Morphology Informed Selections for Subword Vocabulary Size

Anonymous ACL submission

Abstract

Currently, guidance around selection of an optimal or appropriate subword vocabulary size is incomplete and confusing at best. Using a measure of subword-morpheme overlap, our analysis shows that one can find a "sweet spot" for a morphology informed subword vocabulary size. This sweet spot exhibits some variation with respect to text complexity and the morphological characteristics of a language. However, it is relatively constant with respect to corpus size.

1 Introduction

It is now a best practice in neural machine translation (NMT) to encode input data using a subword (e.g., byte-pair encoding, or BPE) vocabulary (Sennrich et al., 2016; Tan et al., 2020). This encoding enables open-vocabulary translation and limits the size of a vocabulary corresponding to a large corpus of text. BPE subword methods, for example, iteratively replace the most frequent pair of character or character sequences in a corpus with a single new character sequence to generate a fixed size vocabulary of subwords capable of tokenizing the corpus. A practitioner can thus specify the size of the subword vocabulary.

Currently, guidance around selection of an optimal or appropriate subword vocabulary size is incomplete and confusing at best. Certain researchers propose simple heuristics based on NMT experiments in certain select languages (Gowda and May, 2020). Others recommend performing a sweep over subword vocabulary sizes (Ding et al., 2019) or other computationally intensive trial and error methods to select subword vocabulary size. Still others suggest that specific numbers of BPE merges exhibit similarities across languages, which could motivate consistent choices for subword vocabulary size in a multilingual context (Gutierrez-Vasques et al., 2021).

In this work, we add another perspective and attempt to bring some clarity to the selection of subword vocabulary sizes for NMT and other Natural Language Processing (NLP) experiments. The usage of subword vocabularies is most often motivated by a desire to enable open vocabulary methods while, at the same time, limiting vocabulary size for the purpose of corpus tokenization. Thus, capturing the true vocabulary of the corpus, or rather morphology of the language, is still the end goal of such approaches. In this paper, we show that the overlap between subwords and morphemes follows a predictable pattern as a function of subword vocabulary size (at least for corpora in a certain domain and of a certain size). That is, a subword vocabulary size can be predictably selected based on the criteria that the subword vocabulary should have a maximum overlap with a corresponding language morphology. We calculate such overlaps for 27 languages to motivate practitioners to consider and experiment with morphology informed selections of subword vocabulary size.

2 Related Work

Various attempts have been made to identify optimal subword vocabulary sizes. Gowda and May (2020) perform systematic NMT experiments on four different target languages. In these experiments they use a range of BPE vocabulary between 500 and 64K types. They finally make a recommendation for using a simple heuristic to identify the near-optimal vocabulary size, which is where a mean sentence length measure is small (low numbers of subwords per sentence) and the frequency of subwords in the corpus at the 95% class rank is 100 or higher. Although this gives some clear guidance, such an approach relies on the ability to segment text into sentences (which is not always an easy task for many languages in the world).

In other NMT experiment informed research, Denkowski and Neubig (2017) make a general recommendation of 32K BPE types for NMT systems with a secondary recommendation of 16K for system with less than 1 million parallel sentences. Ding et al. (2019), on the other hand, conduct experiments with 5 different NMT architectures on 4 language pairs and come to the conclusion that a sweep over BPE merge operations from 0-4K or even 0-32K types is useful. This shows how results from NMT-based studies can vary (or even contradict). Further, the authors are not aware of any such works that use language morphology to motivate the selection of subword vocabulary size.

In a different vein of research, Gutierrez-Vasques et al. (2021) utilize information theory as a tool to explore BPE merges. At each merge, they analyze Shannon entropy across 47 languages. This entropy across subword distributions shows a lack of variability, which suggests that a language that is complex at the word level is not as complex at the subword level. However, the "turning point" highlighted in Gutierrez-Vasques et al. (2021) is at around 200 BPE merges, which is significantly less than the number of BPE merges generally recommended in practice for NMT and other NLP experiments. This discrepancy, along with the inconsistency of NMT-based studies of subword vocabulary size, begs the question: is there a linguistically informed way to provide guidance on vocabulary size selections that is more consistent with published NMT studies?

3 Methodology

To provide linguistically informed guidance on the selection of subword vocabulary sizes, we analyze the overlap of subwords and morphemes for a range of subword vocabulary sizes and for a variety of languages. We look for a "sweet spot" where the overlap between subwords and morphemes is a maximum.

For each language, we pre-process the available data to remove blank lines and to lower case all characters. We then use the morphological analysis implemented in the Python polyglot library (Virpioja et al., 2013) to obtain morphemes for each word contained in the corpus. Although this use of polyglot limits the method to the 135 supported languages, vocabulary sizes for languages supported by polyglot are likely a good starting point for vocabulary sizes in related language experiments.

For a range of vocabulary sizes from 0 to 8000, we first train a unigram subword model using that vocabulary size. SentencePiece[1] is used for all experiments, and all experiments define the same random seed to maintain reproducibility. Next, we encode each word in the corpus to get the unique subwords corresponding to the word. These unique subwords are compared with the corresponding morphemes for that work to determine the percentage of these subwords that are also morphemes. This percentage is what we define as the overlap between subwords and morphemes. Because we are doing this at a word level, we then aggregate the overlap metrics for all words in a corpus to get the average overlap for a given vocabulary size.

4 Experiments

In order to evaluate the behavior of subword-morpheme overlap in many languages, we use data from the JHU Bible Corpus (McCarthy et al., 2020). We filtered all of the text files in the corpus down to those that were: (i) supported by polyglot's morphological analysis; and (ii) including the full text of the Bible. This resulted in 63 full Bibles in 27 languages including Latin and Cyrillic writing system scripts. We ran the analysis detailed in Section 3 on each of these full Bible files. The reported maximum overlaps and vocabulary sizes at the maximum overlaps are the averages for the set of full Bibles in each respective language.

We also wanted to analyze the influence of corpus size on the subword-morpheme overlap. To this end, we took a single full Bible from 4 languages (Polish [pol], French [fra], Vietnamese [vie], and Romanian [ron]) and split the Bible into random collections of Bible verses ranging in length from 100 verses to 30,000 verses. We then ran the analysis detailed in Section 3 on each of these collections. The result is a morphology optimal subword vocabulary size as a function of the number of samples/lines in the respective corpus.

Finally, to analyze the influence of text complexity and morphological extremes, we ran the analysis detailed in Section 3 on two various specific pairings of Bible texts. The first of these pairs was the Spanish [spa] Nueva Versión Internacional (NVI) Bible and the Spanish Nueva Versión Internacional Simplificada (NVIs). The NVIs is a "simplified" version of the NVI, and, thus, this pairing should demonstrate how subword-morpheme overlap is influenced by text complexity. The second

[1] https://github.com/google/sentencepiece

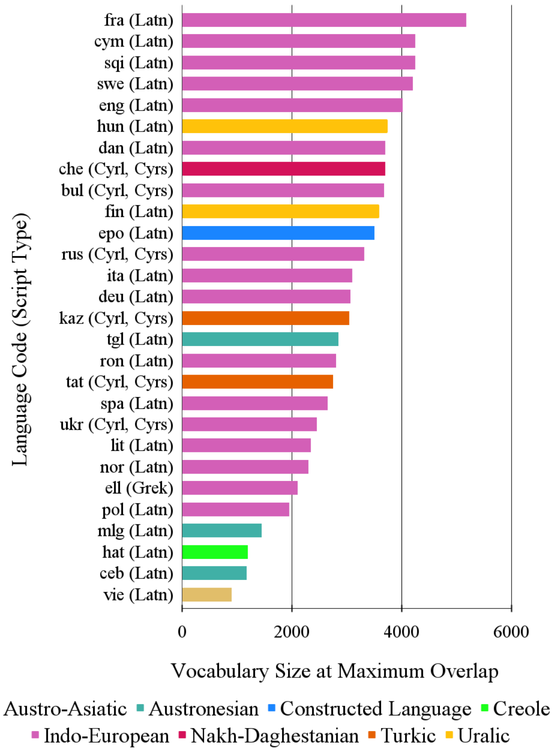

Figure 1: Subword vocabulary sizes for various languages at the maximum overlap of subwords and morphemes. Writing system scripts are shown in parentheses and language families are shown via color.

of these pairs was a Quechua [quc] Bible and the Bible in Basic English (BBE). This basic English text (BBE) should differ significantly, on a morphological level, as compared to the agglutinating language of Quechua.

5 Results and Discussion

Figure 1 shows the subword vocabulary size at the maximum overlap between subwords and morphemes for the 27 languages we considered. This vocabulary size ranges from 900 on the low end (for Vietnamese [vie]) to 5,175 at the high end (for French [fra]). The analysis finds higher vocabulary sizes for corpora with a large number of morphemes (like French, with 4300+ morphemes in the corpus) and lower vocabulary sizes for corpora with a small number of morphemes (like Vietnamese, with only 864 morphemes found in the corpus). In fact, the lowest vocabulary size, for Vietnamese, is consistent with the fact that Vietnamese is known to be an extreme in Austro-Asiatic languages in that it has very little morphology (Noyer, 1998).

To see how the subword-morpheme overlap changes as a function of subword vocabulary size,

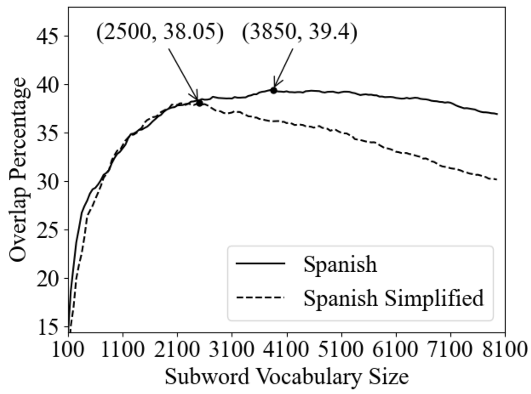

Figure 2: The percentage overlap between subwords and morphemes as a function of subword vocabulary size for the Nueva Versión Internacional Bible ("Spanish" in the figure) and the Nueva Versión Internacional Simplificada Bible ("Spanish Simplified" in the figure).

see Figure 2 and Figure 3. All of these curves indicate a general trend: the subword-morpheme overlap rises gradually to a peak and then starts to decrease. This trend makes sense in terms of the subword merges. Starting from characters at the low end of vocabulary size, the overlap with rise gradually as these characters and sets of characters are merged into morphemes. However, eventually the merges will start merging two morphemes or a non-morpheme set of characters with a morpheme. These latter merges with decrease the overall overlap.

Of course the shape of the subword-morpheme overlap curve will change as a function of both text complexity (influencing the total number of morphemes) and morphological characteristics of a language (influencing the number of subwords per word and per sentence). Figure 2 shows that the overlap curve for a simplified Spanish corpus (the NVIs) peaks earlier than the corresponding non-simplified version. Figure 3 shows how the long words of the agglutinating language of Quechua cause the overlap curve to peak earlier and decrease more rapidly than the curve for the English BBE translation. This is because Quechua has both fewer total morphemes and longer words than English.

Finally, Figure 4 shows how the vocabulary size at the maximum subword-morpheme overlap changes as a function of corpus size. This variability over corpus size is shown for a subset of the languages represented in Figure 1. One can see that the vocabulary size rises quickly to a relatively

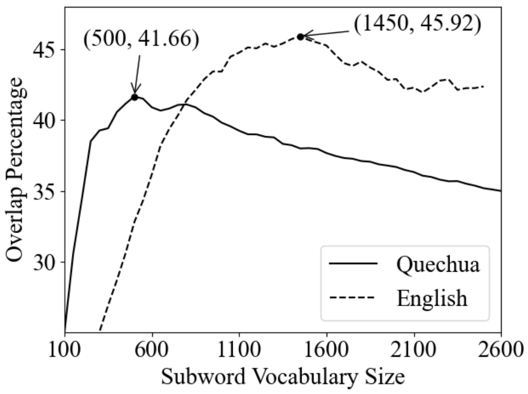

Figure 3: The percentage overlap between subwords and morphemes as a function of subword vocabulary size for the Quechua [quc] and the Bible in Basic English (BBE).

constant value as a function of corpus size. This suggests that, if a practitioner finds a morphology informed choice of subword vocabulary size (at least for unigram subwords), the choice of vocabulary size can be re-used for experiments with a variety of corpus sizes. In fact, a practitioner could look at the language represented in Figure 1, find a language similar to the language they are using in their experiments, and select a vocabulary size similar to that of the related languages. Related languages can be found using language classifications, such as those in the Ethnologue (David M. Eberhard et al., 2021). Such a process may be a good starting point for vocabulary size selections.

6 Conclusions and Future Work

Using a measure of subword-morpheme overlap, our analysis shows that one can find a "sweet spot" for a morphology informed subword vocabulary size. For Bible data, this vocabulary size shows little variation with corpus sizes greater than 15,000 samples, although it does exhibit some variation with respect to text complexity and general morphological characteristics of a language. We acknowledge that the results presented here are very limited in terms of domain (the Bible) and this kind of subword-morpheme analysis may produce different results in other domains or with different corpus sizes. In any event, we submit that such a morphology informed analysis could serve as a starting point for vocabulary size in NMT or other NLP experiments. In future work, we would like to more fully explore data from other domains and the variation in downstream NLP task performance

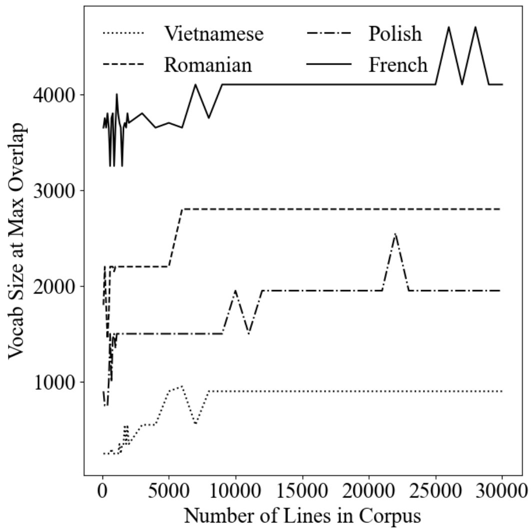

Figure 4: The "sweet spot" vocabulary size (at maximum subword-morpheme overlap) as a function of corpus size for 4 languages (Polish [pol], French [fra], Vietnamese [vie], and Romanian [ron]). The vocabulary size at the maximum overlap stays relatively constant (i.e., it plateaus) after the corpus size reaches around 10-15K samples. From vocabulary sizes from 100 to 2000 we use an interval of 100, and from 2000+ we sample with an interval size of 1000.

with morphology informed vocabulary sizes.

References

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World*, twenty-fourth edition. SIL International, Dallas, Texas.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver. Association for Computational Linguistics.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *MTSummit*.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. *arXiv preprint arXiv:2004.02334*.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Rolf Noyer. 1998. Vietnamese'morphology'and the definition of word. *University of Pennsylvania Working Papers in Linguistics*, 5(2):5.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21.

Sami Virpioja, Peter Smit, and Stig-Arne Grönroos abd Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. In *Aalto University publication series*. Department of Signal Processing and Acoustics, Aalto University.