Towards Text to Audio Machine Translation

Hargen Zheng
Halıcıoğlu Data Science Institute
University of California, San Diego
San Diego, CA 92092
yoz018@ucsd.edu

Nathaniel del Rosario
Halıcıoğlu Data Science Institute
University of California, San Diego
San Diego, CA 92092
nadelrosario@ucsd.edu

Chuong Nguyen
Computer Science and Engineering Dept.
University of California, San Diego
San Diego, CA 92092
chn021o@ucsd.edu

Ziyue Liu
Electrical and Computer Engineering Dept.
University of California, San Diego
San Diego, CA 92092
zil085@ucsd.edu

Adam Tran
Department of Mathematics
University of California, San Diego
San Diego, CA 92092
ant010@ucsd.edu

Abstract

TODO. Will add after we have obtained all model results and drawn more insights from the project.

1 Introduction

With the globalization of the modern world, the problem of machine language translation has become all the more relevant. In order to establish clear communication with foreign partners it is important to have reliable and efficient translation. Furthermore, as auditory creatures, conversion of text to audio is important. In an ever so connected digital world, many situations exist where language translation is very powerful such as news outlets, social media, and academic texts online.

We propose a sequence to sequence based translation model for language translation using multiple baseline models including Google's T5 small and normal as well as incorporating Bark, a model capable of doing text-to-audio translation as well. Additionally, we attempt a naive approach for a baseline with an ordinary encoder-decoder architecture for the sequence to sequence task. This involves tokenize words and embedding their meanings to later be decoded through a transformer. The next method involves transfer learning with pretrained models as well as retraining layers of the model on Mandarin since they were not trained on the language before. We expand on this transfer learning by further incorporating knowledge distillation, which is the process of transferring knowledge / learned material from a larger model to a smaller model. These procedures are especially important given the memory usage limit constraint for this project.

2 Related Work

Before the emergence of translation using Large Transformer Models, various models and algorithms were employed by tech companies and researchers to address text-to-text translation. Among these approaches is the Phrase-Based Statistical Machine Translation (PBSMT) technique [19]. While

Preprint. Under review.

these algorithms varied, they shared a common characteristic: many earlier machine learning techniques, including PBSMT, attempted text-to-text translation through literal word-for-word translation without considering contextual meanings or positional contexts. This approach was soon recognized as ineffective and often led to mistranslations.

Figure 1: Bad Translation

The above figure is a prime example of computer mistranslating text from a language to another. From the above figure, the computer was supposed to translate a text in Japanese, which read as "It's dangerous, so don't go in". Yet, the computer translated to "Because you are dangerous, you must not enter".

Translating text from one language to another poses significant challenges, and it is crucial not to underestimate the complexity of this task. The reason for this complexity lies in the fact that words can have multiple meanings, and the meaning of a word can vary depending on the context in which it is used. For instance, consider the word "rose" in English, which can be employed in different contexts. In one sentence, it may be used as a noun to refer to a flower, as in "I am gifting rose flowers to my girlfriend on Valentine's Day." In another sentence, it may function as a verb to describe an action, as in "I quickly rose from my seat to greet my girlfriend arriving from San Diego." Despite both sentences containing the word "rose," its meaning differs based on the context in which it is used. Such variations in meaning are common in English due to the presence of words with multiple meanings. However, this may not be the case in other languages. When translating text word-for-word without considering the context, it is possible for the computer to select the correct words but convey the incorrect message or meaning intended by the user. This can lead to machine translation systems like Phrase-Based Statistical Machine Translation (PBSMT) producing inaccurate and misleading translations frequently and rapidly [9][17][1].

In response to the drawbacks associated with word-for-word translation, researchers have initiated investigations into methods aimed at integrating sentence-level context into translation models. One such approach involves the utilization of Long Short Term Memory (LSTM) [11][5], which employs activation gates to discern which preceding information should be retained for future reference and which should be discarded. This model effectively mitigates the memory and context limitations of earlier models. However, it comes with the drawback of requiring extensive memory storage and exhibiting slow processing speeds. Another variant of LSTM, known as the Neural Turing Machine, endeavors to store previous context or information in an external memory; nevertheless, this approach remains inefficient in terms of computational time and memory usage. To tackle this challenge, researchers have proposed the incorporation of attention mechanisms [15], which prioritize keywords that contribute most to the sentence context while disregarding stop words such as "is," "are," "or," and "and". This introduction of "Attention is All You Need" gave way to the rapid rise of Transformers. One of those was Long Short Term Masking Transformer (LSTMT) [16]. This approach considers not only the positional context and word-to-word relationships but also the preceding word context, enabling more accurate prediction and alignment with texts in different languages. The effectiveness of LSTMT has been demonstrated in its ability to handle polysemous words, preserve original meanings, and facilitate text-to-text translation tasks. Given the success of Transformer in natural language processing, particularly in text-to-text translation, researchers are motivated to further enhance the transformer architecture by increasing the number of

layers [16]. Advancements in GPU and TPU technology, along with the availability of extensive and robust datasets, have significantly simplified the training of large and deep Transformer models. The capacity to expand architectures like the Transformer with additional layers enables the extraction and learning of more features from text inputs, resulting in highly accurate translation of text to another language. So much so that the error values made by these Deep layer Transformer models match or even outperform the translations done by humans.

While Deep Layer Transformers boast high performance accuracy, they typically consist of billions or even trillions of parameters that need to be trained. This necessitates infinitely large datasets for training. Moreover, due to the sheer number of parameters, it is impractical to construct and train these Transformer models from scratch on a standard computer. The high computing unit requirements and the need for extensive dataset collections come with a hefty price tag for building and maintaining the model. As a result, only a handful of organizations possess the computing and financial resources to meet these demands, thereby limiting access to a small community. In order to mitigate this inequality, Google introduced a transfer learning model called T5 [7], which demonstrates robustness and proficiency across various NLP tasks, including text-to-text translation. The T5 model has exhibited exceptional performance, often surpassing baseline models trained specifically for NLP tasks, with minimal additional training data required for self-supervised pre-training and fine-tuning [8]. Therefore, we have opted for T5 small and base models as our primary architecture for training and conducting text-to-text translation. We believe that with minor adjustments and additional training, these models can achieve state-of-the-art accuracy in text-to-text translation.

Yet, text-to-text translation is only the first half of our project. We believe that text-to-text translation is wonderful idea, but we believe that we can create a more useful tool for everyone by combining text-to-text translation with text-to-audio conversion.

With the globalization of the modern world, the problem of machine language translation has become all the more relevant. In order to establish clear communication with foreign partners it is important to have reliable and efficient translation. Furthermore, as auditory creatures, conversion of text to audio is important. Many situations exist where text to audio is applicable such as with navigation systems, audio books, and situations that require accommodations in accessibility. Broadly speaking, the combination of Language Translation and Text to Audio Translation can have large, positive impacts within the areas of business, tourism, and media.

For text-to-audio conversion we plan to apply Non-Autoregressive Vector Quantized Variational Autoencoder (VQ-VAE) Model and an Auto-Regressive Transformer model [6] which has been shown to outperform conventional Transformer Text-to-Speech model.

3 Dataset

Google T5 model is trained on 4 languages – English, French, Romanian, German, but not Chinese. Therefore, the model has learned a lot of useful representation of the English language. We want to apply the model, including our baseline from scratch model to take in English as a source language and output Chinese as a target language, so we look for datasets that has English-Chinese language pairs as training data.

At this point, we have two datasets in deposit. We plan to debug the training loop and get the code running with the smaller news commentary dataset [14], which has $69.2k$ English-Chinese language pairs. By inspection, the dataset has pretty good English-Chinese language pairs and we believe it is a moderate size dataset to check if our model is learning. Additionally, since the dataset does not come with a validation or test set, we perform $90 - 10$ train-test split in order to evaluate our model performance without overfitting to the training dataset. We chose to start working on the project with a smaller-sized dataset because larger dataset would take longer to train and thus taking much longer time to debug our code.

After we have all code running on the smaller dataset, we plan to scale up to a much larger dataset – wmt19 [3]. Since the dataset is loaded from Hugging Face, it would be easy to scale up with another larger dataset and train the model with the code we have. The dataset has $19M$ training language pairs and $3.98k$ validation language pairs. This much larger dataset would, hopefully, enable our model to learn much more useful hidden, or latent, representations of the input text and thus helping the decoder network better understand the context and do the translation job.

4 Methods

4.1 Baseline from Scratch

A naive yet traditional approach to machine translation task is by leveraging the power of sequence-to-sequence model as shown below:

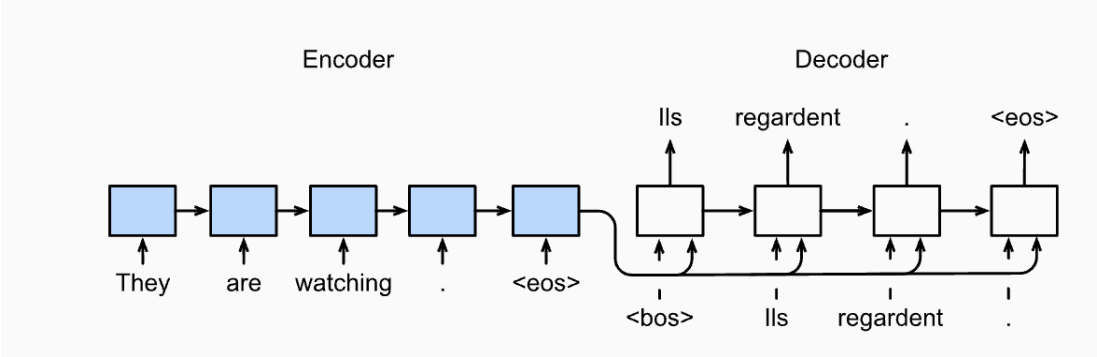

Figure 2: Sequence-to-Sequence Model Architecture [20]

With the input text from source English text, we tokenize the words with `spacy` tokenizer. Then, we embed the tokenized input into high dimensional embedding space - more hyperparameter choices would be detailed in the final report. Similarly, we tokenize Chinese text and embed it into the same high dimensional embedding space. With these, we train a transformers model that has a encoder block and a decoder block to generate neural machine translation sequence. A linear layer is followed by the last layer of decoder network – the feedforward layer expands the dimensions of output layer into number of words in the Chinese language after translation. Positional encoding is incorporated into the Transformer architecture to capture long-term dependencies inherent in the given pair of texts – a feature that is especially relevant for tasks like machine translation, as S(subject)-O(object)-V(verb) sequence might be very different from one langauge to another. By capturing the long-term dependencies within the input English text, we expect our model to generate much better output in the target Chinese translation.

4.2 Google T5 Model Transfer Learning

The baseline structure of the Google T5 model comprises a standard transformer model with both encoder and decoder stacks.

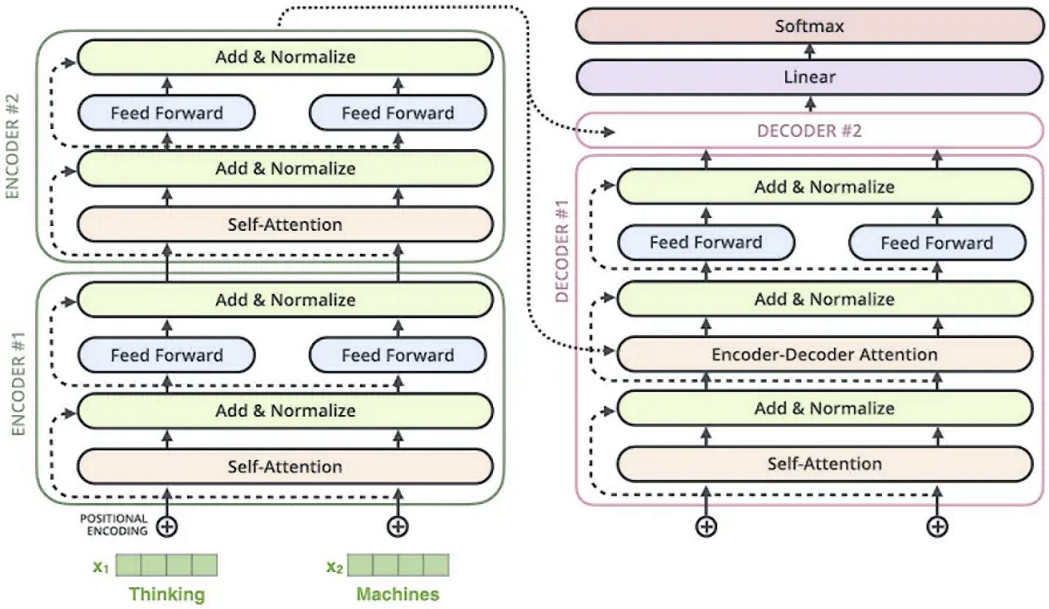

Figure 3: T5 architecture [10]

The encoder follows a structure similar to Bert, featuring 12 blocks, each consisting of self-attention, optional encoder-decoder attention, and a feed-forward network, all with a sub-layer dimensionality of 768. Similarly, the decoder mirrors the encoder's structure, with the addition of a self-attention layer enabling attention to past output. Consequently, the Google T5 model boasts a total of 220 million parameters. There is a smaller version of the model with only one-fourth of the parameters called T5 Google Small. It has 12 layers, and each stack has only 6 layers of transformers with an output sub-layer dimensionality of 512.

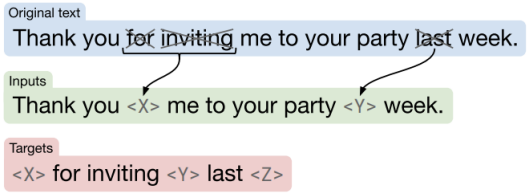

Figure 4: Unsuperised denoising Objective [10]

As demonstrated in Figure 3, The Google T5 models were trained using an unsupervised denoising training techniques, where random words were dropped, and the model was tasked with correctly recovering the missing word in the sentence. The model was fine-tuned for machine translation from English to German, French, and Romanian by concatenating datasets in English and the target language and performing the denoising unsupervised learning as described above.

Inspired by a GLUE (General Language Understanding Evaluation) score of 83.8, we were motivated to delve deeper into the capabilities of the Google T5 model in understanding two additional languages - Chinese and English. In contrast to Romanian and French, and German and English, these languages share more similarities with each other and with our particular task. Thus, this is a good way to

explore both its Google T5 baseline and Google T5 small's ability to perform machine transformation as well as finishing task of Chinese-to-English translation.

4.3 Knowledge Distillation

Besides basic fine-tuning of hyperparameters such as learning rate and dimension of embedding space, we also want to leverage the state-of-the-art machine translation model to provide a ceiling of performance on our dataset. Also, published by facebook, now Meta, in 2022, No Language Left Behind model (NLLB-200) allows for single sentence translation among 200 languages, including both English and Chinese [2]. We want to apply techniques such that our smaller model could learn from this state-of-the-art model and obtain a better performance in the neural machine translation task. Since we have memory constraint, we decided to use the model with 600 million parameters, instead of 1.3 billion parameters one. This would be feasible given the computational resources and GPU memory that are available to us.

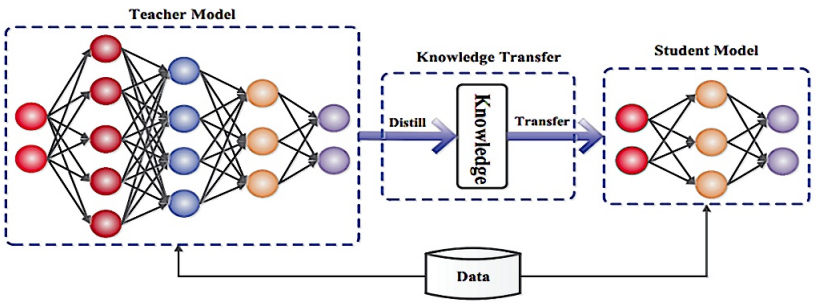

Figure 5: The teacher-student framework for knowledge distillation [4]

To achieve that, we apply a technique called knowledge distillation. Knowledge distillation enables us to transfer knowledge from a large model to a smaller one because we believe larger model is overparametrized in a way that it has more knowledge capacity than the task it serve as a result of the large number of model parameters. We believe NLLB-200 could be a good Teacher Model, that, through knowledge distillation, could transfer the knowledge of performing neural machine translation from English text to Chinese text to our baseline models (baseline from scratch and Google T5). As a result, with a smaller model, similar to the result of pruning connections through iterative pruning, we expect our model to achieve a higher performance in the translation task.

Due to the time constraint, we will not implement customized knowledge distillation from scratch. KD-Lib [12] enables us to easily perform knowledge distillation through the library, thus significantly reduced the amount of time we need for implementation. Moreover, we might also look into TextBrewer [18] – another knowledge distillation library we can leverage to ease the implementation.

4.4 Text-to-Audio

Originally, we wanted to train a separate network in parallel that is capable of generating audio in Chinese given Chinese characters. However, the inference time for audio generating takes much longer than we expected and thus makes it infeasible to train a network or even fine-tune through transfer learning, given the amount of time we have for the project period. To still enable this feature to output audio of target language given input text in source language, we leverage the power of the `bark` model that was released in 2023 by Suno [13], which supports Chinese inputs. As a result, we would use the `bark` model to transform the translated text into audio form.

5 Results

TODO

6 Discussion

TODO

7 Further Research

TODO

8 Contribution

To be summarized towards the end of the project.

Acknowledgments

We appreciate TA Aishwarya Manjunath and Professor Garrison W. Cottrell for their feedback on our project proposal. We'd also like to thank TA Eric Yang Yu's constructive feedbacks during office hours. The support aforementioned people provided enables us to truly apply what we have learned in the CSE 151B: Deep Learning course and conduct a more comprehensive project.

References

[1] Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. Phrase-based statistical machine translation with pivot languages. In *Proceedings of the 5th International Workshop on Spoken Language Translation: Papers*, pages 143–149, 2008.

[2] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[3] Wikimedia Foundation. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news.

[4] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[5] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[6] Tomoki Hayashi and Shinji Watanabe. Discretalk: Text-to-speech as a machine translation problem. *arXiv preprint arXiv:2005.05525*, 2020.

[7] Mihir Kale and Abhinav Rastogi. Text-to-text pre-training for data-to-text tasks, 2021.

[8] Antonio Mastropaolo, Nathan Cooper, David Nader Palacio, Simone Scalabrino, Denys Poshyvanyk, Rocco Oliveto, and Gabriele Bavota. Using transfer learning for code-related tasks. *IEEE Transactions on Software Engineering*, 49(4):1580–1598, 2023.

[9] Johanna Monti, Anabela Barreiro, Brigitte Oroliac, Fernando Batista, et al. When multiwords go bad in machine translation. In *Workshop Proceedings for: Multi-word Units in Machine Translation and Translation Technologies (Organised at the 14th Machine Translation Summit)*, pages 26–33. The European Association for Machine Translation, 2013.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[11] Beibei Ren. The use of machine translation algorithm based on residual and lstm neural network in translation teaching. *Plos one*, 15(11):e0240663, 2020.

[12] Het Shah, Avishree Khare, Neelay Shah, and Khizir Siddiqui. Kd-lib: A pytorch library for knowledge distillation, pruning and quantization. *arXiv preprint arXiv:2011.14691*, 2020.

[13] suno ai. bark, 2023.

[14] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation, 2019.

[17] Hua Wu and Haifeng Wang. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21:165–181, 2007.

[18] Ziqing Yang, Yiming Cui, Zhipeng Chen, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu. Textbrewer: An open-source knowledge distillation toolkit for natural language processing. *arXiv preprint arXiv:2002.12620*, 2020.

[19] Richard Zens, Franz Josef Och, and Hermann Ney. Phrase-based statistical machine translation. In *KI 2002: Advances in Artificial Intelligence: 25th Annual German Conference on AI, KI 2002 Aachen, Germany, September 16–20, 2002 Proceedings 25*, pages 18–32. Springer, 2002.

[20] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. https://D2L.ai.