Comparative Analysis of Languages To English Through Machine Translation

Anonymous ACL submission
https://github.com/StevenKoniaev/comp550_sub

Abstract

In this study we analyze the complexity and relationship between different languages and English. We do this by analyzing the effectiveness of Machine Translation from a target language to English. We examine a variety of Romance Languages, encompassing French, Dutch, German, Italian, and Romanian in our analysis. Furthermore, to increase diversity, our investigation extends to Korean and Arabic. Each language is taken from the IWSLT2017 Machine Translation task, with the goal of translating a phrase of an initial language to another, in our case English. We explore common challenges in Machine Translation, including model complexity, data cleaning, and computational limits. Our results show that despite computational limitations, we produce results which are helpful to draw conclusions about semantic and syntactic relationship between different languages. Our conclusions align with modern day NLP beliefs about difficulty of translations between different languages to English as the most linguistically similar languages have the strongest translations. We find that the best translations come from French, Dutch, German, Romanian, Italian, Arabic, and Korean respectively.

1 Introduction

Machine Translation is a classical Seq2Seq task in Natural Language Processing with a variety of aspects to explore. Being from distinct backgrounds, we are motivated by looking at the comparisons that can be drawn from how different languages compare to English in the context of Machine Translation. More specifically, our goal is to look at translations between a target language and English and compare them. We want to know what sort of hierarchies form, and which languages perform similarly or irregularly and what are the limitations in translating some languages given our chosen method.

The data set chosen, IWSLT2017 has translations of several languages from the same set of English phrases which makes it ideal for these experiments. From this dataset we picked French, Dutch, Korean, Arabic, German, Romanian and Italian to be the target languages.

Intuitively, our Hypothesis is that since Latin based languages share the most roots with English and thus perform the best in this task. Dutch and German are linguistically most similar to English, so we expect these languages to perform the best of the ones chosen. We expect Arabic to perform the worse due the language beings structured very differently from English, reading from right to left.

We evaluate our results based on BLEU score, a popular metric used for Machine Translation, while also providing some examples of translations. There are some limitations in our study. The first is the computational resources available for us to use, we therefore make some simplifications that aid in generating quicker and better results. This includes generous data filtering and using smaller Seq2Seq models. For the model itself, we wanted a modern method that still performs well with limited computational power. Because of this we decided to use a GRU Network instead of more complex architectures like Transformers or LSTMs. There are several options and decisions to be made about the model which we discuss thoroughly in this study. Our main model consists of a single layer GRU in the encoder with a GRU and Bahdanau attention mechanism in the Decoder.

Another issue we face is that we do not have the knowledge of all of these languages which makes it difficult to see how well the annotations and translations capture the meaning of the original phrase.

Overall, we attempt to use neural machine translation as a medium to study the relationships between different languages and English as we explore the difficulties of translating a target language to English.

2 Related Work

There are several key papers that we referenced in approaching this study. Our first inclination was to look at Vaswani et al.'s *Attention is All You Need*. The paper which first introduced the concept of the Transformer architecture. The authors had used transformer models to carry out English-French and English-German translation experiments, achieving at the time a state of the art BLEU score of 28.0 for the German translation, and a score of 41.0 for the French task, surpassing much of the previous attempts in the literature. We were interested in this approach at first as Large Language Models are built from this architecture, but we quickly realized that the computational cost of this method would be far too high. However, we still use a type of attention based off the hidden states from a GRU, and we use the same dataset as a comparison. However, since we generously filter the data we expect our results to be much better as we are not expecting OOV items or items that have a very low frequency in the corpus.

In COMP 550, we saw the concept of recurrent neural networks which are very powerful for Seq2Seq tasks, making it a natural candidate for machine translation. We looked at *Neural Machine Translation by Jointly Learning to Align and Translate* by Dzmitry Bahdanau et al and *Effective Approaches to Attention-based Neural Machine Translation* by Minh-Thang Luong et al. Both of these papers introduced methods for using attention in LSTMs and achieve more precise results by attending to relevant words. Bahdanau allowed a model to soft-search for parts of a source sentence that are relevant to predicting a target word with the use of a variable-length vector for a encoder-decoder architecture. Luong iterated on this by a local attention mechanism, whereby only a subset of source words are considered at a time instead of the entire source sentence.

We differ from Bahdanau's and Luong's paper by employing smaller networks that rely on the GRU architecture. However, we use the same attention mechanism as introduced by Bahdanau. Our breadth of languages explored also differs from both Luong's and Bahdanau's papers as we perform Machine Translation on a much wider pool of languages compared to all of these papers. Reducing data size and cleaning helped in approaching our computational needs. We explore variations on the models we used such as the choice of hidden dimension size, word embedding size, and bidirectionally of the encoder.

3 Methods

A dataset was chosen with a diverse amount of languages for Machine Translation. We used the popular IWSLT2017 for Machine Translation. From this dataset we picked French, Dutch, Korean, Arabic, German, Italian, and Romanian. Each dataset has equivalent English sentences which makes it more interesting to analyze and compare translations. We then proceed to make some simplifications in the data due to the limited amount of computational resources available.

3.1 Dataset

The dataset which we used is taken from the IWSLT 2017 Evaluation Campaign, introduced by Cettolo et al. in their paper *Overview of the IWSLT 2017 Evaluation Campaign*. We gather the data from Hugging Face, and import them as Pandas dataframe which we clean, vectorize, and use alongside Pytorch to implement the model.

The motivation behind the IWSLT 2017 Evaluation Campaign was to evaluate models built to carry out three distinct tasks under the field of machine translation; many-to-many machine translation systems, resolution of anaphora typically in human dialogue, and automatic transcription of spoken natural language. We, in particular, focused on the task of machine translation. The motivation behind this dataset and the task is, in part, taken from the task of translating TED talks, which has continued to receive attention since its introduction to IWSLT back in 2010. Much of the data came from older TED talks, supplied by the WIT^3 website, the Web Inventory of Transcribed and Translated Talks, containing transcriptions and translations of TED talks[6]. In addition, more recent TED talks were added to the IWSLT 2017 dataset. The IWSLT 2017 dataset itself does not contain entire TED talks, but rather several sentence-long snippets, both in the origin language and a translation to the target language. The dataset is organized in a fashion where each file contains several sentences, both in an origin language and a translation of that sentence into a target language.

3.2 Data Cleaning

We remove all phrases with more than 20 words. Then we further removed pairs of phrases where

a word in a target language is seen twice or less. We do this for English phrases as well but on a larger scale, removing sentences with words that are present 5 or less times in the corpus. Finally, we remove all phrases in the test set with words that are not present in the target language's and English vocabulary in order to avoid OOV items.

These simplifying adjustments to the dataset was used for various reasons. We wanted to decrease the size of the vocabulary and avoid OOV items, this method accomplishes this and allows our model to focus on words that appear frequently while lowering the computational requirements of our experiments. It also makes the distribution of words in each language more uniform which also assists in training smaller sized models.

3.3 Evaluation

We evaluate each Seq2Seq model by generating the corresponding BLEU score relative to the test set. The BLEU score is a popular metric and tries to measure the precision of the predicted phrase through matching n-grams and a brevity penalty. Although it is not a perfect evaluation method it works quickly and gives a general sense of the quality of translation making it a popular choice in literature.

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} W_n \log p_n\right) \quad (1)$$

Where p_n is the geometric average of n-gram precisions and BP is the brevity penalty.[3]

$$p_n = \frac{\sum_{C \in \text{Candidates}} \sum_{\text{ngrams} \in C} \text{Count}_{\text{Clip}}(\text{ngram})}{\sum_{C' \in \text{Candidates}} \sum_{\text{ngrams}' \in C'} \text{Count}(\text{ngram}')} \quad (2)$$

We use NLTK's implementation of the BLEU score and use the standard 4-gram comparison. We use a smoothing function to to avoid the edge case of having a phrase with 0 matching (or existing) n-grams.

3.4 Model

The architecture for Translation chosen was an Encoder Decoder Model relying on the GRU unit and additionally has Bahdanau Attention. We use PyTorch to implement the models.

`GRU Unit:`

The GRU is an RNN and a simplified version

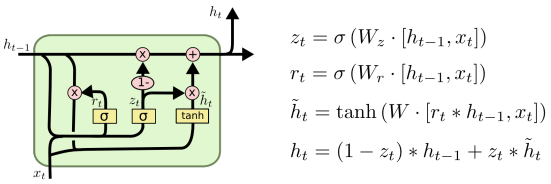

Figure 1: GRU Architecture [2]

of an LSTM cell. An LSTM has three gating components (input, forget, output). The GRU has two gating components, the reset gate and the update gate. The reset gate determines how much of the previous hidden state should be lost, while the update gate determines how much of the new input to use. In the diagram above, z_t represents the update gate output. r_t represents the reset gate output. A weighted sum is taken between the previous hidden state and the proposed one, which is determined by the update gate. An LSTM has an additional memory cell, used for long range dependencies which is not present in the GRU. The GRU combines the memory cell and hidden cell together and reduces the amount of parameters needed.

`Bahdanau Attention:`

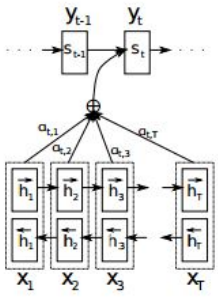

Figure 2: Bahdanau Attention Mechanism [4]

We generate keys and queries from a feedforward network using the encoder outputs and hidden outputs from the encoder. Bahdanau's attention is additive, we follow his implementation and add the keys, queries, and send them through another feedforward network to produce values. These values are normalized through $tanh$, then softmaxed, This gives us the attention scores. We then multiply by the original keys to get the new inputs to use in the Decoder GRU.

3.5 Hyperparameters and Design Decisions

We tried a variety of methods in order to try to optimize the model's performance relative to the computational requirements needed. We present a

discussion on several hyperparameters and model design decisions and how it altered performance.

1. Bidirectional Encoder
   In the literature bidirectional encoders are favored as they provide more information overall by having outputs and hidden states associated with moving forward and backwards through the sequence. We use both a Forward and Bidirectional GRU to see how each language responds to such a change in architecture.

2. Number of GRU Layers
   Adding additional layers had a very weak impact on gaining performance quickly. Each layer essentially doubles the parameters associated with the GRU which nearly doubled the time to train a similar model of a single layer. Our computational constraints allowed us to only train a one layer GRU.

3. Word Embedding Size
   We tested different common embedding sizes ranging from 256, 512, 768, 1024. We saw no noticeable improvements in performance, because of this we left the embedding dimension at 256.

4. Hidden Dimension Size
   This hyper parameter determines the size of each hidden state. We tried values of `256, 512, 768, 1024`. For the unidirectional case we used a hidden state of 1024. For the bidirectional experiments we used 768 to reduce training time. We found that this parameter is quite important in generating good results.

5. Learning Rate
   A low learning rate is needed for this task due to the high dimensional count of the hidden state. However, too high of a learning rate leads to the model beginning to diverge after 30 epochs. We amend this by keeping a higher rate for training the first 20 epochs at 3×10^{-4} learning rate. We decrease this parameter to 1×10^{-4} for the next 5 epochs which helps the model converge to losses below a tenth while keeping training time lower than just using a low learning rate from the start.

4 Results

4.1 Tables

In order to reduce the complexity of the model needed and the Computational Requirements we impose heavier restrictions on the frequency as stated previously. The results show the number of phrase pairs for each Target language to English (Target Language $\rightarrow$ English) as well as the vocabulary size before and after filtering. Finally we provide the BLEU scores associated with each language translation to English.

Language	Vocabulary	Train Size	Test Size
French	46414	139849	5256
Dutch	62870	164993	1288
Korean	195698	151348	5741
Arabic	114155	151431	5755
German	69745	133847	5428
Italian	56717	149417	1047
Romanian	69017	142390	1147
English	42164	-	-

Figure 3: Initial Dataset of Phrases Less Than 20 Words

Language	Vocabulary	Train Size	Test Size
French	15732	88014	3094
Dutch	17392	99567	753
Korean	25756	34887	992
Arabic	29021	64170	2112
German	16571	69029	2692
Italian	18628	90316	627
Romanian	22024	80083	668
English	9989	-	-

Figure 4: Filtered Dataset

Language	Forward	Bidirectional
French	44.52	51.49
Dutch	46.24	47.29
Korean	23.36	22.69
Arabic	29.71	36.05
German	45.37	46.13
Italian	44.77	44.53
Romanian	43.41	45.51

Figure 5: BLEU Score Results

4.2 Example Translations

Given below are some outputs generated.

Language: French

From: vous devez utiliser ces cellules pour réfléchir à ce jeu

Annotated: you should use those cells to think carefully about this game

Predicted: you should use those cells to think about this play

Language: German

From: all das ist mit einem handy möglich.

Annotated: all this is possible with your mobile phone

Predicted: so it is all that is possible with a cell phone

5 Discussion

5.1 Analysis

Despite the computation limitations, the results obtained indicate several things about how these languages relate to English. The sample translations show that meaning can indeed be captured and the Decoder can produce semantically equivalent sentences. The BLEU scores indicate that our initial hypothesis of Latin based languages outperforming non-Latin based languages holds true.

Surprisingly, in our second trial French becomes the highest scoring language when introducing the bidirectional Encoder which disagrees with our Hypothesis that Dutch and German should score higher. Interestingly, the authors of Attention is all you need also scored better on French than German from their experiments. One reason why the performance of German and Dutch change so little in the bidirectional case could potentially mean there is too much "obvious" information provided in the bidirectional case for German and Dutch since their structure is similar to English. Korean begins to perform worse in the bidirectional trial and Arabic achieves a higher score. This could be explained by Arabic needing more context due to the structure of the language being from right to left. Korean's lower score could signal needing higher hidden size dimensions as we reduced the value from 1024 to 768 for the experiment which impacted performance. The lower scores on Korean and Arabic test sets can be due to cultural and contextual phrases spoken in these dialects which may not even exist in English or vice-versa. Korean has a lot of word ambiguity, honorifics, cultural nuances and inequivalent particles to English. Korean also has a subject-object-verb order in its sentence structure while English has subject-verb-object. These reasons all increase the difficulty of translations. Arabic also differs in sentence structure having verb-subject-object ordering. We believe we remove many more complicated words in the initial data filtering process which significantly helped in translating Arabic, since the root based morphology of the language could make more complicated words much more difficult to translate. All the Romance Languages form a cluster and only deviate by a few BLEU score points from each other, highlighting the Latin roots of these languages. Languages which have similar roots not only influence the syntax and semantics but also the progressive development of new phrases with a similar context, this justifies the better performance on languages which share Latin roots as most of the Latin languages share many similar words.

5.2 Extending Experiments, Limitations

Our main limitation was in the amount of computational power available. Because of this we were limited in the amount depth for the neural network and the amount of data to use on it. Due to the limitations of computational power that we had access to, we decided to carry out our experiments using a smaller GRU-based model. We believe we could have converged to better values given more epochs and smaller learning rate. To continue these experiments we would carry out these trials using a more powerful model using things like a transformer, not limiting dataset through length of sentences and trying to train on a higher vocabulary count, and most importantly analyzing more languages, specifically ones of different origins than English.

5.3 Conclusion

We found that the Romance languages perform similarly in Machine Translation given equal computational resources. Eastern languages have a lot of aspects which makes translation difficult to English due to the sentence structure and composition of the words. Our initial hypothesis seems to overall hold true, but we had some interesting results and unexpected performance increases by varying the model used such as French and Arabic who both increased scores significantly by introducing bidirectional encoders.

6 Statement of Contribution

Steven: Code, report. Taran: Report. Priyanshu: Report.

We discussed and all participated in analyzing the results, discussing our approach and implementation details and what we can do with our limitations.

References

[1] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

[2] Olah, Christopher."Understanding LSTM Networks." colah's blog, 27 Aug. 2015, https://colah.github.io/posts/2015-08/Understanding-LSTMs/.

[3] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[4] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[5] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[6] Cettolo, Mauro, Christian Girardi, and Marcello Federico. "Wit3: Web inventory of transcribed and translated talks." Proceedings of the Conference of European Association for Machine Translation (EAMT). 2012.