CSE 582 Final Project: One Model To Rule Them All

Collin Beaudoin

May 4, 2023

Abstract

The Lord of the Rings is one of the most influential tales of all time, selling over 150 million copies and each of Peter Jackson's theatrical renditions grossed nearly $1 billion to date worldwide. The series has inspired many authors and scriptwriters to attempt their rendition of the story, perhaps the most famous rendition is the *Game of Thrones* series. For my novel task, I decided to implement an English-to-Sindarin translator. Sindarin is one of the elf languages from J.R.R. Tolkien's *The Lord of the Rings* series and to the best of my knowledge, this is the first attempt at such a process.

Contents

1 Introduction

The Lord of the Rings has developed a massive cult following over the many decades of its existence. This unwavering support is attributable to J.R.R. Tolkien's strict attention to detail and the expansive world he created. J.R.R. spent the preponderance of his life working on *The Lord of the Rings* world, and his son Christopher undertook his father's works upon his death [1]. J.R.R. imagined at least 15 separate languages for his series; yet only two are deemed complete – Quenya and Sindarin [2]. Quenya is the elder language of the two and offers historical context to the development of Sindarin, but Sindarin is the principal language spoken by the Middle-Earth elves from *The Lord of the Rings* series. Because of Sindarin's heavy usage throughout the series, it has become the focus of fans and scholars alike [3–6]. While there has been a focus on the translation of text and neologisms, there are no known attempts to apply machine learning to the translation process. To create a Sindarin machine learning model, I had to learn the basics of Sindarin, collect writings and their translations across many sites, find a fitting model for the task, and evaluate the model's performance.

2 Approach

2.1 Data

I collect a labeled data set to train an NLP model; while this is time-consuming, it offers higher quality translations which I believe to be of higher value to the Sindarin scholar community. Not tied to a single language, Sindarin closely relates to Celtic languages such as Welsh [7]. Due to the Celtic influence, many things may appear odd to the typical English speaker. For example, the change of tense or plurality may result in a completely different spelling in Sindarin. In English, these changes usually result in augmentation with a prefix or suffix, but in Sindarin, the root can mutate; different prefixes or suffixes may also be affixed. Sindarin takes many forms; the most widely accepted is the canonical translations by J.R.R. Tolkien and his son Christopher. However, the spelling of words and the mutations applied changed throughout periods of J.R.R.'s life, which caused inconsistent spellings of Sindarin phrases across translation samples [8]. Even though spellings of canonical terms may adjust based on the era, they are generally all accepted as correct. In addition to the different eons of Sindarin, there are different dialects. The main dialects are Doriathren, Woodelven, Exilic, and Gondorian. The patois have minor spelling differences from the others but are generally understandable across the regions; each spelling difference is held as canonical Sindarin. Sindarin is considered a complete language, yet many terms are quotidian parts of modern dialogue undeveloped by the Tolkiens' work. In other cases, a term may exist, but the Tolkiens' did not update it, making the expression inconsistent with the standards set by a given era; covering terms that did not exist or adjusting terms to fit closer to the era's standards is commonly referred to as Neo-Sindarin. Neo-Sindarin and its place within the Sindarin community is a topic of hot debate [9, 10]; I wish to avoid conflict over my data set, but there are not enough canonical phrase translations to create a large data set, so I opt to include the debated Neo-Sindarin; for clarity, I list the Sindarin types encompassed in my data set.

- Canon Sindarin (both Tolkiens)
- Doriathren Sindarin
- Woodelven Sindarin
- Exilic Sindarin
- Gondorian Sindarin
- Neo-Sindarin

As an example translation, ”I vrennil vain ben-dihenad” is Neo-Sindarin for the phrase ”the beautiful lady without mercy”. This phrase is the English translation of John Keats' famous poem title *La Belle Dame sans Merci*.

Collection: I base collection decisions on Paul Strack's opinion, Neo-Sindarin can be widely accepted when reviewed [10]; data collection of non-canonical resources such as books, magazines, and websites must have references from several other resources. Where references from communities are low, only group-reviewed translations are acceptable.

- My first source is Pedin Edhellen [11], a book held as one of the best introductory lessons on Sindarin. It offers many terms, and phrases, and gives an in-depth analysis of how Sindarin developed over the years.
- My second source is science-and-fiction [6], a website from the author of Pedin Edhellen. Renk offers translations of many poems and songs, which offers a more complex dialogue.
- My third source is Elm [12], a website hosted by E.L.F. the Elvish Linguistic Fellowship, an international organization focused on scholarly studies of J.R.R. Tolkien and his languages. This source focuses again on songs and poems.
- My fourth source is Real Elvish [13], a website run by Fiona Jallings, an author of Neo-Sindarin guides. She regulates forum submissions for common small talk phrases.
- My fifth source is Tolkien's Languages in *The Lord of the Rings* [4], a website hosted by E.L.F. the Elvish Linguistic Fellowship, which contains translations of the Sindarin used throughout The Lord of the Rings movies.

This collection includes 4005 phrases of Sindarin and Neo-Sindarin, with their English translations. The data set encompasses all English translations of Sindarin statements for the most accurate interpretation. All valid variations of Sindarin spellings of phrases are adopted. As previously discussed, different spellings are usually the sign of different eras or dialects of Sindarin. While the poems and songs contain verbose statements, I break the phrases apart using punctuation as my splitting point. While copying data from sources, I check for prior occurrences before adding the new sample. In the data set, previous phrases may occur within a statement, but I opt to maintain both in the data set. The order in which words appear in Sindarin is different than in English. Removing smaller phrases that are duplicates of samples in verbose statements could cause incorrect translations: translations of terms based on their location. My hope for the maintenance of the smaller phrases is to remove this ambiguity.

2.2 NLP Process

This section describes the process for training and evaluating the NLP model.

Dictionary: NLP tasks typically use lemmatization or stemming while constructing a dictionary of root words (or a close approximation of root words). Stop words are commonly removed from phrases to improve a model's comprehension. No lemmatizing, stemming, or eliminating stop words is applied while building the dictionary of Sindarin terms. Sindarin relies on a complex set of rules when attempting to pluralize or change the tense of a word, making it near impossible to create a reliable rooting of a term. For example, the word "aran" is Sindarin for "king" but the plural form is "erain"; the word "amon" is Sindarin for "hill" but the plural form is "emyn"; both examples are plural, yet they require different changes to do so [2]. Remaining consistent across the languages, no lemmatization, stemming, or removing stop words is applied to the English terms. The English phrases remain in their pure form as the different mutations of Sindarin words change the meaning of a phrase which should reflect in the translation.

Phrases: The dictionary is verbose due to no word or phrase trimming, increasing the search space complexity. Requiring mitigation, a pre-processing step removes Sindarin or English phrases longer than ten words. While there is active research and discussion on the best sentence length for reader understanding, I base my decision on Vigen's analysis of average sentence length in popular books over time [14, 15]. Examining Figure 1, a sentence length of ten covers the majority of sentences for the two most popular fantasy series in the modern era. Extrapolating the statistic, a sentence length of ten hypothetically allows for translation of most current writing samples. Within the data set, the number of usable phrases goes from 4005 phrases to 3786 phrases. Roughly a 9.5% reduction in the data set, so there is a low loss of samples but a reduction in phrase complexity.

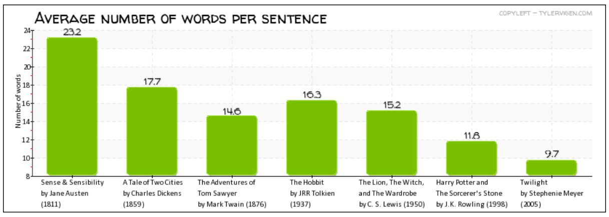

Figure 1: Average words per sentence in popular books throughout the ages [15]

Model: The intention is to create an NLP model that translates English to Sindarin. An encoder and decoder sequence-to-sequence model is selected to implement the translator. Sometimes, it is possible to use just the encoder, but Sindarin usually requires a reordering of terms when translating from English. Due to the reordering, the best option is to allow an encoder to create a representation of a phrase while a decoder translates and orders terms. The encoder design uses an embedding layer followed by a recurrent neural network, specifically the gated-recurrent unit (GRU). GRUs, based on the same concept as long-short-term memory (LSTM), optionally ignore/remove terms from history to create a strong context state. Ignoring words helps the model focus on the most important words within a phrase to build its representation. LSTMs are usually more popular because they can offer more parameters than the GRU, and transformer approaches are more popular due to their ability to represent longer dependencies occurring within a phrase. However, the GRU using a length of roughly ten words per phrase should offer about 90% accuracy [16], meaning there is likely little or no accuracy loss compared to the LSTM or transformer. For my decoder design, I use an attention mechanism in combination with a GRU, based on the suggestions by Robertson [17]. An attention mechanism in the decoder is necessary to resolve the issue of ordering the terms while translating to Sindarin. As an example of the reordering Sindarin requires: the English phrase ”mighty king” roughly translates to ”bell aran” but this is improper Sindarin. The proper translation of ”mighty king” is ”aran vell” where the Sindarin word for king comes before the adjective [2]. The basic structure of the NLP model is in Figure 2. The architecture of a GRU is in Figure 3.

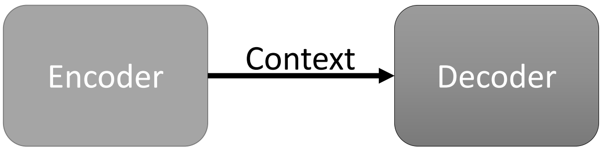

Figure 2: Basic encoder decoder sequence to sequence structure

The GRU in Figure 3 is broken down into four equations, equation 1 is used to decide the probability of basing the representation on either the history or some combination of the history and the new term. The σ represents some activation function, x_t represents the current input, and h_{t-1} represents the historical representation from the previous state. The W_z is the weight associated with the combination of the history and current input, while b_z is some bias factor. Equation 2 represents how much history to account for within the combination of the history and current input. Equation 3 represents the combination of the necessary history and current input, where $tanh$ is the arch-tangent activation function. Finally, equation 4 represents the combined information of current input and history.

$$z_t = \sigma(W_z * [h_{t-1}, x_t] + b_z) \tag{1}$$

$$r_t = \sigma(W_r * [h_{t-1}, x_t] + b_r) \tag{2}$$

$$h_{t*} = tanh(W_h * [r_t * h_{t-1}, x_t] + b_r) \tag{3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_{t*} \tag{4}$$

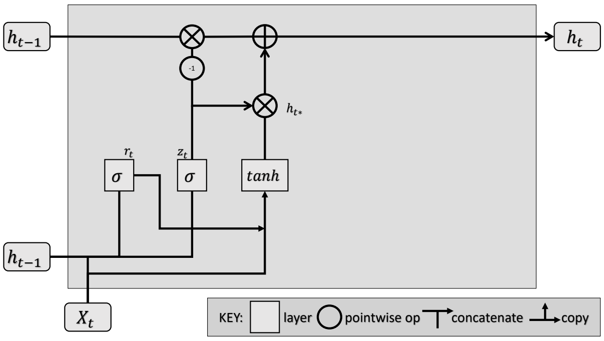

Figure 3: Visualization of the GRU structure

Inserting the GRU within the encoder-decoder structure results in the overview in Figure 4. The encoder takes an English phrase and uses the GRU to create a contextual embedding using important sequential information. The decoder takes this contextual vector to generate and order the Sindarin text for a proper translation.

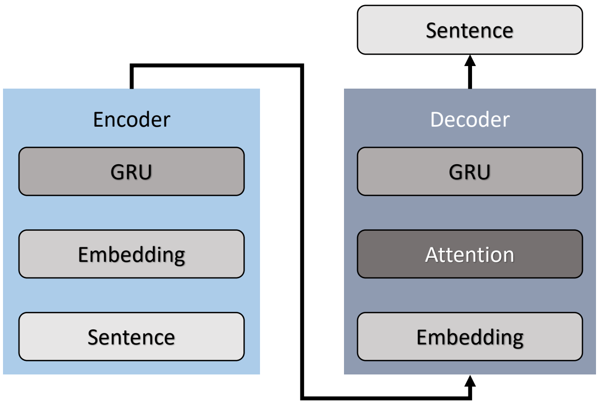

Figure 4: Internal structure of encoder and decoder

Evaluation: Two statistics measure the model's ability to translate. The first is the BLEU score [18]. The BLEU score is not wholely reliable as it might produce low scores for valid translations, but it offers a rough estimate of the model's abilities. The second metric is the word error rate, which attempts to account for substitutions, deletions, and insertions of words between two texts [19]. Both statistics attempt to build an understanding of a model's ability to create a result similar to the "golden" sample. Unfortunately, both scores are likely to show worse performance than what is achieved by the model. In general, this is due to the ambiguous spelling of terms within Sindarin; for example, the model may produce the Exilic spelling of a phrase, but the "golden" sample uses a Woodelven spelling.

3 Results

3.1 Description

Following another suggestion from Robertson, I include the usage of teacher forcing; teacher forcing will take the golden output for a term rather than the potentially incorrect output of the recurrent model when predicting the next word[17], allowing for a quicker convergence of the model but sometimes leading to instability in results [20]. To avoid rapid convergence, I use random chance for fluctuation between the usage of teacher forcing and the usage of the value from the decoder. The model training runs for a set number of iterations; in each iteration, a random pair of phrases from the training data set will train the model. For clarity, I include the hyper-parameters used for training.

- Teacher Forcing: Random 50/50
- Max Phrase Length: 10
- Unique Sindarin Terms: 2275
- Unique English Terms: 1702
- Hidden Dimension: 256
- Total Phrases: 3786
- Optimizer: SGD
- Learning Rate: .01
- Decoder Dropout: .1
- Training Iterations: 75k
- Train/Testing Split: 90/10
- Cost Function: NLL Loss

3.2 Analysis

To ensure that the model trained properly the loss curve over the 75k iterations of training in Figure 5 is included.

From Figure 5, there are improvements in performance over the first 40k iterations; the loss jumps from 4.8 to around .52. However, the loss improvement attenuates from this point on, with only a loss reduction of .03 from .27 at 60k iterations to .24 at 75k iterations. Pointing to minor further improvement of the model, or in other words, the model is at a stable point in training performance. Evaluating overall performance, I perform a corpus BLEU scoring using NLTK's evaluation package and a separate test data set [21]. By default, NLTK uses an equal summation of the 1-gram, 2-gram, 3-gram, and 4-gram BLEU scores, where a 1 is the highest possible BLEU score. Understanding the performance requires a general breakdown of the data within the data set. This breakdown is in Table 1.

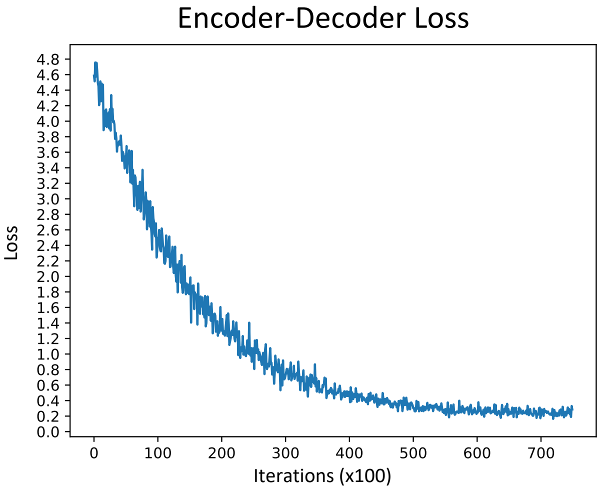

Figure 5: Training loss of encoder-decoder model

Table 1: Table of occurrences of given Sindarin phrase lengths.

Sindarin Phrase Length	**Count of Phrases Matching Length**
1 word	285
2 word	1036
3 word	989
4 word	671
5 word	396
6 word	196
7 word	137
8 word	57
9 word	18
10 word	0

Unfortunately, 2,310 Sindarin phrases are three words or less. Therefore, around 61% of the phrases in the data set will have a 0 value for the 4-gram result, causing low performance for the default evaluation by NLTK. Mitigating the high percentage of 0s in the 4-gram requires a smoothing function. I also include the 3-gram summation and 2-gram summation BLEU corpus scores with the same smoothing function to build an understanding of how the model is performing with the smaller phrases.

From Table 2, the 4-gram corpus summation BLEU score is near 0. Attributable to over half of the data set containing less than four words. In the 3-gram corpus summation performance, the score nearly doubles to .14 but is still low. Referring to Table 1, roughly a third of the data set has two words or less, attributing to the low performance. The 2-gram corpus summation performance is .24, which is a .1 improvement over the 3-gram summation score but is slightly less than a quarter of the highest possible score. To understand why the 2-gram corpus summation score is low, evaluation of the individual 1-gram and 2-gram BLEU scores is necessary, and for the sake of completeness, the individual 3-gram and 4-gram BLEU scores are also present in Table 3.

Starting with the BLEU 1-gram score in Table 3, the score is .428 meaning less than

Table 2: Table of BLEU summation corpus scores for varying n-gram sizes.

BLEU n-gram	**Score**
BLEU 4-gram sum	.08
BLEU 3-gram sum	.14
BLEU 2-gram sum	.24

Table 3: Table of BLEU individual corpus scores for varying n-gram sizes.

BLEU n-gram	**Score**
BLEU 4-gram	.016
BLEU 3-gram	.048
BLEU 2-gram	.143
BLEU 1-gram	.428

half the ”golden” terms are in the model’s translation. The 2-gram BLEU score is .148, which appears low given that it represents how often two words are in the correct order. The low score is likely from incorrect ordering, amplified by incorrect word selection by the model. To further evaluate the model performance, I use the word error rate. Word error rate scores the substitution, insertion, and deletion of words and ranges from 0 to 1, where 0 is the best possible score. The total word error rate of the model is .78. This means over three-quarters of the words vary from the ”golden samples”; the word error rate confirms the 1-gram BLEU score evaluation; the model has difficulty selecting the correct terms. To understand why I randomly review ten test results, evaluating the four incorrect translations from the set.

For the first example: the English phrase is ”Will you teach me?” whereas the ”golden” Sindarin translation is ”nin gollathol?”. The model produces the translation ”nin golthathodh?” so half of the terms are correct. But it turns out ”golthathodh” is the formal version of ”you teach” whereas ”gollathol” is the informal version. The BLEU scores are 0 for the individual 2-gram, 3-gram, 4-gram, and a .5 for the individual 1-gram. In the case of the word error rate, it also receives a .5, as one word is a substitute. All the BLEU scores and the word error rate show poor performance for the model, which creates an accurate translation.

The second English phrase is ”release” whereas the ”golden” Sindarin translation is ”leithio”; the model produces ”leithiad” which results in no correct words. However, ”leithiad” and ”leithio” mean ”release” and are interchangeable. Despite no difference in the terms, other than the spelling, the BLEU scores are all 0, while the word error rate is 1.

The third English phrase is ”Who is Aragorn?” and the ”golden” Sindarin translation is ”Man Aragorn?”. The model produces ”ma Aragorn?” so half of the terms are correct. But ”ma” and ”man” are interchangeable spellings of ”who” or generically ”what”; the ”n” is commonly appended to ”ma” if the following word begins with a vowel but is not necessary. The BLEU scores are 0 for the individual 2-gram, 3-gram, 4-gram, and a .5 for the individual 1-gram. The word error rate is .5 due to the ”substitution” of a word. All the BLEU scores and the word error rate show poor performance for the model, despite the accurate translation where the only difference is an optional character.

The fourth English phrase is ”we go to” where the ”golden” Sindarin translation is ”meninc na”. The model produces ”menif na” so half of the terms are correct. In the case of ”menif” and ”meninc” they both mean ”we go”, but the ”we” each phrase represents is different. In the case of ”menif”, it is for large groups of people, whereas ”meninc” is for just two individuals (like ”you and I”). In summary, the incorrect translation is due to ambiguity in the English language. For BLEU scores, the model obtains 0 for the individual 2-gram, 3-gram, and 4-gram with a .5 for the individual 1-gram. The word error rate is .5 due to the ”substitution” of a word. All the BLEU scores and the word error rate show poor performance for the model.

4 Conclusion

In this work, I successfully implement (to the best of my knowledge) the first English-to-Sindarin NLP translator. The model uses a GRU encoder and a GRU in combination with an attention mechanism decoder. In model evaluation: cumulative BLEU scores, the individual n-gram BLEU scores, and the word error rate all report that the model performs poorly. But when analyzing the data samples, 61% of the phrases are less than four words long, and over a third of the data has less than three words in a phrase. This high sampling rate of small phrases biases the performance of the 3-gram and 4-gram BLEU scores; the 1-gram and 2-gram BLEU scores are promising but lower than expected; examining what is causing the poor performance, I review multiple test samples. The model produces accurate results, but ambiguity in English and Sindarin creates predictions that don’t match the ”golden” phrases. There are a few options for potential future work. The first improvement would be using pre-trained English to Celtic (or Welsh specifically) model and then performing fine-tuning training. As Sindarin is like the Celtic language, this should decrease the time to train the model. Using pre-training should also help resolve some of the ambiguity issues present. The second improvement would be adding additional labels to the current data set to mark determining factors like which form of Sindarin the translation takes, which ”we” is meant by the English and Sindarin, and other descriptive information. The third improvement would be creating a stemming operation to decrease the Sindarin dictionary size. I believe the model with a smaller dictionary would be able to learn how to translate and apply the best mutations to the phrases. If not, a combination of two models, where the first model could offer the root translation with a secondary model that could apply the mutations. The last potential improvement would be the collection of a large unlabeled Sindarin corpus, which would allow for the usage of a GPT-like model.

References

[1] Åke Bertenstam. (2023) A chronological bibliography of the writings of j.r.r. tolkien. [Online]. Available: https://www.forodrim.org/arda/tbchron.html 2

[2] G. Inglorion. (2023) Elvish linguistics unofficial faq. [Online]. Available: http://www.elvish.org/gwaith/faq.htm 2, 4, 5

[3] C. F. Hostetter. (2006) Resources for tolkienian linguistics. [Online]. Available: https://www.elvish.org/resources.html 2

[4] R. Derdzinski. (2023) Tolkien's languages in the fotr movie. [Online]. Available: http://www.elvish.org/gwaith/movie_elvish.htm 3

[5] D. Salo. (2005) Tolkien's languages in the lord of the rings movie. [Online]. Available: http://www.elvish.org/gwaith/movie.htm

[6] T. Renk. (2015) Mae govannen nan mbar thorond! [Online]. Available: http://www.science-and-fiction.org/elvish/index.html#sindarin_poetry 2, 3

[7] P. Strack. (2023) Conceptual development. [Online]. Available: https://eldamo.org/content/words/word-2957722179.html 2

[8] ——. (2023) Conceptual history of elvish. [Online]. Available: https://eldamo.org/general/conceptual-history.html 2

[9] C. F. Hostetter. (2004) The tolkienian linguistics faq. [Online]. Available: https://www.elvish.org/FAQ.html 2

[10] P. Strack. (2023) Neo-sindarin neologisms. [Online]. Available: https://eldamo.org/content/neologism-indexes/neologisms-ns.html?neo 2, 3

[11] T. Renk, "Pedin edhellen a sindarin-course," *Version*, vol. 3. 3

[12] A. Bican. (2003) Elm. [Online]. Available: http://www.elvish.org/elm/others.html 3

[13] F. Jallings. (2023) Real elvish. [Online]. Available: https://realelvish.net/phrasebooks/sindarin/ 3

[14] A. Moore, "The long sentence: a disservice to science in the internet age," *BioEssays*, vol. 33, no. 12, pp. 193–193, 2011. 4

[15] T. Vigen. (2014) Literature statistics. [Online]. Available: https://www.tylervigen.com/literature-statistics 4

[16] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017. 5

[17] S. Robertson, "Nlp from scratch: Translation with a sequence to sequence network and attention," 2017. 5, 7

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. 6

[19] J. Woodard and J. Nelson, "An information theoretic measure of speech recognition performance," 1982. 6

[20] H. Jaeger, "Tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the" echo state network" approach," 2002. 7

[21] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009. 7