Fine-tuning and analyzing BERT for resource constrained neural translation

Shreya Pandit
Department of Computer Science
Boston University
shreyap@bu.edu

Alec Hoyland
Center for Systems Neuroscience
Department of Psychological & Brain Sciences
Boston University
ahoyland@bu.edu

Abstract

We trained a BERT-based encoder and Transformer decoder on a Neural Translation task between the low resource language Latin and English. We achieved a BLEU score of 20.23 in comparison to our German-English translations which use two high resource languages and achieved a BLEU score of 25.30. We found that the attention mechanism in the encoder correctly attends to the Latin text in a human-interpretable manner but noted strong sensitivity to the training corpus' English translations. The translations are over a century old and this affects the connotation of words used. The classical Latin samples are over 2000 years old (for example, the poet Catullus' works are dated between the time period 74BC- 20BC). Despite this, our fine-tuned model was able to leverage non-literal translations to improve clarity in the English translation for modern readers.

1 Introduction

Automated machine translation is an active field of study in natural language processing. Neural networks currently achieve state-of-the-art in machine translation [1, 2]. One popular model architecture is the encoder-decoder framework. The encoder embeds a representation of the input text in a high-dimensional vector space. The decoder produces a translation output, using the embedded text as input.

Training an encoder-decoder model requires a large bilingual corpus; therefore, trained models are sensitive to the quality and subject matter of the training corpora. The model is limited by the vocabulary of the corpora and quality of the translation. This sensitivity is apparent even within corpora from the same language [3]. Regardless of architecture, an encoder trained on an English Wikipedia corpus (https://www.english-corpora.org/wiki/) should be expected to specialize in encoding Standard American English. This is because the English Wikipedia corpus contains an extremely large (ca. 1.9 billion words) and expansive (ca. 4.4 million articles) sampling of text written in Standard American English. Training on another corpus would naturally yield different results, even if the semantic meaning is nearly identical. For example, if the model were trained on the Simple English Wikipedia instead, a Wikipedia in Standard American English but only using simple words and sentence constructs, we might expect a totally different encoding.

In this paper, we use the pretrained BERT encoder [4] and a decoder trained on poetical translations of classical Latin poetry drawn from The Latin Library (https://thelatinlibrary.com/). The corpus was constructed and tokenized using the Classical Language Toolkit (CLTK) [5]. We explore how training on corpora with the same semantic meaning but different diction influences neural machine translation.

2 Related Works

In recent years, neural machine translation (NMT) with deep neural networks has achieved significant improvements over statistical machine translation [6, 7, 8]. In particular, recurrent neural network models have shown success [9, 10]. These models are limited by sequence length, since recurrent neural networks typically generate a sequence of hidden states conditions on the last and the positionally-aligned input.

In contrast, the Transformer architecture disperses with recurrence that relies entirely on an attention mechanism to draw global dependencies between input and output [7, 11]. State-of-the-art in Transformer-based architectures is BERT [4]. BERT is a multi-layer bidirectional Transformer encoder with self-attention. This allows the model

to attend to tokens both before and after the current token. The published model has been pretrained using a masked language model task and a next sentence prediction task, in order to be as general as possible. BERT was trained using the top 100 language Wikipedia corpora and has been released as a multilanguage model.

Though BERT achieves state-of-the-art results for many language tasks including reading comprehension and text classification, the model is not designed specifically for neural machine translation tasks. Since training BERT is a computationally expensive endeavor, recent efforts have focused on utilizing the pretrained BERT encoder on neural machine translation tasks [12, 13, 14, 15].

Previous work has predominately focused on developing pretraining methods for NMT, resulting in multiple variants of BERT. In XLM [16], the model is pretrained on multiple languages without the next-sentence-prediction (NSP) task. ROBERTA [17] is trained with more unlabeled data and similarly skips the NSP task. XLNET [13, 18], a permutation-based modeling approach, introduces a Transformer architecture without fixed-length context. BERTVIZ is a tool for visualizing attention in Transformer models [19].

3 Methods

3.1 Code Availability Statement

All source code is freely available at https://github.com/shreyapandit/bert-nmt-latin.

3.2 The BERT model

BERT makes use of the Transformer architecture, an attention mechanism that learns contextual relations between tokens. While the original Transformer model [11] includes both an encoder and decoder, BERT is only a language model and therefore only includes an encoder mechanism. Figure 1 describes the Transformer architecture, in which the input sequence is embedded in a high-dimensional output space. A matching decoder can then decode the embedded sequence into a new language.

The Transformer in BERT is bidirectional, allowing the model to learn the context of a word based of its surroundings in both directions (forward and backward). The BERT model implemented in this paper is the BASE-MULTILANGUAGE model, with 110,000,000 parameters. BERT was pretrained on 4 cloud tensor-processing units, using an unsupervised masked language modeling task (Figure 2).

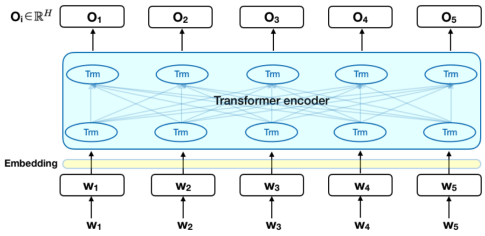

Figure 1: The Transformer architecture. Token representations are embedded in a 768-dimensional space.

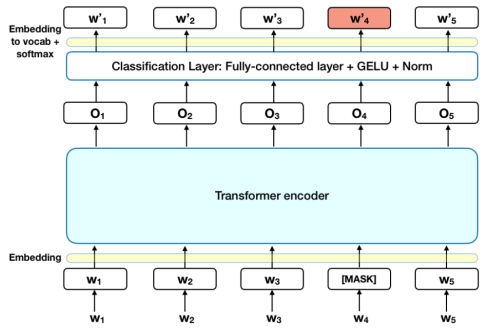

Figure 2: Schematic of the masked language modeling task used to train BERT. For machine translation, the classification layer was stripped and replaced with a decoder.

In this task, random tokens in the sequence were obfuscated or mutated. The model attempted to guess the masked or replaced word.

3.3 Corpora

Training data were acquired using the Classical Language Toolkit (CLTK) [5], from The Latin Library corpus (https://thelatinlibrary.com/), comprised of classical Latin texts and verbose, literary translations in English. CLTK lemmatizes and tokenizes using a PUNKT tokenizer. Classical latin does not include punctuation, however the texts collected at The Latin Library have editorial punctuation. Therefore, sentence separation was determined by an expert linguist. The corpus consists of samples of classical Latin writing with English translations by 19th and 20th century linguists.

The original data were scraped from the Perseus Digital Library at Tufts University (https://www.perseus.tufts.edu/hopper/) and compiled in `.xml` files. The Latin text was segmented by expert linguists into "milestones," which correspond to sentences or paragraphs in English. We used the BEAUTIFULSOUP Python package (https://www.crummy.com/software/BeautifulSoup/) to parse xml attributes and content. We remove markup attributes such as italics and notes and

Length of sentence	Proportion of lines
(10, 127]	0.634777
(127, 251]	0.279557
(251, 375]	0.064279
(375, 499]	0.015708
(499, 624]	0.004007
(624, 748]	0.001161
(748, 872]	0.000293
(872, 996]	0.000131
(996, 1120]	0.000069
(1120, 1245]	0.000019

Table 1: Distribution of length of German lines in test corpus (normalized).

Length of sentence	Proportion of lines
(19=0, 738]	0.960073
(738, 1459]	0.028086
(1459, 2180]	0.005684
(2180, 2901]	0.003131
(2901, 3623]	0.000913
(3623, 4344]	0.000890
(4344, 5065]	0.000659
(6507, 7229]	0.000323
(5065, 5786]	0.000162
(5786, 6507]	0.000081

Table 2: Distribution of lengths of Latin lines in test corpus (normalized).

divide the dataset into training and validation. These data were saved in a Python data frame, producing 2,870 samples each consisting of a full sentence or paragraph from classical Latin texts.

Testing data were compiled by the authors, one of whom is literate in Latin. The sentences came from three sources, designed to explore three "genres" of Latin literature. The first includes samples by the neoteric poet Gaius Valerius Catullus. Each poem is written as a long sentence without punctuation.

The second category is oratory prose, drawn from Marcus Tullius Cicero's Orations Against Catiline. These sentences are all brief rhetorical questions which use formal language.

The third category includes samples from an undergraduate Latin I curriculum, drawn from the first chapter of the textbook, *Ecce Romani* (http://www.tabney.com/ecce1.html). In all cases, gold-standard literal translations were provided by the authors.

Sentences in the Latin corpus were very long.

Latin-En Training	21,544
Latin-En Testing	4,063
De-En Training	160,239
De-En Testing	7,283

Table 3: Number of sentences in each dataset

We split lines with full stop tokens and managed to reduce the skewed distribution. With longer sentences, we encountered CUDA resource allocation errors. BERT has difficulties with long sequences. The implementation of BERT limits sequences to 512 tokens and was pretrained on sentence-level representations, meaning that BERT is poorly-optimized for translating long sequences [4, 13, 18]. Increasing the token limit leads to exponential increases in performance time, so we used the extant hyperparameters of BERT itself as-is.

3.4 Training scheme

Before training, dictionaries were generated from the Latin and English corpora and the training set was tokenized. We used CLTK to lemmatize and tokenize using a PUNKT tokenizer. Tokens outside the training vocabulary were set to "Unknown."

Training was done on an Nvidia Tesla V100 for 100 epochs for both translation (German and Latin) tasks. Validation was checked after each epoch. The Adam optimizer was used with a learning rate of 0.001, weight-decay of 0.0001, and dropout rate of 0.3. Due to resource limitations and time constraints, we were able to run only a handful of the configuration using a free GPU on Google Colab. Google Colab provides a V100 during the night and an Nvidia K89 during daytime. Each run of the model completed within a few hours.

4 Results

We used BERTVIZ to visualize the attention of several words using the pretrained multilingual BERT model. Due to its inflectional syntax, Latin allows for very flexible word ordering. Despite this, some regularities exist, which can be visualized by observing the attention weights.

For example, the word *quam* has many meanings, including "in what way," "in comparison to/than," however in context with *diu*, the meaning of the word pair becomes "so long as." This expression remains in legal contexts today where it implies "so long as [good behavior continues]." BERT pretrained on Latin preserves this meaning, as *quam*

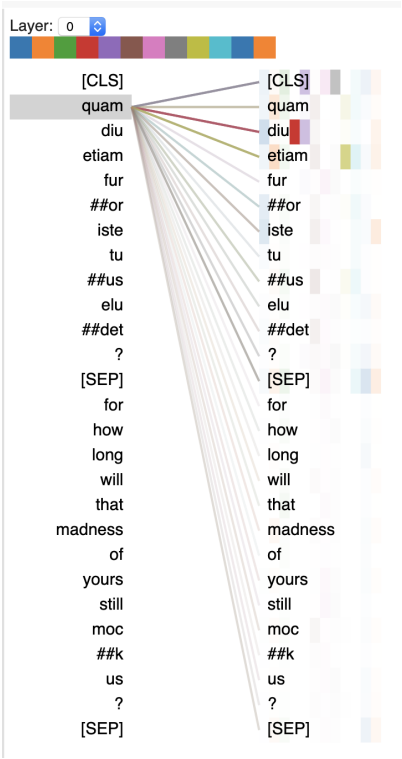

Figure 3: Attention of word *quam* using bertviz

BLEU (De-En)	25.3
BLEU (Lat-En)	20.23

Table 4: BLEU scores for the translation tasks.

strongly attends to *diu* (Figure 3). BERT is also able to capture the meaning of compound words.

The word *etiam* annexes a fact or thought which has already been said. In general, it means "and also/furthermore/besides," but in the context of time, it means "still/even now". BERT is able to disambiguate between these meanings, since *etiam* attends to words indicating temporal context (Figure 4).

In addition, BERT can handle rhetorical "doubling" of words, such as when *iste* ("that") is used in a pejorative context along with other pronouns. While it refers to the *furor* ("madness") of the target of this vitriol (the scheming Lucius Catiline), BERT is able to recognize that *tuus* ("your/yours") has a similar meaning in context, resulting in the phrase "that madness of yours" (Figure 5).

We report BLEU scores [20] for the German and Latin translation tasks (Table 4).

4.1 Examples of translations

The model perform adequately in translating Latin sentences. It performed especially well on shorter

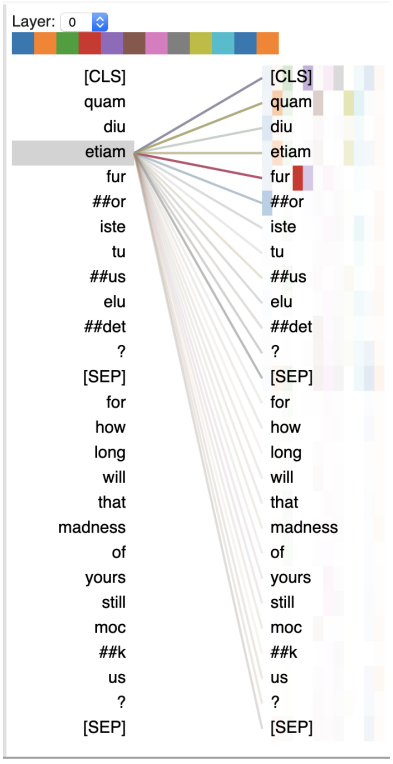

Figure 4: Attention of word *etiam* using bertviz

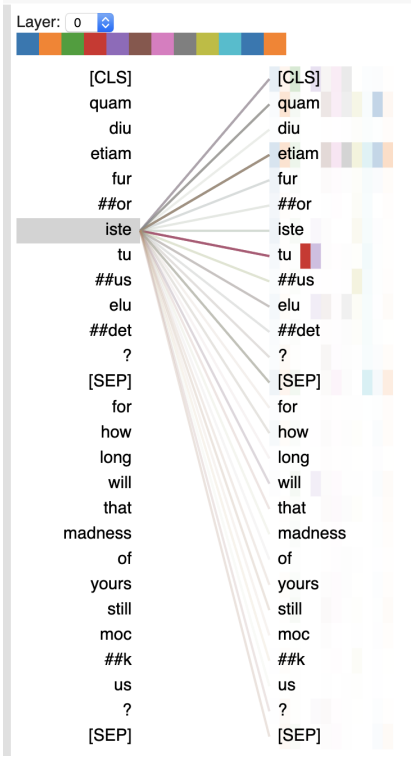

Figure 5: Attention of word *iste* using bertviz

Language Pair	Dict. Tokens
De-En	105,878
Lat-En	8,368

Table 5: Total number of dictionary tokens generated for the two translation tasks.

phrases. For example,

> *de parsimonia ac de pudicitia sua memoratissimus*

was correctly translated as

> regarding his frugality and continence

It is important to note that non-standard English words were used in the translation. In common parlance, "continence" is generally used in a medical context, to mean the voluntary control of bodily functions. However, *The Century Dictionary*, published in 1911, defines "continence" as "moderation or self-restraint.' This definition much more closely matches with the gold-standard meaning of *pudicitia*, meaning "chastity, shamefacedness, or modesty." Here we see a clear example of how a model trained on formal, academic translations of Latin from the turn of the century affects its ability to translate.

The attention mechanism built into the model is very powerful. For example, the word *agmen* usually means a "train" or "flock" of something, but can also mean a disordered marching column of soldiers.

> *[UNK]is auditis laudatoque suasore et iusso ducere qua n[UNK] agmina cuncta ab institute itinere conversa prae vium sequebantur.*

is translated as

> When this proposition had been heard and its author commended and bidden to lead them by the way that he knew, the whole army changed its intended line of march and followed its guide.

The conjugated form *agmina* is the plural. Here, the model correctly identified that while *agmina* is being used in its military sense, that the plural "armies" is not a good translation in English. This is because the Roman understanding of an army is much less unified than the modern definition. The model opted to translate *agmina* as "whole army" to preserve the idea that while they are marching in different columns, that they are part of the same organization. This translation is not literal, but it is more explicable to an English reader not versed in classical Roman history.

5 Model Weakness and our implemented improvement

5.1 Poor performance on low resource classical dataset

- The BERT multilingual model we implemented as a baseline performed very poorly on sentences of medium to longer length, such as those having 250-300 characters. Such sentences are common in poetry across languages. **The BLEU scores achieved running the baseline model on sentences of length more than 300 characters were under 1.0**

- We can attribute this to the way BERT calculates embeddings. BERT seeks to provide a pre trained method for obtaining contextualized word embeddings for each word of our corpus. It is able to do so by training on sentence pairs and predicting masked words in the input. While this approach may lead BERT to recognize certain subtleties regarding different meaning of a word, it doesn't take into account the sentence length (which might affect the word embeddings)

5.2 Our approach to solve this issue

- We filtered the training data to be a subset of the corpus that had less than 300 characters per sentence. We had 21,554 such training sentences in Latin. These sentences now had 40 tokens on average. **We fine-tuned BERT embeddings using these and the translations we achieved a BLEU score of 20.23**

- We also tried to construct our training corpus in a different way, such as trying to separate Latin milestones using punctuation symbols like ".", however this improved the BLUE score only marginally further than what we achieved using shorter sentences as described above. This approach needs further tweaking because Latin does not have the classical punctuation style as English.

6 Conclusion

We trained a BERT-based encoder and Transformer decoder on a translation task between the low resource language Latin and English. We achieved a BLEU score of 20.23 in comparison to our German-English translations which used two high resource languages. We found that the attention mechanism in the encoder correctly attends to the Latin text in a human-interpretable manner but noted strong sensitivity to the training corpus' English translations. The translations are over a century old and this affects the connotation of words used. Despite this, the classical Latin samples are over 2000 years old (for example, the poet Catullus' works are dated between the time period 74BC- 20BC) and the model was able to leverage non-literal translations to improve clarity in the English translation for modern readers.

In addition, we performed experiments to check how different sequence length in the training data affects model performance. We noticed very poor BLEU scores when the tokens per training sample was high (close to 100). We tried various approaches to overcome this issue, such as filtering on sequences of shorter length, and using punctuation as a baseline to separate training samples. Our improvements over the original Latin training corpus structure improved the BLEU score to 20.23, which is comparable to the BLEU scores achieved for high resource language pairs such as German-English (25.3).

For this paper, we limited our training corpus to only classical Latin rhetoric, poetry, and histories. We could consider adding translations of the Bible, however Late Latin and Classical Latin are not perfectly compatible.

We believe that the model can also be improved by tuning model hyperparameters using approaches like Bayesian optimization [21] and Hyperopt [22].

References

[1] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[2] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416, 2018.

[3] Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi, and Simonetta Montemagni. Assessing the readability of sentences: Which corpora and features? 06 2014.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[5] Kyle P. Johnson et al. Cltk: The classical language toolkit. https://github.com/cltk/cltk, 2014–2019.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[8] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[9] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling, 2016.

[10] Oleksii Kuchaiev and Boris Ginsburg. Factorization tricks for lstm networks, 2017.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[12] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders, 2019.

[13] Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li. Towards making the most of bert in neural machine translation, 2019.

[14] Kenji Imamura and Eiichiro Sumita. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong, November 2019. Association for Computational Linguistics.

[15] Stéphane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. On the use of bert for neural machine translation, 2019.

[16] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[18] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019.

[19] Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[21] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180, 2015.

[22] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008, 2015.