SentiInc: Incorporating Sentiment Information into Sentiment Transfer Without Parallel Data

Kartikey Pant(✉), Yash Verma, and Radhika Mamidi

International Institute of Information Technology, Hyderabad, India
kartikey.pant@research.iiit.ac.in, yash.verma@students.iiit.ac.in, radhika.mamidi@iiit.ac.in

Abstract. Sentiment-to-sentiment transfer involves changing the sentiment of the given text while preserving the underlying information. In this work, we present a model SentiInc for sentiment-to-sentiment transfer using unpaired mono-sentiment data. Existing sentiment-to-sentiment transfer models ignore the valuable sentiment-specific details already present in the text. We address this issue by providing a simple framework for encoding sentiment-specific information in the target sentence while preserving the content information. This is done by incorporating sentiment based loss in the back-translation based style transfer. Extensive experiments over the Yelp dataset show that the SentiInc outperforms state-of-the-art methods by a margin of as large as ∼11% in G-score. The results also demonstrate that our model produces sentiment-accurate and information-preserved sentences.

Keywords: Textual style transfer · Sentiment analysis

1 Introduction

Esoteric methods in sequence to sequence tasks use massive amounts of parallel data. However, in many style transfer tasks such as sentiment-to-sentiment transfer, such data is not readily available. Therefore, most of the recent work in language style transfer focuses on solving this task in an unsupervised setting [2,6]. Unsupervised learning involves learning a latent representation of data in a shared latent space, which provides fine control over the latent attributes in the data. Autoencoders have been used for generating sentences with controllable attributes by the disentangled latent representations [1,4]. Most of the work done previously on the task uses adversarial training for learning this latent representation [3,4,12,18].

The task of sentiment-to-sentiment transfer is equivalent to finding target sentence y that maximizes the conditional probability of y given a source sentence

K. Pant and Y. Verma—Contributed equally to the work.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 312–319, 2020.
https://doi.org/10.1007/978-3-030-45442-5_39

x and sentiment s, i.e., $\max p(y|x, s)$. Sentiment-to-sentiment transfer can be seen as a special style transfer task. It involves changing the underlying sentiment of the source text while preserving the underlying non-semantic content.

In this work, we propose an architecture that extends the model proposed by [6], which performs machine translation relying on monolingual corpora in each language. We form language models for each sentiment and use iterative back-translation [10], thereby converting this unsupervised task into a semi-supervised task.

To summarize, the following are the contributions[1]:

- We propose a novel approach for sentiment-to-sentiment transfer incorporating sentiment-specific information in an unsupervised setting.
- The proposed method encodes sentiment-specific information in the target sentence by incorporating sentiment-based loss in the iterative back-translation algorithm.
- The proposed method also gives finer control over the trade-off between content-preservation and sentiment-transfer, thus making it adaptable for various applications.
- We extensively evaluate the performance of our model on a real-world dataset, and results reveal that it outperforms the state-of-the-art methods.

2 Related Work

The most closely related works are on the areas of neural unpaired sentiment-to-sentiment translation and encoding sentiment-specific information in training neural networks.

Generating sentences with controllable attributes has been addressed in [4] by learning disentangled latent representations [1]. Their model builds on variational auto-encoders (VAEs) and uses independency constraints to enforce reliable inference of attributes back from generated sentences. [12] focuses on separating the underlying content from style information, [19] employs an iterative back-translation algorithm by using a pseudo-parallel corpus created using a word-to-word transfer table built by cross-domain word embeddings and style specific language models. Use of back translation to facilitate unsupervised machine translation is addressed in [6]. This model was extended by [13] to address the limitations of attribute transfer by performing attribute conditioning and latent representation pooling.

On the other hand, sentiment-specific information is encoded with lexicon-based approaches [14,16,17], mostly sentiment-polarity pairs, incorporating negation and intensification to compute sentence polarity for each sentence. [15] proposed learning sentiment-specific word embeddings (SSWE) which encoded the sentiment information into a continuous representation of words by incorporating a sentiment-specific loss function in the training process.

[1] Made available at the following Github repository: https://github.com/kartikeypant/sentiinc-sentiment-transfer.

To the best of our knowledge, there is no previous work that facilitates sentiment-to-sentiment transfer while encoding sentiment-specific information, in an unsupervised setting.

3 Sentiment Transfer Using Sentiment-Specific Loss

In this section, we define the task and its preliminaries before presenting the details of learning sentiment-specific information for unsupervised sentiment transfer.

Let x be an input sentence (having a sentiment y), $\tilde{y}$ be the target sentiment, and $\tilde{x}$ be the target sentence. Our task is to map $(x, \tilde{y})$ to $\tilde{x}$, such that maximum amount of original content-based semantic information from x is preserved independent of the sentiment information. For the remainder of the paper, we denote the source space as S, and the target space as T. Let P_s and P_t denote the language models trained on datasets from each domain and $P_{s\to t}$ and $P_{t\to s}$ be the translation models from source to target and vice versa.

3.1 Style Transfer as Unsupervised Machine Translation

Instead of considering words, we use BPE tokens [11] as they reduce the vocabulary size and eliminate the presence of unknown words in the output. The data of both the sentiment domains (positive and negative) is jointly processed to form the BPE tokens which are shared across sentiment domains. Token embeddings [8] are now learned to initialize the lookup tables for encoder and decoder. We accomplish language modeling via denoising autoencoders [6], by minimizing the following loss:

$$L_{lm} = E_{x\sim S}[-logP_{s\to s}(x|C(x))] + E_{y\sim T}[-logP_{t\to t}(y|C(y))] \tag{1}$$

where, C is a noise model with some words dropped and swapped. $P_{s\to s}$ and $P_{t\to t}$ are the composition of encoder and decoder both operating on the source and target sides, respectively.

For converting this unsupervised task to a semi-supervised setting, we use iterative back-translation. We consider $\forall x \in S$, let $v^\star(x) = \arg\max P_{s\to t}(v|x)$ and $\forall y \in T$, $u^\star(y) = \arg\max P_{t\to s}(x, v^\star(x))$ and $(u^\star(y), y)$ which forms automatically-generated parallel sentences. Using this, we train two transfer models by minimizing the following loss:

$$L_{back} = E_{y\sim T}[-logP_{s\to t}(y|u^\star(y))] + E_{x\sim S}[-logP_{t\to s}(x|v^\star(x))] \tag{2}$$

3.2 Sentiment-Specific Loss

For incorporating the sentiment-specific loss into the training, we use pretrained fastText classifiers [5], which provide polarities from -2 to 2, -2 being the most negative. Let the score predicted by fastText classifier be denoted by $\delta(t)$. Now, let $f(t)$ be the indicator of the target sentiment, given a score $+1$ if the

target sentiment is negative and -1 if it is positive. The sentiment-specific loss is modeled as:

$$L_{s(t)} = exp(n^2 \cdot k) \cdot max(0, f(t) \cdot \delta(t)) \tag{3}$$

Here, n denotes the number of epochs and k is a hyperparameter and $exp(n^2 \cdot k)$ is used to make the sentiment-specific information more dominant as the model learns to generate content-preserved sentences with an increase in n.

3.3 Shared Latent Representation

SentiInc uses latent representation for both language modeling and style transfer as it ensures inductive transfer across both tasks. To share the encoder representation, we share all the encoder as well as decoder parameters across the two sentiment domains.

While minimizing the loss function, backpropagation is not performed through the reverse model which generated the data since as observed by [6], no significant improvement were observed by doing so. The final loss function to be minimized is as follows:

$$L = L_{lm} + L_{back} + L_{s(t)} \tag{4}$$

4 Experiments

In this section, we introduce the datasets and briefly describe the evaluation metrics used by contemporary models. We also compare SentiInc's performance to the current state-of-the-art on these datasets.

4.1 Dataset

We conduct experiments on the publicly available Yelp food review dataset as previously used by [7,12,13]. Unlike most work in the area, we operate at the scale of complete reviews and do not assume that every sentence of the review inherits the sentiment of the complete review. We, therefore, relax constraints enforced in prior works [7,12] that discard reviews with more than 15 words and only consider the $10k$ most frequent words. Instead, we consider full reviews with up to 100 words and use byte-pair encodings (BPE) [11] eliminating the presence of unknown words.

Small Yelp Review Dataset (SYelp): It is used by many of the previous works conducted in this area [3,7,12], and contains sentences instead of complete reviews. It is based on the assumption that each sentence in a review inherits the sentiment of the complete review. Finally, the data is encoded in $10k$ BPE Codes.

Big Yelp Review Dataset (BYelp): It contains full reviews instead of individual sentences. Following previous work on reviews spanning multiple reviews [13], we preprocess this data to remove reviews that are not written in English

according to a fastText classifier [5] which achieves competitive performance. Based on the rating associated with the review, we classify it as either positive or negative. Finally, the data is encoded in $60k$ BPE Codes.

4.2 Evaluation Metrics

In this work, we use a combination of multiple automatic evaluation criteria informed by our desiderata. We want our model to generate sentences that have the target sentiment while preserving the structure and content of the input.

Therefore, we evaluate samples from different models along the following two different criteria:

- **Transfer of Sentiment** (Accuracy): We measure the extent to which the sentiment is converted using the pretrained fastText classifier for the polarity of the reviews. The fastText classifier [5] achieves a high accuracy of 95.7% in determining the polarity of the review.
- **Content preservation** (BLEU): We measure the extent to which a model preserves the content present in a given input using n-gram statistics, by measuring the BLEU score [9] between generated text and the input itself.

It has been observed by [7,13] that as the BLEU score increases, the sentiment transfer accuracy decreases. As we want our model to produce sentences that have the target sentiment while preserving the content, we use the Geometric mean (G-score) of Accuracy and BLEU as an evaluation metric to evaluate the overall performance [18].

4.3 Baselines

We compare our proposed model with the following baselines:

1. **Style Embedding** [3]: In this method, the model learns a representation for the input sentence that only contains the content information after which it learns style embeddings in addition to the content representations.
2. **Multi-Decoder** [3]: This method uses a multi-decoder model with adversarial learning which uses different decoders, one for each style, to learn generation of sentences in each corresponding style.
3. **Cross-Alignment Auto-Encoder (CAE)** [12]: This model uses refined alignment of latent representations in hidden layers.
4. **Retrieval, DeleteOnly, and DeleteAndRetrieve** [7]: DeleteOnly works on extracting content words by deleting phrases associated with the sentence's original style. Retrieval works on retrieving new phrases associated with the target style. DeleteAndRetrieve is a neural model that smoothly combines the DeleteOnly and Retrieval method into a final output.

5 Results and Analysis

Table 1 compares the results of the baselines with our model on SYelp dataset where accuracy evaluates transfer of sentiment, BLEU evaluates semantic content preservation, and G-score is the geometric mean of accuracy and BLEU. Our models outperform the current state-of-the-art by a G-score of 11%.

In SYelp dataset, StyleEmbedding achieves a high BLEU score of 67.63; however, it is unable to transfer the sentiment significantly, showing a low sentiment transfer accuracy of 14.5%. MultiDecoder achieves a BLEU score of 40.22 showing sentiment transfer accuracy of 50.40%. CAE achieves a BLEU score of 20.28 and shows a sentiment transfer accuracy of 73.7%, obtaining a G-score of 38.66. Retrieval achieves a much lower BLEU score of 2.62; however, it shows a high sentiment transfer accuracy of 83.8%. DeleteOnly and DeleteAndRetrieve show competitive performances among the baselines, having BLEU scores of 37.45 and 35.55 respectively and sentiment transfer accuracy of 82.6% and 84% respectively.

Table 1. Results for the SYelp dataset.

Model	Accuracy	BLEU	G-score
StyleEmbedding [3]	14.5%	67.63	31.31
MultiDecoder [3]	50.4%	40.22	45.02
CAE [12]	73.7%	20.28	38.66
Retrieval [7]	83.8%	2.62	14.83
DeleteOnly [7]	82.6%	35.45	54.11
DeleteAndRetrieve [7]	84.0%	35.55	54.64
SentiInc w/o Sentiment Loss	74.1%	57.53	65.29
SentiInc	73.7%	59.56	**66.25**

We also compare SentiInc with and without sentiment loss. With sentiment loss, it shows the maximum G-score of 66.25, and obtains sentiment transfer accuracy of 73.7% and BLEU score of 59.56. This denotes that SentiInc produces better sentiment-accurate and content-preserved sentences than all the baselines. Without sentiment loss, we obtain a G-score of 65.29, observing a drop of 0.96 on the SYelp dataset.

The ablation study conducted on the BYelp dataset shows a 1.76 increase in G-score upon incorporation of sentiment loss. We also observe a decrease in the trade-off between BLEU and sentiment transfer accuracy with respect to the baseline without the sentiment loss. This shows a direct effect of the sentiment loss in reducing the limitations by the BLEU-accuracy trade-off, which makes the target sentences content-preserved while maintaining sentiment accuracy.

6 Conclusion and Future Work

In this paper, we focus on unpaired sentiment-to-sentiment translation and proposed our model SentiInc based on back-translation and sentiment analysis. SentiInc incorporates sentiment-based loss that enables training through only mono-sentiment data. Our experiments on review datasets (with varied maximum sentence length) show that our method substantially outperforms the state-of-the-art models in overall performance. We further show that by incorporating sentiment-loss into the back-translation based model, it is possible to decrease the limitations of the trade-off between content preservation and sentiment transfer accuracy. In the future, we would like to experiment with converting offensive text and hate-speech into polite forms and increasing the degrees of polarity in the sentiment transfer.

References

1. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: interpretable representation learning by information maximizing generative adversarial nets (2016). https://arxiv.org/abs/1606.03657
2. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
3. Fu, Z., Tan, X., Peng, N., Zhao, D., Yan, R.: Style transfer in text: exploration and evaluation. arXiv e-prints arXiv:1711.06861, November 2017
4. Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., Xing, E.P.: Toward controlled generation of text. arXiv e-prints arXiv:1703.00955, March 2017
5. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
6. Lample, G., Ott, M., Conneau, A., Denoyer, L., Ranzato, M.: Phrase-based & neural unsupervised machine translation. CoRR abs/1804.07755 (2018). http://arxiv.org/abs/1804.07755
7. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: a simple approach to sentiment and style transfer. CoRR abs/1804.06437 (2018). http://arxiv.org/abs/1804.06437
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. CoRR abs/1310.4546 (2013). http://arxiv.org/abs/1310.4546
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, pp. 311–318. Association for Computational Linguistics, Stroudsburg (2002). https://doi.org/10.3115/1073083.1073135
10. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. CoRR abs/1511.06709 (2015). http://arxiv.org/abs/1511.06709
11. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. CoRR abs/1508.07909 (2015). http://arxiv.org/abs/1508.07909

12. Shen, T., Lei, T., Barzilay, R., Jaakkola, T.: Style transfer from non-parallel text by cross-alignment. arXiv e-prints arXiv:1705.09655, May 2017
13. Subramanian, S., Lample, G., Smith, E.M., Denoyer, L., Ranzato, M., Boureau, Y.L.: Multiple-attribute text style transfer. arXiv e-prints arXiv:1811.00552, November 2018
14. Taboada, M., Brooke, J., Tofiloski, M., Voll, K.D., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011). http://dblp.uni-trier.de/db/journals/coling/coling37.html#TaboadaBTVS11
15. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification, vol. 1, pp. 1555–1565 (2014). https://doi.org/10.3115/v1/P14-1146
16. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. J. Am. Soc. Inf. Sci. Technol. **63**(1), 163–173 (2012). https://doi.org/10.1002/asi.21662
17. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. CoRR cs.LG/0212032 (2002). http://arxiv.org/abs/cs.LG/0212032
18. Xu, J., et al.: Unpaired sentiment-to-sentiment translation: a cycled reinforcement learning approach. CoRR abs/1805.05181 (2018). http://arxiv.org/abs/1805.05181
19. Zhang, Z., et al.: Style transfer as unsupervised machine translation. arXiv e-prints arXiv:1808.07894, August 2018