Peter Loux, and Nitish Patil

Neural Machine Translation of Old English

Abstract--**No automatic system yet exists for the translation of Old English into Modern English. We expand a small pre-existing dataset of sentence pairs. We then explore three different methods of neural machine translation, a basic RNN model, an RNN with an attention layer, and a transformer. On the expanded dataset, we find that the transformer model performs best, with a BLEU score of 0.70, while the other models are ineffective. Our findings indicate that the transformer model is the most promising approach to this problem, and that significant further expansion of the dataset is necessary to improve the performance of the model.**

I. Introduction

Old English is the earliest form of the English language, spoken in England and southern and eastern Scotland in the early Middle Ages. It was brought to England by Anglo-Saxon settlers in the mid-5th century. The Norman Conquest of 1066 saw the replacement of Old English by Middle English, the language used by Chaucer. Modern English evolved from Middle English in the late 15th century. Shakespeare's plays, which many modern readers have difficulty understanding, are written in Early Modern English. Old English is for the most part unintelligible to modern English speakers, but it is still studied by linguists and historians [15]

Old English was originally written in a runic alphabet; however, the Anglo-Saxons later adopted the Latin alphabet. Six letters do not exist in Old English, except potentially as loans; these are j, w, v, k, q, and z. However, Old English has four letters which do not occur in Modern English: æ (ash), þ (thorn), ð (eth), and ƿ (wynn) [15]. There are approximately 3000 surviving Old English texts [22], totalling about 3 million words [23]. From a database [24], it is estimated that there are 15,000 distinct word roots in the language. When conjugations, plurals and spelling variants are accounted for, this number rises to 130,000 total words. Although some words have survived into Modern English, most have not. The grammar of Old English is most similar to German or Icelandic. An example of an archaic feature of Old English is the inflection of nouns for case. Old English nouns have five cases (nominative, accusative, genitive, dative, instrumental), which, accounting for singular and plural forms, means each noun may have up to ten different spellings. Additionally, nouns are gendered, like many modern European languages. [9] These grammatical features make translation of Old English into Modern English more difficult than performing direct word-to-word replacement.

Many Old English texts have been lost, and many of the surviving texts are fragmentary or badly damaged. These texts have been digitized and transcribed and are generally available in the public domain in their untranslated form. For many years Old English has been translated by hand. Peter, one of the authors, has a Masters of Fine Arts in Writing, and has taken graduate coursework in Old English. At the time (2011), the Old English manuscripts were printed in books. The translator would photocopy or type the original manuscript, look up each word individually, and write the definition or definitions above it. Then from this, the most likely interpretation of the sentence would be written in Modern English. Peter used this process to translate Beowulf, and although rewarding, it was labor-intensive and tedious. Today, online texts are available, as well as online dictionaries are available [26], although the latter can only process one word at a time, requiring the translator to still be familiar with Old English grammar and syntax. As of 2022, Old English is not offered as a language option on sites such as Google Translate [25].

The Old English corpus includes religious texts, legal documents, histories, scientific texts, riddles, and poetry, most famously the epic poem Beowulf [17]. Currently, only dedicated scholars are able to decipher these texts, and casual readers must either purchase professional translations, or rely on translations available online, which is incomplete, and of varying quality.

II. Related Word

One of the few recent works concerning both machine learning and Old English is [33]. Although direct translation is not attempted, the authors use quantitative profiling to confirm the single authorship of Beowulf, as well as the attributing the authorship of several other poems to Cynewulf, two topics that have been debated by scholars for several hundred years.

Attempts have been made to decipher ancient languages by comparing them to similar languages. [20] proposed using Basque to interpret Iberian, a lost and hitherto undecipherable language. [21] While Old English is not a lost language, the paucity of sentence pairs, led us to consider this approach.

Other experiments in low-resource neural machine translation include [17], [8], [4], and [18]

A number of works deal with the preservation of endangered languages, for instance Cherokee [19], rare Eastern European languages [13], or Estonian and Slovak [2]

Our initial RNN model is based on an English-German translation model developed by Jason Brownlee [7]. This model was further adapted to Old English [11].

The attention layer was added, following the guidance of [1] A helpful description of attention used in the presentation was found in [5]

A number of GitHub projects dealing the conjunction of NLP and Old English were discovered, in addition to the sources previously mentioned. These are [27, 28, 29, 30, 31]

A transformer model for translation was implemented using code from [6]

Texts used for the training, test and validation sets were obtained from [10, 14]

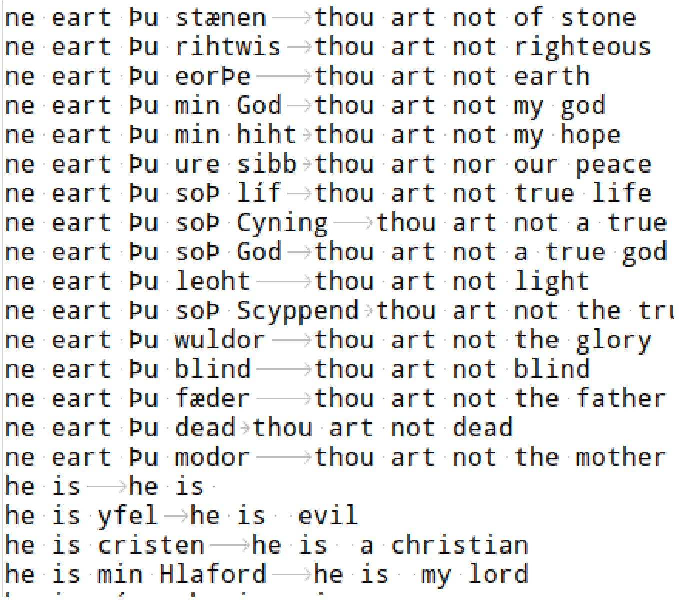
Of course, this section would not be complete without mention of the original transformer paper [3]. The transformer is a neural network architecture which has revolutionized the field of natural language processing. It is a self-attention mechanism which allows the model to learn long-range dependencies. The transformer has been applied to a wide variety of tasks, including machine translation, text generation, and image captioning.

III. Proposed Method

We propose to compare three different methods of neural machine translation of Old English into Modern English. The first model uses an RNN (Recurrent Neural Network). The second model is similar, with the addition of an attention layer. The third model is a transformer.

A. Dataset

We begin with a toy dataset of 385 sentence pairs, and expand it to over 1000 sentence pairs. The initial dataset came from the Homilies of Aelfric [10]. The Homilies are chosen because they are in the public domain, and feature side-by-side Modern and Old English versions. They also have a single translator, Benjamin Thorpe, and a similar subject matter (religion) which lends consistency. However, the initial dataset was heavily simplified.



```
ne eart Þu stænen	thou art not of stone
ne eart Þu rihtwis	thou art not righteous
ne eart Þu eorÞe	thou art not earth
ne eart Þu min God	thou art not my god
ne eart Þu min hiht	thou art not my hope
ne eart Þu ure sibb	thou art nor our peace
ne eart Þu soÞ líf	thou art not true life
ne eart Þu soÞ Cyning	thou art not a true
ne eart Þu soÞ God	thou art not a true god
ne eart Þu leoht	thou art not light
ne eart Þu soÞ Scyppend	thou art not the tru
ne eart Þu wuldor	thou art not the glory
ne eart Þu blind	thou art not blind
ne eart Þu fæder	thou art not the father
ne eart Þu dead	thou art not dead
ne eart Þu modor	thou art not the mother
he is	he is
he is yfel	he is  evil
he is cristen	he is  a christian
he is min Hlaford	he is  my lord
```

Figure 1. Old Dataset

This simplification led to inflated BLEU scores and facile results. A cursory inspection of this dataset shows that using it for training will produce a model capable of translating sentences from the original dataset, but not much else, including other sentences from the Homilies.

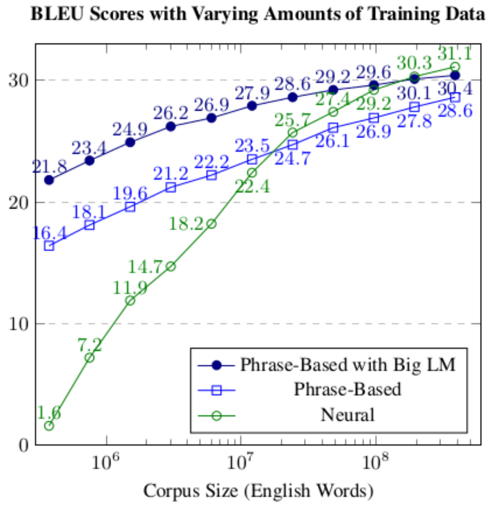
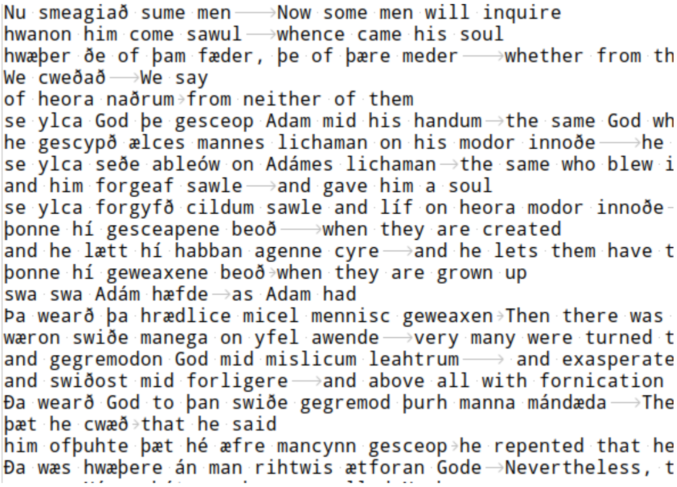


Figure 3: BLEU scores for English-Spanish systems trained on 0.4 million to 385.7 million words of parallel data. Quality for NMT starts much lower, outperforms SMT at about 15 million words, and even beats a SMT system with a big 2 billion word in-domain language model under high-resource conditions.

Figure 2 BLEU Scores with Varying Amounts of Training Data

400,000 words is considered an absolute minimum for machine translation. The original dataset contains only 3,000. Our expanded dataset contains about 12,000. NMT begins to surpass SMT (Statistical Machine Translation) in performance at 15 million words. [12]. While this information is from 2017, hence pre-Transformer models, it is clear that we are still sorely lacking in this department.

Expanding the dataset was a somewhat challenging process. Due to the extremely long sentences of the Homilies, we needed to break them into smaller independent clauses. This had to be done one at a time, manually, by someone with at least a passing familiarity with the language.



```
Nu smeagiað sume men → Now some men will inquire
hwanon him come sawul → whence came his soul
hwæþer ðe of þam fæder, þe of þære meder → whether from th
We cweðað → We say
of heora naðrum → from neither of them
se ylca God þe gesceop Adam mid his handum → the same God wh
he gescypð ælces mannes lichaman on his modor innoðe → he
se ylca seðe ableów on Adámes lichaman → the same who blew i
and him forgeaf sawle → and gave him a soul
se ylca forgyfð cildum sawle and líf on heora modor innoðe
þonne hí gesceapene beoð → when they are created
and he lætt hí habban agenne cyre → and he lets them have t
þonne hí geweaxene beoð → when they are grown up
swa swa Adám hæfde → as Adam had
Þa wearð þa hrædlice micel mennisc geweaxen → Then there was
wæron swiðe manega on yfel awende → very many were turned t
and gegremodon God mid mislicum leahtrum →  and exasperate
and swiðost mid forligere → and above all with fornication
Ða wearð God to þan swiðe gegremod þurh manna mándæda → The
þæt he cwæð → that he said
him ofþuhte þæt hé æfre mancynn gesceop → he repented that he
Ða wæs hwæþere án man rihtwis ætforan Gode → Nevertheless, t
```

Figure 3. New Dataset

Approximately 100 sentences of the new dataset are from a different source which also had side-by-side translation [14]. These sentences were chosen because they are in the public domain, and because they are from a variety of sources, including the Lord's Prayer to the Magna Carta, a text on the treatment of colds, and excerpts from the Anglo-Saxon Chronicle.

B. Preprocessing

All three models use a similar preprocessing pipeline. The dataset begins in the format of Old English->tab->Modern English. Punctuation is removed, and all words are lowercased. The pickle API is then used to serialize the dataset.

C. Models

1) RNN Model

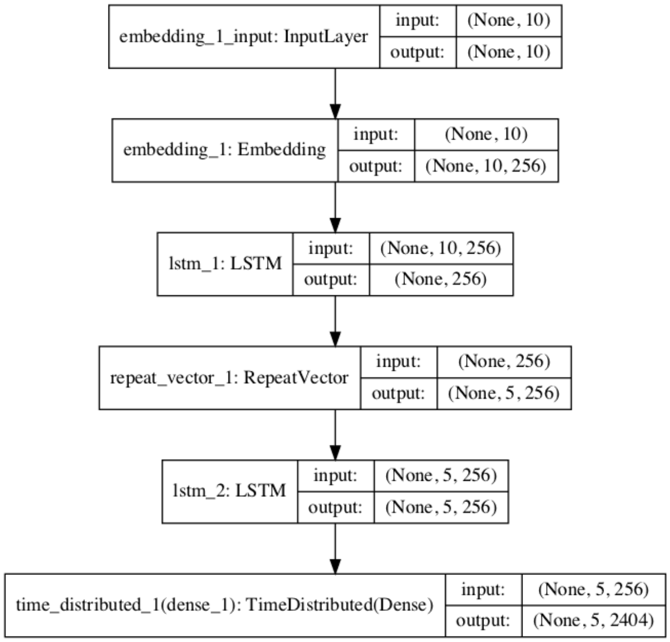
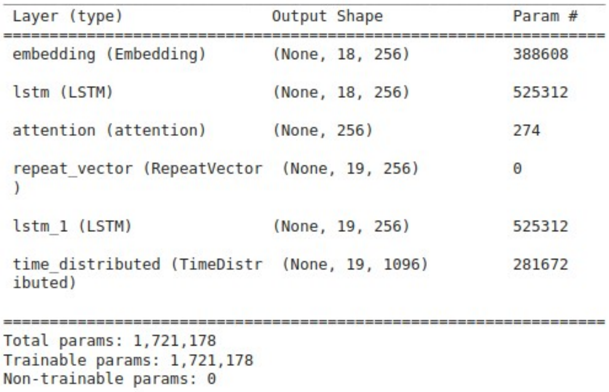


Figure 4. Model Graph for NMT

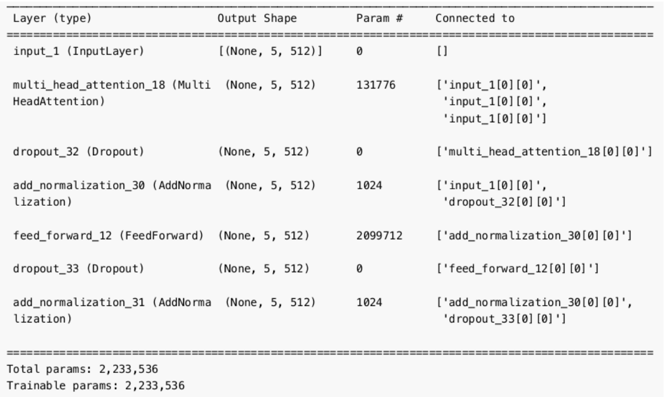
2) RNN with attention model



```
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 18, 256)           388608

 lstm (LSTM)                 (None, 18, 256)           525312

 attention (attention)       (None, 256)               274

 repeat_vector (RepeatVector  (None, 19, 256)          0
 )

 lstm_1 (LSTM)               (None, 19, 256)           525312

 time_distributed (TimeDistr  (None, 19, 1096)         281672
 ibuted)

=================================================================
Total params: 1,721,178
Trainable params: 1,721,178
Non-trainable params: 0
```

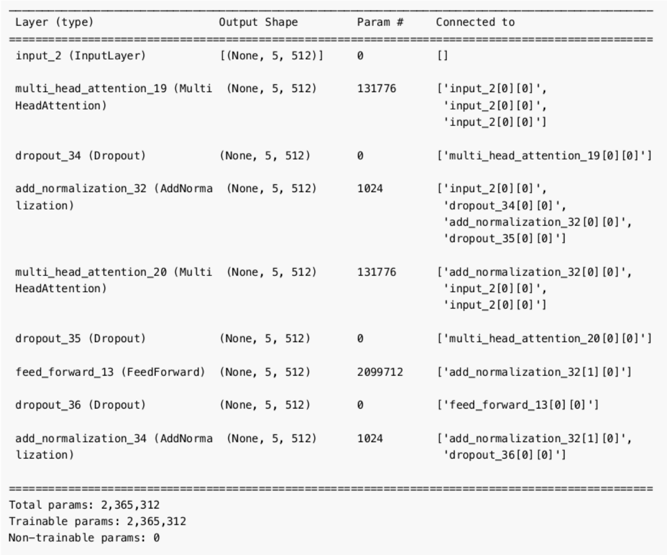
Figure 5. RNN with Attention

3) Transformer Model



```
__________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 input_1 (InputLayer)           [(None, 5, 512)]     0           []

 multi_head_attention_18 (Multi  (None, 5, 512)      131776      ['input_1[0][0]',
 HeadAttention)                                                   'input_1[0][0]',
                                                                  'input_1[0][0]']

 dropout_32 (Dropout)           (None, 5, 512)       0           ['multi_head_attention_18[0][0]']

 add_normalization_30 (AddNorma  (None, 5, 512)      1024        ['input_1[0][0]',
 lization)                                                        'dropout_32[0][0]']

 feed_forward_12 (FeedForward)  (None, 5, 512)       2099712     ['add_normalization_30[0][0]']

 dropout_33 (Dropout)           (None, 5, 512)       0           ['feed_forward_12[0][0]']

 add_normalization_31 (AddNorma  (None, 5, 512)      1024        ['add_normalization_30[0][0]',
 lization)                                                        'dropout_33[0][0]']

==================================================================================================
Total params: 2,233,536
Trainable params: 2,233,536
```

Figure 6. Transformer Encoding Figure



```
________________________________________________________________________________________________________
 Layer (type)                   Output Shape          Param #     Connected to
========================================================================================================
 input_2 (InputLayer)           [(None, 5, 512)]      0           []

 multi_head_attention_19 (Multi  (None, 5, 512)       131776      ['input_2[0][0]',
 HeadAttention)                                                    'input_2[0][0]',
                                                                   'input_2[0][0]']

 dropout_34 (Dropout)           (None, 5, 512)        0           ['multi_head_attention_19[0][0]']

 add_normalization_32 (AddNorma  (None, 5, 512)       1024        ['input_2[0][0]',
 lization)                                                         'dropout_34[0][0]',
                                                                   'add_normalization_32[0][0]',
                                                                   'dropout_35[0][0]']

 multi_head_attention_20 (Multi  (None, 5, 512)       131776      ['add_normalization_32[0][0]',
 HeadAttention)                                                    'input_2[0][0]',
                                                                   'input_2[0][0]']

 dropout_35 (Dropout)           (None, 5, 512)        0           ['multi_head_attention_20[0][0]']

 feed_forward_13 (FeedForward)  (None, 5, 512)        2099712     ['add_normalization_32[1][0]']

 dropout_36 (Dropout)           (None, 5, 512)        0           ['feed_forward_13[0][0]']

 add_normalization_34 (AddNorma  (None, 5, 512)       1024        ['add_normalization_32[1][0]',
 lization)                                                         'dropout_36[0][0]']

========================================================================================================
Total params: 2,365,312
Trainable params: 2,365,312
Non-trainable params: 0
```

Figure 7. Transformer Decoding Figure

D. *All Models:*

All models:

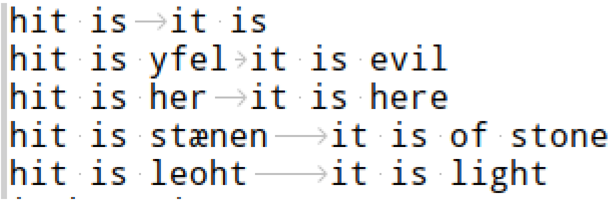
Optimizer	Adam
Loss	Categorical Cross entropy
Initial Learning Rate	0.1
Epochs	200 (with early stopping)
Batch Size	64

Transformer Model:

Number of self-attention heads	8
Dimensionality of the linearly projected queries and keys	64
Dimensionality of the linearly projected values	64
Dimensionality of model layers' outputs	512
Dimensionality of the inner fully connected layer	2048
Number of layers in the encoder stack	6
Beta_1	0.9
Beta_2	0.98
Epsilon	10^{-9}
Dropout rate	0.1
Warmup steps	4000

E. *Evaluation Procedure*

The primary evaluation for the models was BLEU-1 score [16], which simply counts the number of words in the predicted sentence that are also in the target sentence. BLEU-2, BLEU-3, and BLEU-4, which count the number of word pairs, triplets, and quadruplets, respectively, were also used. Finally, the translation output was manually inspected for quality. This manual inspection proved important as a model can achieve a relatively high BLEU-1 score by a variety of deceptive means. For example, if the dataset consists of only sentences with the following repetitive pattern, by guessing a translation of "it is" for every sentence, the model will achieve a BLEU-1 score of approximately 0.6.



```
hit is → it is
hit is yfel → it is evil
hit is her → it is here
hit is stænen → it is of stone
hit is leoht → it is light
```

Figure 8. Blue dataset

F. *Results:*

The three models are evaluated on the test set, which consists of 20% sentences randomly chosen from the dataset after the model has been trained on 80%. Both the original and expanded datasets were used. The results are shown in the table below.

Model	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4
RNN	Original	0.92801	0.91139	0.90214	0.82847
RNN	Expanded	0.15338	0.07536	0.05497	0.01679
Attention	Original	0.88198	0.85759	0.84303	0.75634
Attention	Expanded	0.13885	0.06732	0.04847	0.00000
Transformer	Original	0.65194	0.00000	0.00000	0.00000
Transformer	Expanded	0.72826	0.00000	0.00000	0.00000

G. Figures:

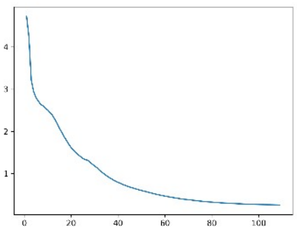


Figure 9. RNN Old Dataset Training Loss

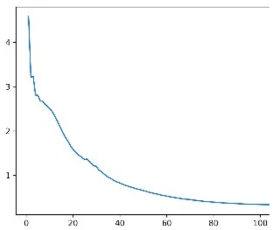


Figure 10. RNN Old Dataset Validation Loss

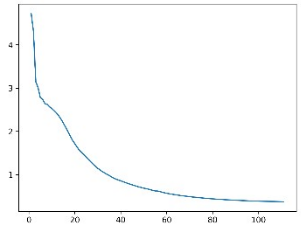


Figure 11Attention Old Dataset Training Loss

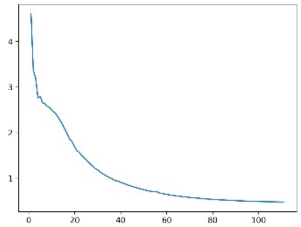


Figure 12Attention Old Dataset Validation Loss

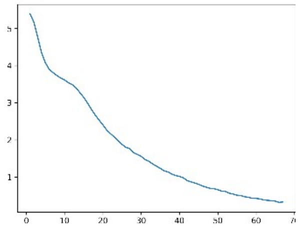


Figure 13. Transformer Old Dataset Training Loss

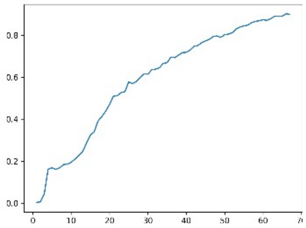


Figure 14 Transformer Old Dataset Accuracy

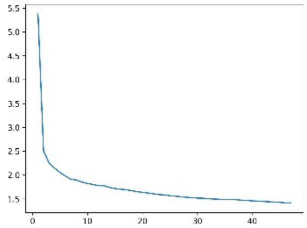


Figure 15. RNN New Dataset Training Loss

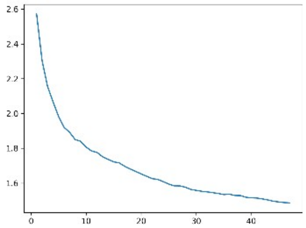


Figure 16. RNN New Dataset Validation Loss

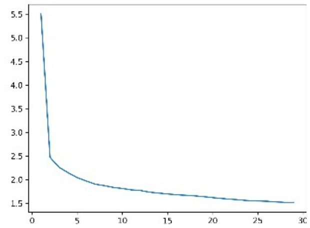


Figure 17. Attention New Dataset Training Loss

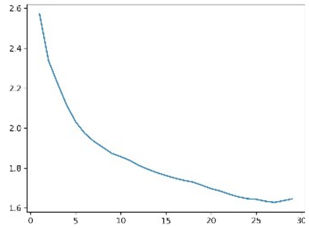


Figure 18. Attention New Dataset Validation Loss

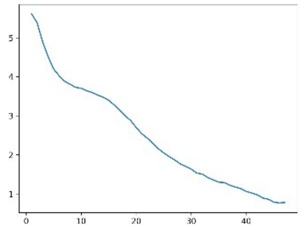


Figure 19.Transformer New Dataset Training Loss

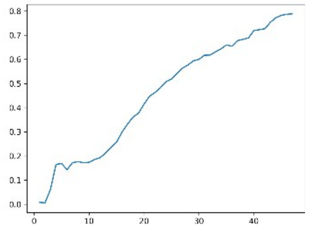


Figure 20.Transformer New Dataset Accuracy

IV. Conclusions

Adding an attention layer showed little benefit from the RNN model. Although the transformer model performed more poorly than either on the original dataset, this dataset is of little serious use. Notably, the transformer model was the only one which could cope reasonably with the expanded dataset. This is likely due to the fact that the transformer model is designed to handle long sequences, and the expanded dataset contains many long sentences.

For practical application, expanding the size of the dataset far beyond even our extended version, as well as deriving sentences from a broad variety of sources is clearly necessary.

A. Contributions of Members

Both members of the group worked equally on the code. Nitish developed the version with added attention layer, while Peter focused on the transformer version. Additionally, Nitish is to be credited with standardizing the three programs, generating usable testing output, and implementing features such as the early stopping monitor. Peter, due to his prior familiarity with Old English, was responsible for expanding the dataset. Both group members also collaborated equally in terms of experimental design and analysis, writing this report, and preparing the presentation.

B. Code

The code for this project is available at https://github.com/Ploux/oe-nmt. The three NMT models are named translate.py, attention.py, and transformer.py. They require the corpus dataset, corpus.tsv in order to run. Additionally, the models are accessible at the following colab links, in which case the corpus will be automatically downloaded. A GPU runtime is recommended.

Translate:
https://colab.research.google.com/drive/1SYTCc2L1kXTizm-SaeHoSbuC4t8x_9OD

Attention:
https://colab.research.google.com/drive/162H4r-QJFdkRN6kPWlGpB48nl9VTTasB

Transformer:
https://colab.research.google.com/drive/1g9SCvSoQmHn28Niiqz06n9a0MJRs6jpa

V. References

[1] https://www.analyticsvidhya.com/blog/2019/11/comprehensive-guide-attention-mechanism-deep-learning/

[2] Tom Kocmi, Ondřej Bojar, "*Trivial Transfer Learning for Low-Resource Neural Machine Translation*" - https://arxiv.org/abs/1809.00357

[3] Vaswani, Ashish, et al., Attention Is All You Need, https://arxiv.org/abs/1706.03762

[4] Lample, Ott, et al, Phrase-Based & Neural Unsupervised Machine Translation. 2018. https://arxiv.org/abs/1804.07755

[5] Philipp Koehn, Draft of Chapter 13: Neural Machine Translation, Statistical Machine Translation, 2017. https://arxiv.org/abs/1709.07809att

[6] Stefania Cristina, Mehreen Saeed, Building Transformer Models with Attention, 2022, MachineLearningMastery.com

[7](https://machinelearningmastery.com/develop-neural-machine-translation-system-keras/)

[8] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou. Word Translation Without Parallel Data, ICLR 2018. https://arxiv.org/abs/1710.04087

[9] https://en.wikipedia.org/wiki/Old_English_grammar

[10] https://www.gutenberg.org/files/38334/38334-h/38334-h.htm

[11] https://github.com/kel-c-lm/translator

[12] Philipp Koehn, Rebecca Knowles. Six Challenges for Neural Machine Translation, First Workshop on Neural Machine Translation, 2017. https://arxiv.org/abs/1706.03872

[13] A Mosolova, K. Smaili. The only chance to understand: machine translation of the severely endangered low-resource languages of Eurasia Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022) October 2022, https://aclanthology.org/2022.loresmt-1.4.

[14] https://oldenglish.info/textselect.html

[15] https://en.wikipedia.org/wiki/Old_English

[16] https://en.wikipedia.org/wiki/BLEU

[17] https://en.wikipedia.org/wiki/Old_English_literature

[18] Rico Sennrich, Biao Zhang. Revisiting Low-Resource Neural Machine Translation: A Case Study. 2019.
https://arxiv.org/abs/1905.11901

[19] S. Zhang, B. Frey, M. Bansal. ChrEn: Cherokee-English Machine Translation for Endangered Language Revitalization, 2020, https://arxiv.org/abs/2010.04791

[20] Luo, Hartmann, et al. Deciphering Undersegmented Ancient Scripts Using Phonetic Prior.
Transactions of the Association for Computational Linguistics (2021) 9: 69–81.
https://doi.org/10.1162/tacl_a_00354

[21] Jiaming Luo, Yuan Cao, Regina Barzilay. Neural Decipherment via Minimum-Cost Flow: from Ugaritic to Linear B https://arxiv.org/abs/1906.06718

[22] https://open.psu.edu/databases/psu00859

[23] https://www.libraries.rutgers.edu/databases/dictionary-old-english-web-corpus

[24] https://github.com/iggy12345/OE_Sentence_Generator

[25] https://translate.google.com/

[26] https://www.oldenglishtranslator.co.uk/

[28]
https://github.com/chadmorgan/OldEnglishPoetryCorpus

[29] https://github.com/iafarhan/Machine-Translation-for-Endangered-Language-Revitalization

[30] https://github.com/nbsnyder/OldEnglishNLP

[31] https://github.com/sharris-umass/oenouns

[32] https://github.com/old-english-learner-texts/old-english-texts

[33] Neidorf, L., Krieger, M. S., Yakubek, M., Chaudhuri, P., & Dexter, J. P. (2019). Large-scale quantitative profiling of the Old English verse tradition. Nature Human Behaviour. doi:10.1038/s41562-019-0570-1