UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Elizaveta Korotkova

Style Transfer via Zero-Shot Monolingual Neural Machine Translation

Master's Thesis (30 ECTS)

Supervisor: Mark Fishel, PhD

Tartu 2019

Style Transfer via Zero-Shot Monolingual Neural Machine Translation

Abstract: Text style transfer is the task of altering stylistic features of a text while preserving its meaning and fluency. It can be viewed as a sequence-to-sequence transformation task, but the scarcity of directly annotated parallel data makes it unfeasible for most settings. We propose an approach to style transfer that builds on the idea of zero-shot machine translation. It performs style transfer within a neural machine translation model, without requiring any parallel style-adapted texts, relying instead only on regular language-parallel data. The method is applicable to multiple languages within a single model. We outline the method, describe our experiments with it, and, finally, present a thorough automatic and manual evaluation of the approach in comparison to a baseline and independently, showing that our zero-shot model outperforms the supervised baseline on several aspects according to human judgments, and is reliable for a number of style transfer aspects.

Keywords: natural language processing, style transfer, neural machine translation, zero-shot machine translation

CERCS: P176 Artificial intelligence

Teksti stiiliülekanne ühekeelse masintõlke abil

Lühikokkuvõte: Teksti stiiliülekande ülesanne on muuta teksti stiilseid omadusi, säilitades samal ajal selle tähenduse ja soravuse. Seda saab vaadata järjendite teisenduse ülesandena, kuid otseselt märgendatud paralleelandmete nappus muudab selle enamiku juhtumite jaoks võimatuks. Käesolevas töös me kirjeldame stiiliülekande lähenemise, mis põhineb zero-shot masintõlke ideel. Meie pakutud mudel sooritab stiiliülekande neuromasintõlke abil, ilma nõudmata paralleelseid stiili-kohandatud tekste, toetudes hoopis regulaarse keele paralleelsetele andmetele. Meetodit saab rakendada mitmele keelele ühes ja samas treenitud mudelis. Meie kirjeldame pakutud metodoloogiat, oma katseid sellega ning esitame põhjaliku automaatse ja käsitsi teostatava hindamise, kus võrdleme seda juhendatud masinõppe põhise baasmudeliga. Me näitame, et meie mudeli tulemuste hinnangud ületavad juhendatud baaslähenemist mitmes aspektis vastavalt inimeste arvamustele ning on usaldusväärsed mitme stiiliülekande aspekti suhtes.

Võtmesõnad: loomuliku keele töötlus, teksti stiiliülekanne, neuromasintõlge

CERCS: P176 Tehisintellekt

Contents

1 Introduction **4**

2 Related Work **6**
2.1 Style Transfer with Parallel Data 6
2.2 Style Transfer with Non-Parallel Data 7

3 Methodology **10**
3.1 Machine Translation 10
3.2 Transferring Style via Zero-Shot Monolingual Translation 12
3.3 Datasets 13
3.4 Evaluation of Style Transfer 14

4 Experiments **15**
4.1 Data and Languages 15
4.2 Technical Details 15
4.2.1 Data Preprocessing 15
4.2.2 Model Hyperparameters 18
4.3 Performance as a Machine Translation System 18
4.4 Evaluation 19
4.4.1 Automatic Evaluation 21
4.4.2 Manual Evaluation 23
4.5 Cross-lingual Style Transfer 29

5 Conclusion **31**

References **36**

Appendix **37**
II. Licence 38

1 Introduction

Style transfer is the task of modifying a text to demonstrate features of a certain style while maintaining the meaning of the original text. The same semantic content or intent may be phrased in many different ways. For example, the sentences *Hi, guys* and *Hello, ladies and gentlemen* may both be used as a greeting, but are suitable for different settings. The notion of "style" can be understood very broadly, and can range from idiolects of individual authors to aspects such as sentiment or political views of the writer.

It is important for natural language generation systems to produce output that is understandable to the user and suitable for the specific situation at hand. A complex text may need to be simplified (for example, compare Wikipedia articles in English to articles in Simple English, the latter better suited for readers with poorer command of the language). An informally written text may need to be transformed into a more formal and polite one. Modifying stylistic attributes of large volumes of text manually is costly and tedious, and, in most situations, not feasible. Thus, there are attempts to perform the task automatically, making use of the recent advancements in deep learning techniques.

Supervised deep neural networks need massive amounts of parallel training data. With style transfer, however, this requirement becomes a problem: there are extremely few resources where one can find pairs of texts with the same semantic content but rendered in different styles. Most of the recent research in style transfer is focused on finding ways to leverage non-parallel data to train neural style transfer systems.

This thesis follows that line of work. We propose a simple approach for style transfer, based on the idea of zero-shot machine translation [Johnson et al., 2017], but applied *monolingually*. The method does not require any style-parallel corpora, instead relying on multilingual corpora with distinct style characteristics. It also does not require any highly specialized neural network architectures, being implemented using an open-source neural machine translation framework.

Research questions we answer include:

1. Is it possible to use only language-parallel instead of style-parallel corpora to perform style transfer?
2. How does such a model compare to a supervised model on the same task?
3. Which aspects of style does the model learn to transfer well?
4. Do the results depend on the language of the text?

The main contributions of this work are the following:

1. an easily implementable method for text style transfer, which does not require any parallel data between styles,

2. support of multiple languages within the same model,

3. a thorough automatic and manual evaluation of the model's performance.

Section 2 presents a review of related work on style transfer approaches. Details of our proposed approach are outlined in Section 3. Details of experiment setup, model specifics and evaluation results can be found in Section 4. The thesis ends with conclusions in Section 5.

2 Related Work

In the following chapter, we briefly outline the dominating approaches to style transfer encountered in recent literature.

An increasing amount of research is being performed in the area of text style transfer using neural networks, having been largely inspired at first by the massive body of work on image style transfer in the computer vision community, most notably [Gatys et al., 2016].

2.1 Style Transfer with Parallel Data

One possible approach to text style transfer is to use directly annotated, style-parallel data, such that a single training example consists of two sentences, each conveying the same semantic content, but in different styles.

An example of this approach is the paper by Jhamtani et al. [Jhamtani et al., 2017], in which authors attempt to automatically transform text from modern English to English in the style of Shakespeare. They use a publicly available dataset compiled by Xu et al. [Xu et al., 2012], which consists of line-by-line paraphrases of 16 plays by Shakespeare into modern English, gathered from the educational site Sparknotes[1]. 18,395 sentences were used for training, with 1218 and 1462 sentences in the validation and test splits, respectively. A sequence to sequence neural model is used for the transformation, with a bidirectional LSTM encoder, LSTM decoder and attention. The model is aided by a pointer network component which enables direct copying of words from input into output sentences, and by modifying pre-trained word embeddings using a dictionary which maps between words used by Shakespeare and modern English words.

Carlson et al. [Carlson et al., 2017] utilize 34 different versions of the Bible in a similar way. Each version is treated as having a distinct writing style: some are written in simplified English, some demonstrate a purposefully archaic style, etc. The corpus is parallel, as it is straightforward to align the different versions by verses. Eight versions out of 34 are made publicly available. The public versions amount to 1.7 million potential source and target verse pairs (that is, when any source version can be paired with any target version), and the full corpus to over 33 million pairs. Authors train two models: a Moses [Koehn et al., 2007] statistical machine translation model to serve as a baseline, and a sequence to sequence model with multi-layer RNN encoder and decoder with attention.

Rao and Tetreault [Rao and Tetreault, 2018] also make use of parallel corpora, focusing on formality transfer and taking another approach to obtaining data. They present the Grammarly's Yahoo Answers Formality Corpus. It is created on the basis of the Yahoo

[1] www.sparknotes.com

Answers L6 corpus[2], which contains textual data gathered from the question answering forum Yahoo Answers[3]. The formality corpus consists of 115,000 informal sentences, half of those sampled from the Entertainment & Music domain of the L6 corpus, and the other half from the Family & Relationships domain. For each of those sentences, a formal rewrite is manually created by workers on Amazon Mechanical Turk. Analysis of these manual rewrites shows that when formalizing text, workers often choose to paraphrase, alter punctuation and expand contractions. Several models are trained on the obtained parallel corpora: rule-based, phrase-based, and several modifications of neural machine translation systems. We use one of these models as a baseline to which to compare our system's performance, as it employs an approach similar to ours, albeit fully supervised, and is applied to a comparable task.

Parallel data for style transfer are scarce. The existing approaches to compiling style-parallel corpora are not scalable: the datasets are based on very limited and highly specialized sources, such as Shakespeare's plays or the Bible. Building datasets manually specifically for the purpose of style transfer research is also costly, and it can be further argued that such manually constructed rewrites may not fully reflect characteristics of spontaneously generated text. Thus, this approach can hardly be applied outside of carefully tailored domains.

2.2 Style Transfer with Non-Parallel Data

Due to style-parallel data being hard to obtain, especially if a massive amount of it is needed for training a neural model, another major line of work on style transfer is concerned with training such models using no parallel data at all, but only non-parallel corpora of texts belonging to different styles.

In this case, a key point is often disentangling a latent representation of text content from its style characteristics, while building decoders that can generate text in style of choice based on the content encoding. Adversarial training can be used to separate style attributes from content ([Zhao et al., 2017], [Shen et al., 2017], [Fu et al., 2017]).

Shen et al. [Shen et al., 2017] build an encoder that attempts to dispose of style information and a style-dependent decoder to reproduce the encoded content in a different style, and use a discriminator aligning transferred sentences to match sentences of the target style as a population. They assume that the two corpora of different styles have the same distribution of content.

Fu et al. [Fu et al., 2017] also experiment with adversarial training, and explore multi-task learning. Their two approaches are a multi-decoder model, where the encoder captures content with several decoders generating text in different styles further on, and introducing style embeddings, learned jointly with the model, with which to augment the

[2] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[3] https://answers.yahoo.com/answer

encoded text representations so that only a single decoder is necessary.

A more transparent approach of Li et al. [Li et al., 2018] makes use of the fact that style attributes are often concentrated only in certain distinctive phrases of a sentence. Their method extracts style-independent content by deleting words and phrases which are associated with the original style, retrieves other words from similar sentences which are associated with the target style, and uses a neural model to combine these two into a fluent output sentence.

Prabhumoye et al. [Prabhumoye et al., 2018] use machine translation as a first step in their style transfer system. Their assumption is that translating a sentence into another language and then translating it back will reduce original style characteristics of the text, yielding a more neutrally rephrased version of it. The pipeline consists of back-translating a sentence, encoding the resulting sentence (assuming it has lost its style attributes to some extent), and then using separate decoders to produce output sequences belonging to different styles.

Subramanian et al. [Subramanian et al., 2018] argue that learning disentangled representations of content and style is not necessary and not realistic. They train what is essentially an unsupervised machine translation model and use it to perform back-translation into the desired style, attempting to "overwrite" the original style characteristics with the target ones without separating style from content.

Zhang et al. [Zhang et al., 2018] also tackle the lack of parallel data by treating style transfer as an unsupervised machine translation task and implementing back-translation. Two style-specific language models and word-to-word transfer information are used to build two statistical machine translation models, with which a pseudo-parallel corpus is generated. This corpus is used to train NMT-based style transfer systems.

The non-parallel data approaches are typically evaluated on the task of transferring sentiment (e.g. by [Shen et al., 2017], [Prabhumoye et al., 2018] [Li et al., 2018], [Zhang et al., 2018], [Subramanian et al., 2018]), less commonly transferring between romantic and humorous image captions (e.g. [Li et al., 2018], [Zhang et al., 2018]), the author's gender (e.g. [Subramanian et al., 2018], [Prabhumoye et al., 2018]), their age (e.g. [Subramanian et al., 2018]), or even such attributes as political affiliation ([Prabhumoye et al., 2018]).

A drawback of many style transfer methods which do not require parallel data is their additional constraints, such as the assumption that the content distribution of different corpora is the same. In practice, in some tasks, for instance, formality transfer, it is likely that texts belonging to different styles will be generated under somewhat differing circumstances, and touch upon different topics. Disentangling content representation from style indicators, as pointed out by some researchers, is not always feasible, and many style transfer systems tend to sacrifice content for style. Our method, on the other hand, does not attempt to learn content and style representations separately, having a shared encoder and decoder for all styles and languages, and is strongly incentivized

to reproduce the original meaning, since the task it learns to perform directly and in a supervised manner is translation.

3 Methodology

In this chapter, we describe the methods we employ to perform zero-shot style transfer. Section 3.1 outlines the modern neural machine translation techniques we use. Section 3.2 details how we apply those to style transfer. In Section 3.3 we specify the kind of data needed for our method, and, finally, Section 3.4 describes our evaluation methods.

3.1 Machine Translation

Machine translation is a sequence to sequence mapping task. Using deep neural networks to perform this task has become the norm in the field, as neural models [Sutskever et al., 2014] surpassed the previously existing statistical machine translation approaches soon after their introduction [Bojar et al., 2016].

For sequence to sequence transformations, the elements of a discrete high-dimensional input sequence are typically mapped into a space of continuous low-dimensional representations [Bengio et al., 2003], or word embeddings. These are further encoded into a vector by some kind of neural network layers, such as recurrent (most typically LSTM) or convolutional. This encoded representation is then used by a decoder to produce an output sequence.

The limitations of the fixed-length vector encoding have been partially overcome with introduction of the attention mechanism [Bahdanau et al., 2015], which learns to determine on which elements of the input sequence should the output elements be conditioned, and to what extent. Currently, the leading architectural approach is the Transformer [Vaswani et al., 2017], which disposes of the recurrent or convolutional layers completely, and relies solely on this attention mechanism in encoder and/or decoder. The architecture includes several similarly constructed layers, and having several attention heads with different learned weights allows the network to learn several complementary attention patterns. For an illustration of the Transformer architecture, see Figure 1.

Apart from intrinsic architectural choices, there exist other additional modifications that influence the quality of machine translation. One such modification that we employ is using word factors [Sennrich and Haddow, 2016]. Factors are labels appended to each input token, which provide the encoder with additional information, for example, part of speech tags. We use them to propagate information about the desired output language and style to the decoder.

Domain adaptation remains one of the challenges of neural machine translation [Koehn, 2017]. Notably in the context of our work, Sennrich et al. use side constrains in order to translate into polite or impolite German [Sennrich et al., 2016] (which can be treated as cross-lingual formality transfer), while we rely on multilingual encoder

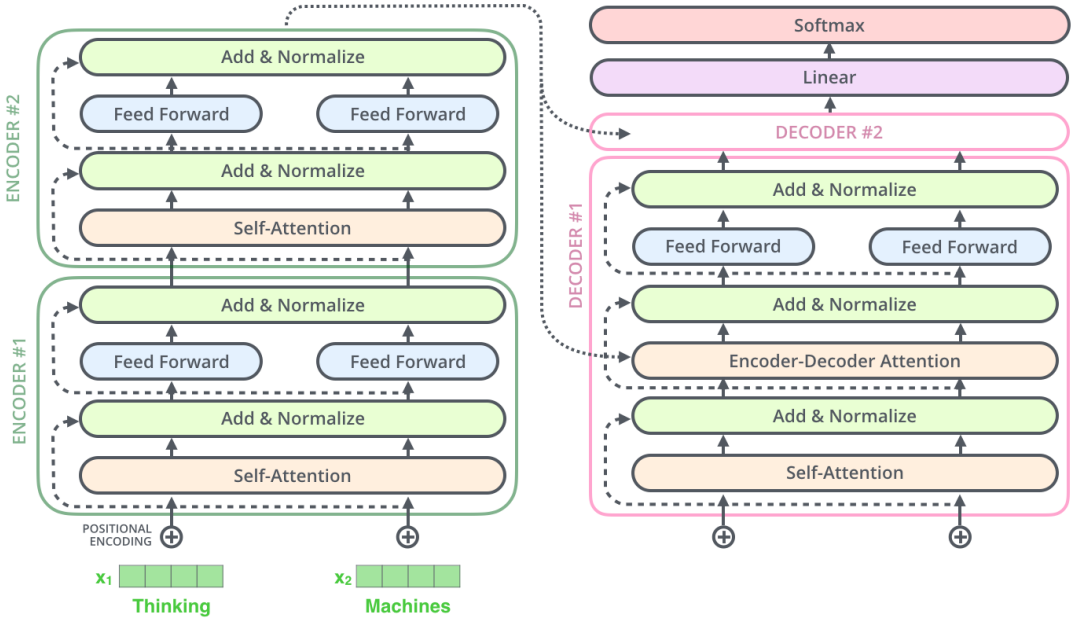


Figure 1. A schematic illustration of the Transformer architecture by Jay Alammar[4]. The figure shows a Transformer model with 2 stacked encoders and 2 decoders. The stacked encoder and decoder layers have the same structure, but learn different weights. Self-attention allows to take into account other words of the sequence when encoding a certain word. Multiple parallel attention heads (not shown) allow to learn multiple attention patterns. Stacking encoder and decoder layers makes it possible to learn representations on a high level of abstraction.

representations and use the system monolingually at inference time.

3.2 Transferring Style via Zero-Shot Monolingual Translation

The proposed approach is based on the idea of zero-shot machine translation presented by Johnson et al. [Johnson et al., 2017]. They show that after training a single model to translate from Portuguese to English as well as from English to Spanish, it can also translate Portuguese into Spanish, without seeing any translation examples for this language pair.

We use the same zero-shot effect to achieve *monolingual translation*. The model is trained to translate on bilingual examples in both directions, and is then used to perform translation into the same language as the input (see illustration of data flow in Figure 2).

With regular sentences monolingual translation does not seem useful, as its behaviour mainly consists of copying. However, what is added to the model is the notion of style. Assuming we have several language-parallel corpora, each having its distinct style characteristics (e.g. formal vs. informal texts, positive vs. negative sentiment, etc.), we present the model with such training examples that source and target sentences are always written in *different languages*, but are of the *same style*. At inference time, we attempt to use the same model to translate inside *one language*, but *between different styles*.

Apart from zero-shot multilingual translation, this approach also relates to multilingual multi-domain NMT [Tars and Fishel, 2018]: it is possible to switch between different domains or styles at runtime, and when the source and target language are the same, the model is thus performing "monolingual domain adaptation" or style transfer.

To create a multilingual multi-domain NMT system we use the self-attention Transformer architecture [Vaswani et al., 2017]. Instead of specifying the output language with a token inside the input sequence, as Johnson et al. [Johnson et al., 2017] did, we follow Tars and Fishel [Tars and Fishel, 2018] and use word features (or *factors*), appended to each token of the input. On one hand, this provides a stronger signal for the model, and on the other – allows for additional parametrization, which in our case is the text domain/style of the corpus, in addition to the target language.

As a result, a pre-processed English-Latvian training set sentence pair *"Hello!"*–*"Sveiki!"* looks like:

En: `hello|2lv|2os !|2lv|2os`
Lv: `sveiki !`

Here `2lv` and `2os` specify Latvian and OpenSubtitles as the output language and domain; the output text has no factors to predict. At application time we simply use the same

[4] `http://jalammar.github.io/illustrated-transformer/`

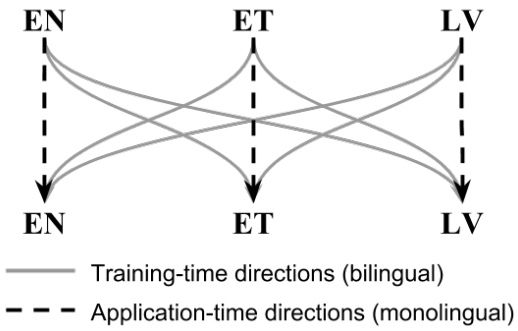


Figure 2. Schematic illustration of zero-shot monolingual translation. The model is trained on bilingual data in all translation directions (English-to-Estonian, Estonian-to-English, English-to-Latvian, etc.) and then applied in monolingual directions only (English-to-English, etc.), without having seen any sentence pairs for them. The illustration is simplified, as it does not show the style (text domain) parametrization.

input and output languages, for example the input *"we are"* looks like the following, after pre-processing:

En: `we|2en|2os are|2en|2os`

The intuition behind our approach is that a multilingual shared encoder produces semantically rich latent sentence representations [Artetxe and Schwenk, 2018], which provide a solid ground for the effective style transfer on top.

Note that such a model will be, indeed, performing zero-shot style transfer, as it has encountered neither examples of monolingual translation (which, by itself, should be quite a simple task, as it could well constitute simple copying of input), nor examples of style transfer, since we simply do not have those, and each pair of training sentences comes from one corpus (or style) only.

3.3 Datasets

For training a style transfer system of the proposed architecture, several corpora are needed, which have to:

1. be *language-parallel* (that is, a single example is a pair of texts which are translations of each other),
2. differ significantly in their distinct style characteristics,

3. include many examples (preferably hundreds of thousands) so as to make possible training of a strong neural machine translation system.

In contrast to obtaining massive amounts of *style-parallel* data, these requirements are quite easily satisfied by several existing and publicly available datasets (see, for example, resources of the OPUS project[5]), typically used for building machine translation systems. These parallel corpora are obtained from a wide range of sources, such as translations of legal documents and proceeding records, software package localization files, corresponding Wikipedia articles in different languages, movie subtitles, or crawled from websites. This variety of settings in which the texts were produced entails significant difference in their styles. It has to be taken into account, however, that these corpora differ also in their quality (the most common issue is poor alignment).

3.4 Evaluation of Style Transfer

Proper methodology for evaluation of style transfer in text is currently largely an open question. There are no firmly established automatic metrics, although recently there have been attempts to devise such metrics [Mir et al., 2019]. Thus, it is typical in style transfer research to rely on both on automatic metrics and on human judgments.

We follow this trend and evaluate our system both automatically and manually. We compare our method's performance to that of a benchmark model applied to a similar task. To quantitatively evaluate style transfer strength automatically, we follow several existing works ([Shen et al., 2017], [Li et al., 2018] and others) and use a convolutional neural network classifier of the architecture proposed by Kim [Kim, 2014], which we train to distinguish between styles. The proportion of texts in a test set which are classified into the target style is treated as a measure of style transfer strength.

However, such automatic evaluation fails to take into account several other important aspects of output quality. It is not only important to alter stylistic features of a text, but also to preserve the meaning of the original sentence and to produce readable and fluent output. To evaluate our system on these aspects, we turn to help of human annotators. We ask them to grade how fluent the outputs are, how well they preserve the meaning of the original texts, whether they transfer style in the desired direction, if at all, what types of changes there are compared to the source texts, and which outputs, ours or the baseline they generally prefer.

Next we present the technical details, the experiment setup and the data we used for training the model used in the experiments.

[5] `http://opus.nlpl.eu/`

4 Experiments

This section describes the experimental choices we made with regards to data and model specifics, and presents an automatic and manual evaluation of the model's performance.

4.1 Data and Languages

Based on a manual estimation of corpora quality and the need for stylistically distinct and sufficiently large corpora, four datasets were chosen for the experiments: Europarl [Koehn, 2005], JRC-Acquis and EMEA [Tiedemann, 2012], and OpenSubtitles2018 [Lison and Tiedemann, 2016].

The Europarl parallel corpus is extracted from the proceedings of the European Parliament, and includes versions in 21 European languages. JRC-Acquis is a collection of legislative text of the European Union in 22 languages, written between 1950s and the present time. EMEA is comprised of documents from the European Medicines Agency and exists in 22 languages. The biggest dataset, OpenSubtitles2018, is a collection of film and television subtitles in 62 languages.

The assumption is that the Europarl and JRC-Acquis corpora should contain very formal, polite and precise language, EMEA even drier language of pharmaceutical documentation, and OpenSubtitles2018 more diverse, informal and emotional style. Table 1 shows several examples of sentences typical for these four corpora. It should be noted that OpenSubtitles2018 is very heterogenous, and can itself be split into subgenres, such as comedy, drama, documentary, etc., which will also demonstrate somewhat differing writing styles. However, many stylistic traits are similar across the whole corpus, which means that these common traits can be treated as a single style.

Three languages are used in our experiments: English, Estonian and Latvian. All three have different characteristics: Latvian and (especially) Estonian are morphologically complex and have relatively loose word order, while English has a stricter word order and is more analytical.

4.2 Technical Details

4.2.1 Data Preprocessing

The data preprocessing pipeline consists of the following steps:

1. Cleaning. Sentence pairs are discarded if:
    - at least one sentence is longer that 100 tokens,
    - at least one sentence is an empty string,
    - at least one sentence does not contain any alphabetic characters,

Table 1. Examples of English sentences typical to Europarl, JRC-Acquis, EMEA and OpenSubtitles2018 corpora

Europarl:	• Some are open, others are not yet open, other negotiations are stalled, and it is plain that Turkey is of great strategic importance to the Union. • Hence I conclude, Madam President. • (The sitting was closed at 11.55 p.m.) • But the recent setbacks in the democratic opening are, indeed, reason for serious concern.
JRC-Acquis:	• The Joint Committee shall be notified beforehand of any changes. • The Annex and Protocols to the Agreement shall form an integral part thereof • The Member State may decide that participation by flock owners in the breeding programme referred to in paragraph 1 shall be voluntary until 1 April 2005. • The product group shall not include cordless or battery operated vacuum cleaners and central vacuum cleaning systems.
EMEA:	• The replacement insulin acts in same way as naturally produced insulin and helps glucose enter cells from the blood. • In both studies, the main measure of effectiveness was the number of children who had to be admitted to hospital because of RSV infection. • Each film-coated tablet contains 0.17 mg soya-lecithin. • No data on the interaction of lacosamide with alcohol are available.
OpenSubtitles2018:	• The vital organs are not affected. Show me his wounds! • Yeah, 'cause there's plenty of this. • I didn't come out of my room for days. • - This is gonna get me that scholarship. • Dude, you're in the wrong uniform.

Table 2. Training set sizes (number of sentence pairs)

	EN ↔ ET	EN ↔ LV	ET ↔ LV
Europarl	0.64M	0.63M	0.63M
JRC-Acquis	0.68M	0.69M	1.5M
EMEA	0.91M	0.91M	0.92M
OpenSubtitles2018	3M	0.52M	0.41M

- the number of tokens in one of the sentences is at least 9 times larger than in the other sentence.

2. Tokenization with Moses tokenizer [Koehn et al., 2007]. During tokenization, punctuation marks, as well as contractions, are separated from words by spaces, so that neighboring words and punctuation marks may later be recognized as separate units. (E.g. `"Darling, it's ridiculous!"` → `"Darling , it 's ridiculous !"`.)

3. True-casing. The true-caser[6] attempts to determine the proper capitalization of words based on frequency of their occurrence with different capitalization, and replaces all occurrences of the token with the same token in its right capitalization form. This step is necessary so that the same word will not appear in the vocabulary as two separate instances (capitalized and uncapitalized). (E.g. `"Hello , I am Gary"` → `"hello , I am Gary"`.) The true-casing model was trained jointly on training data for all languages.

4. Subword segmentation. Tokens are separated into smaller units based on their frequency, so that the more frequent tokens will remain unchanged, and the less common ones will be split into several parts. The goal of text segmentation is to reduce the number of unseen words, i.e words that do not occur in the training data: a previously unseen word may be constructed of known subwords. For segmentation, SentencePiece [Kudo and Richardson, 2018] was used, with a joint model for all languages and 32,000 subword units learned. Initially existing whitespaces are marked with special meta symbols, so that the text can be later restored without ambiguities. (E.g. `"this is figurative"` → `"_this _is _fi gur ative"`.)

For Europarl, JRC-Acquis and EMEA all data available for English-Estonian, English-Latvian and Estonian-Latvian language pairs were used. From OpenSubtitles2018 we take a random subset of 3M (out of over 12M) sentence pairs for English-Estonian. This

[6] https://github.com/TartuNLP/truecaser

is done to balance the corpora representation and to limit the size of training data. Sizes of training corpora after cleaning are shown in Table 2. After duplication to accommodate both translation directions in each language pair, the total size of the training corpus is 22.9M sentence pairs. Validation set consists of 12K sentence pairs, 500 for each combination of translation direction and corpus. We also keep a test set of 24K sentence pairs, 1000 for each translation direction and corpus.

4.2.2 Model Hyperparameters

We trained a Transformer NMT model using the Sockeye framework [Hieber et al., 2017], mostly following the so-called *Transformer base model*: we used 6 layers, 512 positions, 8 attention heads and ReLU activations for both the encoder and decoder; Adam optimizer was used. Source and target token embeddings were both of size 512, and factors determining target language and style had embeddings of size 4. Batch size was set to 2048 words, initial learning rate to 0.0002, reducing by a factor of 0.7 every time the validation perplexity had not improved for 8 checkpoints, which happened every 4000 updates. The model converged during the 17th epoch, when validation perplexity has not improved for 32 consecutive checkpoints. The parameters of a single best checkpoint were used for all translations, with beam size set to 5.

4.3 Performance as a Machine Translation System

First, the model was evaluated in the context of machine translation, as the translation quality can be expected to influence other tasks the model performs.

We use public benchmarks for Estonian-English and Latvian-English translations from the news translation shared tasks of WMT 2017 and 2018 [Bojar et al., 2017, Bojar et al., 2018]. The BLEU scores for each translation direction and all included styles/domains are shown in Table 3.

Table 3. BLEU scores of the multilingual MT model on WMT'17 (Latvian↔English) and WMT'18 (Estonian↔English) test sets. EP denotes translation with Europarl as target style factor, JRC, OS and EMEA – JRC-Acquis, OpenSubtitles2018 and EMEA, respectively

	to EP	to JRC	to OS	to EMEA	Best score at WMT
EN→ET	**20.7**	19.9	20.6	18.6	**25.8**
ET→EN	24.7	23.6	**26.1**	23.8	**31.8**
EN→LV	15.7	15.3	**16.3**	15.0	**21.6**
LV→EN	18.3	17.8	**19.0**	17.5	**22.9**

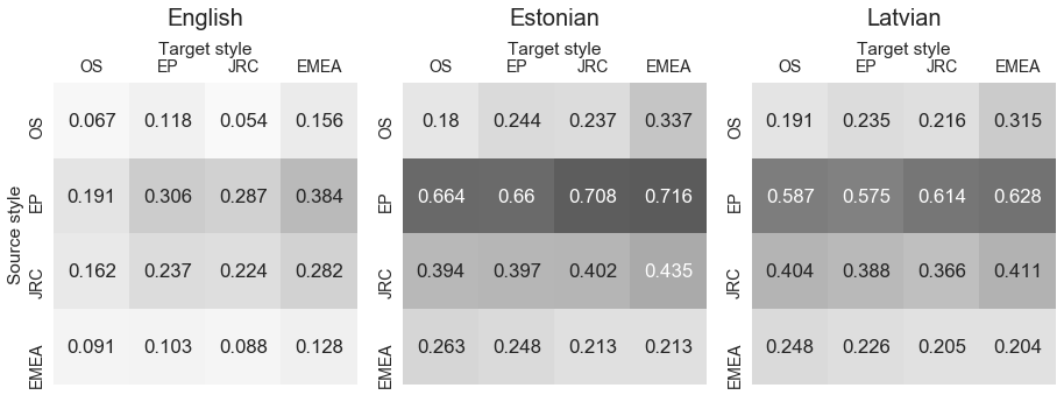


Figure 3. Proportions of sentences with token-wise edit distance ≥ 3 from the original when translated monolingually from and into different styles

The BLEU scores for translation from and into Latvian are noticeably below English-Estonian scores, which is likely explained by smaller datasets that include Latvian. Also, translation into English has higher scores than from English into Estonian/Latvian, which is also expected.

Our scores are lower than the scores best systems submitted to the WMT achieve on the same test sets. However, they are still reasonably high to yield good enough automatic translations. We believe this lag does not present a problem, as our system was trained without any special modifications (such as leveraging back-translated data or fine-tuning using a corpus of the same domain as the test data), and would still be competitive in the corresponding WMT competitions.

An interesting side-effect we have observed is the model's resilience to code-switching in the input text. The reason is that the model is trained with only the target language (and domain), and not the source language, as a result of which it learns to perform language normalization. For example, the sentence "Ma tahan two saldējumus." (*"Ma tahan"* / *"I want"* in Estonian, *"two"* and *"saldējumus"* / *"ice-creams"* in genitive, plural in Latvian) is correctly translated into English as "I want two ice creams.".

4.4 Evaluation

Next we move on to evaluating the model in terms of its performance in the context of style transfer.

At first, we examined how often the sentences change when translated monolingually. While it is to be expected that the model's behaviour when generating output text in the same language as the input should mainly consist of copying, the assumption is that passing modified style factors should prevent the model from simply copying the source sequences and incentivize it to match its output to certain style characteristics typical for

Table 4. Examples of style transfer in English

Translation into informal style (OpenSubtitles)	
I could not allow him to do that.	I couldn't let him do that.
That is correct.	That's right.
I will put you under arrest.	I'll arrest you.
He will speak with Mr. Johns.	He'll talk to Mr. Johns.
I will reimburse you.	I'll pay you back.
I did not wish to offend you.	I didn't want to offend you.
I did not participate in the game.	I didn't take part in the game.
Heed my advice.	Take my advice.
You shall not pass!	You won't pass!
I do not want to continue.	I don't want to go on.
Why though?	But why?
One does not simply walk into Mordor.	One doesn't just walk into Mordor.
Halt right there.	Stop right there.
Translation into formal style (Europarl)	
Yeah, like I said.	Yes, as I said.
How come you think you're so dumb?	Why do you think you are so stupid?
I've been trying to call.	I have tried to call.
It's a pretty important part of the deal.	It is a fairly important part of the deal.
This modification is definitely useful.	This amendment is certainly useful.
Is this thing completely necessary?	Is this entirely necessary?
I'm finished here.	I am done here.
I'd like to hear it from you though!	However, I would like to hear it from you!
Weird flex, but okay.	Strange flex, but all right.
This is gonna work.	This will work.
Sorry I'm late.	I am sorry I am late.

different corpora. Figure 3 shows the proportions of sentence pairs in the 1000-sentence test sets where there was a significant difference between translations into different styles. We can observe that English texts change less often than Estonian or Latvian, while Europarl sentences are changed more often than those of other corpora.

To assess whether these changes actually correspond to the model's capability for transferring style, we performed automatic and manual evaluation of style transfer quality. As BLEU scores for multilingual translation into Latvian were considerably below those for English and Estonian, we limit further evaluation to English and Estonian, assuming that translation quality would significantly influence performance on other tasks as well. We also use only two target style factors in the evaluation: OpenSubtitles2018 and Europarl. We expect these two corpora to demonstrate characteristics of informal and formal style, respectively.

For an idea of what our systems tends to do when translating monolingually into the styles of these two corpora, see Table 4.

4.4.1 Automatic Evaluation

We compared the system's outputs to those of a benchmark model built by Rao and Tetreault [Rao and Tetreault, 2018]. They train machine translation systems on parallel informal texts and their formal rewrites, which were manually created specifically for the GYAFC corpus. This approach, while relying on style-parallel data, is similar to ours in two important aspects: performing style transfer via monolingual translation and being applied to the task of formality transfer. Moreover, the GYAFC dataset includes output texts produced by Rao and Tetreault's models on the test split, which allows for straightforward comparison.

In all following evaluations, the model compared to is NMT Combined, which showed the best overall scores in Rao and Tetreault's experiments. We compare its outputs to the test split of the GYAFC dataset translated by our system. The corpus contains two domains: Entertainment & Music and Family & Relationships. In both domains, the informal sentences were translated with our zero-shot model into the style of Europarl, and formal into the style of OpenSubtitles2018.

We train a convolutional neural network classifier, using Lee's implementation[7] of the architecture proposed by Kim [Kim, 2014]. As training data, we use the train split of the GYAFC corpus from both Entertainment & Music and Family & Relationships domains.[8] The total number of sentences in the train split is 209,000, of which 40,000

[7] `https://github.com/DongjunLee/text-cnn-tensorflow`

[8] There exists a separate corpus of human judgements of text formality (`http://www.seas.upenn.edu/~nlp/resources/formality-corpus.tgz`) released by Pavlick and Tetreault [Pavlick and Tetreault, 2016], which may seem like a more fair choice of data to train a classifier than the very same data used for training the benchmark style transfer system. However, the Answers genre of this PT16 corpus, gathered from Yahoo Answers question answering forum similarly to the GYAFC dataset, is

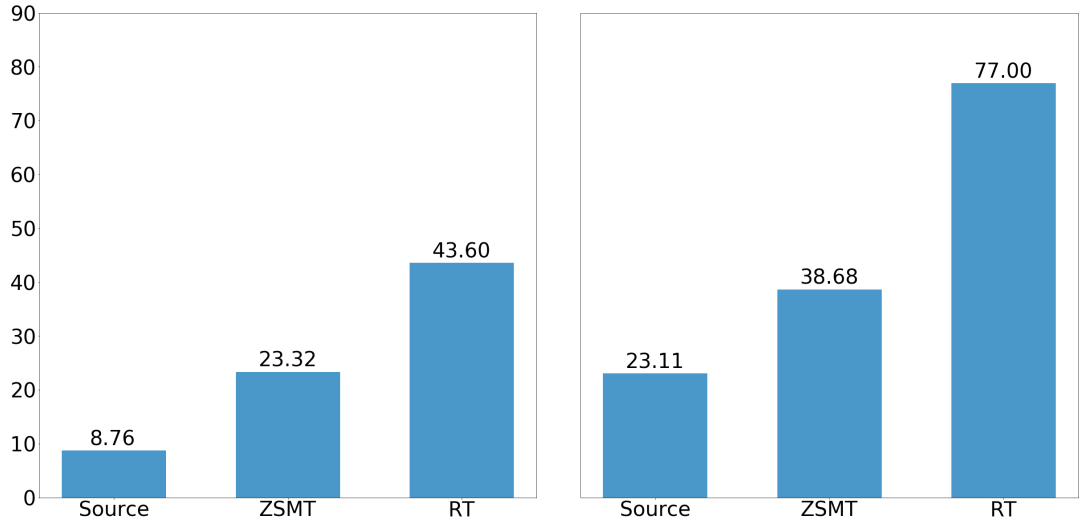
are used for validation. We train a convolutional neural network with filter sizes 3, 4, and 5, with 256 filters per size, with ReLU activation and max-pooling. No pre-trained word embeddings are used, instead, embeddings of dimension 300 are learned during training. Adam optimizer is used, with initial learning rate 5×10^{-4}. The convolutional layer is followed by one feed-forward layer. Dropout rate is set to 0.5. The classifier achieves validation accuracy of 0.79.

A note on automatic classification. The achieved classification accuracy of 0.79 is far from perfect, which somewhat weakens the validity of measures obtained with the classifier. However, classifying by formality turns out to be quite a hard task. For comparison, when training a model of similar architecture to distinguish between sentences of the Europarl and OpenSubtitles2018 corpora, one can easily achieve validation accuracy over 0.95 with little to no hyperparameter tuning. However, the two corpora vary a lot not only in their style, but also in the topics they cover, as one would expect. The neural network mostly learns to rely on the content of the sentences, and sentences from the Europarl corpus translated with our model into the informal style of OpenSubtitles2018 are still classified as belonging to Europarl and vice versa, even when there are noticeable differences in formality. Similarly, the formal test sets of the GYAFC corpus are overwhelmingly classified as OpenSubtitles2018, presumably because the topics covered on the Yahoo Answers forum are much more similar to those of films than of parliament proceedings. We believe this does not imply that OpenSubtitles2018 and Europarl cannot serve as sources of informal and formal texts, respectively, or that the style transfer systems do not function as intended, but rather that we need to take extra care with automatic classification, and possibly look into ways to force neural classifiers to disregard the semantics of the text. For these reasons, we make sure that the classifier is trained on texts which are similar in their content to the test sets.

For classifier scores on formal to informal and informal to formal transfer, see Figures 4a and 4b, respectively.

It can be seen that the zero-shot machine translation model achieves lower scores than Rao and Tetreault's benchmark, but there is noticeable movement in the desired direction: when modifying style to be more informal, after applying our model, 15% more sentences were considered to be informal than in the source set (16% in Family & Relationships domain and 13% more in Entertainment & Music domain). In the formal direction the figures are 16% overall, and 16% and 15% for Family & Relationships and Entertainment & Music, respectively.

relatively small (4977 sentences in total) and unbalanced (1373 sentences have positive average formality scores, and 3363 sentences negative). Moreover, Rao and Tetreault report that outputs of classifiers trained on the PT16 corpus applied to style transfer of the GYAFC corpus do not correspond well enough to human judgments. Our own experiments showed that validation accuracy of classifiers trained using the PT16 corpus was lower than that of the classifier eventually used, trained on GYAFC training data. For these reasons, we chose not to use the PT16 corpus in further experiments.



(a) Formal → informal (b) Informal → formal

Figure 4. Percentages of sentences classified as the target style in original GYAFC test sets of the opposite style, and test sets translated into the target style by Rao and Tetreault's (RT) model and by zero-shot machine translation (ZSMT) model. Higher scores for ZSMT/RT models indicate better performance in terms of style transfer

It should be noted that the classifier was trained on the GYAFC dataset, which means that it learns to recognize the kind of formality and informality that is understood as such in the GYAFC corpus. Moreover, the translated texts are similar to those which the RT model was trained on, but are out-of-domain for ZSMT model, as it has not encountered any user generated texts from the Internet in training. Finally, the RT model is trained on these in-domain parallel data, while ours has no style-parallel data at all. With these restrictions, our system still manages to show a noticeable improvement in style transfer scores when compared to source texts.

4.4.2 Manual Evaluation

Manually, we also compare our system to Rao and Tetreault's supervised machine translation benchmark [Rao and Tetreault, 2018]. We evaluate a subset of the test split of the GYAFC corpus, randomly choosing 25 sentences from each domain (Entertainment & Music and Family & Relationships) and style transfer direction (informal to formal and formal to informal), 100 sentences in total.

The evaluators are presented with the original sentence and two transferred versions of it: the output of Rao and Tetreault's NMT Combined model and of our model, in random order. Thus, they do not know which output was produced by which system. About each of the outputs, they are asked the following questions:

Table 5. Fluency and meaning preservation of Rao and Tetreault's (RT) and zero-shot machine translation (ZSMT) systems, as judged by humans

	RT	ZSMT
Fluency (1-4)	3.67	**3.82**
Meaning preservation (1-4)	3.61	**3.85**

1. How fluent is the sentence? (On a scale of 1 to 4, where 1 means "unreadable", and 4 "perfectly fluent")
2. How similar is the meaning of the sentence to the original? (On a scale of 1 to 4, where 1 means "completely different or nonsensical", and 4 means "the same")
3. Does the sentence sound:
    - more formal than the original,
    - more informal than the original,
    - or neither?
4. What differences are there between the transformed sentence and the original? (With options such as "no difference", "lexical", "grammatical", "word order", "punctuation", "contractions", "missing or added words, phrases or clauses". The evaluators are given examples of different types of changes before answering the questions.)

One more question concerns both outputs. The evaluators are asked which of the two outputs is better in general, that is, alters style in the right direction, while at the same time preserving meaning of the original text and being fluent. There is also an option "No preference", for cases where both outputs are equally good or equally bad.

3 people participated in the evaluation. All of them are fluent, but non-native speakers of English.

Average fluency and meaning preservation scores of the two systems are shown in Table 5. It can be seen that while both systems score very high on both aspects, zero-shot machine translation is also better in both.

We also assessed how often annotators found that the style was altered in the right direction. Each annotator's vote was counted as a 1 if they thought that a sentence became more formal, and as -1 if it became more informal after translation. The average of votes for each sentence was calculated. If the average is negative, we consider the sentence more informal than the original text, if it is positive, the sentence is considered more formal, and if it equals 0, no change in style has occurred. Then we calculate the proportion of sentences in which the direction of style transfer corresponded to the

Table 6. Proportions of sentences where direction of style transfer was found by human annotators to be correct, wrong, and not present, in outputs of Rao and Tetreault's (RT) and zero-shot machine translation (ZSMT) systems

	RT	ZSMT
Right direction	**0.71**	0.58
Wrong direction	**0.04**	0.05
No change in style	0.25	0.37

intended one, was wrong, or not present. The percentages are given in Table 6. Our system tends to be more conservative: it has more sentences where humans detected no changes in style at all. Rao and Tetreault's system shows the correct behaviour (style transfer in the intended direction) more often. However, the overall pattern is similar for both systems.

Next, the overall preference was assessed. In 49% of cases, outputs of our system were preferred by the average annotator. The RT system was preferred in 27% of cases, and in 24% of cases no preference was expressed.

While being more conservative in terms of transferring style, our system appears to win in both fluency and meaning preservation, which could, in part, lead to it being generally preferred more often by human annotators. (If a text shows strong characteristics of the intended style but fails to communicate the content of the source, it may be less likely to be preferred than a text which differs from the original less but preserves the semantics.)

Finally, we examined what types of changes were reported by the annotators. The most frequent type for our system was contractions (e.g. *I have just been* vs. *I've just been*), reported by at least one annotator in 38 cases. The next most frequent change is lexical substitutions (e.g. *'cause* vs. *because*, or *sure* vs. *certainly*) with 33 cases, then comes punctuation (e.g. *no!!!* vs. *no.*) with 22 cases and missing or added words or phrases with 20 cases. Less common are grammatical changes (e.g. replacing *no one's gonna* with *no one will*, or *method of production* with *production method*), reported in 8 sentences, and changes in word order, reported in 5 sentences. In 27 sentences out of 100 the annotators did not see any difference from the source sentence. See Table 7 for a comparison to the RT system. All modifications except for contractions are made by the benchmark model more often than by ours, and our model leaves more sentences unchanged.

For lack of an Estonian-language benchmark to compare our system to, we evaluated its performance independently. Manual evaluation was performed on a subset of 100 sentences, 50 of them selected randomly from the OpenSubtitles test set and the other 50 from Europarl. Each sentence was translated into the opposite style. The resulting pairs

Table 7. Proportion of output sentences of the two systems in which varying types of changes were reported by human annotators (RT – Rao and Tetreault's benchmark, ZSMT – zero-shot machine translation system)

	ZSMT	RT
Contractions	0.38	0.36
Lexical substitutions	0.33	0.46
Punctuation	0.22	0.49
Missing or added words or phrases	0.20	0.39
Grammatical variations	0.08	0.16
Word order	0.05	0.07
No difference	0.27	0.10

were presented to participants, who were asked the following questions about each of them:

1. Do the sentences differ in any way?
2. How fluent is the translated sentence? (On a scale of 1 to 4, where 1 is "unreadable", and 4 is "perfectly fluent")
3. How similar are the sentences in meaning? (On a scale of 1 to 4, 1 meaning "completely different or nonsensical", and 4 "the same")
4. Does the translated sentence sound:
    - more formal than the original,
    - more informal,
    - or neither?
5. What differences are there between the sentences? (With options such as "grammatical", "lexical", "missing words or phrases", "word order", "the use of formal 'you'")

For completeness, the same independent evaluation was performed in English as well. 3 people participated in each of the evaluations. All annotators of Estonian are native or natively bilingual speakers of Estonian, while all annotators of English are fluent, but non-native speakers of English.

Table 8. Fluency and meaning preservation, as judged by humans

	English	Estonian
Fluency (1-4)	3.84	3.64
Meaning preservation (1-4)	3.67	3.35

Inter-annotator agreement. In evaluation of fluency, all three human evaluators gave the translated sentences the same score in 41 out of 55 cases in English (not taking into account sentences which were simply copied from their originals), and in 51 out of 68 cases in Estonian. In evaluation of direction of style transfer, all three evaluators agreed in 16 cases and at least two agreed in 43 cases in English, and in Estonian in 19 cases all three agreed and in 59 at least two.

Of the 100 translated sentences, 45 were marked by all participants as being the same as their original sentences in the English set and 32 in Estonian. The remaining 55 and 68, respectively, were used to quantify style transfer quality.

Being a reasonably strong MT system, our model scores quite high on fluency (3.84 for English, 3.64 for Estonian) and meaning preservation (3.67 for English, 3.35 for Estonian). Average scores are given in Table 8.

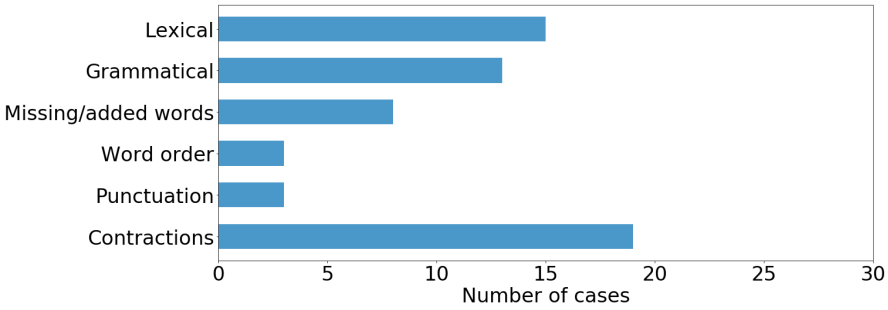
We evaluated the style transfer itself in the same way as in comparison to the baseline. For each pair of sentences, the average score given by three evaluators was calculated, in which the answer that the translated sentence is more formal counts as +1, more informal as -1, and neither as 0. Of the 55 sentences in English that were different from their source sentences, in 33 cases the sign of the average human score matched the desired one, in 7 it did not, and in 15 no change in style was observed by humans. In Estonian 36 sentences showed the right direction of style transfer, 10 wrong, and 22 no change.

In English sentences where the direction of style transfer was found to be correct (Figure 5a), changes in use of contractions were reported in 19 cases, lexical changes in 15 cases , grammatical in 13, missing or added words or phrases in 8 cases.

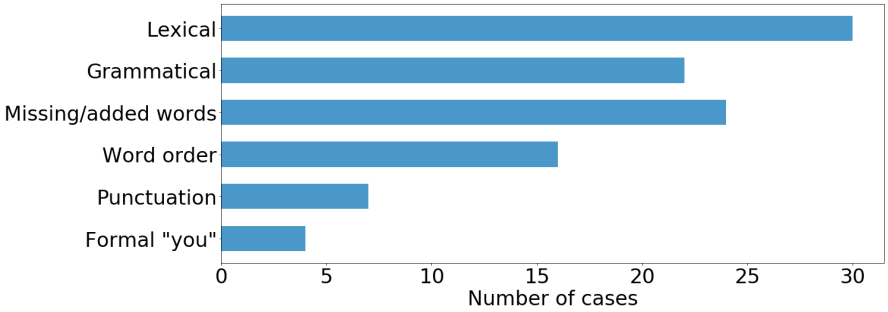
In Estonian correctly transferred sentences (Figure 5b), the most frequently reported were lexical substitutions (30 cases), followed by missing of added words or phrases (24 cases), changes in grammar (22 cases) and in word order (16 cases).

We can observe that, as purely formal measures of the extent of monolingual changes suggested (see Figure 3), Estonian is generally modified more than English. The types of changes the system introduces are also different: apart from obvious ones, such as the use of polite "you", which exists in Estonian but not in English, and contractions, which exist in English but not in Estonian, lexical and grammatical changes are more frequent in Estonian. Word order is also altered in Estonian more often: Estonian word order is less strict, and the model explores this avenue of stylistic modification.

Unlike many style transfer models which produce text with strong style characteristics



(a) English



(b) Estonian

Figure 5. Number of sentences in which evaluators reported different types of changes

(e.g. with strong positive or negative sentiment), often at the cost of preserving meaning and fluency, our model gravitates towards keeping the meaning and fluency of the original sentence intact and mimicking some of the desired stylistic traits. However, it mostly succeeds in transferring style in the desired direction, and, most importantly, is generally preferred by human evaluators to a baseline trained in a supervised manner and using costly manually constructed data, while performing zero-shot style transfer without any parallel examples.

4.5 Cross-lingual Style Transfer

Being able to translate between languages and also to modify the output to match the desired style allows the model to essentially perform domain adaptation. When translating from a language which has no formal "you" (English) into one that does (Estonian or Latvian), it will quite consistently use the informal variant when the target style is OpenSubtitles and the formal when the target style is Europarl (*you rock* $\rightarrow$ *sa rokid/te rokite*). The model is also quite consistent in use of contractions in English (*es esmu šeit* $\rightarrow$ *I am here/I'm here*). Grammatical constructions can change as well: *need on Kadri lapsed.* can be translated as either *these are Kadri's children.* or *these are the children of Kadri*. This modification is not consistent, but is sometimes dropped in favor of lexical change: *need on Matti lapsed.* $\rightarrow$ *those are Matt's kids./these are Matt's children.* Word order may change: *Where is Anna's bag?* is *Kus on Anna kott?* in the more formal variant, and *Kus Anna kott on?* in the more informal. More examples can be found in Table 9.

Table 9. Examples of domain adaptation (or "cross-lingual style transfer")

Source	Formal	Informal
You are great!	Te olete suurepärane!	Sa oled suurepärane!
You got cash?	Kas sul on raha?	Sul sularaha on?
No biggie.	Ei midagi suurt.	Pole midagi.
I don't care, and you shouldn't either.	Mind ei huvita ja ka teie ei peaks seda tegema.	Mind ei huvita ja sina ka mitte.
I wouldn't like to.	Ma ei tahaks seda teha.	Mulle ei meeldiks.
Where is my pen?	Kus on minu pastakas?	Kus mu pastakas on?
Are you doing okay?	Kas teil läheb hästi?	Kas sinuga on kõik korras?
Pole probleemi.	There is no problem.	No problem.
See on lihtne.	It is simple.	It's easy.
Ma ei saanud aru.	I did not understand.	I didn't get it.
Kas teil läheb hästi?	Are you doing well?	Are you all right?
Kus sa olid?	Where have you been?	Where were you?

5 Conclusion

In this thesis, we proposed a simple method for text style transfer, which does not require parallel data between styles, and builds on the idea of zero-shot neural machine translation, applying it to styles instead of languages. The architecture allows to incorporate multiple languages within the same model and can be easily implemented using open-source software and publicly available data.

We built a style transfer system of the proposed structure, using three languages and four text styles. We performed an extensive evaluation of the model on two of those languages (English and Estonian), assessing the model's ability to modify the level of formality in text. While showing modest automatic scores, our system outperformed a benchmark trained using parallel data on several aspects in manual evaluation. While being more conservative than the benchmark and keeping the output text similar to the input more often, our approach scored higher than the benchmark on fluency and meaning preservation, and was generally preferred by human annotators in 49% of cases, compared to the benchmarks's 27%.

We are now able to answer the research questions posed in the introduction:

1. It is possible to leverage regular language-parallel corpora to perform style transfer without using parallel corpora. We conclude so from automatic and manual evaluation results, as well as illustrative examples shown in Table 4.

2. The model trained on such data scores lower than the supervised baseline on automatic metrics, showing, however, a noticeable movement in the desired direction compared to the original texts, but outperforms it on several aspects in manual evaluation. Our model shows higher fluency and meaning preservation scores, and is overall preferred to the baseline by annotators, presumably due to the quality of output text compensating for its relatively conservative style transfer.

3. The model uses aspects such as contractions, lexical substitutions, grammatical constructions and word order to transfer style.

4. The results shown by the model depend on the language of the text. The model makes different choices in modifying certain aspects of the text, for example, learning contractions, which are specific to English, and leveraging free word order of Estonian. The model's performance depends on the machine translation quality it shows: better BLEU scores for translation into English than into Estonian do translate into higher fluency and meaning preservation scores for style transfer in English. English is also the least strongly modified language in general, as indicated by Figures 3 and 5.

In the future, we plan to apply the proposed style transfer method to other style transfer tasks apart from formality transfer, and experiment with augmenting the model

with pseudo-parallel data generated by back-translation.

References

[Artetxe and Schwenk, 2018] Artetxe, M. and Schwenk, H. (2018). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

[Bojar et al., 2017] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

[Bojar et al., 2016] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 131–198. Association for Computational Linguistics.

[Bojar et al., 2018] Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

[Carlson et al., 2017] Carlson, K., Riddell, A., and Rockmore, D. N. (2017). Zero-shot style transfer in text using recurrent neural networks. *CoRR*, abs/1711.04731.

[Fu et al., 2017] Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2017). Style transfer in text: Exploration and evaluation. *CoRR*, abs/1711.06861.

[Gatys et al., 2016] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Hieber et al., 2017] Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017). Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

[Jhamtani et al., 2017] Jhamtani, H., Gangal, V., Hovy, E. H., and Nyberg, E. (2017). Shakespearizing modern language using copy-enriched sequence-to-sequence models. *CoRR*, abs/1707.01161.

[Johnson et al., 2017] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

[Koehn, 2005] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

[Koehn, 2017] Koehn, P. (2017). Neural machine translation. *CoRR*, abs/1709.07809.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

[Kudo and Richardson, 2018] Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

[Li et al., 2018] Li, J., Jia, R., He, H., and Liang, P. (2018). Delete, retrieve, generate: A simple approach to sentiment and style transfer. *CoRR*, abs/1804.06437.

[Lison and Tiedemann, 2016] Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[Mir et al., 2019] Mir, R., Felbo, B., Obradovich, N., and Rahwan, I. (2019). Evaluating style transfer for text. *CoRR*, abs/1904.02295.

[Pavlick and Tetreault, 2016] Pavlick, E. and Tetreault, J. (2016). An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

[Prabhumoye et al., 2018] Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. (2018). Style transfer through back-translation. *CoRR*, abs/1804.09000.

[Rao and Tetreault, 2018] Rao, S. and Tetreault, J. R. (2018). Dear sir or madam, may I introduce the YAFC corpus: Corpus, benchmarks and metrics for formality style transfer. *CoRR*, abs/1803.06535.

[Sennrich and Haddow, 2016] Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. *CoRR*, abs/1606.02892.

[Sennrich et al., 2016] Sennrich, R., Haddow, B., and Birch, A. (2016). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40. Association for Computational Linguistics.

[Shen et al., 2017] Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. S. (2017). Style transfer from non-parallel text by cross-alignment. *CoRR*, abs/1705.09655.

[Subramanian et al., 2018] Subramanian, S., Lample, G., Smith, E. M., Denoyer, L., Ranzato, M., and Boureau, Y. (2018). Multiple-attribute text style transfer. *CoRR*, abs/1811.00552.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

[Tars and Fishel, 2018] Tars, S. and Fishel, M. (2018). Multi-domain neural machine translation. In *Proceedings of The 21st Annual Conference of the European Association for Machine Translation (EAMT'2018)*, pages 259 – 268, Alicante, Spain.

[Tiedemann, 2012] Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Xu et al., 2012] Xu, W., Ritter, A., Dolan, W., Grishman, R., and Cherry, C. (2012). Paraphrasing for style. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 2899–2914.

[Zhang et al., 2018] Zhang, Z., Ren, S., Liu, S., Wang, J., Chen, P., Li, M., Zhou, M., and Chen, E. (2018). Style transfer as unsupervised machine translation. *CoRR*, abs/1808.07894.

[Zhao et al., 2017] Zhao, J. J., Kim, Y., Zhang, K., Rush, A. M., and LeCun, Y. (2017). Adversarially regularized autoencoders for generating discrete structures. *CoRR*, abs/1706.04223.

Appendix

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Elizaveta Korotkova**,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Style Transfer via Zero-Shot Monolingual Neural Machine Translation,
   (title of thesis)

   supervised by Mark Fishel.
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Elizaveta Korotkova
16/05/2019