Group 9 Basic Working Systems

Noah Bright
University of Pittsburgh
nmb84@pitt.edu

Xiaoyan Wei
University of Pittsburgh
xiw249@pitt.edu

Annanya
University of Pittsburgh
an1@pitt.edu

Jayden Serenari
University of Pittsburgh
jserenari@pitt.edu

1 Introduction

While machine translation is a branch of NLP with a good deal of development, style transfer is a much less developed branch. One of the challenges in translating any language is the maintenance of nuance in delivery of semantic meaning, so it is important to develop models with sensitivity to style. It is also important to understand style when models tend to perform asymmetrically across styles - e.g., when models that transcribe speech perform better on male speech than female speech.

Much of the initial work in the field of style transfer begins with developing a latent representation of an input sentence, i.e. a mapping from a style-inflected sentence to a version of the sentence that retains the same semantic meaning, but with style removed. Style is then reapplied to this latent representation as desired. Empirically, it has been found that one way to develop this latent representation is through a pair of translations, one to another language, then back to the original.

In this work, we propose to combine style transfer and machine translation. The back translation approach relies on style getting lost in translation. We fine tune a model to translate a single time across languages, and produce a latent representation in the target language that maximizes the confusion of a classifier trained to recognize a given style. As a baseline, we use a latent representation created by translation to and from the target language, as has been done in the past, and then translated another time into the target language.

In the back-translation framework, it is hypothesized that using an intermediate language more distant from the original language will maximize the removal of style, possibly as the expense of semantic meaning as well. In our experiments, we use English as a source language and Japanese as the target. These languages are extremely unlike one another, so we expect style to be thoroughly washed out.

The seminal work using this approach is that of (Prabhumoye et al., 2018), in which English sentences were translated from English to French and back to English to generate a latent representation. In our approach, we propose to remove one of these translations by fine-tuning a model to translate English to Japanese without style in a single step. The fine tuning is performed in an adversarial/reinforcement learning framework. A loss function that rewards high confusion on a style classifier and similarity to the original model is optimized to create representations that are simultaneously without style and retain the original semantic meaning of the source sentence.

Due to data availability, the form of style we investigate is sentiment. This is a difficult type of style to study, because sentiment and semantic meaning are more deeply intertwined that semantics and, say political slant. The sentences, "This is great, where is it from?", and, "This is awful, where is it from?", have different sentiments and semantics. Nonetheless, the output generated from both these sentences would, ideally, simply be, "どこで手に入れた？" (where did you get this?).

The difficulty associated with the semantic meaning is only relevant after the latent representation is generated - fine-tuning to create a style-purging translator is style-agnostic. The model trained to reapply style is the cross-aligned autoencoder of (Shen et al., 2017). Due to the lack of parallel corpi with and without particular style, following an approach such as theirs is necessary. At this stage, the sentiment data is used to train the autoencoder, but in principle, any dataset can be used with whatever style. When reapplying style to the example sentence above, we expect an outcome such as, "これはおいしいよ。どこで手に入れたの？" (this is delicious, where did you get it?) for positive sentiment, and "これはまずいよ。どこで手に入れた？" (this is awful, where did you get it?),

for negative sentiment.

2 Related Work

(Prabhumoye et al., 2018) pioneered the approach of using backtranslation to create the latent representation conditioned on in the style reapplication step. In their case, the latent and style-reapplied sentences are both in English, so it is possible to train decoders to condition on the latent representation and decode, effectively on parallel text. Their intermediate language representation was in French, which is similar to English. Because of this, translation is inherently easier than in the case of English to Japanese. Language similarity is one aspect we seek to explore in this study. The back translation method of style removal effectively relies on detail getting lost in translation. For the purpose of style transfer within a single language, an interesting variable to study is what languages produce the least style-inflected latent representations, while retaining semantics. These two factors are always at odds with one another. In our work, we are forced to use the target language to produce the intermediate representation. Because English and Japanese are inherently unlike, a single translation comes with a higher semantic cost. The baseline using several translations therefore becomes unreliable, necessitating the finetuning approach.

The gold data for a project like this would be something like parallel text with and without the desired styles. This would allow for the use of a very direct reconstruction loss to optimize in training. (Shen et al., 2017) developed a cross aligned autoencoder model that is designed to transfer style without parallel text. After a rigorous statistical analysis and proof of concept, they develop and show their model outperformed the contemporary state of the art. Due to the availability of sentiment datasets, this becomes a powerful tool for expanding the range of languages that we can perform style transfer experiments in.

3 Datasets

As a source of English sentences with sentiment, the Yelp (Asghar, 2016) dataset is used. While the sentiment is not necessary for us to train our reinforcement model, it helps form a basis for those of us who can not speak Japanese. On the other hand, we do note that reviews tend to have strong sentiment and are therefore harder to train on.

For the reapplication of style step, the CHABSA dataset is used. It contains sentences from Japanese business transactions, and are labeled with the sentiment as to whether or not the transaction is discussed favorably or not.

4 Model setup

To fine tune an English-to-Japanese translation model, a reward function that maximizes a neutral and semantically correct output is designed. We use a Japanese sentiment classifier (fine tuned from BERT) with emphasis on the neutral score, and a small negative penalty for positive/negative scores.

$$Reward = C_{neutral} - \lambda * (sum(C_{pos}, C_{neg})/2)$$

The models chosen were sourced from HuggingFace, with *Helsinki-NLP/opus-tatoeba-en-ja* being the translation model and *christian-phu/bert-finetuned-japanese-sentiment* being the sentiment model. These models were perfect for our use case, and were two of very few Japanese models available.

We also had a Japanese to English model for backtranslation: *jbochi/madlad400-3b-mt*. This was a multilingual model with 3b parameters- too large for us to use for any training.

5 Results

Latent representations were evaluated for human readability on a scale of 0-4. 0 indicating a completely incoherent translation, and 4 indicating an essentially perfect translation.

5.1 Baseline Model

The average readability score among 92 sentences was 2.477. The baseline model greatly struggled with staying coherent for inputs that exceeded even a few sentences, sentences with low frequency words like "Halloween" and "vegetarian" produced incomprehensible outputs, and almost all idioms led to an overtly incorrect output.

5.2 Reinforcement learning fine-tuned model

Reinforcement learning of LLMs usually has two components: the reward function and the divergence from a reference model. This divergence has many formulas to capture the difference between two models, and we will be using KL Divergence- which is the most popular. More divergence is worse, and because KL divergence is asymmetric,

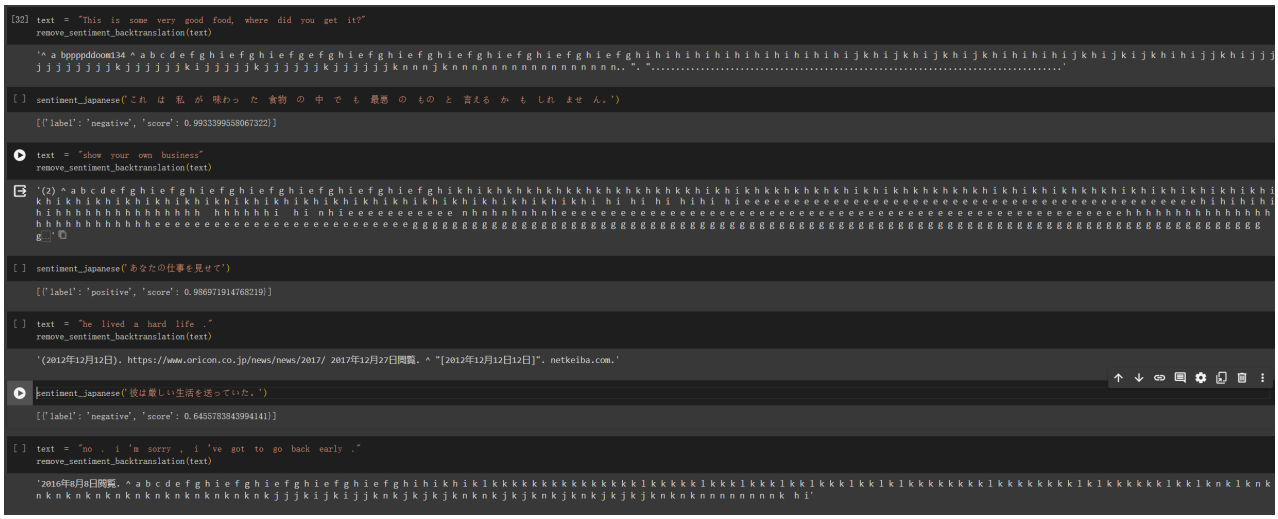

Figure 1: result

negative divergence is much worse than positive divergence.

The training of the RL model did not go well, KL Divergence steadily- nay, *rapidly* decreased into the negative number range, and scores remained steady around 0.3 or lower throughout the first epoch, and never showed any signs of increasing- and on some runs even decreasing gradually. Numerous reward functions were evaluated, but none changed the performance much.

On another note, is divergence worthwhile here? The model will **must** diverge in order to become more neutral in sentiment. But this would leave us open to the woes of reinforcement learning, in which the model finds some work around. Yes, model, "–[oooo" is *very* neutral.

5.3 Style transfer

Due to lack of available resources, the style transfer step could not be performed. Were a GPU available in a Python 2.7 enabled environment, this step could be performed using the CHABSA dataset, and conditioning on the latent representations produced by the fine-tuning.

6 Future Work

As extensions to this project, using languages more or less similar to English (or whatever source language) as the target can be explored, and the precise effect on accuracy of the latent representation quantified.

The proposed loss function uses the KL divergence to maintain semantic meaning retention of the fine-tuned translation model. A more direct measure of semantic meaning can be used in the loss function, something that more directly computes semantic similarity to the source sentence, as opposed to the weights of the model, which can only act as a proxy for similarity to the original model's performance.

7 Ethical Concerns

Depending on the training data, application of style presents an opportunity for a model to learn potentially inflammatory or stereotypical language. In the case of sentiment, if a training set contains negative sentiments where slurs are used, the model could learn that slurs are an effective way to achieve the desired style. On the other hand, if a group is represented as having a stereotypical "good" quality, the model can learn to produce positive sentiment by stereotyping that group. The training data used in the cross-alignment step must be carefully chosen for these reasons. This also extrapolates across other style areas, such as gender, political slant, and almost anything associated with identity groups.

8 Division of Labor

Noah proposed the project, wrote the drafts of the basic working systems/final report, suggested the loss function to optimize in the RL step, adapted the code for the style transfer step from (Shen et al., 2017), and evaluated translated sentences for readability/semantic loss.

Jayden handled training and usage of all of the models, and (unsuccessfully) worked on the training of the RL model. Xiaoyan helped with transla-

tion and data curation, and Annanya helped with evaluation.

References

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. *CoRR*, abs/1804.09000.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6833–6844, Red Hook, NY, USA. Curran Associates Inc.