Non-parallel Style transfer in French
Machine Learning for Natural Language Processing 2020

Amine TAZI
ENSAE
amine.tazi@ensae.fr

Mathilde KAPLOUN
ENSAE
mathilde.kaploun@ensae.fr

Abstract

Non-parallel Style transfer between french news articles and Proust's books. (See our code at [1])

1 Problem Framing

The main aim of this project was to verify if it was possible to do style transfer between french news articles and Proust's texts using non parallel data. Among other things this involved checking if these two writing styles were different enough, as this is a necessary condition to be able to transfer texts from a style to another.

2 Data Selection

We decided to use Proust's novel which were readily available in an R package [2] which is possible to import directly into Python . For the news articles, we selected articles from classical formal news publishers (Yahoo, Le Parisien, Le Figaro ...) from a dataset [3] of french news articles.

When it comes to embedding, for the classifier we used Camembert's tokenizer and french model to do the embedding. For the style transfer model, we used a regular expression based tokenizer and a limited size vocabulary embedding ($\approx$ 35000 most frequent words in Proust's texts and in the news articles) taken from a torchtext.Fasttext vocabulary embedding.

3 Experiments Protocol

3.1 Classifier

We first tried to develop a classifier (based on CamemBert (Martin et al., 2020)) able to distinguish between the two styles of texts, which are the syles of news articles and Proust's text. This classifier would help ensure that the two styles are different enough for a simple classifier to distinguish between them. It would also give us a metric to automatically evaluate our style transfer model as there seems to be no definite classical metrics to evaluate sequence to sequence models trained with non parallel data. Human evaluation is often used, but was too time-consuming in our case.

The classifier directly uses the embedding procured by CamemBert as input. We only added two classification layers linked by a ReLu function in order to fine-tune the model to our task. We retrain both these two layers and the CamemBert weights in order to produce the best results possible.

3.2 Style transfer

After that, using (Shen et al., 2017) as inspiration we worked on building a style transfer model much simpler than the one presented in the article (without adversarial training). The main idea behind the architecture of our model is an RNN based seq2seq model, where the initial hidden state of the encoder RNN depends on the the style of our entry texts, and where the initial hidden state of the decoder RNN depends on the embedding of the content of the input texts (this embedding is taken from the output hidden state of the encoder) and on the target styles of our output texts.

We finally used the classifier to evaluate our transfer model by comparing the style labels of the output texts generated by our style tranfer model to the target style labels of these outputs. We also compared these results to those of a dummy model based on synonyms.

3.3 Dummy model

The dummy model is very simple : for each word of a given sentence, we take all their synonyms

[1] https://github.com/Ellana42/NLP_Ensae
[2] https://cran.r-project.org/web/packages/proustr/index.html
[3] https://webz.io/free-datasets/french-news-articles/

from the nltk french version of wordnet [4]. We filter on part-of-speech (POS) with the help of the CamemBert POS tagger [5]. Then, of the resulting list of synonyms we take the one most common in the target style (in Proust books for instance), and we replace it in the original sentence. This keeps very well the structure of the sentence, since stopwords and very common words are not modified. However, the synonyms are often badly chosen, and even quite outlandish, since they are only based on the word and not the context. We should avec chosen an context based embedder in order to improve the quality of the results.

4 Results

The classifier we developed worked well, it had an accuracy of 93 % on our validation dataset and a Matthew's correlation coefficient of 0.9 on the test, which is really good. This could be indicative of some underlying properties of the data giving " hints " to our classifier (for e.g the size of the sentences). However, it performed well on hancrafted examples so we kept it.

The performance of the dummy model is very poor, as expected. The classifier never predicts the right class (accuracy close to 0). This was to be expected since the dummy model has a very small impact on the original sentence, so it makes sense that the classifier would still systematically classify them in their original category. That shows very clearly that our simple method doesn't outperform and therefore renders obsolete our more complex transfer model.

As for our main model we get a 35% accuracy. This is quite close to the results of (Shen et al., 2017), who got 40% accuracy. However, this result is deceiving. Our model generates sentences which emulate the style of the target, mostly through vocabulary. However, it completely fails to translate both grammatical structure and meaning. The classifier is fooled by the vocabulary, but fails to recognise the loss in meaning. In order to improve this evaluation, we should have used a model more regarding concerning meaning (by for instance adding adversarial training as described in the article)

5 Discussion/Conclusion

The simplified version of the parallel style classifier had low-quality results, which was to be expected as the author's more advanced model had mixed results in the first place. Parallel style transfert in general seems very difficult to implement properly, and this difficulty was exacerbated by our choice to use french data. All tools surrounding french NLP are both rarer and less used, which means we had a lot more difficulty learning about them and making them work than we would have otherwise. Especially for embeddings and our work with synonyms we would have liked to work with Word2Vec for instance. This was a very interesting experience, but it would be more satisfying repeating it in English, and seeing how much the results could be improved.

[4] https://www.nltk.org/howto/wordnet.html

[5] https://huggingface.co/gilf/french-camembert-postag-model

.

References

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suá rez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.