Active Learning to Improve Low-Resource Neural Machine Translation

Colton Bishop
Princeton University
cmbishop@princeton.edu

Pulkit Singh
Princeton University
pulkit@princeton.edu

Abstract

This paper explores and evaluates active learning approaches that seek to select optimal sentences with which to train a neural machine translation system with limited data. These approaches leverage multiple features (namely, the perplexity of candidate sentences with respect to a language model generated from the examples the model has seen, the distribution of the attention mechanism on candidate sentences, the frequency with which the candidate sentences' unigrams and bigrams appear in previously seen examples, and sentence features such as length and word repetition). The proposed approaches are extrinsically evaluated and their translation performance is compared with a baseline random sentence selection strategy. This paper also explores how varying degrees of resource-scarcity impact these approaches.

1 Introduction

It has been demonstrated that the performance of neural machine translation models in particular degrade significantly in low-resource contexts with respect to other systems such as rule-based or statistical machine translation.[1] However, in recent years the neural approach to machine translation has been shown to greatly surpass prior approaches to machine translation. For languages with very limited translated data available, as is the case with many indigenous languages, there are often few computational linguists able to design sophisticated rule-based systems. Thus, example-based approaches like neural machine translation systems (which require no concrete linguistic expertise and can be more easily generalized to unseen languages) provide a hope for effective translation to these unsupported languages. Furthermore, recent work in the last few years has shown that neural networks properly constructed for low-resource conditions are able to surpass the alternative methods with far less data than was previously though: with as few as 100,000 sentences.[2] Data collection initiatives in indigenous language spaces have not yet risen to meet the opportunities created by these new machine translation breakthroughs,[3] but when they do, it is of the utmost importance have effective data collection strategies (or active learning strategies) so that the impact of translation labor is maximized.

Project Goal

This paper proposes and evaluates several feature-based approaches to active learning that seek to improve the performance of a neural machine translation (NMT) system in low-resource conditions. It seeks to understand which individual features are most important when it comes to selecting the most useful sentences to translate and it studies how the impact of each of these features changes with respect to the amount of translation data available. A combined approach that leverages multiple weighted features and seeks to learn optimal weightings for each of these features is explored. Each approach is evaluated extrinsically by assessing the performance of a neural machine translation system that is trained on data collected using said approach.

[1] Liu, M., Buntine, W., Haffari, G. (n.d.). Learning to Actively Learn Neural Machine Translation. Retrieved from https://www.aclweb.org/anthology/K18-1033.

[2] Sennrich, Rico, Zhang, Biao. (2019, May 28). Revisiting Low-Resource Neural Machine Translation: A Case Study. https://arxiv.org/abs/1905.11901.

[3] Font Llitjos, A., Aranovich, R. Building Machine translation systems for indigenous languages . Language Technologies Institute: Carnegie Mellon University . http://www.cs.cmu.edu/ aria/Papers/FontAranovich_CILLA2_mapuche_quechua(2).pdf

2 Related Work

The task of active learning for machine translation has been approached in many different ways. Some approach this problem as a machine learning classification problem. (Peris, 2018) frame the problem as follows: given "an unbounded stream of source sentences" determine which sentences are worth translating by a human agent and which are not.[4] They also raise the idea of exploiting information from the attention mechanism to evaluate candidate sentences.

However, as (Liu, 2018) explain in their paper "Learning to Actively Learn Neural Machine Translation," traditional active learning heuristics for machine translation such as the one presented above are significantly limited when there is very little initial bitext.[5] Methods that rely exclusively on information learned from already translated text are limited. One potential avenue to approach this problem is presented by (Sennrich, 2016), who demonstrate that when the amount of bilingual text is limited, leveraging monolingual text can greatly improve the effectiveness of active learning.[6]

In (Bishop, 2020), an independent work project developed by Colton Bishop this past semester, a low-resource optimized sequence-to-sequence neural network and online web framework for the collection of indigenous language translation data is created and evaluated. As a component of this project, two active learning methods that approached the task as a classification problem were implemented and intrinsically evaluated. One of these methods approaches the problem by leveraging monolingual (untranslated) text.[7] This paper will advance this preliminary research which will be explained in more detail in the following Preliminary Research section below.

[4]Peris, A., Casacuberta, F. (n.d.). Active Learning for Interactive Neural Machine Translation of Data Streams. https://www.aclweb.org/anthology/K18-1015.

[5]Liu, M., Buntine, W., Haffari, G. Learning to Actively Learn Neural Machine Translation. Retrieved from https://www.aclweb.org/anthology/K18-1033.

[6]Sennrich, R., Haddow, B., Birch, A.. Neural Machine Translation of Rare Words with Subword Units. https://www.aclweb.org/anthology/P16-1162/.

[7]Bishop, C. (2020) Resource-Scarce Machine Translation: Designing a System to Facilitate the Data Collection and Machine Translation of Indigenous and Resource-Scarce Languages. Princeton Fall 2019 Independent Work Project

3 Approach

3.1 Preliminary Research: Insights from Independent Work

The two basic approaches to active learning explored in the aforementioned paper both frame the problem as a classification task on a stream of incoming sentences, following (Peris, 2018). There are two methods explored: a frequency-based approach and a perplexity-based approach.

The frequency-based approach analyzes a large corpus of monolingual English text to learn the distribution of words and phrases. It then cross-references this knowledge with the corpus of translated text to figure out which of the highest priority (most common) words and phrases are not sufficiently represented in our translated corpus. In the evaluation of a candidate sentence, if the sentence has a high enough number of underrepresented words or bigrams with respect to a set threshold, it is classified as useful and selected for translation. The intuition here is that this approach helps the model acquire the most important (by frequency) words and phrases as quickly as possible to best improve its applicable coverage of a language.

The perplexity-based approach, on the other hand, begins by analyzing the English corpus that is already translated to construct a language model. It leverages the perplexity metric, which is a measurement of how well a language model can predict a sample of text, which is in this case a candidate sentence. It evaluates the perplexity of new sentences with respect to this language model and if it is above a certain threshold, this sentence is classified as useful and selected for translation. The intuition here is that the sentences with the highest perplexity according to the learned language model will "surprise" our translator the most, and thus they would be the most important sentences to translate.

In (Bishop, 2020), these methods were only evaluated intrinsically by studying properties of the actively learned corpora of examples, with metrics such as percentage of unigrams and bigrams in an unseen text that are present in the actively learned corpora or the perplexity on an unseen text with respect to a language model generated from the actively learned corpora.

3.2 Designing Improved Approaches

This paper will seek to leverage key insights from this independent research project and from other related work to design an improved approach to active learning.

We observe that one major limitation to the frequency-based approach described above is that the algorithm only considers the frequency of words and phrases in English, rather than also analyzing data from the minority language. This can create an imbalance in the text that is selected for translation, which can subsequently result in problematic ethical issues and social issues; as indigenous languages are often much less strictly defined and grammatically regulated by a centralized governing body compared to dominant languages, emerging translators and data sets that are skewed too heavily toward dominant languages with large monolingual texts to analyze run the risk of disenfranchising indigenous language authorities. Thus, the approaches of this paper will limit their scope to information learned from bitext only and will not assume or leverage monolingual text data.

We next observe that treating active translation as a classification task, as (Peris, 2018) and both the frequency and perplexity-based approaches above do, limits the extent to which we can compare the usefulness of useful sentences. In other words, while it helps us determine which sentences should be translated and which should not, it does not tell us which of the sentences are the most important relative to the other important sentences. Thus, the sentences selected for translation become much more heavily influenced by the order in which they are evaluated. To lessen this influence, the approaches in this paper approach active learning as a comparative ranking problem: given a batch of candidate sentences, rank the sentences in order from most important to least important.

We next observe that the use of perplexity as the sole metric for evaluation of candidate sentences, as it is in the perplexity-based approach described above, is very limited. In practice, while perplexity is often a useful metric, the perplexity of many sentences evaluates to infinity whether or not they have one unseen bigram or ten unseen unigrams. As a result, this metric is especially flawed in very low-resource contexts and the early stages of active learning. Thus, this paper will seek to develop metrics that more effectively capture information like this. In general, this paper seeks to develop approaches that leverage a more diverse set of features to evaluate sentences. It will draw on features exploited successfully in related work (such the attention mechanism and perplexity).

Finally, during testing and development, these approaches will be refined and optimized extrinsically (by seeking to maximize how much these approaches improve the translation of an NMT system) rather than seeking to intrinsically maximize certain properties of the training corpus, as was demonstrated by (Bishop, 2020).

3.3 Feature Selection

In designing the features to consider in the evaluation of candidate sentences, we selected a set of metrics that intend to expose key sentence properties with regard to the current state of the system. The following features were selected:

- Sentence Length
- Frequency of Seen N-grams in Training
- Count of Unseen N-grams
- Count of Repeated Unigrams
- Sentence Perplexity
- Distribution of the Attention Mechanism

This paper explores and evaluates three feature-based active learning approaches using different combinations of the above features: an N-gram approach, an attention-based approach, and an aggregate weighted approach.

4 Implementation

4.1 The N-Gram Approach

As has been demonstrated by (Bishop, 2020), language models generated from the training examples that the NMT system has already seen shed important light into the value of candidate

sentences. However, perplexity alone is not adequate or nuanced enough of an approach to evaluate sentences. This approach seeks to complicate and advance the perplexity-approach by incorporating other information into the analysis: namely, how often each n-gram in the sentence appeared during training and a count of how many unseen n-grams are present in the sentence. We also introduce smoothing into the language model. For both the frequency of seen n-grams and the count of unseen n-grams, we consider only unigrams and bigrams.

We sought to design a score function that can convert each of these features into a metric that serves as a assigned score for candidate sentences. We aimed to capture the following intuitions. We want to weigh the number of unseen unigrams (u_1) and unseen bigrams (u_2) heavily such that sentences with many of these receive a higher score. If the model has seen many of the sentence's unigrams and bigrams before (s_1 and s_2, respectively) we would like to punish the model slightly; many previously seen n-grams are not ideal, as translating them may be a waste of valuable time and resources, but they also shouldn't strongly impact the score relative to the count of completely unseen n-grams. Finally, we would like to incorporate the perplexity metric such that if the perplexity (p) of any sentence falls below a certain perplexity threshold (t), a poor score is given and that sentence is not selected. To capture these intuitions, we devised the following score equation:

$$\text{score} = \mathbb{1}_{p>t}\left(\frac{u_1+u_2}{log(s_1+s_2)}\right)$$

This component was implemented in Python. Unlike in the previous independent work of (Bishop, 2020), which implemented its own language model, this paper's approach employs Python's Natural Language Toolkit to write new functions for the creation and updating of the language model with sentences the model was exposed to. This allowed for improved language model smoothing. Upon learning the translation for a new sentence, this language model is updated along with the dictionary used to store the frequencies of seen unigrams and bigrams. A batch of candidate sentences are then ranked according to this scoring function and sorted according to score.

4.2 The Attention-Based Approach

The second approach this paper explores exploits the attention mechanism of the downstream neural architecture (which was build by following an online Tensorflow tutorial[8]) to assign a score to candidate sentences. In general, the attention mechanism is the component of a network's architecture that is responsible for managing and quantifying the interdependence between the input and output elements.[9]

To visualize this, we look to the following representations of attention: the first on a sentence that is correctly translated by the model and the second on a sentence that is incorrectly translated by the model.

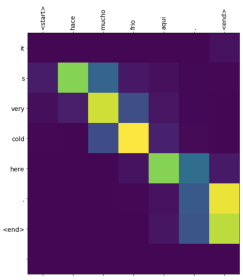

Figure 1: Attention on a Successfully Translated Sentence

We see above that the attention on this correctly translated sentence is fairly evenly distributed over each of the input elements. This uniformity is not present in the below representation of attention on a incorrectly translated sentence. We see that there is poor distribution of attention and that the model seems to be focusing disproportionately on certain input elements (like the period).

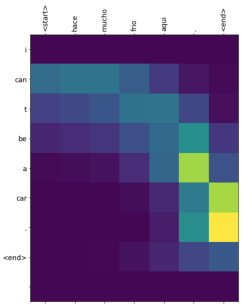

Figure 2: Attention on an Unsuccessfully Translated Sentence

[8]Neural machine translation with attention : TensorFlow Core. (2019). Retrieved from https://www.tensorflow.org/tutorials/text/nmt_with_attention.

[9]Loye, G. (2019, September 17). Attention Mechanism. Retrieved from https://blog.floydhub.com/attention-mechanism/.

In general, we observe that for sentences which the model is able to effectively translate, the distribution of attention over input and outputs elements is more uniform, indicating that the model is focusing on all parts of the input to predict the output. We assume that those sentences which the model has the most challenge focusing on (those sentences whose attention is least uniformly distributed) are most important to see as an example. Thus, the attention-based approach seeks to capture this by assigning each candidate sentence a score (using the equation below) based on how unevenly distributed the attention is on that sentence when the NMT system attempts to translate it.

$$\text{score} = \sum_{i=1}^{n} \left\| \mu - a_i \right\|$$

We calculate this score by first learning the mean maximum attention (μ) on the input elements (an average of the brightest squares in each of the rows seen in Figure one). We next iterate over all n input elements to see how far the maximum attention on each (a_i) diverges from the mean.

This component was also implemented in Python using the same sequence-to-sequence neural network (without resource-scarce optimizations) that was implemented in (Bishop, 2020).

4.3 The Aggregate Approach

The aggregate approach seeks to combine the intuitions of both the attention-based approach and the n-gram approach presented above. It also introduces a new feature (sentence length, L) that in testing improved performance. The intuition is that longer sentences are rewarded because they are more likely to be complex and informative to the model. Two weights (w_1 and w_2, respectively) are introduced to weigh the modified n-gram based component and the attention-based component. The final score equation for the aggregate approach is thus as follows:

$$\text{score} = \mathbb{1}_{p>t}\left(\frac{L * w_1(u_1+u_2)}{log(s_1+s_2)} + w_2\|\mu - a_i\|\right)$$

To find appropriate weights, we implemented a simple algorithm: the weights w_1 and w_2 always add to one. They are initialized equally and the performance of the aggregate approach with these weights is evaluated. Next, a small random change is made to each of these weights and they are then re-normalized to sum to one. Performance is evaluated again. If it improved, we repeat the above steps with the new weights. Otherwise, we repeat with the former weights (that had performed better). This continues until the BLEU score does not improve for a set threshold of times.

5 Evaluation

To evaluate these approaches, we performed an extrinsic evaluation by observing how an NMT system (Bishop, 2020) trained on corpora actively learned through each approach performs under various degrees of resource scarcity compared with a baseline random sentence selection (RSS) approach. Each approach sorted a bulk of candidate sentences with respect to each of their respective scores, and the top sentences of each sorted batch was used to train the system. The metric we chose to evaluate models is the BLEU score, a widely accepted though imperfect metric to evaluate translations. We wrote the evaluation script using NLTK's 'sentence_bleu' library and tested on several thousand unseen sentences with known translations.

5.1 Evaluating the N-Gram Approach

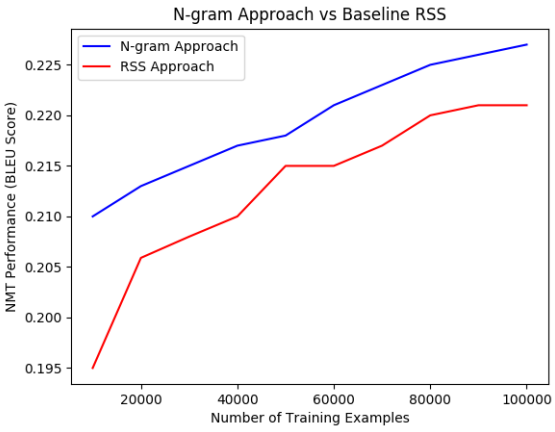

Figure 3: Results of N-gram Approach

We observe that the N-gram approach performs consistently better than the random sentence selection approach, particularly in low resource contexts (less than 20,000 sentences). This seems to confirm that our intuitions about which sentences are more valuable (those most perplexing,

those with the most unseen bigrams, and those with the least seen bigrams) were reasonable.

5.2 Evaluating the Attention-Based Approach

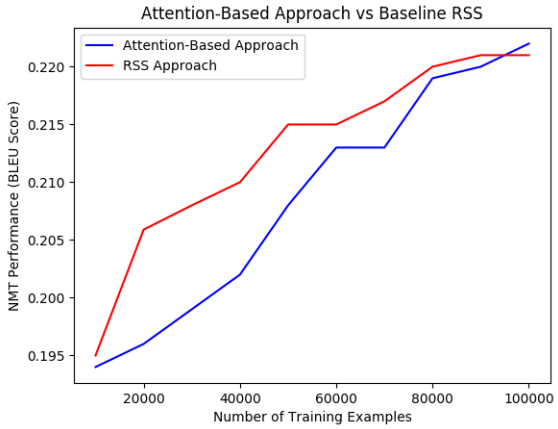

Figure 4: Results of Attention-based Approach

When comparing the performance of the attention-based approach to a random sentence selection strategy approach, we observe that surprisingly the attention-based approach does not outperform the random sentence selection strategy until there are over 90,000 example sentences learned. In fact, it performed significantly worse, especially when the model was trained with between 10,000 and 60,000 example sentences. We hypothesize that this indicates the attention mechanism is less helpful at earlier stages of the model's learning process, or perhaps not helpful at all.

5.3 Evaluating the Aggregate Approach

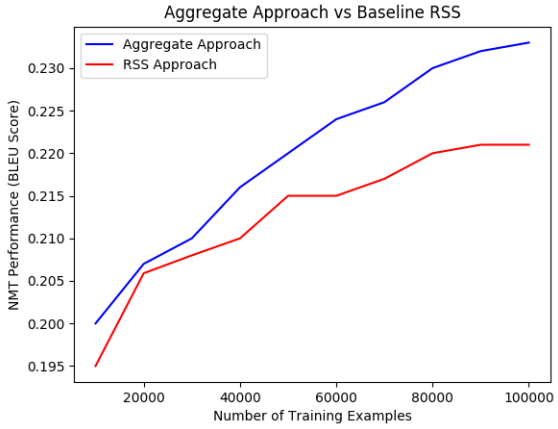

Figure 5: Results of the Aggregate Approach

We observe that at lower resource levels (less than 30,000 sentences) the weighted aggregate approach performs very similarly to the baseline. However, when there is more data, this method significantly outperforms the random sentence selection approach, reaching a BLEU score of .23: the highest that any of our approaches achieved.

6 Conclusions

According to our results, we observe that different active learning approaches perform better at different levels of resource scarcity. We observe below a visual comparison of each approach's performance at different levels of resource scarcity. One important note as we reflect on the performance of these approaches under these data constraints is that the entire spectrum of simulated data amounts can be considered 'low-resource.' When it comes to training effective neural machine translation systems, even the maximum amount of data considered in these experiments (100,000 sentences) is relatively scarce and 'low-resourced' when compared to the many millions of sentences from which accurate NMT systems like Google Translate's learn.

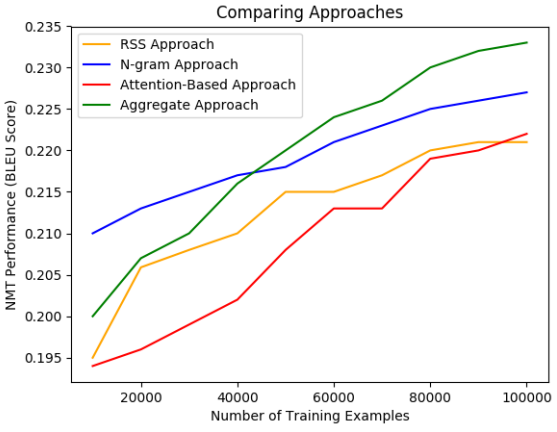

Figure 6: Comparing Approaches

We see that the N-gram method seems to be superior in very low resource contexts, outperforming all other approaches until over 40,000 sentences are learned. The weighted aggregate approach seems to perform the best overall, outperforming the baseline and attention-based approach between 10,000 and 40,000 sentences and outperforming all other approaches when there are over 40,000 sentences. We believe that

the aggregate approach's seeking of improved weights (which gradually improved the BLEU score) as well as its consideration of sentence length is what makes it superior overall.

With the unexpectedly poor performance of the distribution of the attention mechanism as a singular score, we were curious to see how removing attention from the aggregate approach impacted performance. We changed the weighting such that the weight on attention made it almost negligible, and performed the experiments again to see if performance was improved. The results of this are illustrated below.

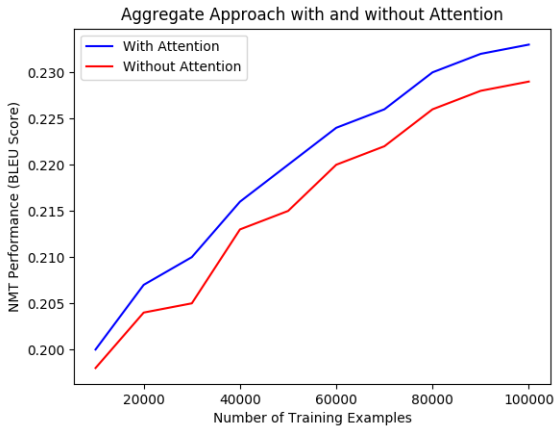

Figure 6: Exploring the Impact of Attention

Surprisingly, we see a decrease in performance. We observe that while attention was not overly helpful as the singular scoring metric, when its influence on the aggregate approach was almost removed the aggregate approach did not perform as well. We observe that this difference in performance is most notable above 20,000 sentences. We hypothesize that perhaps the distribution of the attention is a less helpful metric at the lower data levels but more informative for how sentences impact model performance at the higher levels.

7 Future Work

We would love to continue forward with this research and ultimately launch a system that aids communities in gathering a practically usable corpus of data with which to train language resources. Concretely, we would like to develop more sophisticated scoring equations than the makeshift ones we put into practice within this paper. We would also like to incorporate more valuable features and develop a more effective algorithm or approach to learning the right weights for features that we consider.

Acknowledgments

This was a project we were both really passionate about and we are grateful to Karthik for his advising and enthusiasm for this work, as well as to all the COS 484 staff for helping us build up the NLP arsenal that allowed us to tackle this problem thoughtfully!

Contributions

We did a lot of our initial research into related work individually, but brainstormed, designed, and implemented each of the approaches together. Colton contributed insight into NLTK and language modeling tools and Pulkit contributed a lot of insight into attention and machine learning!

References

Bishop, C. 2020 *Resource-Scarce Machine Translation: Designing a System to Facilitate the Data Collection and Machine Translation of Indigenous and Resource-Scarce Languages.*, volume 1. Princeton University Fall 2019 Independent Work Project

Font Llitjos, A., Aranovich, R *Building Machine translation systems for indigenous languages . Language Technologies Institute: Carnegie Mellon University.*, volume 1. http://www.cs.cmu.edu/ aria/Papers/FontAranovich_CILLA2_mapuche_quechua(2).pdf

Liu, M., Buntine, W., Haffari, G. *Learning to Actively Learn Neural Machine Translation.*, volume 1. https://www.aclweb.org/anthology/K18-1033.

Loye, G. 2019 *The Attention Mechanism.*, volume 1. https://blog.floydhub.com/attention-mechanism/.

Peris, A.,Casacuberta, F. *Active Learningfor Interactive Neural Machine Translation of Data Streams.*, volume 1. https://www.aclweb.org/anthology/K18-1015.

Sennrich, R., Haddow, B.,Birch, A. *Neural Machine Translation of Rare Words with Subword Units.*, volume 1. https://www.aclweb.org/anthology/P16-1162/.

Sennrich, Rico, Zhang, Biao. 2019 *Revisiting Low-Resource Neural Machine Translation: A Case Study.*, volume 1. https://arxiv.org/abs/1905.11901.