A. Project Team members

1. Maggie Sellers
2. Rachael Dewey

B. Project Topic Introduction

The goal of our project is to use Neural Machine Translation (NMT) to explore whether translations to English from other Germanic languages are more accurate than translations to English from Romance languages. For the purposes of this exploration, we have selected three languages to work with in addition to English: German, French, and Italian. We hypothesize that translations to English from other Germanic languages will be more accurate than translations from Romance languages; therefore, we believe that translations from German to English will be the most accurate, as they are both Germanic languages that share the same root language of Proto-Germanic. Of the two Romance languages, we believe that translations from French will be the most accurate. This is because the Norman William the Conqueror conquered England in the 11th century and as a result, the English language has a great deal of French influence. Finally, English and Italian share the most distant linguistic relationship, which we believe will result in the least accurate translations.

In practice, using a single NMT model for each language pair is an extremely inefficient practice. Multilingual NMT, which translates across multiple languages within a single model, vastly increases the model efficiency; however, using a single model to handle all languages is also often impractical. Therefore, it is vitally important to understand the relationship between historical language family relationships and NMT accuracy in order to determine how languages should be supported by multilingual NMT models when working with a large number of disparate languages across several language families.

C. Prior work

Maggie Sellers: I read the paper *Interactive neural machine translation* by Álvaro Peris, Miguel Domingo, and Francisco Casacuberta, from Volume 45 in September 2017 of Computer Speech & Language. This discussed the role of human input over the course of a neural machine translation algorithm's execution to improve the accuracy of the translation.

Rachael Dewey: I read a number of different papers related to Neural Machine Translation (NMT), the various different NMT models, and the mutual intelligibility of romance and germanic languages. *A Gentle Introduction to Neural Machine Translation* by Jason Brownlee provides important information on NMT and the basics of the various NMT models. *Mutual intelligibility between closely related languages in Europe* by

Charlotte Gooskens, Vincent J. van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt discusses the historical background of the mutual intelligibility among romance and germanic language families. *Multilingual Neural Machine Translation with Language Clustering* by Xu Tan, Jiale Chen, Di He, Tao Qin, and Tie-Yan Liu explores the use of multilingual NMTs using languages clustered according to language family.

Next: Indicate which of the ideas was selected to continue for the project.

We decided to build an RNN model using the basic encoder-decoder structure for our final model.

D. Data sources

Maggie Sellers: I propose the dataset from http://www.manythings.org/anki/. This dataset derives from the Tatoeba Project. The Tatoeba Project is an open-source database of sentences and translations compiled by its members. According to their website, "Tatoeba provides a tool for you to see examples of how words are used in the context of a sentence. You specify words that interest you, and it returns sentences containing these words with their translations in the desired language."

Rachael Dewey: The Stanford NLP team offers a medium-sized dataset of English-German translations, a large English-Czech dataset, and a small English-Vietnamese dataset. (https://nlp.stanford.edu/projects/nmt/)

Next: Indicate which of the data sources was selected to continue for the project.

We decided to use the manythings dataset. All the data for each language in this dataset was in one file, making reading in the data much easier. The manythings data was also in the format of a tab-separated values file, which made preprocessing much, much easier. Finally, the manythings dataset contained sentences that are similar to what a layperson would use in an everyday setting. This means that a model created from this data may be more useful.

E. Approach

For our data, we utilized the text files deu.txt, fra.txt, and ita.txt, all of which we downloaded from http://www.manythings.org/anki/. These files consisted of several sentences in English tab-separated from their translations according to the Tatoeba Project in their respective languages. Each line in these files contained one sentence and its translation.

We utilized Python via Jupyter notebooks and Google Collaboratory to build our NMT models. For our Python packages, we used string, statistics, re, nltk, numpy, pandas, keras, matplotlib, sklearn, collections, google.colab, unicodedata, io, tqdm, tensorflow, time, os, and rouge.

We used several different approaches to our model as our understanding of the data, Neural Machine Translation, and Recurrent Neural Networks (RNNs) evolved over the course of this project. For our initial baseline model, we built a Long Short Term Memory (LSTM) encoder-decoder RNN model. This was due to the fact that encoder-decoder models are more effective than traditional RNNs at dealing with input and output sequences of various lengths. Encoder-Decoder RNN models consist of two RNNs. The first RNN encodes the variable-length source sequence to a fixed-length vector, while the second RNN decodes the fixed-length vector back into a variable-length target sequence. We initially selected the LSTM cells for the encoder-decoder RNN model in order to avoid issues with vanishing gradients that are prevalent within traditional RNNs. We built one model for each of the three languages: German, French, and Italian.

We ran into a multitude of issues when building the baseline model with our original datasets. We were eventually forced to find new datasets to work with in order to build a functioning baseline model. However, when it came to building our final model using the new datasets, we found that we had a great deal of difficulty making any significant improvements to our baseline model via tuning the hyperparameters of the model. After performing additional exploration and research, we decided to attempt a GRU encoder-decoder model, as GRU cells often perform better on smaller training datasets, which we were limited to for the purposes of computational efficiency. Additionally, we found that, with sentence sequences that were much shorter than the original datasets, the vanishing gradient issue was not very prevalent, and the more sophisticated LSTM cells that are required for modeling long-distance relations were no longer necessary for the new datasets.

We used Bilingual Evaluation Understudy (BLEU) and Recall Oriented Understudy for Gisting Evaluation (ROUGE) to evaluate the quality of our NMT models. We chose to use BLEU as a primary evaluation metric for a number of reasons. BLEU is designed specifically to evaluate Machine Translations by comparing strings of words and can be thought of as a measure of similarity. In our case, the BLEU scores are a number between 0 and 1 that suggests the quality of the machine translation as compared to human

translation (i.e. the predicted to test data set). BLEU scores that are closer to 1 indicate that the machine translation more closely reflects the human translation. Although BLEU is not necessarily an effective metric to directly compare across languages, it was useful for interpreting the usefulness of each model and measuring improvement between the base and final models.

We also elected to utilize ROUGE as a primary evaluation metric. We used ROUGE-N, which measures unigram and bigram overlap between the predicted and reference English translations. Specifically, we used ROUGE-1 (unigram) precision and recall. The ROUGE-1 precision can be interpreted as the percentage of unigrams in the predicted sequences that also appear in the reference sequences, while ROUGE-1 recall can be interpreted as the percentage of unigrams in the reference sequences that also appear in the predicted sequences. In general, ROUGE is considered more useful for interpreting the performance of models across different languages than BLEU.

Project Deliverable	Contributor
Project Proposal	The selected project topic was proposed by Maggie Sellers. Each group member completed half of the Project Proposal Submission.
Data Preprocessing and Exploratory Data Analysis	Maggie implemented the preprocessing and exploratory analysis for the German and Italian datasets, while Rachael did the same for the French dataset.
Baseline Model	Since we experienced difficulty building a working model with our initial datasets, Maggie found new datasets and updated our preprocessing and exploratory analysis for the new datasets. Maggie implemented the baseline models for the German and Italian datasets, while Rachael did the same for the French dataset. Rachael selected the metrics and implemented them for all three models, as well as developed the written explanation for the metrics.
Final Submission and Report	Maggie implemented the final model for the German dataset, while Rachael implemented the Italian and French final models. Rachael completed the report sections B as well as E-G, while Maggie and Rachael both worked on sections C-D.
Final Presentation	Rachael created and built the slide deck. Both contributors will present half of the material for the final presentation.

F. Results

Exploratory Analysis

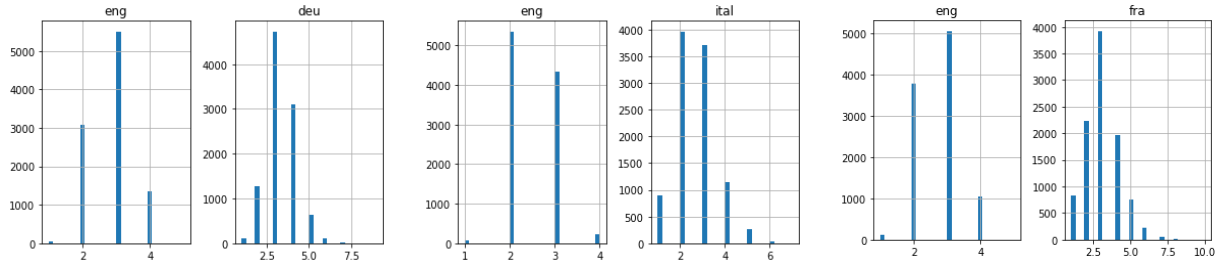
During the Exploratory Analysis phase, we compared the total number of tokens, number of unique tokens, average sentence length, and distribution of sentence length for each dataset. From the exploratory analysis, it can be seen that all of the language datasets are comparable in terms of number of total tokens and number of unique tokens. Across the three datasets, there is an average of 26,641 English tokens and 29,985 target language tokens, with an overall average of 28,313 tokens. Additionally, there is an average of 1,918 unique English tokens and 3,943 unique target variable tokens, with an overall average of 2,931 unique tokens. Finally, the average English sentence length is 2.6641 and the average target language sentence length is 2.9985, for an overall average sentence length of 2.8313.

Exploratory Data Analysis

Dataset	Language	Number of Tokens	Unique Tokens	Average Sentence Length
German	English	28171	2276	2.8171
	German	33356	3646	3.3356
Italian	English	24736	1415	2.4736
	Italian	26023	3665	2.6023
French	English	27016	2063	2.7016
	French	30576	4518	3.0576

From the distributions of sentence lengths for each dataset, shown in the figure below, it is clear that these datasets utilize mostly short sentences, within a range of 1 to 10 words per sentence. The English language sentences, in particular, range only between 1 and 4 words per sentence.

German **Italian** **French**



LSTM vs GRU Model Comparison

When comparing our baseline models, which contain LSTM cells, to our final models, which contain GRU cells, it is clear that the final models perform much better on our datasets. We used the BLEU metric to compare the effectiveness of the baseline to the final models.

Baseline vs Final Models BLEU Comparison

Dataset	Baseline Model (LSTM Cells)	Final Model (GRU Cells)	Percent Improvement
German	0.62017	0.77630	25.175%
Italian	0.70711	0.79305	12.154%
French	0.64550	0.74848	15.954%

From the above figure, it can be observed that the BLEU score for the German model improved 25.175 percent from the baseline to the final model, from a BLEU score of 0.62017 to a BLEU score of 0.77630. The Italian model improved from 0.70711 to 0.79305, a 12.154 percent increase. Finally, the BLEU score for the French model increased 15.954 percent from 0.64550 to 0.74848. Therefore, we decided to retain the GRU cells for our final model and use this model to compare translation accuracy across the three languages.

Language Translation Comparison

Ultimately, we were most interested in comparing the results of the three language models (German, Italian, and French) in order to answer our research question and determine whether close language family relationships are a reliable indicator of more effective Neural Machine Translation models. In addition to the BLEU scores listed in the above table, we also used ROUGE-N scores to compare across the three final models.

Language Translation ROUGE Comparison

Dataset	ROUGE-1 Precision	ROUGE-1 Recall	ROUGE-2 Precision	ROUGE-2 Recall
German	0.25936	0.13999	0.08120	0.04034
Italian	0.34046	0.18993	0.14288	0.06200
French	0.21183	0.12837	0.08838	0.04283

For both the BLEU scores and the ROUGE scores, the same pattern emerges: contrary to our expectations, the Italian translation model performs the best out of all of the models, with an average BLEU score of 0.79305 and ROUGE-1 precision of 0.34046 and recall of 0.18993. This is followed by the German model, with an average BLEU score of 0.77360 and ROUGE-1 precision of 0.25936 and recall of 0.13999. Finally, the French model has the lowest average BLEU score of 0.74848, ROUGE-1 precision of 0.21183, and ROUGE-1 recall of 0.12837.

G. Summary

Our project set out to explore the relationship between historical linguistics and Neural Machine Translation by examining the performance of NMT models to translate within and across the Germanic and Romance language families. From our prior work, we learned that understanding this relationship is important for the building of multilingual NMTs. We utilized English, a member of the Germanic language family, as our target language. We chose three source languages: German, a member of the Germanic language family, as well as Italian and French, both members of the Romance language family. Through our research, we explored the linguistic histories of our selected languages. We hypothesized that translations to German from English would be the most accurate, as they share the closest linguistic relationship as members of the same language family. Of the two Romance languages, we hypothesized that translations to English from French would be the most accurate, as they share a closer historical relationship than do English and Italian.

We used Python, through Jupyter notebooks and Google Collaboratory, to build models for each language pair we explored. For our baseline models, we built Encoder-Decoder Recurrent Neural Networks with Long Short Term Memory Cells. For our final model, however, we used Encoder-Decoder RNNS with Gated Recurrent Unit cells, in order to improve performance and computational efficiency. Using the BLEU and ROUGE metrics, we found that the Italian model performed the best, followed by German, and then finally French. This did not support our hypothesis that languages that are more closely related linguistically will perform better in Neural Machine Translation. One reason for the high performance of the Italian model might be that the Italian source language data has a lower average sentence length and smaller maximum sentence length than the other source language data. Further exploration into the relationship between historical linguistics and Neural Machine Translation could calculate similarity measures between languages and use larger datasets and more languages to improve results. Future work could also explore the performance of multilingual Neural Machine Translation on different language clusters beyond what has already been examined in existing research.

H. References

Brownlee, J. (2019, August 7). *A Gentle Introduction to Neural Machine Translation*. Machine Learning Mastery. https://machinelearningmastery.com/introduction-neural-machine-translation/.

Brownlee, J. (2019, August 7). *Encoder-Decoder Recurrent Neural Network Models for Neural Machine Translation.* Machine Learning Mastery. https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/.

Brownlee, J. (2019, August 14). *Encoder-Decoder Long Short-Term Memory Networks.* Machine Learning Mastery. https://machinelearningmastery.com/encoder-decoder-long-short-term-memory-networks/.

Eriksson, Meredith. (2019, August 28). *Which Languages Are The Closest To English?* Babbel Magazine. https://www.babbel.com/en/magazine/languages-closest-to-english.

freeCodeCamp.org. (2017, October 26). *An intro to ROUGE, and how to use it to evaluate summaries.* freeCodeCamp.org. https://www.freecodecamp.org/news/what-is-rouge-and-how-it-works-for-evaluation-of-summaries-e059fb8ac840/.

Gooskens, C., van Heuven, V. J., Golubović, J., Schüppert, A., Swarte, F., & Voigt, S. (2017). Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism*, *15*(2), 169–193. https://doi.org/10.1080/14790718.2017.1350185

Khandelwal, R. (2020, January 13). *Intuitive explanation of Neural Machine Translation.* Medium. https://towardsdatascience.com/intuitive-explanation-of-neural-machine-translation-129789e3c59f.

Kostadinov, S. (2019, November 10). *Understanding Encoder-Decoder Sequence to Sequence Model.* Medium. https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346.

Kostadinov, S. (2019, November 10). *Understanding GRU Networks.* Medium. https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be.

Malik, U. (n.d.). *Solving Sequence Problems with LSTM in Keras: Part 2*. Stack Abuse. https://stackabuse.com/solving-sequence-problems-with-lstm-in-keras-part-2/.

Peris, Á., Domingo, M., & Casacuberta, F. (2017). Interactive neural machine translation. *Computer Speech & Language*, *45*, 201–220. https://doi.org/10.1016/j.csl.2016.12.003

Prateek, J.. Experienced in machine learning. (2020, May 12). N*eural Machine Translation: Machine Translation in NLP.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/01/neural-machine-translation-keras/.

Olah, C. (2015, August 27). *Understanding LSTM Networks.* Understanding LSTM Networks -- colah's blog. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Tan, X., Chen, J., He, D., Xia, Y., Qin, T., & Liu, T.-Y. (2019). Multilingual Neural Machine Translation with Language Clustering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. https://doi.org/10.18653/v1/d19-1089

Taneja, A. (n.d.). *Attention for Neural Machine Translation (NMT).* LinkedIn. https://www.linkedin.com/pulse/attention-neural-machine-translation-nmt-ajay-taneja/.

Vagadia, H. (2020, December 3). *Seq2Seq Models: French to English translation using encoder-decoder model with attention*. Medium. https://medium.com/analytics-vidhya/seq2seq-models-french-to-english-translation-using-encoder-decoder-model-with-attention-9c05b2c09af8.