otherwise. As a defence, the author produces a *transformed* text T^*. She[2] succeeds in *style transfer* if $P(T^*) \neq C_1$, and succeeds in *imitating* another class $C_2 \neq C_1$ if $P(T^*) = C_2$. Style transfer and imitation are thus assimilated in two-class settings.

Adversary models: The adversary can access labeled *profiler training data*, which he uses to train the author profiler P. The labels include the author's true class C_1. The adversary also accesses a text written by the author; this being originally unknown to the adversary. He profiles the text with P and receives P's prediction of the text's author class. In the baseline scenario the text is T, i.e. the original unmodified text. In style transfer scenarios it is T^*, i.e. T transformed by the author.

We distinguish between two adversary types. The *weak* adversary has no access to the author's style transfer technique, and trusts the profiling result. The *strong* adversary knows the particular style transfer technique used by the author, and can use the same technique to transform any other text he accesses. To thwart style transfer he can use *adversarial training*, i.e. re-training the author profiler with transformed training data.

Assumptions: The author can either perform random transformations, or select transformations to avoid or target a specific style. For the latter, she needs a *surrogate profiler*, which is a ML classifier trained on *surrogate training data*. We distinguish between different author types by the surrogate profiler's relation to the adversary's profiler P.

If the surrogate profiler is the same as P, the author has *query access*. Alternatively, the surrogate profiler can differ from P in model architecture or training data, giving her *data access* or *architecture access*, respectively. Finally, the weakest author only has access to a surrogate profiler that is distinct from P in both architecture and training data. This *surrogate access* represents the most realistic use scenario. We summarize the different access variants in Table 2.

4 Design of ParChoice

Figure 1 shows an overview of the ParChoice pipeline. It consists of two stages: (i) *paraphrase generation*, which takes an input document and generates a set of *paraphrase candidates* (4.1); and (ii) *paraphrase selection*, which selects the candidate closest to the target writing style (4.2). In this section, we explain each stage.

2 For notational convenience, we denote the author as "she" and the adversary as "he".

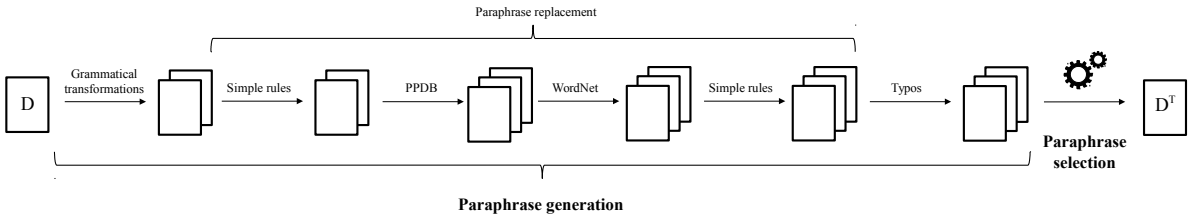

Fig. 1. ParChoice pipeline

4.1 Paraphrase generation

Our paraphrase generation stage consists of four modules, which we discuss in 4.1.1–4.1.3. We generate the *Cartesian product* of all transformations, with the aim of producing maximally varied paraphrases. We call this approach *combinatorial paraphrasing*.

4.1.1 Grammatical transformations

Grammar is a crucial aspect of writing style, and especially important for maintaining content-neutrality in stylometry [29, 30]. Yet, prior style transfer approaches have not systematically applied grammatical transformations (Section 2). A possible reason for this has been the lack of available techniques.

We used a recent tool developed by Gröndahl and Asokan, obtaining the same model used in the original work presenting it [23]. This technique is based on three tasks: (i) generating an abstract representation of the sentence (*"EAT"*) derived from its *dependency parse* [69]; (ii) transforming EAT according to targeted grammatical features; and (iii) generating an English sentence from the transformed EAT via a NMT network. The NMT network consists of an encoder and a decoder, both of which are LSTMs. It has been trained to translate EATs to English on a large corpus, consisting of 8.5 million sentences derived from multiple open-source corpora. For further details, we refer to the original paper [23]. All our transformations target only the *main verb* of the sentence. We explain the transformations below, and Table 3 shows examples.

Sentence (category)	**Transformations (category)**
John saw Mary. (active)	Mary was seen by John. (passive)
John didn't see Mary. (negated, active)	Mary wasn't seen by John. (negative, passive)
	I don't think (that) John saw Mary. (affirmed, active)
	I don't think (that) Mary was seen by John. (affirmed, passive)
Did John see Mary? (question, active)	Was Mary seen by John? (question, passive)
	Is it (true) that John saw Mary? (declarative, active)
	Is it (true) that Mary was seen by John? (declarative, passive)

Table 3. Examples of grammatical transformations (added prefixes in italics).

Voice: We produced both active and passive variants of transitive verbs, which take both a subject and a direct object in the active voice. The direct object of an active clause is expressed as the subject in a passive clause, and the subject of an active clause is expressed in the passive via the preposition *by*.

Negation: In addition to using the negative particle *not/n't*, an affirmative sentence can be negated by embedding it in a clause that states its falsity. We produced the affirmed version of an originally negated sentence, and wrapped it in the context: *I don't think (that) (...)*. The non-contracted variant *I do not think (that) (...)* was later automatically produced in the paraphrase replacement stage (Section 4.1.2).

Questions: To paraphrase a polar (*yes-no*) question, we first transformed it to a declarative variant, which we then appended to the prefix *Is it (true) that (...)*. For negative questions, we additionally generated the affirmed declarative variant embedded in *Is is not (true) that (...)* and *Isn't it (true) that (...)*.

4.1.2 Paraphrase replacement

After grammatical transformations we applied paraphrase replacement using *simple rules* and two external paraphrase corpora: *PPDB* [18] and *WordNet* [50]. We used the order *simple* → *PPDB* → *WordNet* → *simple*.

The first application of simple rules increased the range of inputs for PPDB, and their re-application at the end further expanded the range of paraphrases considered for selection. PPDB was applied before WordNet since it retained semantics better (Section 6.1.1), and thus this order was less likely to propagate errors.

Simple rules: We manually programmed a small set of rules to produce simple transformations. First, we take the presence or absence of *commas* as having only a marginal effect on interpretation, and hence allowed commas to be optionally removed.

Second, we treated the following as paraphrases:[3]

{*not*, *n't*}{*am*, *'m*}{*are*, *'re*}{*have*, *'ve*}
{*nobody*, *no-one*}{*anybody*, *anyone*}
{*somebody*, *someone*}

Third, we replaced equivalent *modal auxiliaries*. However, some are only equivalent in either an *affirmed* or a *negated* context, but not both. The context is negated if the auxiliary precedes a negation (*not*/*n't*), and affirmed otherwise. We therefore distinguished between equivalent modal auxiliary groups in affirmed and negated contexts. We additionally appended *ought* with the preposition *to* (following the negation in negated contexts), and conversely removed *to* if *ought* was replaced with another auxiliary. The following sets display the equivalent modal auxiliary groups we used:

Affirmed context
{*might, may, could, can*} {*should, ought, must*}
{*will, shall*}
Negated context
{*can, could*} {*should, ought, must*} {*will, shall*}

PPDB: The Paraphrase Database (PPDB) is a parallel corpus of paraphrases, annotated with additional semantic and syntactic information [18, 55]. PPDB-paraphrases have been used for author profiling [59], indicating that they are relevant for writing style. However, they have been less prevalent in rule-based style transfer than WordNet (Table 1). This may be due to difficulties in their direct application, which we discuss below and help overcome in ParChoice.

Syntactic context	Phrase	Paraphrase
[NN]	restriction	constraint
[VB]	co-operate	collaborate
[S/VP]	i am sorry to have to	i regret to
[S/S]	i am sorry	i regret

Table 4. Example PPDB entries.
(NN: noun, VB: verb, S: sentence, VP: verb phrase)

We restricted our use of PPDB to the *equivalent* class, comprising 245691 paraphrase pairs. We derived these from the "*PPDB-TLDR*" version.[4] However, simply replacing a phrase with a random PPDB-paraphrase easily leads to ungrammaticality due to context effects. We remedied this problem with a *grammatical filter* that only allowed entries that fit the *syntactic context* specified in the PPDB-entry.

Examples of PPDB-entries are shown in Table 4. Single-word paraphrases include the part-of-speech (POS) tag, and multi-word paraphrases contain the syntactic context in the format [X/Y]. X describes the original phrase, and Y the phrase immediately following it in the original context. Phrases are higher-level syntactic objects than words, and receive their grammatical status from their *head* word [11]. For example, the final row of Table 4 is interpreted as *i am sorry* being paraphrasable as *i regret*, when followed by a sentence.

We obtained the POS-tags and phrase structure of the original sentence with the *Spacy* parser [31].[5] For each word n-gram in the sentence, we detected the *largest phrase immediately following it*, and used this to restrict paraphrase replacement. For single-word paraphrases we used the POS tag instead. This *grammatical filtering* algorithm drastically reduced ungrammatical paraphrases produced via PPDB-replacement, and we believe it to be useful in future work on automatic paraphrasing extending beyond style transfer.

WordNet: As a manually built knowledge base of *word senses*, WordNet [50] represents possible word meanings along with multiple semantic properties, including *synonyms*. Word senses are stored as uninflected *lemmas*. While WordNet has commonly been used in rule-based style transfer [36, 40, 45, 46, 48], the lemma format is a major limitation in its direct application.

In contrast to prior studies, we *inflected* the WordNet lemmas, significantly increasing their range of ap-

3 We only produced the contraction *'ve* after a pronoun in {*I, you, we, they*}, and the contracted negation *n't* after an auxiliary in {*is, are, was, were, have, has, had, wo* (variant of *will*), *must, should, need, ought, could, can, do, does, did*}.

4 http://paraphrase.org/#/download.

5 https://spacy.io/

plication. We created a dictionary from lemmas to their surface manifestations with different POS and inflection tags, deriving these from a large text corpus.[6] We then inflected synonyms of the original word based on its POS and inflection tag, and produced paraphrase candidates with each inflected synonym. For *word sense disambiguation* [52], we used the *simple Lesk* algorithm from Python's WSD library.[7]

4.1.3 Typos

Typographical errors have demonstrated success at evading text classification [25, 42]. However, given the vast number of possible misspellings and their varying effects on readability, introducing them randomly is not justifiable. We used a simple typo algorithm of optionally removing an apostrophe, as in *you're→youre*. Additionally, we introduced typos that appear in the *target corpus* of the surrogate training data. For obtaining these we used the Python port of SymSpell,[8] applying it to the target corpus and storing a dictionary from correct spellings to possible misspellings. We additionally spell-checked the original sentence and included original typos to this typo dictionary. This allowed either retaining original typos or correcting them, depending on their effects on paraphrase selection.

Unlike other paraphrasing mechanisms, typos are *reversible* via spell-checking. The full paraphrase generation pipeline is not reversible: since a paraphrase could correspond to a very large number of possible inputs, it is practically impossible to find the correct input from the paraphrase alone. However, as typos are an important aspect of stylistic variation [1, 7, 27], removing them at pre-processing can potentially reduce the accuracy of author profiling. Our empirical results (Section 6.1.2) indicate that their effectiveness at style transfer varies across datasets, but is usually not the most important factor.

4.2 Paraphrase selection

We selected the paraphrase candidate by a *surrogate profiler*, which is a local author profiler trained on *surrogate training data* (Section 3, Table 2).

In sentence-level experiments (Section 6.1), we chose the candidate that was assigned the *highest difference between target and source class probabilities* by the surrogate profiler. While this metric assimilates to the lowest source-class probability (or the highest target-class probability) in two-class settings, in multi-class settings these probabilities come apart. This allows us to perform *imitation* instead of mere style transfer (Sections 3, 6.1.2).

In document-level experiments (Section 6.2) we replicated a genetic algorithm -based paraphrase selection method from our best-performing rule-based baseline technique: Mutant-X [45]. This selection performs only style transfer instead of imitation. We first produced a set of candidates by paraphrasing a random sentence in the document. Following Mutant-X, we then ranked the candidates based on the probability of misclassification (using query access to the author profiler), and METEOR score with the original document. The best-performing candidates were then used as inputs for further iterations of the same process. Sections 5.3.2 and 5.3.3 further discuss the parameters used for Mutant-X and the document-based ParChoice variant.

5 Experiments

In this section we describe the datasets (5.1), evaluation setup (5.2), and style transfer techniques (5.3) used in our experiments. We compared ParChoice with *three encoder-decoder baseline techniques* on *four two-class sentence-level datasets*. We additionally applied ParChoice to *multi-author imitation* on a *five-class sentence-level dataset*. Finally, we compared ParChoice with *two rule-based baseline techniques* on *four multi-class document-level datasets*.

5.1 Datasets

The encoder-decoder baseline techniques [58, 63, 64] use *sentences* as inputs, and are only applicable to *two-class* data. The rule-based baselines [36, 45] are tailored for *multi-author datasets* of *multi-sentence documents*. We used datasets specifically tailored for the baselines. Par-

6 We used the POS tagger of NLTK [6]. We obtained the inflected lemmas from the same corpus of 8.5 million sentences that was used for training the NMT network used for grammatical transformations [23] (Section 4.1.1). As the inflected variant of a lemma, we chose the most common surface form associated with a lemma-tag combination in the tagged corpus.

7 https://github.com/alvations/pywsd

8 https://github.com/wolfgarbe/SymSpell

Dataset	Classes	Training set size		Test set size
		Profiler	Surrogate	
YG	female, male	2577862	386678	10000
BA	adult, teen	2637850	395676	10000
AB	Alice, Bob	25319	3797	1176
TO	Trump, Obama	20860	3128	4668

Table 5. Sentence-based datasets (size: number of sentences).

Choice is applicable to all of them, which illustrates its flexibility across different use cases.

5.1.1 Sentence-based datasets

We used four two-class corpora, one labeled by *gender*, one by *age*, and two by *author identity*. In the gender and age experiments, we replicated the setups of two encoder-decoder baselines [58, 64]. We divided each dataset to a larger *profiler training set*, a smaller *surrogate training set* (15% of profiler training set size), and a *test set*. The original datasets are available on GitHub.[9] Size-related information is presented in Table 5.

Yelp Gender (YG): This dataset consists of restaurant reviews labeled by *gender* [61]. It contains two training sets, which results in partially divergent data/architecture access between the baselines. We discuss this in more detail in Section 5.3.1.

Blog Age (BA): The blog dataset [62] contains blog posts labeled by authorship, gender, and age. We experimented on age profiling.[10]

Alice-Bob (AB): We extracted two authors from BA: a female in the age range $13-19$, and a male in the age range $23-40$. We call them *Alice* and *Bob*, respectively.

Trump-Obama (TO): This dataset includes speeches by Barack Obama and Donald Trump [64]. We were able to improve profiling accuracy markedly by truncating the larger class (Obama) to the same size as the smaller. We used this balanced variant in our experiments.

Multi-class: We created a five-class dataset by appending AB with three additional authors from BA, introducing $7899+7270+6766$ additional training sentences.

5.1.2 Document-based datasets

For document-level experiments, we used the Brennan-Greenstadt corpus (BG) and the Extended-Brennan-Greenstadt corpus (EBG) [7]. These corpora have been manufactured specifically for the purposes of stylometry, and contain multiple documents by different authors. BG contains 12 authors and EBG 45 authors. We used the full BG, and replicated the test settings of Mahmood et al. [45] by using subsets of 5 and 10 authors from EBG. We additionally experimented on the whole EBG, giving us four datasets altogether: *BG*, EBG_5, EBG_{10}, and EBG_{45}. Following Mahmood et al. [45], we used 12 documents from each author for training the profiler, and the rest for testing.

5.2 Evaluation setup

We measured the effectiveness of each technique on two fronts: *semantic retainment* and *style transfer*.

5.2.1 Semantic retainment evaluation

We measured semantic retainment using both automatic and manual metrics, and conducted a user study with independent evaluators for comparing ParChoice to the encoder-decoder baselines.

Automatic evaluation: We calculated the *METEOR* score [5] between the original and transformed test sets. METEOR is based on n-gram overlap, and additionally considers synonyms and paraphrases. We used the METEOR implementation of the *nlg-eval*[11] package.

Manual evaluation: We manually examined a subset of test set transformations, assessing whether they constituted acceptable paraphrases or had errors. For sentence-based data, we evaluated 100 random sentences from each two-class dataset (50 from each direction). For document-based data, we manually evaluated those sentences that were transformed by all the

9 YG: https://github.com/shrimai/Style-Transfer-Through-Back-Translation
BA/AB/TO: https://github.com/rakshithShetty/A4NT-author-masking/blob/master/README.md

10 While we also experimented with gender, classification was heavily biased toward the *female* class on original data, not achieving $> 40\%$ accuracy on the *male* class with any of our three classifiers. Since the same classifiers worked much better with other datasets, this problem seemed to be due to the data itself rather than classifier architectures. These results are in line with those of Shetty et al. [64] who also received much lower classification accuracy on gender than on age in the blog dataset. We therefore focused on age in our experiments.

11 https://github.com/Maluuba/nlg-eval

techniques compared (81 altogether), and combined this evaluation with the rate of transformed sentences per document for each technique. This yields an estimation of how likely the techniques are to retain the original meaning of a sentence in a document, by not transforming it or by generating an appropriate paraphrase.

User study: We conducted a user study on 20 participants to evaluate transformations by ParChoice and the encoder-decoder baselines. Of the participants, 25% were native English speakers. 55% were female and 45% male. 60% were $20-30$ years old, and 40% were 30 or older. 90% had a university degree, most often on the Master's level (65%). In an online questionnaire, each participant was allocated a set of 20 sentences drawn from the YG dataset. They were shown the original sentence along with transformed versions by all four imitation techniques (in a random order). All users were given *different* sentences to increase variation, resulting in $20 \times 20 \times 4 = 1600$ evaluations altogether. To ensure the relevance of the evaluation, we only used imitations that were *non-identical* with the original sentence, i.e. not exact reproductions. Participants compared each variant to the original sentence, and rated it on a $0-5$ scale based on similarity of meaning.

5.2.2 Style transfer evaluation

For sentence-level datasets, we trained three author profilers that represent the state-of-the-art in stylometry (Section 2). In document-level tests we used the best-performing profiler setup from Mahmood et al. [45].

Profilers: We adopted the most commonly used deep learning text classification techniques: *long short-term memory networks* (LSTM) [28] and *convolutional neural networks* (CNN) [41]. We used implementations from Shetty et al. [64] (LSTM) and Prabhumoye et al. [58] (CNN), which also form parts of our baseline style transfer techniques (Section 5.3.1). Both use words as input features. The source codes are available on GitHub.[12]

Writeprints features have exhibited strong performance in stylometry (Section 2) [1–3, 15, 47, 53, 74]. For our third sentence-based author profiler, we collected *static* Writeprints features [74] and trained a *multilayer perceptron* (MLP) profiler that we call *WP*. Static features apply to any user and are thus more general than *dynamic* features, which include user-specific information. We used the following features: number of words, average word length, number of short words (≤ 3), number of characters, digit percentage, uppercase character percentage, spacial character percentage, number of letters, number of digits, common character bigram/trigram percentages, number of hapax- and dislegomena, number of function words, number of POS tags, and number of punctuation markers. Based on a comparative evaluation between five ML architectures (MLP, logistic regression, naive Bayes, decision trees, and support vector machines), MLP fared the best on our datasets with the static Writeprints features. WP has a single hidden layer of 100 nodes.

For document-based datasets, we replicated the test settings of Mahmood et al. [45], who used a *random forest classifier* trained on static Writeprints-features. This profiler had the highest performance in their comparative evaluation with other architectures.

Proofreading: From 100 sentences of YG, we manually produced additional *proofread* transformations. The purpose of the proofreading was to ensure semantic retainment while changing as little of the transformation as possible. All corrections were made to the direction of the original sentence; i.e. we did not produce any novel paraphrases. This test evaluates how well the style transfer techniques are able to perform if the author secures semantic retainment by correcting the output. It thus resembles semi-automatic style transfer frameworks like Anonymouth [47] or AuthorWebs [14].

Adversarial training We tested the effectiveness of thwarting style transfer via adversarial training on the AB and five-class datasets, with the LSTM profiler. With each style transfer technique, we appended transformed variants to the original training set, and retrained the LSTM with this adversarial training set. We then measured profiling accuracy on both the original (non-transformed) test set and the transformations performed by the technique used for adversarial training.

5.3 Style transfer techniques

We review the technical details of the baseline techniques (5.3.1–5.3.2) and ParChoice-variants (5.3.3).

12 https://github.com/shrimai/Style-Transfer-Through-Back-Translation
https://github.com/rakshithShetty/A4NT-author-masking/blob/master/README.md

Name	Architecture access	Data access
CAE	-	BA/AB/TO
BT	CNN	YG/BA/AB/TO
A^{4}NT	LSTM	BA/AB/TO
ParChoice-CNN	CNN	YG/BA/AB/TO
ParChoice-LSTM	LSTM	BA/AB/TO
ParChoice-WP	WP	YG/BA/AB/TO
ParChoice-LR$_d$	-	YG/BA/AB/TO
ParChoice-LR$_s$	-	-
ParChoice-random	-	-

Table 6. Architecture/data access of sentence-based techniques.

5.3.1 Encoder-decoder Baselines

The main idea behind the encoder-decoder baseline techniques is to generate a style-neutral encoding of the original sentence, which functions as the input to a style-specific decoder. Some models are available as pre-trained, and the rest we trained ourselves. Pre-trained models and the code for training the baselines are available on the respective projects' GitHub pages (linked below). The baselines partly differ in their access assumptions (cf. Table 2), as summarized in Table 6.

Cross-aligned autoencoder (CAE): The CAE technique is a style-specific autoencoder that uses a method called *cross-alignment* for calibrating encoding distributions [63]. The *encoder* produces a latent content variable from the input sentence, and the *decoder* generates the target sentence from this content variable together with a target style feature. CAE does not have query access to any of our author profilers, but has data access to every dataset except YG (explanation below). We trained CAE for every dataset, using the project's code from GitHub.[13]

Back-translation (BT): As an alternative to cross-alignment, BT produces the latent content variable with a pre-trained MT system [58]. The original English sentence is first translated to French, which is then encoded with the French-English encoder. An English decoder then produces the target sentence from the encoding, and separate decoders are trained to target specific styles. Style-specificity is enforced by a CNN, which we also use as one of our author profilers (Section 5.2.2).

BT trained on YG is provided on the project's GitHub page.[14] It has been trained with two separate datasets: one for the *classifier* and another for the *decoder* (Section 5.1). We used the decoder training set to train the other baselines, and the classifier training set for training the profilers. Therefore BT has data access to YG, but CAE and A^{4}NT do not. With other datasets we trained BT using the same training data for both the classifier and the generator.

A^{4}NT: *Adversarial Author Attribute Anonymity Neural Translation* (A^{4}NT) [64] is a style transfer technique based on *generative adversarial networks* (GANs). A GAN consists of two neural networks, where one (the *classifier*) is trained to classify outputs generated by the other (the *generator*), which in turn is trained to deceive the classifier [21]. The A^{4}NT-generator is trained to produce sentences classified as the target style by an LSTM, which we also use as one of our author profilers (Section 5.2.2). During training, semantic retainment is regulated by the *reconstruction probability* of the original sentence via a reverse transformation.

We used pre-trained A^{4}NT-models for BA and TO,[15] and trained A^{4}NT ourselves for YG and AB. While we always used the full dataset for training the initial classifier and generator, hardware limitations required us to truncate YG during the GAN-training phrase. We used a subset of 100000 sentences for this.

5.3.2 Rule-based baselines

We used two rule-based baselines for the document-level tests. The first [36] exhibited leading performance on the PAN2016 Author Obfuscation task [57], and the second [45] achieved state-of-the-art results on the EBG corpus (Section 5.1.2). Following Mahmood et al [45], we call these *PAN2016* and *Mutant-X*. We implemented both with code from the original projects (links below).

PAN2016:[16] This technique [36] uses multiple hand-crafted rules along with word replacement from WordNet and PPDB. Unlike ParChoice, PAN2016 does not conduct either inflection or grammatical filtering to increase the readability of the output. In addition to synonyms, WordNet-replacement also uses *hypernyms* and *definitions*. Additional hand-crafted rules include e.g. replacing or injecting stopwords, replacing or removing punctuation, and expanding contracted forms.

13 https://github.com/shentianxiao/language-style-transfer

14 https://github.com/shrimai/Style-Transfer-Through-Back-Translation

15 https://github.com/rakshithShetty/A4NT-author-masking/blob/master/README.md

16 https://bitbucket.org/pan2016authorobfuscation/authorobfuscation/src/master/

PAN2016 contains its own stylometric feature set (similar to Writeprints), and calculates the *average* of these features from the training corpus. It then alters the test document to shift its features closer to this average. Hence, it relies on *data access* to the original training set, but does not require query access to the profiler.

Mutant-X:[17] This technique [45] replaces original words with their *word embedding neighbours* obtained from a pre-trained Word2Vec [49] model. The neighbour order is further modified to shift words of opposite sentiment (e.g. *good* and *bad*) away from each other [72]. Mutant-X repeats random word replacement multiple times, applying a *genetic algorithm* [20] to keep the best performing variants after each iteration. Performance is measured as the weighted combined effect of METEOR score and how much original class probability is taken down. For calculating the latter, Mutant-X uses *query access* to the author profiler. We used the same hyperparameters as Mahmood et al. [45]: maximally 5% of document words changed per run; 100 runs per iteration; maximally 25 iterations; and 0.25/0.75 weights for METEOR and class probability, respectively.

5.3.3 ParChoice

We implemented ParChoice in Python 3.

Sentence-based variants: To provide a maximally close comparison to the encoder-decoder baselines, we replicated the data/architecture access of BT and A^{4}NT by using the CNN and LSTM profilers for paraphrase selection, respectively (ParChoice-CNN and ParChoice-LSTM). We also experimented with black-box access to the WP profiler (ParChoice-WP). As a separate *surrogate profiler*, we used a *logistic regression* classifier with word unigrams as input features. We trained two versions of the surrogate profiler: one with *data access* to the targeted profiler's training data (ParChoice-LR$_\mathrm{d}$), and another trained on separate *surrogate training data* (ParChoice-LR$_\mathrm{s}$). For consistency across experiments, we always used 15% of the author profiler training data size as the surrogate training data size (Section 5.1; Table 5).[18] Finally, we also experimented on *random paraphrase selection* without any surrogate profiler (ParChoice-random). Rows 4 – 9 of Table 6 summarize sentence-based ParChoice variants and their data/architecture access.

Genetic algorithm: On document-level data, we used the same genetic algorithm as Mutant-X (Section 5.3.2), except that instead of changing random words in the document, each run randomly paraphrased *one sentence* in the document. With the hyperparameters used, this meant that for 25 iterations, 100 new candidates of the document were produced by paraphrasing a single random sentence. The best candidates were then selected for further iterations, as in Mutant-X. We replicated Mutant-X's paraphrase selection using a combination of METEOR and *query access* to the targeted author profiler, with 0.25/0.75 weights, respectively (Section 5.3.2).

ParChoice + Mutant-X: We additionally combined the two best-performing rule-based techniques: ParChoice and Mutant-X. We first ran ParChoice, and then applied Mutant-X only to those documents that had not yet succeeded in style transfer. This combination thus maximized the use of ParChoice, and applied Mutant-X when ParChoice alone was insufficient.

Hyperparameters: Initial manual evaluation indicated that most semantic problems occurred in long sentences with multiple transformations. This motivated an upper limit to transformations per sentence. A limit based on sentence length is problematic for short sentences, where even minor transformations are percentually large. Instead, we used a constant edit distance limit of 10, which allows large changes in short sentences but limits them in long sentences. For computational efficiency, we also limited the number of PPDB- and WordNet replacements to 1000 per sentence. Comparison with larger values indicated that further increasing this number had little to no effect on performance.

6 Evaluation results

We present the results on experiments on the sentence-level (6.1) and document-level (6.2). Raw data and example transformations are provided in Appendix A.

17 https://github.com/asad1996172/Mutant-X/

18 In ParChoice-LR$_\mathrm{s}$ we also used the surrogate training data for obtaining typos (Section 4.1.3). To maintain consistency in surrogate data sizes across all experiments, we used a different YG surrogate dataset for ParChoice-LR$_\mathrm{s}$ than the decoder training set used by A^{4}NT and CAE, even though both are distinct from the classifier training set (Section 5.3.1, Table 6). The ParChoice-LSTM variant was trained on the decoder training set, as it replicates the access properties of A^{4}NT (Table 6).

Technique	YG	BA	AB	TO
CAE	19.63	21.49	6.81	4.40
BT	20.88	17.91	5.64	4.62
A^{4}NT	44.95	48.98	22.98	19.81
ParChoice-CNN	46.09	50.70	48.61	51.20
ParChoice-LSTM	45.33	48.94	41.44	48.86
ParChoice-WP	45.20	48.72	41.06	49.84
ParChoice-LR$_d$	46.00	48.59	43.02	49.17
ParChoice-LR$_s$	45.89	48.83	45.82	50.27

Table 7. METEOR scores between original and transformed sentences with all sentence-based style transfer techniques.

Technique	YG	BA	AB	TO
CAE	2%	15%	1%	0%
BT	3%	3%	0%	0%
A^{4}NT	31%	44%	16%	6%
ParChoice-LR$_s$	54%	59%	60%	61%

Table 8. Manual evaluation: rate of acceptable paraphrases from 100 sentences in each dataset (50 to both directions), transformed with each sentence-based technique.

Technique	Mean	Median	≥ 4	5
CAE	0.8	0	5%	2%
BT	0.9	0	8%	3%
A^{4}NT	1.7	1	20%	9%
ParChoice-LR$_s$	2.7	3	41%	24%

Table 9. User study results: grade statistics from human evaluations of meaning similarity from 400 sentences from YG, transformed with each sentence-based technique (grade scale $0-5$).

6.1 Sentence-level experiments

ParChoice exhibits a *higher semantic retainment* than any encoder-decoder baseline (6.1.1). Its style transfer performance is higher than that of A^{4}NT, which is the only encoder-decoder baseline that achieves non-negligible semantic retainment (6.1.2).

6.1.1 Semantic retainment

METEOR and manual evaluation scores are presented in Tables 7–8, and user study results in Table 9.

METEOR: ParChoice and A^{4}NT always clearly outperformed CAE and BT in METEOR. Especially on the smaller datasets (AB, TO), CAE and BT attained very poor scores (< 10) that imply almost no semantic overlap with the original sentences. A^{4}NT performed comparably to ParChoice in the large datasets (YG, BA), but never exceeded ParChoice-CNN, which was the highest performing ParChoice-variant. However, in the small datasets (AB, TO) A^{4}NT's scores dropped sharply. Different ParChoice-variants performed comparably.[19] Unlike A^{4}NT, ParChoice achieved high scores (~ 50) on *both* large and small datasets.

We also compared METEOR scores with each ParChoice-module applied alone in the ParChoice-LSTM variant. All achieved scores between 50 and 67. Simple rules remained the highest, as expected due to the small extent of paraphrases they produce. Typos had the largest range of variation ($50-67$), which demonstrates their dependency on the extent to which possible typos are available in the target class training data (Section 4.1.3). Grammatical transformations, PPDB, and WordNet performed similarly (in the range $55-63$), WordNet being systematically slightly higher than the rest. A likely reason for this is METEOR's bias toward WordNet synonyms as opposed to the kinds of paraphrases produced with grammatical transformations or PPDB. In contrast, in manual evaluation PPDB fared better than WordNet.

Manual evaluation: Table 8 presents our manual evaluation on 100 sentences from each two-class dataset. For practical reasons we limited our manual ParChoice-evaluation to only one variant. We chose ParChoice-LR$_s$ for two reasons. First, compared to other variants, it had neither the highest nor lowest overall METEOR score (Table 7), which indicates that the manual evaluations are not likely to either over- or underestimate the general performance of ParChoice. Second, it implements the most realistic access assumptions out of all (non-random) ParChoice-variants (Section 3, Table 6).

CAE and BT produced practically *no acceptable paraphrases*. This was especially true in the small datasets (AB, TO), where imitations bore no resemblance to the original sentence and simply repeated certain words prevalent in the target corpus.

With A^{4}NT, sentence *reproduction* was much more common than anywhere else, but actual paraphrases remained rare. For example, A^{4}NT's acceptable paraphrase rate in BA decreases from 44% to only 2% when reproductions are excluded. A^{4}NT also generated a

19 ParChoice-random performed the best overall, but we exclude it here because of its lack of targeted paraphrase selection. The METEOR scores of ParChoice-random were 46.43 (YG), 49.72 (BA), 50.54 (AB), and 49.88 (TO).

Technique	YG			BA			AB			TO		
	LSTM	CNN	WP	LSTM	CNN	WP	LSTM	CNN	WP	LSTM	CNN	WP
CAE	0.31	0.30	0.21	0.15	0.16	0.13	0.74	0.77	0.55	0.34	0.20	0.28
BT	0.33	0.34	0.20	0.15	0.17	0.09	0.77	0.84	0.59	0.49	0.33	0.27
A^{4}NT	0.16	0.17	0.10	0.10	0.11	0.05	0.19	0.22	0.14	0.53	0.07	0.33
ParChoice-CNN	0.39	0.49	0.19	0.21	0.35	0.08	0.26	0.39	0.16	0.13	0.21	0.12
ParChoice-LSTM	0.44	0.40	0.22	0.37	0.27	0.11	0.47	0.32	0.33	0.47	0.09	0.17
ParChoice-WP	0.26	0.23	0.48	0.16	0.17	0.34	0.21	0.22	0.60	0.15	0.06	0.59
ParChoice-LR$_d$	0.39	0.40	0.16	0.28	0.29	0.11	0.26	0.25	0.29	0.32	0.09	0.17
ParChoice-LR$_s$	0.36	0.36	0.15	0.22	0.21	0.09	0.21	0.22	0.20	0.23	0.07	0.14
ParChoice-random	0.12	0.12	0.08	0.04	0.05	0.01	0.10	0.11	0.09	0.09	0.03	0.09

Table 10. Author profiler accuracy decrease in sentence-based datasets, best (highest) scores framed.

large number of *omissions*, reproducing only part of the original sentence without any changes or additions. Such pure omissions were rare with other techniques. These characteristics are likely due to A^{4}NT's training function, which includes a *reconstruction loss* [64] (Section 5.3.1). In the small datasets (AB, TO) A^{4}NT's semantic retainment starkly declined to almost non-existent.

ParChoice clearly stood out by producing many acceptable paraphrases. In contrast to the baselines, ParChoice's performance was similar across all four datasets. The most prevalent ParChoice-transformation was paraphrase replacement from PPDB or WordNet, most commonly targeting a single word. Typos and grammatical transformations were rare in the manually evaluated test sets, but some were encountered. When PPDB- and WordNet-replacement could be distinguished, PPDB-replacement fared better in semantic retainment. Most ParChoice-errors were caused by contextually unsuitable WordNet synonym choice (e.g. *man→mankind*). Such errors are due to faulty word sense disambiguation (Section 4.1.2), improving which is an important challenge for future research.

User study: Table 9 presents the user study results. ParChoice clearly outperformed all baselines. As expected, CAE and BT performed especially poorly. The majority of ParChoice-imitations were on the upper half of the six-point scale ($3-5$), whereas the majority of all baseline imitations had the lowest grades ($0-1$).

6.1.2 Style transfer

We present style transfer results on two-class settings and multi-class author imitation.

Two-class tests: To evaluate style transfer success, we measured *accuracy decrease*: i.e. how much original accuracy dropped after style transfer. Table 10 provides these results. Table 11 shows profiling accuracies on original test sets with the three profilers (Section 5.2.2).

Profiler	YG	BA	AB	TO
LSTM	0.83	0.62	0.91	0.82
CNN	0.82	0.63	0.93	0.64
WP	0.75	0.59	0.82	0.74

Table 11. Original author profiling accuracies.

A^{4}NT's performance was the weakest everywhere except TO. CAE and BT achieved almost full imitation in AB and the Obama→Trump direction of TO. This was unsurprising since they simply repeated words unrelated to the original (Section 6.1.1). However, both showed a clear bias toward the Trump class, and failed in the Trump→Obama direction.

In large datasets (YG, BA), *all* (non-random) variants of ParChoice *outperformed all baselines*. This happened even under the weakest access assumptions, i.e. ParChoice-LR$_s$. In small datasets (AB, TO), ParChoice retained similar performance but baselines increased theirs. ParChoice-LR$_d$ achieved 5% better average accuracy decrease than ParChoice-LR$_s$, and query access to the author profiler (CNN or LSTM) increased it $10\%-20\%$. Query access to WP allowed effective black-box evasion of WP, but did not reach the performance of other variants with other profilers. ParChoice-random expectedly took accuracy down the least.

We additionally applied each ParChoice module separately on the ParChoice-LSTM variant, and compared accuracy decrease (Table 12) and prediction overlap (Table 13) between the modules on the LSTM profiler. All techniques had at least a minor impact. PPDB and WordNet were the most effective overall and had similar success ($13\%-20\%$). Grammatical transformations outperformed them in AB, but were less successful oth-

Technique	YG	BA	AB	TO	Avg.
Grammatical	0.12	0.07	0.14	0.08	0.10
Simple rules	0.01	0.04	0.03	0.05	0.03
PPDB	0.19	0.18	0.13	0.25	0.19
WordNet	0.15	0.20	0.13	0.23	0.18
Typos	0.02	0.10	0.15	0.08	0.09
All	0.44	0.37	0.47	0.47	0.46

Table 12. Profiler accuracy decrease with ParChoice modules individually and together (LSTM profiler, ParChoice-LSTM variant).

	Gram.	Simple	PPDB	WordNet	Typos
Gram.	100%	93%	80%	80%	89%
Simple	93%	100%	83%	84%	95%
PPDB	80%	83%	100%	81%	82%
WordNet	80%	84%	81%	100%	83%
Typos	89%	95%	82%	83%	100%

Table 13. Prediction overlap between ParChoice modules (combined from all datasets, LSTM profiler, ParChoice-LSTM variant).

erwise. Typos had the largest variation, ranging from almost nonexistent in YG (2%) to being the most effective in AB (15%). Simple rules were expectedly the least effective when applied alone ($1\% - 5\%$). The most effective techniques (PPDB and WordNet) were also the most diverse, having the least prediction overlap with other techniques ($80\% - 84\%$) and each other (81%).

Proofreading: To emulate a scenario where the author manually checks the results of style transfer, we proofread 50 transformed sentences from both the *male* and *female* test sets of YG. For CAE and BT proofreading often required reproducing the original sentence due to very poor semantic retainment (Section 6.1.1). Proofreading negatively impacted style transfer with all techniques, but markedly less with ParChoice than the baselines (Table 14).

Multi-class tests: We evaluated five-class author imitation to every direction on a blog author dataset (Section 5.1). Since ParChoice-CNN and ParChoice-LSTM had the best overall performance in two-class settings,

Profiler	CAE	BT	A^{4}NT	ParChoice-LR$_s$
LSTM	0.11	0.14	0.09	0.31
CNN	0.12	0.11	0.09	0.26
WP	0.09	0.09	0.08	0.10

Table 14. Profiler accuracy decrease on a YG-subset after manual proofreading; best (highest) scores framed.

Technique	Profiler	Source decrease	Target increase	Imitation
ParChoice-LSTM	LSTM	0.39	0.29	15/20
	CNN	0.28	0.20	1/20
ParChoice-CNN	LSTM	0.21	0.13	0/20
	CNN	0.40	0.27	7/20

Table 15. Five-class author imitation in the blog author dataset with ParChoice (query access to LSTM/CNN): average source class accuracy decrease, target class accuracy increase, and imitation success (imitation threshold: majority of source class documents assigned to target class).

we experimented on these variants on the CNN and LSTM profilers. We discarded WP here because it failed to achieve high profiling accuracies on the original five-class test sets. On each author's test set, we considered both the *accuracy decrease of the original class*, and the *accuracy increase of the target class*. Additionally, if the majority of the source author test sentences was assigned to the target author, we considered imitation to succeed for that source-target pair. Table 15 summarizes the results from all 20 imitation directions. ParChoice-LSTM's overall performance was superior to ParChoice-CNN's. ParChoice-LSTM reached the threshold 75% of the time on the LSTM profiler.

6.2 Document-level experiments

We compare ParChoice to the rule-based PAN2016 [36] and Mutant-X [45] baselines (Section 5.3.2) in semantic retainment (6.2.1) and style transfer (6.2.2). ParChoice is markedly better in semantic retainment than either baseline, but Mutant-X remains stronger in style transfer. A combination of ParChoice and Mutant-X outperforms prior rule-based schemes in both style transfer and semantic retainment.

6.2.1 Semantic retainment

We use three measures of document-level semantic retainment: (i) METEOR score, (ii) rate of sentences transformed per document, and (iii) manual evaluation of paraphrases in *transformed* sentences. Originally misclassified documents were discarded from the evaluation, since they were left unchanged.

METEOR: All techniques achieved relatively high METEOR scores (Table 16), with ParChoice outperforming PAN2016 (by $19 - 34$ points) and Mutant-X

Technique	METEOR			
	BG	EBG_5	EBG_{10}	EBG_{45}
PAN2016	43.17	45.93	46.96	47.15
Mutant-X	56.71	57.23	53.46	56.75
ParChoice	77.40	65.41	66.26	69.34
ParChoice+Mutant-X	61.61	60.25	62.31	65.86

Table 16. METEOR scores between original and transformed documents from rule-based techniques (originally misclassified documents discarded).

Technique	Acceptable Paraphrases	Error count			
		A	W	C	G
PAN2016	44%	0	32	12	1
Mutant-X	32%	19	62	7	4
ParChoice	60%	1	20	12	2
ParChoice+Mutant-X	53%	2	40	16	1

Table 18. Manual evaluation results from rule-based techniques (calculated from 81 sentences transformed by all techniques). Error types: A=antonym; W=word; C=context; G=grammar.

Technique	Transformed sentences			
	BG	EBG_5	EBG_{10}	EBG_{45}
PAN2016	98%	98%	98%	96%
Mutant-X	71%	72%	79%	73%
ParChoice	6%	25%	11%	7%
ParChoice+Mutant-X	43%	41%	22%	14%

Table 17. Average transformed sentence rates per document (originally misclassified documents discarded).

Technique	Corpus			
	BG	EBG_5	EBG_{10}	EBG_{45}
PAN2016	50%	13%	32%	76%
Mutant-X	82%	60%	87%	96%
ParChoice	36%	47%	84%	91%
ParChoice+Mutant-X	82%	77%	97%	100%

Table 19. Successful style transfer rates of rule-based techniques (originally misclassified documents discarded).

(by $8-21$ points). Combining ParChoice with Mutant-X also always retained a higher score than Mutant-X alone (by $9-3$ points).

Transformed sentence rates: PAN2016 transformed almost all sentences per document, Mutant-X a clear majority ($>70\%$), and ParChoice only $6\%-25\%$ (Table 17). This is a major divergence among the techniques, as it makes ParChoice much more likely to retain semantics by focusing transformations only on a minority of sentences. While the transformed sentence rate increased when Mutant-X was applied after ParChoice, it remained significantly lower than with Mutant-X alone.

Manual evaluation: To ensure fair comparison, we manually evaluated each technique on the *same* sentences. Across all four datasets there were 81 sentences that all techniques had made changes to. Table 18 shows manual evaluation results on these sentences.

We provide an error analysis, categorizing errors to four types based on decreasing severity:

- **Antonym**: opposite meaning
- **Word**: non-antonym but different meaning
- **Context**: possible paraphrase but wrong context
- **Grammar**: correct paraphrase but wrong grammar

Mutant-X produced 19 antonyms across all 81 sentences, including e.g. *more→less* and *easier→harder*. While Mutant-X uses sentiment-based neighbour ranking [72] to avoid antonyms (Section 5.3.2), sentiment is only one possible source of antonymity. In contrast, antonyms were absent in PAN2016, and only one was found in ParChoice: *defended→opposed*. The most likely source for this error was PPDB. Other word errors were the most common in all techniques, but much rarer in ParChoice than elsewhere. In ParChoice the rate of context errors was the same as in PAN2016, but their percentage of all errors larger. These included e.g. *issue→number*, which would be correct in a magazine-related context but not otherwise. All techniques produced a few $(1-4)$ purely grammatical errors, mostly due to inflection.

Overall, Mutant-X made markedly more semantic errors than either PAN2016 or ParChoice. This result seems to contrast PAN2016 having a lower METEOR score. This may be due to PAN2016 adding words (such as *Additionally*) before sentences, which usually does not negatively impact semantics but brings n-gram overlap (and hence METEOR) down. Applying ParChoice before Mutant-X significantly reduced the most fatal errors (antonym or word error) produced by Mutant-X.

6.2.2 Style transfer

Table 19 summarizes document-level style transfer performance. We define the *successful style transfer rate* as the frequency at which an originally correctly classified document was incorrectly classified after style transfer. Mutant-X had superior performance to PAN2016 and ParChoice across all datasets, although the differ-

Technique	Original profiler	New profiler
PAN2016	13%	20%
Mutant-X	60%	17%
ParChoice	47%	20%

Table 20. Style transfer success rates on EBG_5 with the original (queried) profiler and a new re-trained random forest profiler.

Technique	Direction	Original	Transformed
CAE	**A→B**	0.88 → 0.78	0.18 → 0.90
	B→A	0.94 → 0.79	0.16 → 0.87
BT	**A→B**	0.88 → 0.85	0.18 → 0.78
	B→A	0.94 → 0.88	0.11 → 0.64
A^{4}NT	**A→B**	0.88 → 0.90	0.74 → 0.86
	B→A	0.94 → 0.85	0.70 → 0.64
ParChoice-LSTM	**A→B**	0.88 → 0.90	0.36 → 0.88
	B→A	0.94 → 0.88	0.52 → 0.92
ParChoice-random	**A→B**	0.88 → 0.79	0.74 → 0.47
	B→A	0.94 → 0.72	0.89 → 0.38

Table 21. LSTM author profiler accuracy without→with adversarial training, on both original and transformed test sets of AB.

ence to ParChoice was only minor in EBG_5 and EBG_{10} ($3\% - 5\%$). However, when Mutant-X was applied after ParChoice, we reached the highest result on all datasets. Given that this combination also achieved significantly higher semantic retainment than Mutant-X alone (Section 6.2.1), it constitutes the state-of-the-art solution to document-level style transfer on these datasets.

However, we note a problem in assuming *query access* to the adversary's random forest profiler. The profiler involves randomness in training, and *re-training* it on the same training data mostly *undoes* the effect of style transfer on query-access techniques. Table 20 shows this effect on EBG_5. As PAN2016 does not use query access, it is not vulnerable to this problem. Query access can thus result in highly *profiler-specific* transformations, which are *non-transferable* across profilers.

Problematically, most style transfer research has so far relied on query access. In particular, with the exception of PAN2016 [36], all our baseline techniques were originally applied in query access settings [45, 58, 63, 64]. Since query access is not a realistic requirement for the author (Section 3), and given its lack of transferability to different profilers, we recommend moving beyond query access in style transfer research. While we were able to do so successfully on sentence-based data (Section 6.1), on the document-level it remains a challenge for future work.

6.3 Thwarting style transfer via adversarial training

To test the effectiveness of adversarial training for countering style transfer, we applied it to every technique on the AB dataset with the LSTM profiler. We chose this setting because it achieved a high original test set accuracy as well as the highest overall accuracy change (Section 6.1.2, Table 10). Table 21 presents the results.

Profiling accuracies on original test sets remained high, but were taken down in all cases except two (A^{4}NT and ParChoice-LSTM in the Alice→Bob direction, which increased by 2%). Accuracy on the transformed test set increased drastically with most techniques, and mostly undid the effect of style transfer.[20] The important exception was ParChoice-random, which we discuss below.

Our results illustrate a major problem in style transfer techniques that rely on the author having query access to the adversary's profiler, or a surrogate profiler that accurately approximates its performance. Crucially, such access assumptions go *both ways*: if the author can access/approximate the adversary's profiler locally, the adversary can also access/approximate the author's local profiler for adversarial training. This problem is fundamental to all of our techniques *except ParChoice-random*, which performs paraphrase selection independently of the profiler.

In striking contrast to other techniques, adversarial training with ParChoice-random reduced profiling accuracy on *both* original and transformed sentences. Hence, while ParChoice-random expectedly performed the least effectively in style transfer (Section 6.1.2), it was the only technique that could effectively resist adversarial training as a counter-measure. One possible reason for this is the large range of stylistic variants produced by ParChoice's paraphrase generation stage, which results in author-specific stylistic markers becoming less prominent in the adversarial training set.

Adversarial training is also expected to be more challenging when the number of classes is increased.

20 Apart from ParChoice-random (discussed separately), the only exception was A^{4}NT in the Bob→Alice direction, where profiling accuracy decreased by 6% in transformed sentences. Since A^{4}NT's original performance was not strong on AB to begin with, and original profiling accuracy also decreased in the Bob→Alice direction (by 9%), we do not take this exception to affect the overall conclusion.

Author	Original	Transformed
A1	$0.75 \rightarrow 0.73$	$0.27 \rightarrow 0.61$
A2	$0.81 \rightarrow 0.76$	$0.44 \rightarrow 0.71$
A3	$0.64 \rightarrow 0.60$	$0.26 \rightarrow 0.50$
A4	$0.71 \rightarrow 0.74$	$0.29 \rightarrow 0.66$
A5	$0.64 \rightarrow 0.72$	$0.35 \rightarrow 0.65$

Table 22. LSTM author profiler accuracy without→with adversarial training, on both original and transformed five-class test sets with randomized target authors on ParChoice-LSTM.

We produced ParChoice-LSTM imitations with random target classes from the five-class training set, and appended them to the original five-class training data. We trained the LSTM profiler on this adversarial training data, and tested it against ParChoice-LSTM imitations with randomly selected targets. Results are shown in Table 22. As predicted, the effectiveness of adversarial training was reduced, although still clearly present. However, the adversary could adopt a two-step profiling procedure of first detecting the imitation target by classifying the transformed document, and then producing the adversarial training set using this target.

On the document-level, the effect of adversarial training is difficult to evaluate due to the techniques' brittleness to re-training the profiler. While we were able to bring style transfer performance down on the EBG_5 dataset via adversarial training, this also happened with regular re-training (Table 20). As discussed in Section 6.2.2, we believe this is due to the high profiler-specificity of techniques relying on query access on the document-level, which remains an important problem for future research.

7 Conclusions and future work

We presented ParChoice: a novel style transfer technique based on combinatorial paraphrase generation and style-specific paraphrase selection. ParChoice considerably improves on the state-of-the-art in retaining semantic content. In sentence-level experiments, it outperformed the only encoder-decoder baseline technique with competitive semantic retainment (A^4NT). On the document-level, combining ParChoice with the best-performing rule-based baseline (Mutant-X) increased state-of-the-art performance in both style transfer and semantic retainment. We thus endorse ParChoice as the most viable style transfer technique overall. On data where ParChoice has limited coverage, we recommend combining it with word embedding -based replacement.

An important extension of the present work is applying style transfer to more complex profiling schemes. In particular, *abstaining classifiers* [13, 16] can be used to detect whether the probability of *any* target class is too low for prediction to be justified. They could be used to filter potential cases of style transfer, as these are less likely to give clear target class predictions. These datapoints could then be subjected to further scrutiny. Abstaining classifiers have been demonstrated to be beneficial for stylometry, and for detecting manually transformed documents [66]. Their effectiveness against automatic techniques remains to be studied.

Another major conclusion we draw is that style transfer needs to properly address possible countermeasures. The main challenge is achieving strong style transfer performance whilst preventing the adversary from replicating it by adversarial training. We demonstrated this to be a realistic concern, but propose that its effectiveness can partly be hindered via increased randomization in paraphrasing. We will continue to explore this issue in future work.

Acknowledgements

We thank Andrei Kazlouski and Sam Spilsbury for their help in implementation and running experiments, and Mika Juuti for valuable discussions related to the project. Tommi Gröndahl was funded by the Helsinki Doctoral Education Network in Information and Communications Technology (HICT).

References

[1] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information and System Security*, 26(2):1–29, 2008.

[2] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, pages 461–475, 2012.

[3] Mishari Almishari, Ekin Oguz, and Gene Tsudik. Fighting Authorship Linkability with Crowdsourcing. In *Proceedings of the second ACM conference on Online social networks*, pages 69–82, 2014.

[4] Douglas Bagnall. Author identification using multi-headed recurrent neural networks. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation Forum*, 2015.

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing, 2009.

[7] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security*, 15(3):12:1–12:22, 2011.

[8] Michael Brennan and Rachel Greenstadt. Practical Attacks Against Authorship Recognition Techniques. In *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence*, pages 60–65, 2009.

[9] Marcelo Luiz Brocardo, Issa Traore, Isaac Woungang, and Mohammad S. Obaidat. Authorship verification using deep belief network systems. *Communication Systems*, 30(12), 2017.

[10] Aylin Caliskan and Rachel Greenstadt. Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In *Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing (ICSC)*, pages 121–125, 2012.

[11] Andrew Carnie. *Constituent Structure*. Oxford University Press, Oxford, 2008.

[12] Daniel Castro-Castro, Reynier Ortega Bueno, and Rafael Mu noz. Author masking by sentence transformation – notebook for PAN at CLEF. In *Working notes of CLEF2017*, 2017.

[13] C.K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16:41–46, 1970.

[14] Siobahn Day, James Brown, Zachery Thomas, India Gregory, Lowell Bass, and Gerry Dozier. Adversarial Authorship, AuthorWebs, and entropy-based evolutionary clustering. In *Proceedings of the 25th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–6, 2016.

[15] Iqbal Farkhund, Hamad Binsalleeh, Benjamin C.M. Fung, and Mourad Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1–2):56–64, 2013.

[16] César Ferri, Peter Flach, and José Hernández-Orallo. Delegating classifiers. In *Proceedings of the 21th International Conference on Machine Leaning (ICML'04)*, pages 106–110, 2004.

[17] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 663–670, 2018.

[18] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 758–764, 2013.

[19] Zhenhao Ge and Yufang Sun. Domain specific author attribution based on feedforward neural network language models. In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 597–604, 2016.

[20] David Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, 1989.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[22] Jack Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007.

[23] Tommi Gröndahl and N. Asokan. EAT2seq: A generic framework for controlled sentence transformation without task-specific training. *arXiv preprint arXiv: 1902.09381*, 2019.

[24] Tommi Gröndahl and N. Asokan. Text analysis in adversarial settings: Does deception leave a stylistic trace? *ACM Computing Surveys*, 52(3):45:1–45:36, 2019.

[25] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISec'11)*, pages 2–12, 2018.

[26] Michael Haardt. *GNU Diction*, 2005 (accessed February 24, 2020). https://www.gnu.org/software/diction/.

[27] Thanh Nghia Ho and Wee Keong Ng. Application of stylometry to DarkWeb forum user identification. In *Proceedings of Information and Communications Security*, pages 173–183, 2016.

[28] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[29] Charles D. Hollingsworth. *Syntactic Stylometry: Using Sentence Structure for Authorship Attribution*. PhD thesis, University of Georgia, 2012.

[30] Charles D. Hollingsworth. Using dependency-cased annotations for authorship identification. In *Proceedings of the International Conference on Text, Speech and Dialogue*, pages 314–319, 2012.

[31] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1373–1378, 2015.

[32] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[33] Patrick Juola. Large-scale experiments in authorship attribution. *English Studies*, 93(3):275–283, 2012.

[34] Patrick Juola. Stylometry and immigration: A case study. *Journal of Law and Policy*, 21(2):287–298, 2013.

[35] Mika Juuti, Bo Sun, Tatsuya Mori, and N. Asokan. Stay on-topic: Generating context-specific fake restaurant reviews. In *Proceedings of the 23rd European Symposium on Research in Computer Security (ESORICS)*, pages 132–151, 2018.

[36] Georgi Karadzhov, Tsvetomila Mihaylova, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. The case for being average: A mediocrity approach to style masking and author obfuscation. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*, pages 173–185, 2017.

[37] Foaad Khosmood. Comparison of sentence-level paraphrasing approaches for statistical style transformation. In *Proceedings of the International Conference on Artificial Intelligence*, 2012.

[38] Foaad Khosmood and Robert Levinson. Automatic natural language style classification and transformation. In *Proceedings of the 2008 BCS-IRSG Conference on Corpus Profiling*, page 3, 2008.

[39] Foaad Khosmood and Robert Levinson. Toward automated stylistic transformation of natural language text. In *Proceedings of the Digital Humanities*, pages 177–181, 2009.

[40] Foaad Khosmood and Robert Levinson. Automatic synonym and phrase replacement show promise for style transformation. In *Proceedings of The Ninth International Conference on Machine Learning and Applications*, pages 958–961, 2010.

[41] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[42] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text againstreal-world applications. In *Proceedings of Network and Distributed Systems Security (NDSS)*, 2019.

[43] Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. Content preserving text generation with attribute controls. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2018.

[44] Nathan Mack, Jasmine Bowers, Henry Williams, Gerry Dozier, and Joseph Shelton. The best way to a strong defense is a strong offense: Mitigating deanonymization attacks via iterative language translation. *International Journal of Machine Learning and Computing*, 5(5):409–413, 2015.

[45] Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. A girl has no name: Automated authorship obfuscation using Mutant-X. In *Proceedings on Privacy Enhancing Technologies (PETS)*, pages 54–71, 2019.

[46] Muharram Mansoorizadeh, Taher Rahgooy, Mohammad Aminiyan, and Mahdy Eskandari. Author obfuscation using WordNet and language models – notebook for PAN at CLEF 2016. In *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, 2016.

[47] Andrew W.E. McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. Use fewer instances of the letter i: Toward writing style anonymization. In Simone Fischer-Hübner and Matthew Wright, editors, *Privacy Enhancing Technologies. Volume 7384 of Lecture Notes in Computer Science*, pages 299–318. 2012.

[48] Tsvetomila Mihaylova, Georgi Karadjov, Preslav Nakov, Yasen Kiprov, Georgi Georgiev, and Ivan Koychev. SU@PAN'2016: Author obfuscation – notebook for PAN at CLEF 2016. In *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, 2016.

[49] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, page 3111–3119, 2013.

[50] George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[51] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 300–314, 2012.

[52] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[53] Rebekah Overdorf and Rachel Greenstadt. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. In *Proceedings on Privacy Enhancing Technologies (PETS)*, pages 155–171, 2016.

[54] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, pages 2473—-479, 2012.

[55] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations,word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 425–430, 2015.

[56] Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efstathios Stamatatos, and Matthias Hagen. Who wrote the web? Revisiting influential author identification research applicable to information retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, pages 393–407. Springer International Publishing, 2016.

[57] Martin Potthast, Matthias Hagen, and Benno Stein. Author obfuscation: Attacking the state of the art in authorship verification. In *CLEF 2016 Working Notes*, 2016.

[58] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. Style Transfer Through Back-Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 866–876, 2018.

[59] Daniel Preotiuc-Pietro, Jordan Carpenter, and Lyle Ungar. Personality driven differences in paraphrase preference. In *Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science*, pages 17–26, 2017.

[60] Press Freedom Index. RSF Index 2018: Hatred of journalism threatens democracies. https://rsf.org/en/rsf-index-2018-hatred-journalism-threatens-democracies (May 1st 2018).

[61] Sravana Reddy and Kevin Knight. Obfuscating gender in social media writing. In *Proceedings of Workshop on Natural Language Processing and Computational Social Science*, pages 17–26, 2016.

[62] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[63] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2017.

[64] Rakshith Shetty, Bernt Schiele, and Mario Fritz. A^{4}NT: Author attribute anonymity by adversarial training of neural machine translation. In *Proceedings of the 27th USENIX Security Symposium*, pages 1633–1650, 2018.

[65] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[66] Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. Breaking the closed-world assumption in stylometric authorship attribution. In Gilbert Peterson and Sujeet Shenoi, editors, *Advances in Digital Forensics X*, pages 185–205, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[67] Kl Surendran, O.P. Harilal, Hrudya Poroli, Prabaharan Poornachandran, and N.K. Suchetha. Stylometry detection using deep learning. In *Proceedings of Computational Intelligence in Data Mining*, pages 749–757, 2017.

[68] Neal Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6), 2017.

[69] Louis Tesnière. *Èléments de syntaxe structurale*. Klincksieck, Paris, 1959.

[70] Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 979–988, 2018.

[71] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In *Proceedings of the International Conference on Neural Information Processing (NIPS)*, pages 7287–7298, 2018.

[72] Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. Refining word embeddings using intensity scores for sentiment analysis. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 671–681, 2018.

[73] Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5902–5911, 2018.

[74] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework of authorship identification for online messages: Writing style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006.

A Raw results

In this appendix we include the raw results from our sentence-based experiments, as well as example transformations by ParChoice and all baseline techniques.

Two-class experiments: Table 23 displays two-class results for the original test sets and each style transfer direction on every style transfer technique, measured with all three author profilers (LSTM, CNN, WP). Table 24 shows corresponding results for each ParChoice module applied separately, using the ParChoice-LSTM variant and the LSTM profiler.

Five-class experiments: Imitation results for each of the 20 author pairs from the five-class experiments are collected in Table 25. Successful target author imitation results are written in bold (higher target author accuracy than source author accuracy).

Example transformations: Table 26 shows example transformations by ParChoice, the encoder-decoder baselines (rows $1-20$) and the rule-based baselines (rows $21-25$).

Data	Direction	Author profiler	Profiler accuracy for source author									
							ParChoice					
			Original	CAE	BT	A^{4}NT	CNN	LSTM	WP	LR$_d$	LR$_s$	random
YG	f→m	LSTM	0.83	0.55	0.54	0.64	0.41	0.35	0.54	0.43	0.46	0.69
		CNN	0.75	0.44	0.43	0.57	0.21	0.28	0.46	0.28	0.32	0.57
		WP	0.71	0.50	0.55	0.57	0.49	0.44	0.21	0.54	0.54	0.60
	m→f	LSTM	0.82	0.48	0.45	0.70	0.46	0.43	0.59	0.45	0.48	0.72
		CNN	0.88	0.59	0.53	0.73	0.45	0.56	0.71	0.56	0.59	0.82
		WP	0.78	0.56	0.54	0.72	0.61	0.60	0.32	0.64	0.65	0.74
BA	a→t	LSTM	0.54	0.26	0.28	0.35	0.38	0.16	0.37	0.26	0.33	0.52
		CNN	0.57	0.38	0.32	0.40	0.19	0.25	0.35	0.24	0.33	0.49
		WP	0.45	0.35	0.40	0.38	0.42	0.36	0.16	0.40	0.41	0.48
	t→a	LSTM	0.70	0.69	0.66	0.69	0.45	0.34	0.55	0.42	0.48	0.65
		CNN	0.68	0.56	0.60	0.64	0.36	0.46	0.56	0.44	0.51	0.66
		WP	0.73	0.58	0.61	0.70	0.61	0.59	0.35	0.57	0.59	0.68
AB	A→B	LSTM	0.88	0.18	0.18	0.74	0.54	0.36	0.63	0.56	0.62	0.74
		CNN	0.92	0.22	0.16	0.80	0.46	0.50	0.58	0.56	0.64	0.76
		WP	0.73	0.45	0.26	0.67	0.50	0.30	0.12	0.37	0.45	0.60
	B→A	LSTM	0.94	0.16	0.11	0.70	0.76	0.52	0.78	0.75	0.78	0.89
		CNN	0.93	0.10	0.01	0.61	0.62	0.72	0.84	0.80	0.78	0.87
		WP	0.90	0.09	0.19	0.68	0.81	0.68	0.31	0.68	0.78	0.85
TO	T→O	LSTM	0.86	0.95	0.64	0.34	0.76	0.43	0.76	0.59	0.66	0.81
		CNN	0.60	0.88	0.60	0.46	0.36	0.52	0.57	0.53	0.55	0.58
		WP	0.75	0.61	0.85	0.46	0.63	0.58	0.14	0.58	0.60	0.67
	O→T	LSTM	0.77	0.00	0.01	0.23	0.61	0.27	0.58	0.40	0.51	0.65
		CNN	0.67	0.00	0.01	0.68	0.50	0.57	0.58	0.57	0.58	0.63
		WP	0.73	0.31	0.09	0.36	0.62	0.57	0.16	0.56	0.61	0.64

Table 23. Author profiling accuracies in two-class sentence-based datasets: best (lowest) results framed.

Technique	Direction	YG	BA	AB	TO
Original (no transformations)	$\mathbf{0 \to 1}$	0.83	0.54	0.88	0.86
	$\mathbf{1 \to 0}$	0.82	0.70	0.94	0.77
Grammatical transformations	$\mathbf{0 \to 1}$	0.72	0.48	0.69	0.80
	$\mathbf{1 \to 0}$	0.70	0.62	0.85	0.68
Simple rules	$\mathbf{0 \to 1}$	0.81	0.51	0.85	0.82
	$\mathbf{1 \to 0}$	0.82	0.66	0.91	0.72
PPDB	$\mathbf{0 \to 1}$	0.65	0.37	0.73	0.64
	$\mathbf{1 \to 0}$	0.62	0.51	0.84	0.50
WordNet	$\mathbf{0 \to 1}$	0.65	0.36	0.77	0.65
	$\mathbf{1 \to 0}$	0.70	0.48	0.80	0.52
Typos	$\mathbf{0 \to 1}$	0.79	0.39	0.79	0.82
	$\mathbf{1 \to 0}$	0.82	0.65	0.74	0.65

Table 24. Author profiling accuracies with individual ParChoice modules (LSTM profiler, ParChoice-LSTM variant). Class 0: {female, adult, Alice, Trump}; Class 1: {male, teen, Bob, Obama}

Source author	Technique	Author profiler	Target author; profiler accuracies with imitated test sets									
			A1		A2		A3		A4		A5	
			s	t	s	t	s	t	s	t	s	t
A1	ParChoice-LSTM	LSTM	(0.75)	–	0.27	**0.51**	0.28	**0.35**	0.26	**0.32**	0.27	**0.30**
		CNN	(0.82)	–	0.51	0.28	0.52	0.25	0.45	0.26	0.52	0.21
	ParChoice-CNN	LSTM	(0.75)	–	0.43	0.32	0.43	0.22	0.42	0.14	0.43	0.14
		CNN	(0.82)	–	0.37	**0.39**	0.37	0.35	0.38	0.19	0.37	0.33
A2	ParChoice-LSTM	LSTM	0.42	0.40	(0.81)	–	0.47	0.21	0.41	0.31	0.44	0.15
		CNN	0.53	0.33	(0.82)	–	0.59	0.13	0.48	0.28	0.59	0.12
	ParChoice-CNN	LSTM	0.61	0.18	(0.81)	–	0.62	0.12	0.61	0.13	0.62	0.07
		CNN	0.46	0.29	(0.82)	–	0.46	0.26	0.46	0.19	0.46	0.25
A3	ParChoice-LSTM	LSTM	0.25	**0.57**	0.24	**0.51**	(0.64)	–	0.24	**0.37**	0.25	**0.27**
		CNN	0.30	**0.48**	0.39	0.30	(0.76)	–	0.36	0.32	0.38	0.24
	ParChoice-CNN	LSTM	0.38	0.30	0.36	0.28	(0.64)	–	0.43	0.15	0.43	0.14
		CNN	0.28	**0.43**	0.29	**0.42**	(0.76)	–	0.30	0.27	0.28	**0.41**
A4	ParChoice-LSTM	LSTM	0.30	**0.47**	0.28	**0.44**	0.30	0.29	(0.71)	–	0.28	**0.32**
		CNN	0.50	0.35	0.52	0.23	0.53	0.11	(0.79)	–	0.46	0.24
	ParChoice-CNN	LSTM	0.55	0.18	0.52	0.23	0.57	0.16	(0.71)	–	0.55	0.12
		CNN	0.31	0.29	0.31	**0.34**	0.31	**0.33**	(0.79)	–	0.31	**0.42**
A5	ParChoice-LSTM	LSTM	0.36	**0.42**	0.35	**0.36**	0.35	**0.36**	0.33	**0.41**	(0.64)	–
		CNN	0.57	0.30	0.63	0.15	0.71	0.12	0.56	0.26	(0.75)	–
	ParChoice-CNN	LSTM	0.52	0.21	0.51	0.20	0.51	0.14	0.51	0.14	(0.64)	–
		CNN	0.45	0.34	0.48	0.27	0.45	0.27	0.46	0.21	(0.75)	–

Table 25. Five-class author imitation results (s = source author accuracy, t = target author accuracy). Successful imitation in bold (target author accuracy > source author accuracy).

YG (female→male)	Original	the dinner portions are huge .
	CAE	the drinks are $ 00 .
	BT	the rooms are great .
	A^{4}NT	the dinner portions are ultra .
	ParChoice	the supper shares are tremendous .
BA (adult→teen)	Original	you feel like killing them but then again they are protected .
	CAE	you feel like then you are them .
	BT	eddy you want to see them , but now they are protégés .
	A^{4}NT	you feel like killing them but then again they are .
	ParChoice	you feel like popping them but then again theyre been safeguarded .
AB (Alice→Bob)	Original	we are so useless when it comes to bugs- its ridiculous !
	CAE	so yeah we went to <unk> it 's <unk> ... .
	BT	i was fattoria nous , , and de ... ..
	A^{4}NT	we are so far when it comes to me !
	ParChoice	were so ineffectual when it is about microbes its farcical .
TO (Obama→Trump)	Original	i can tell you .
	CAE	we 're going .
	BT	i 's n't .
	A^{4}NT	i can you disagree
	ParChoice	you might well be told by me .
EBG$_5$	Original	They also tend to be somewhat adapted to fire of varying frequencies.
	PAN2016	They also tended to be somewhat tailor to fire of varying frequencies.
	Mutant-X	They also tend to be somewhat adapted to fireball of aforementioned frequencies.
	ParChoice	they also tend to be somewhat adjusted to a fire of differing frequencies .
	ParChoice+Mutant-X	they also tend to continue somewhat adjusted to a fire of differing frequencies .

Table 26. Style transfer examples (ParChoice variant in YG/BA/AB/TO: ParChoice-LR$_s$).