Stylistic Transfer in Natural Language Generation Systems Using Recurrent Neural Networks

Jad Kabbara and **Jackie Chi Kit Cheung**
School of Computer Science
McGill University
Montreal, QC, Canada
jad@cs.mcgill.ca jcheung@cs.mcgill.ca

Abstract

Linguistic style conveys the social context in which communication occurs and defines particular ways of using language to engage with the audiences to which the text is accessible. In this work, we are interested in the task of stylistic transfer in natural language generation (NLG) systems, which could have applications in the dissemination of knowledge across styles, automatic summarization and author obfuscation. The main challenges in this task involve the lack of parallel training data and the difficulty in using stylistic features to control generation. To address these challenges, we plan to investigate neural network approaches to NLG to automatically learn and incorporate stylistic features in the process of language generation. We identify several evaluation criteria, and propose manual and automatic evaluation approaches.

1 Introduction

Linguistic style is an integral aspect of natural language communication. It conveys the social context in which communication occurs and defines particular ways of using language to engage with the audiences to which the text is accessible.

In this work, we examine the task of stylistic transfer in NLG systems; that is, changing the style or genre of a passage while preserving its semantic content. For example, given texts written in one genre, such as Shakespearean texts, we would like a system that can convert it into another, say, that of simple English Wikipedia. Currently, most knowledge available in textual form is locked into the particular data collection in which it is found. An automatic stylistic transfer system would allow that information to be more generally disseminated. For example, technical articles could be rewritten into a form that is accessible to a broader audience. Alternatively, stylistic transfer could also be useful for security or privacy purposes, such as in author obfuscation, where the style of the text is changed in order to mask the identity of the original author.

One of the main research challenges in stylistic transfer is the difficulty in using linguistic features to signal a certain style. Previous work in computational stylistics have identified a number of stylistic cues (e.g., passive vs active sentences, repetitive usage of pronouns, ratio of adjectives to nouns, and frequency of uncommon nouns). However, it is unclear how a system would transfer this knowledge into controlling realization decisions in an NLG system. A second challenge is that it is difficult and expensive to obtain adequate training data. Given the large number of stylistic categories, it seems infeasible to collect parallel texts for all, or even a substantial number of style pairs. Thus, we cannot directly cast this as a machine translation problem in a standard supervised setting.

Recent advances in deep learning provide an opportunity to address these problems. Work in image recognition using deep learning approaches has shown that it is possible to learn representations that separate aspects of the object from the identity of the object. For example, it is possible to learn features that represent the pose of a face (Cheung et al., 2014) or the direction of a chair (Yang et al., 2015), in order to generate images of faces/chairs with new

Proceedings of EMNLP 2016 Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, pages 43–47,
Austin, TX, November 5, 2016.

poses/directions. We plan to design similar recurrent neural network architectures to disentangle the style from the semantic content in text. This setup not only requires less hand-engineering of features, but also allows us to frame stylistic transfer as a weakly supervised problem without parallel data, in which the model learns to disentangle and recombine latent representations of style and semantic content in order to generate output text in the desired style.

In the rest of the paper, we discuss our plans to investigate stylistic transfer with neural networks in more detail. We will also propose several evaluation criteria for stylistic transfer and discuss evaluation methodologies using human user studies.

2 Related Work

Capturing stylistic variation is a long-standing problem in NLP. Sekine (1997) and Ratnaparkhi (1999) consider the different categories in the Brown corpus to be domains. These include *general fiction*, *romance and love story*, *press: reportage*. Gildea (2001), on the other hand, refers to these categories as genres. Different NLP sub-communities use the terms *domain*, *style* and *genre* to denote slightly different concepts (Lee, 2001). From a linguistic point of view, domains could be thought of as broad *subject fields*, while genre can be seen as a category assigned on the basis of external criteria such as intended audience, purpose, and activity type. Style conveys the social context in which communication occurs and define particular ways of using language to engage with the audiences to which the text is accessible. Some linguists would argue that style and domain are two attributes characterizing genre (e.g., (Lee, 2001)) while others view genre and domain as aspects representing style (e.g., (Moessner, 2001)).

The notion of genre has been the focus of related NLP tasks. In genre classification (Petrenz and Webber, 2011; Sharoff et al., 2010; Feldman et al., 2009), the task is to categorize the text into one of several genres. In author identification (Houvardas and Stamatatos, 2006; Chaski, 2001), the goal is to identify the author of a text, while author obfuscation (Kacmarcik and Gamon, 2006; Juola and Vescovi, 2011) consists in modifying aspects of the texts so that forensic analysis fails to reveal the identity of the author.

In (Pavlick and Tetreault, 2016), an analysis of formality in online written communication is presented. A set of linguistic features is proposed based on a study of human perceptions of formality across multiple genres. Those features are fed to a statistical model that classifies texts as having a formal or informal style. At the lexical level, Brooke et al. (2010) focused on constructing lexicons of formality that can be used in tasks such as genre classification or sentiment analysis. In (Inkpen and Hirst, 2004), a set list of near-synonyms is given for a target word, and one synonym is selected based on several types of preferences, e.g., stylistic (degree of formality). We aim to generalize this work beyond the lexical level.

A similar work is that of Xu et al. (2012) which propose using phrase-based machine translation systems to carry out paraphrasing while targeting a particular writing style. Since the problem is framed as a machine translation problem, it relies on parallel data where the source "language" is the original text to be paraphrased–in that case, Shakespeare texts–and the "translation" is the equivalent modern English version of those Shakespeare texts. Accordingly, for each source sentence, there exists a parallel sentence having the target style. They also present some baselines which do not make use of parallel sentences and instead rely on manually compiled dictionaries of expressions commonly found in Shakespearean English. In a more recent work, Sennrich et al. (2016) carry out translation from English to German while controlling the degree of politeness. This is done in the context of neural machine translation by adding side constraints. Specifically, they mark up the source language of the training data (in this case, English) with a feature that encodes the use of honorifics seen in the target language (in this case, German). This allows them to control the honorifics that are produced at test time.

3 Proposed Approach

Recently, RNN-based models have been successfully used in machine translation (Cho et al., 2014b; Cho et al., 2014a; Sutskever et al., 2014) and dialogue systems (Wen et al., 2015). Thus, we propose to use an LSTM-based RNN model based on the encoder-decoder structure (Cho et al., 2014b)

to automatically process stylistic nuances instead of hand-engineering features. The model is a variant of an autoencoder where the latent representation has two separate components: one for style and one for content. The learned *stylistic* features would be distinct from the *content* features and specific to each style category, such that they can be swapped between training and testing models to perform stylistic transfer. The separation, or *disentanglement*, between stylistic and content features is reinforced by modifying the training objective from (Cho et al., 2014b) that maximizes the conditional log-likelihood (of the output given the input). Instead, our model is trained to maximize a training objective that also includes a cross-covariance term dedicated for the disentanglement.

At a high level, our proposed approach consists of the following steps:

1. For a given style transfer task between two styles A and B, we will first collect relevant corpora for each of those styles.

2. Next, we will train the model on each of the styles (separately). This would allow the system to *disentangle* the content features from the stylistic features. At the end of this step, we will have (separately) the features that characterize styles A and those that characterize style B.

3. During the testing phase, for a transfer, say, from style A to style B, the system is fed texts having style A while the stylistic latent variables of the model are fixed to be those learned for style B (from the previous step). This would force the model to generate text using style B. For a transfer from style B to A, the system is fed texts having style B and we fix the stylistic latent variables of the model to be those learned for style A.

We intend to apply the model to datasets with reasonably differing styles between training and testing. Examples include the complete works of Shakespeare[1], the Wikpedia Kaggle dataset [2], the Oxford Text Archive (literary texts) [3], and Twitter data. A future research direction would be to further improve the system to process texts that have differing but similar styles.

4 Evaluation

We first present a simple example that shows the input and output of the system during the testing phase. Assuming the system was trained on texts taken from Simple English Wikipedia, it would learn the stylistic features that are particular to that genre. During the testing phase, if we feed the system the following sentence taken from Shakespeare's play As You Like It (Act 1, Scene 1):

> *As I remember, Adam, it was upon this fashion bequeathed me by will but poor a thousand crowns, and, as thou sayest, charged my brother on his blessing to breed me well. And there begins my sadness.*

we expect the system to produce a version that might be similar to the following:

> *I remember, Adam, that's exactly why my father only left me a thousand crowns in his will. And as you know, my father asked my brother to make sure that I was brought up well. And that's where my sadness begins.*

We see three main criteria for the evaluation of stylistic transfer systems: **soundness** (i.e., the generated texts being textually entailed with the original version), **coherence** (e.g., free of grammatical errors, proper word usage, etc.), and **effectiveness** (i.e., the generated texts actually match the desired style). We propose to evaluate systems using both human and automatic evaluations. Snippets of original and generated texts will be sampled and reviewed by human evaluators, who will judge them on these three criteria using Likert ratings. This type of evaluation technique is also used in related tasks such as to evaluate author obfuscation systems (Stamatatos et al., 2015). A future research direction is

[1] http://norvig.com/ngrams/shakespeare.txt

[2] https://www.kaggle.com/c/wikichallenge/Data

[3] https://ota.ox.ac.uk/

to investigate automatic evaluation measures similar to ROUGE and BLEU, which compare the content of the generated text against human-written gold standards using word or n-gram overlap.

5 Conclusion

We present stylistic transfer as a challenging generation task. Our proposed research will address challenges to the task, such as the lack of parallel training data and the difficulty of defining features that represent style. We will exploit deep learning models to extract stylistic features that are relevant to generation without requiring explicit parallel training data between the source and the target styles. We plan to evaluate our methods using human judgments, according to criteria that we propose, derived from related tasks.

References

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 90–98. Association for Computational Linguistics.

Carole E Chaski. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8:1–65.

Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. 2014. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Sergey Feldman, Marius A Marin, Mari Ostendorf, and Maya R Gupta. 2009. Part-of-speech histograms for genre classification of text. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4781–4784. IEEE.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202.

John Houvardas and Efstathios Stamatatos. 2006. N-gram feature selection for authorship identification. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 77–86. Springer.

Diana Zaiu Inkpen and Graeme Hirst. 2004. Near-synonym choice in natural language generation. In *Recent Advances in Natural Language Processing*, volume 3, pages 141–152.

Patrick Juola and Darren Vescovi. 2011. Analyzing stylometric approaches to author obfuscation. In *IFIP International Conference on Digital Forensics*, pages 115–125. Springer.

Gary Kacmarcik and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 444–451. Association for Computational Linguistics.

David Y.W. Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning and Technology*, 5(3):37–72, September.

Lilo Moessner. 2001. Genre, text type, style, register: A terminological maze? *European Journal of English Studies*, 5(2):131–138.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393.

Adwait Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. *Machine learning*, 34(1-3):151–175.

Satoshi Sekine. 1997. The domain dependence of parsing. In *Proceedings of the fifth conference on Applied natural language processing*, pages 96–102. Association for Computational Linguistics.

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

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of babel: evaluating genre collections. In *LREC*. Citeseer.

Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. 2015.

Overview of the pan/clef 2015 evaluation lab. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 518–538. Springer.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Wei Xu, Alan Ritter, William B Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *24th International Conference on Computational Linguistics, COLING 2012*.

Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. 2015. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pages 1099–1107.