A Setting Transformer: Write in Harry Potter's Setting

Yipeng Lin
lorenlin@umich.edu

Abstract

Writing style transfer have been a heated research area in natural language processing. With the help of writing style transfer, people can mimic other people's tone when writing. Current researcher have been laying emphasize on building writing style transfer models and ways to evaluating those models. Some have look into mechanisms like encoder-decoder, deep neural networks, and machine translation (Toshevska & Gievska, 2021). However, the text style transformation works well mostly on text with strong styles. Research into setting transfer can be more helpful for altering the text style. In this project, we explore possibilities to make setting transfer come true, especially bringing *Harry Potter*'s setting into mundane writings. We have fine-tuned a BERT model with unique words that may contribute to *Harry Potter*'s setting. We then make a pipeline that can take in a sentence, lay masks on words, then use the fine-tuned BERT model to predict the masked words with *Harry Potter*'s setting. We have also evaluate on the model using a trained *Harry Potter* setting classifier and human annotation on text naturalness. Our results shows that the model can be used to transfer settings with a high naturalness, but the transferred style is not very intense. Further work can be done to improve the setting transfer pipeline. We purpose to use some techniques to identify key words in a sentence and mask those words for model to predict. This will increase the setting intensity. Also, more works could be done to look into how to preserve the text content while transfer the settings.

1 Introduction

Writing style transfer is currently used to correct users' writings. It can be further used to mimic other people's tone. Currently, researchers in this field have explored deep neural networks, encoder-decoder strategy, and machine translation (Toshevska & Gievska, 2021). Some researchers have also been research into the evaluation of text style transformation as it is hard to define the style of an author (Toshevska & Gievska, 2021). According to our research writing style transfer is only very perceivable when the author has a strong writing style like Shakespeare or Jin Yong. It would be much easier to identify a writing style if the setting of a specific book can be transferred.

In this project, a model will be trained to transform the mundane writing into writing that is based on setting in the magic world of world famous *Harry Potter* written by J.K. Rowling. Words and phrases will be replaced with those in the book to make the normal story happen in the magic world.

In this project, unique words are firstly be extracted from *Harry Potter* to be set as learning target. Then the data is fed into a BERT model to do the fine tune. Finally, the words in predict data are randomly replaced by output from BERT model. The performance is evaluated by a classifier to identify whether the style has been learnt and human evaluation to test if the transformed sentences can be read naturally.

The setting intensity classifier model we trained have very high F1 score and macro-F1 score to be 0.97 and 0.97 on a data set containing half the *Harry Potter* text and half the normal texts. It also runs over baselines we set which are randomly guessing and count unique words from *Harry Potter* text in it.

The setting transfer model also provide a not very bad performance on transforming the settings. Even though it doesn't proceed the baseline model in the setting transfer intensity, it works much more better in making the text readable comparing to the baseline model. Also, the setting transfer intensity of the setting transfer model is not that low, which is that 3 out of 10 sentences can be identified from *Harry Potter* settings.

This project can help people transform their writing into a different background settings, to make it more attractive. By doing so, writers or designers can gain inspirations from transforming texts from daily life to a new settings. Also, this project is fun and can be appealing to *Harry Potter* fans.

From this project, readers can get to know some approaches to make setting transformation come true. They can also learn from this project about what they can further complete to improve setting transformation quality.

2 Data

2.1 Main Train Data

J.K. Rowling's *Harry Potter* will be used in this task as the target style data. To be specific, the total of 7 *Harry Potter* book written by J.K. Rowling, excluding *the Cursed Child*, will be used. In total, there are 95678 sentences and 25715 unique words.

The source data is obtained from a GitHub repository ("Harry Potter source file", n.d.). We download the data and combine 7 books into a single txt file. The text are firstly tokenized into sentences. Then each sentence is tokenized into words. So that it can be fit into Bert model. No other prepossessing will be needed for the data.

2.2 Frequency Compare Data

To extract unique words in *Harry Potter* settings, word frequencies in it will be compared to ordinary word frequencies. The ordinary word frequencies are calculated by University of Leeds based on an Internet corpus in 2005 containing around 160 million words (of Leeds, n.d.). The frequency data contains 333333 words together with their frequency. In this project, we uses the top frequent 20000 words as common words. After comparing with the unique words in Harry Potter, 12675 words are identified as unique words that contributes to Harry Potter's settings.

2.3 Ordinary Text Data

To train the classifier, a text data set with realistic descriptions will be needed to train with *Harry Potter* text. We choose to use an amazon reviews data set that contains 400000 amazon reviews(face, n.d.). As reviews contains mostly things people talk about in daily life, it would fit well to be distinguished with fantasy literature settings. In this project, we used the first 50000 reviews to balance the ratio of sentences from Harry Potter and reviews.

3 Related Work

3.1 A Review of Text Style Transfer using Deep Learning

This paper summarizes and reviews the field of the text style transfer. It describes this field from multiple aspects, from linguistic background to data sets, evaluation methods, and possible models to use. It has a very detailed classification of text style transfer task types (Toshevska & Gievska, 2021).

3.2 Sequence to Sequence Learning with Neural Networks

In Ilya et als' paper, they describe a kind of neural network that can take in a sequence of words and output another sequence of words using encoder and decoder structure with long short-term memory (Sutskever et al., 2014). The sample it uses is the task of translating English into French, which is a different task. However, the structure of the network could be referenced and adjusted to be implemented in this project. The training workload of that network is also very heavy, I will try to find some way to improve the training procedure.

3.3 Contextual Text Style Transfer

In Chen et als' paper, they describe a method to transform sentences into a desired style without changing the semantic meaning of it (Cheng et al., 2020). The way they construct the context-aware model to receive non-parallel model is refreshing, it provides me with thoughts on overcoming difficulties to get parallel corpus of *Harry Potter* as it is originally written in modern English and doesn't really has a style in writing but a style in settings.

3.4 Revisiting few-sample BERT Fine-tuning

In this paper, authors show some sample on BERT fine-tuning and describe some problems when fine-tuning BERT (Zhang et al., 2020). They have also given some suggestions and practices to solve the instability in fine-tuning. This should be very helpful to this project as fine-tuning will also be used in this project. I'm also concerning about the data size of my project.

3.5 Evaluating Style Transfer for Text

Authors of this paper proposed three dimensions to evaluate a text style transfer task: style transfer intensity, content preservation, and naturalness. They conducted evaluation between automated evaluation methods and human evaluation methods (Mir et al., 2019). Their work provide me with what kinds of evaluation method I can use to evaluation my model.

4 Methodology

4.1 Extract Unique Words

First thing first, a data analysis on what are the unique word that corresponding to *Harry Potter* settings is conducted. The word frequency in *Harry Potter* is compared to modern language word frequency extracted from the Internet according to University of Leeds (of Leeds, n.d.). As has mentioned in the data part, we selected the most frequent 20000 words as common words. Then for each word in *Harry Potter*, we check if those words are in the 20000 top normal word frequency. If the word frequency from *Harry Potter* falls outside the scale, it will be considered to be unique word that contribute to the *Harry Potter* settings. Then those unique words are stored in a set for quick lookup.

4.2 Setting Transfer

4.2.1 Model

A pre-trained BERT model, distilbert-base-uncased[1] from Hugging Face, is fine-tuned to make its prediction more biased to the unique words under *Harry Potter* settings. The model is chosen because its smaller and faster comparing to the full BERT model.

4.3 Data processing

For data input, it will take in sentences with masks. All unique words that is figured out in the previous part are masked. Then the sentences with masks are converted to token ids and fed into BERT model to adjust and learn the masked words in the context by the BERT model.

4.4 Training

Here we use Trainer[2] library from Hugging Face to automatically fine-tune the model. We set the trainer to have batch size 32 and other parameters are default. The model is trained for 5 epochs.

4.4.1 Text transform

When putting into usage, the input sentence will be tokenized again. By setting the replace probability, masks will be put on words randomly. Note that masks won't be masked next to each other. This is to prevent generating weird words with combined affix. The sentences are then predicted by the BERT model to fill in the gap. From the output words, another parameter is set to select randomly from the most possible words from the BERT model output. The involvement of this parameter can add to the variety of the predicted words. By doing so, the context can be preserved while making the

[1] https://huggingface.co/distilbert-base-uncased

[2] https://huggingface.co/docs/transformers/main_classes/trainer

sentence look like it happens in *Harry Potter* settings.

4.5 Classifier

When evaluation, a classifier is used to judge if the style transfer works. It takes in a sentence and outputs a binary value judging whether the text is under the settings of *Harry Potter*. This is achieved by labeling sentences from *Harry Potter* as 1 and labeling ordinary sentences as 0.

4.5.1 Model

A pre-trained BERT model, microsoft/MiniLM-L12-H384-uncased[3], is utilized to do the text classification. The reason we choose this model is because this one is tested in previous works to show high accuracy for text classification as well as a rather small size, making it fast to train.

4.5.2 Data setup

To train the classifier, we extract sentences from both *Harry Potter* and Amazon review data set. 50000 sentences are collected from Amazon review data set. The data then are combined with data from *Harry Potter* and labeled with 0 if it's from amazon review data set and 1 if it's from *Harry Potter* data set. From the mixed data set, I extract 20000 lines of data in total, with half and half ratio of either data source, to keep the balance of 0 and 1 data.

4.5.3 Data split

Then the 20000 lines of data are split into training set, test set and development set. Test set and development set each occupies 10% of the 20000 lines, which are 2000 lines each. The rest are set as train set, which is 16000 lines.

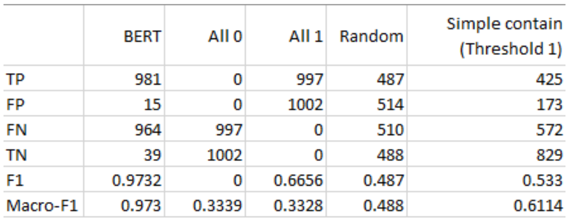
4.5.4 Training

We also use Trainer[4] library from Hugging Face here. We load the pre-trained BERT model as BertModelForSequenceClassification. We set the trainer to have batch size 8, and evaluate every 500 steps. The model is trained for 3 epochs.

5 Evaluation and Results

5.1 Classifier

Firstly, we should take a look at the classifier performances. Here we have three baselines. The first one is to identify all as 0 or 1. The second one is to randomly select between 0 and 1. The third one is to utilize the unique word for *Harry Potter* settings we derived from previous part, if sentence contains more than certain amount of those words, it will be counted as from *Harry Potter* settings. From the figure, it can be seen that when the threshold when there are more then threshold number of words from unique words sets in the sentence, the macro-f1 score is the highest. So we take 1 word as the threshold.



	BERT	All 0	All 1	Random	Simple contain (Threshold 1)
TP	981	0	997	487	425
FP	15	0	1002	514	173
FN	964	997	0	510	572
TN	39	1002	0	488	829
F1	0.9732	0	0.6656	0.487	0.533
Macro-F1	0.973	0.3339	0.3328	0.488	0.6114

Figure 1: Baseline scores

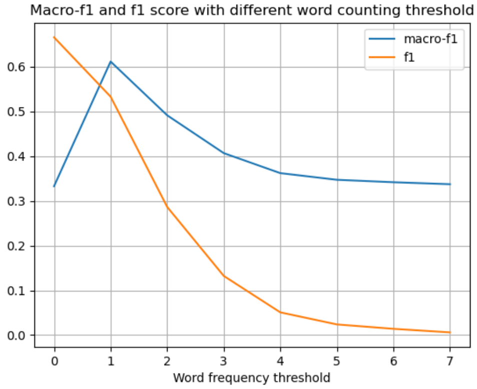


Figure 2: Simple countain scores with different threshold

The experiments on different baselines indicates that our classifier model works much more better then those baselines and the accuracy of

[3] https://huggingface.co/microsoft/MiniLM-L12-H384-uncased

[4] https://huggingface.co/docs/transformers/main_classes/trainer

the classifier model is able to be used for later evaluation.

5.2 Setting Transfer

Learnt from Remi et als' paper, we plan to evaluate model from three perspectives: style transfer intensity, content preservation, and naturalness (Mir et al., 2019). However, the content preservation isn't very applicable here as our model mainly focus on transforming the settings, the content can be altered significantly.

5.2.1 Setting transfer intensity

For setting transfer intensity, the setting classification model is used to take input of texts and output of binary result whether it resembles the target style and contain enough element from the *Harry Potter* settings. The classification model we trained, as mentioned in previous part, have 0.97 F1 score. With this model, we tested the setting transfer model with different word replacing probabilities(for each single word, it will be of this probability to be turned into a mask). The data set we test on is from Amazon reviews. We take 100 lines of reviews, Note that there won't be two masks set side by side to prevent affix issues as mentioned in Method part, this means even if the replacing probability of each word is 1, the total replaced words in the sentence will only count up to 50%.

We set our baseline to be randomly replacing words in a sentence to be randomly chosen words from *Harry Potter*'s unique words.

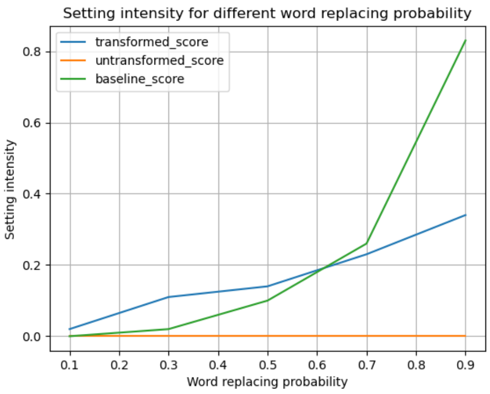


Figure 3: Setting intensity

Here the setting intensity is calculated as the ratio of sentences identified as from *Harry Potter* settings among total sentences input.

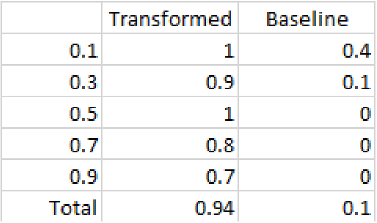
It can be seen from the figure that the setting transfer model can actually transfer the settings. The more words replaced, the higher the setting transfer intensity is.

However, the setting transfer model score are relatively lower comparing to the baseline model. This means our model doesn't have better performance comparing to randomly replacing words. We will discuss this in discussion part.

5.2.2 Naturalness

We conducted human evaluation on output naturalness. We labeled text as 1 if the text can be read, understood, and doesn't contain grammar mistakes and 0 if the text can't be understood or contain grammar mistakes.

We labeled 50 reviews on both transformed texts and baseline texts, 10 from each replacing frequency in 0.1, 0.3, 0.5, 0.7, 0.9. The result gives:



	Transformed	Baseline
0.1	1	0.4
0.3	0.9	0.1
0.5	1	0
0.7	0.8	0
0.9	0.7	0
Total	0.94	0.1

Figure 4: Naturalness scores

The score is calculate as the ratio of labeled as natural among total number.

It can be seen that the setting transform model provides more natural outputs of the text.

5.2.3 Results

Please see Appendix for detailed results.

6 Discussion

6.1 Classifier

The classifier works perfectly well in this case. It has 0.97 F1 score and 0.97 Macro-F1 score. These high score indicates that the model can classify texts very accurately, making it able to distinguish between texts from *Harry Potter* settings and Amazon reviews. It also proceed the baseline of randomly guessing labels and doing counting of unique words from *Harry Potter*. Both of these have only 0.488 and 0.6114 Macro-F1 score. The performance of the classifier far exceed what we expected. We originally expect the classifier to have a F1-score around 0.8. We are kind of doubting why it has such a high score. We have the following guesses:

1. As we train and test the classifier on amazon reviews, and we haven't test it on other data set against texts from *Harry Potter*, it is hard to say whether it learnt the settings in *Harry Potter* or text styles from the amazon reviews.

2. Amazon reviews also have some kind of pattern or style in it, making it easier for classifier to identify.

3. The model may learn the word frequency in *Harry Potter* to distinguish between two data source.

6.2 Setting transfer model

6.2.1 Setting transfer intensity

Here we define the setting intensity to be the ratio of sentences identified as from *Harry Potter* settings among all sentences. By looking at this score, we can know how well the model can transfer the settings.

If we only look at the model itself, we can see that with the increased replacing probability, the model can generate contents that are more resembling the settings from *Harry Potter*. When the replacing probability goes up to 0.9, the setting intensity can goes up to 0.33, meaning 3 out of 10 sentences are judged as from *Harry Potter* settings, which is rather good according to our expectations.

However, when we get the results from the baseline, from which we randomly replace words with random words from *Harry Potter*'s unique word list, it turns out the baseline generates better contents that resembles the *Harry Potter* settings. This makes us think whether something goes wrong with the classifier model or the setting transfer model is not working very well. We think they both are contributing to the result.

As for the classifier model, it can just distinguish whether text is from *Harry Potter*, it doesn't care about who well the text contents are or if the text is readable. It could possibly find out some specific words that exists in *Harry Potter* settings or identified some patterns from the text. For baseline, it replace more words under the same replacing rate as we restricted the setting transfer model to not place masks on two neighbor words. And the classifier model may consider the baseline text has more words from *Harry Potter* and give it a higher score.

As for the setting transfer model, we consider there are problems in the pipeline of converting the settings. Currently, we randomly choose word to lay masks on them. Not only will the cover up some important information the should be preserved in the text, but also not helpful to transfer the setting is masks are laid on some connection words, which doesn't count towards *Harry Potter* settings.

6.2.2 Naturalness

However, just the setting intensity score won't count. We also need to look at the naturalness score. We can see that both model have naturalness goes down with increasing replacing probability. This is understandable as the more text it replaced, the more possible that the text structure and logic are broken. Comparing the setting transfer model and the baseline, it can be seen that our setting transfer model works much more better than the baseline. The baseline model's text start to not making any sense when replacing probability goes to 0.3. In the mean time, the setting transfer model still has very high understandable rate to be 0.7 even the replacing rate is 0.9. This should be thankful to the BERT model. It helps prevent placing words that are not logical or has grammar mistakes into the masks.

6.2.3 Conclusion

In all, the setting style transfer model is still usable as for the task to transfer the settings. Even though the model doesn't perform well on setting transfer intensity comparing to baseline, it outperform the baseline in naturalness score. Both evaluations are important to the model and it is necessary to keep the balance of these two aspects.

7 Conclusion

In this project, we explore possibilities to conduct setting transformation on mundane texts to *Harry Potter* settings. We trained two models to achieve this, which are a setting transfer model and a setting classifier. We build a setting transfer pipeline to embed the setting transfer model to conduct setting transformation. We then evaluate the setting transfer pipeline with the classifier we trained to test the setting transfer intensity and also with human annotation to test the naturalness of the transferred text.

Even though the classifier we trained performed well, the performance of the text setting transfer model is not that satisfying. It is lacked behind the baseline when it comes to setting transfer intensity. However, the current setting transfer model can generate much more natural sentences than the baseline, which makes the model usable for our audiences.

In the future, researchers can explore more on different pipelines to do text setting transformation as our pipeline is rather arbitrary. Also, researchers can also inspect on how to preserve the meanings of the sentence while doing the setting transformation, which is rather challenging.

The code to this project can be found in this Github repository [5].

8 Other Things We Tried

Initially, we tried to do text style transformation. We tried to transform text according to J.K. Rowling's writing styles. We look into many papers to figure the whole process out. We learnt that parallel corpus can be obtained by using the translation, which is using existing translation model to translate text into another language and than translate it back to obtain a text without any writing style. We implemented the function to use google translate API to translate text from *Harry Potter* to any specific language and translate them back. When we notice the speed of the API is not very fast, we further implemented a multi-thread function to accelerate the process. After all these works, we realized that J.K. Rowling's style is not very strong and it is hard to distinguish text comparing to Shakespeare and Jin Yong. So we cease exploring the style transfer model and move to research into the setting transfer model.

9 What We Would Have Done Differently

9.1 Unique word generation

In our implementation, we simply take the top 20000 words from the most frequent words and take the words in *Harry Potter* if the words are not in the top 20000 frequent words. However, this is actually not proper as a word is rare doesn't mean that word should be count as word that contribute to a setting. Instead, a better way to do this will be analyze each word frequency in the book and word frequency list. Words that appear more often or less often should be counted as unique words. The judgement can be using standard deviation.

9.2 Classifier training data

Currently, the classifier is trained on Amazon review data and *Harry Potter*. Just by using Amazon review data as normal text would cause problem that the model may be trying to identify patterns from Amazon reviews. We think those reviews also have a style in it and can affect the classification result and are not generalized to be applied to other sentences. To address this, data from different data set can be combined into training data, for example, email data set,

[5] https://github.com/10RE/SI630-repo/

article data set, conversation data set. This will reduce the effect of style in Amazon review.

9.3 New setting transfer pipeline

Due to the time issue, we only implemented a simple setting transfer pipeline, which is randomly apply mask on words and replacing them with random top frequent words from the model. This can result in a bad setting transfer performance. If the mask is put on a connection words like "to", the model will predict "to" as well, and there is no setting transfer. Also, if the mask is laid on some keywords in the sentence, then the content meaning of the sentence would be changed. If we have more time, we would try to analyze the sentence first and figure out which words can be laid mask on and then do the setting transformation. Other approach can also work, our work here can serve as a baseline model for further implementation.

References

Cheng, Y., Gan, Z., Zhang, Y., Elachqar, O., Li, D., & Liu, J. (2020). Contextual text style transfer. *CoRR*, *abs/2005.00136*. https://arxiv.org/abs/2005.00136

face, H. (n.d.). Ag_news. https://huggingface.co/datasets/ag_news

Harry potter source file. (n.d.). https://github.com/neelk07/neelkothari/tree/master/blog/static/data/text

Mir, R., Felbo, B., Obradovich, N., & Rahwan, I. (2019). Evaluating style transfer for text.

of Leeds, U. (n.d.). Word frequency on internet. http://corpus.leeds.ac.uk/internet.html

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks.

Toshevska, M., & Gievska, S. (2021). A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, 1–1. https://doi.org/10.1109/tai.2021.3115992

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2020). Revisiting few-sample BERT fine-tuning. *CoRR*, *abs/2006.05987*. https://arxiv.org/abs/2006.05987

10 Appendix

10.1 Results

Here are some examples under different replacing frequency:

Original text One of the best game music soundtracks - for a game I didn't really play: Despite the fact that I have only played a small portion of the game, the music I heard (plus the connection to Chrono Trigger which was great as well) led me to purchase the soundtrack, and it remains one of my favorite albums.

1. 0.1

   Transformed text one of the best game musics - for a game you didn ' t really play : despite the fact that i have only played a small portion during the game , the music i heard (plus the roar to chrono trigger which was great as well) led me onto purchase the soundtrack , and it remains one of my favorite albums .

   Baseline One of the best game music soundtracks - for a doxys I didn't really play: Despite the crumpet that I have ferrets played a small portion of the game, the music I heard (plus the connection to Chrono Trigger which was great as winking led me to purchase puppeteers soundtrack, transfiguring it remains one of my favorite albums.

2. 0.3

   Transformed text one of the best game music soundtracks - a game i didn ' t really love : despite the fact that i personally only played a demo portion of the game , the screams i heard (plus the intro to chrono trigger blasts was great as well) led eagerly to purchase theophone , and it remains one of your favorite soundtracks .

   Baseline One cardigan felicis best game abruptly droplets - proprietor a dogged savin didn't delirious play: Despite the fact that I worryingly only reedy strung small portion of the game, the music I heard (plus the connection to Chrono lethally which was great as well) transylvania me to averted insinuations soundtrack, and iscent remains one coldly my favorite albums.

3. 0.5

   Transformed text one amongst the best rune music soundtracks - for a werewolf i didn ' t wanna play : despite the restriction that i had only sold a sizable portion of the soundtrack , the wand i craved (plus the connection to chronoquist which was confiscated as parchment) led me to purchase the recorder , and thus remains reminiscent of my favorite albums .

   Baseline salazar blemishes the best keener anted flapping impregnated for outbursts game mimsy gnashing eighteenth chinned Despite purr agleam that matured have only played a steamy twitching of couktve abouts the music I heard splay the connection to Chrono Trigger madder victorious kinder as clapped led me to purchase the anticlimax and it remains one drawstring sniggered favorite embarked.

4. 0.7

   Transformed text one of the best rune music re - is a potion i didn ' t wanna play : despite the fact that i have only mastered a tiny portion to the potion , the music i heard (the connection to muriel trigger twitching was great as possible) led me the purchase the soundtrack , and viktor remains worthy of my favorite curses .

   Baseline One of whiling withering repressive porion attest ceilinged dribbled beribboned absences youngsters despard really dred Despite bunnies hedges signalled I seater gernumbli blueish shamefacedly small superstitions snuffed the moleskin raining marginally wizardishness realizes fiendfyre intently midnightblue strangles Chrono surmised spurt satisfactorily lingerec crumble hopped culminating soften supremely ensured deflate whirled squire it remains skulked of gendy favorite shoving.

5. 0.9

   Transformed text one of my best trance music meta - wrote a instrument i cannot really play : cursing the fact that i could only mastered a demo portion of that game , haunting music he heard (the wand to summon trigger twitch was described as gunfire) led gaga to recapture the mandolin , and it remains worthy of her favorite albums .

   Baseline marshaled polyjuice miscalculated eyewitness subsuming celestina soundtracks - expec unconcerned banish flagstones agonising hammocks airy paused redoubled cleaved shred semiconscious antidote embargo banquets disservice feelin wispy covertly the perpetrate the determinedly hauled mended tarnished vided longbow cealed glitters enchantingly toughest napkin keening plinths serpents angrily stupidest brutality churlish fogging straighten weirdest insignificant bartemius moanin buckbeak.