[Paper Review]
Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates

Paul Jason Mello
Department of Computer Science and Engineering
University of Nevada, Reno
pmello@unr.edu

Abstract

"Subword units are an effective way to alleviate the open vocabulary problems in neural machine translation (NMT).While sentences are usually converted into unique subword sequences, subword segmentation is potentially ambiguous and multiple segmentations are possible even with the same vocabulary. The question addressed in this paper is whether it is possible to harness the segmentation ambiguity as a noise to improve the robustness of NMT. We present a simple regularization method, subword regularization, which trains the model with multiple subword segmentations probabilistically sampled during training. In addition, for better subword sampling, we propose a new subword segmentation algorithm based on a unigram language model. We experiment with multiple corpora and report consistent improvements especially on low resource and out-of-domain settings." [1]

1 Summary

This 2018 paper provides a novel way to improve neural translation models (NMT) through extensive sampling methodologies and harnessing the ambiguity input and target sequence segmentation as a natural type of noise to introduce stochasticity during training. The researchers also introduce a novel subword segmentation algorithm with roots in the unigram language modeling.

2 Main Contributions

The main contributions of this work are two fold. The introduction of a subword regularization technique aimed to train models to probabilistically sample subwords and a new subword segmentation algorithm to model noise generated during segmentation. Both of these approaches are agnostic of model architecture and can be used to universally improve subword segmentation tasks by an average of $\sim 2\%$ across machine translation languages tasks.

2.1 Key Contributions

The key contributions consist of the following.

- Subword regularization is a novel approach introduced in this work aimed at efficiently sampling segmentations. The key to understanding this approach is the aggregation of segmentation candidates. This means that all possible word segmentations are generated

through sampling, then the segmentation candidates with the highest probability of leading to a strong subword segmentation, based on repeated occurrences, is selected. They improve the efficiency of this sampling approach by using a forward-DP and backward A* algorithm followed by recursively sample from every branch on the lattice. To generate the possible candidates for segmentation, they utilize a "Forward Filtering and Backward Sampling" (FFBS) approach. Together, with sampling from a multinomial distribution and linear time throughput, they provide a highly efficient sampling technique for subword regularization.

- The next notable contribution is that of a subword segmentation algorithm to model the natural distribution of language, utilizing the ambiguity of subword segmentation as a form of noise. Utilizing an underlying language model, the researchers emulate random noise and segmentation errors cause from translation between input and output task. This provides a robust training process for modeling and is used in tandem with subword regularization sampling where they notably define their method to be "on-the-fly". This "on-the-fly" method improve model learning. Despite this, later in the paper they mention their model suffers from poor generalizability on data which is not seeing during training. This becomes a particular weakness in out-of-distribution generalization.

2.2 Innovative Aspects

A few innovations exist in this work. However, only two stand out. That of subword ambiguity as a form of training noise and biased sampling improving language modeling.

- Utilizing subword ambiguity as a form of noise is a very interesting approach to the task of segmentation. The authors proposed generating synthetic subword sequences from an underlying language model which intrinsically emulates naturally distributed noise in language. As a form of data augmentation, this technique is strong for its added stochasticity to model training. This approach is also unique, to my understanding, in the NLP setting. The nuance here is that the randomness introduced by the underlying language model is not truly random and thus has an implicit bias in the noise it generates.
- As a minor side note, but something I consider is an important contribution, the results that biased samples help improve the language modeling are surprising. I believe this to be a product of utilizing language models to produce synthetic data which models a true distribution of language which, through human error, likely institutes common errors. These errors help the trained model handle representations of subword candidates in a robust fashion since these errors are more likely to occur since they represent a "true" distribution and may help the model to learn better contextual representation and nuance.

3 Strengths and Weaknesses

This paper proposes utilizing random errors in language modeling as noise for subword candidates during the sampling process. This is a key strength of this paper, as the intuition to turn stochastic error into useful noise has, to the best of my knowledge, not been studied in the language modeling setting before 2018. Another strength of this work is training across multiple datasets and sizes to provide robust coverage of context and vocabulary sizes. More evaluations and consistent improvements across the board provide sound evidence of this methods capabilities. Despite these strengths, a significant weakness of this approach lies in its inability to handle out-of-distribution generalization.

3.1 Strengths

As previously described, the strengths include utilizing implicit noise in language modeling as robustness for subword candidates. The researchers notes that other model types, such as lattice sequence models can not handle the target side ambiguity that this approach leverages by treating it as noise. They run their techinque on many datasets with various sizes and note they use a standard moses tokenizer to preprocess data, which provides the ability to compare between similar models easily. Finally, they also demonstrate strong understanding of the problem space in their derivation of the optimal loss function. Through tests they demonstrate the optimal loss function for sampling

words would grow exponentially. For this reason the researchers approximate the optimal loss to provide a linear time system.

3.2 Weaknesses

The most notable weakness of this paper is the inability of this method to generalize beyond the training data. This is a very important weakness as it highlights its inability to work in real world environments, prevents scalability, and is especially notable given the "on-the-fly" approach the authors point out as a main contribution of this work.

3.3 Areas of Improvements

I would argue that the utilization of language and segmentation ambiguity as natural noise for the system, while beneficial, should be coupled or replaced with an explicit or traditional noise injection technique. This approach introduces biases which, while shown to improve language modeling, does not provide the benefits of genuine stochasticity. Alternative approaches like a denoising score matching objective or the removal or addition of words in the sentence would give simple and effective results to improve NMT models. Any of these approaches would improve the models generalizability. Given the biased nature of the method proposed in the paper and the state of the AI field focusing on bias reduction in model architectures in 2024, this approach appears to leverage the biases and demonstrate interesting results.

4 Discussion

Overall this paper provides limited improvements over prior state of the art, but a very interesting take on leveraging translation errors as a type of data augmentation, namely noise. Their focus on sampling subwords to efficiently find the best possible prediction is a best-of approach which always leads to improvements at the cost of model diversity and likely contributes to the failure of out-of-distribution generalization which prevents it from being realized in modern systems. In this regard, the most beneficial component is in the theory of leveraging randomness in the system and the usage of efficient graph structures when sampling from all possible combinations segmentation candidates.

5 Conclusion

In conclusion, this paper presents a simple but effective improvement over the prior state of the art to the tune of $\sim 2\%$ through biased, extensive, and efficient sampling. They use the randomness of subword ambiguity between input and target sequences as a form of noise which prior model types can not do. Finally, their approach to generating subwords candidates and selecting the probabilistically max result provides a best-of solution that achieves an improvement over the commonly used byte pair encoding [2] (BPE) baseline.

References

[1] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates, 2018.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.