Overcoming Vocabulary Challenges in Natural Language Processing

Elizabeth Salesky elizabeth.salesky@jhu.edu
Johns Hopkins University, Baltimore, MD, USA

Abstract

Natural language processing (NLP) applications such as machine translation (MT) have the potential to overcome language barriers and reduce the digital divide, improving information access for speakers of all languages. Despite recent advances it remains challenging to build language models which generalize across the world's 7,000 languages and 150 scripts—a challenge which begins from one of the very first steps, designing the model vocabulary. Language models are typically defined over a finite set of inputs, even if designed to have an "open vocabulary" through common tokenization techniques like subword segmentation. Current tokenization approaches are primarily designed around how computers handle text storage (using bytes and Unicode) rather than how languages and scripts are structured, which limits their effectiveness and adaptability across settings. These challenges are exacerbated in multilingual models where balancing computational costs and coverage results in a **"vocabulary bottleneck."**

This thesis focuses on techniques to build adaptive and robust open-vocabulary language models. We address three core challenges pertaining to model vocabularies within the context of machine translation: optimizing vocabulary size, enhancing model robustness, and overcoming the vocabulary bottleneck in multilingual models. We first demonstrate the importance of tuning the size of subword model vocabularies and propose a method to do so more efficiently through online incremental vocabulary expansion. We then propose a line of work in which we overcome the vocabulary bottleneck by replacing the embedding matrix with representations built on visually rendered text (pixels). Our vocabulary-free approach improves model robustness across a linguistically diverse set of languages and also improves cross-lingual transfer both within and across scripts, particularly for under-resourced languages. Finally, we present a new benchmark for a real-world task, visually-situated translation of text in natural images, where direct multimodal models can reduce error propagation and improve robustness compared to cascaded approaches.

1 Summary

Machine translation (MT) is a branch within natural language processing (NLP) focused on using computational methods to translate text or speech from one language into another. Machine translation has long served as a testing ground for new methods in NLP due to its extensive history within the field, the availability of data for benchmarking new approaches, and its direct relevance to practical applications. While the quality of MT has advanced significantly since neural models were introduced (Luong and Manning, 2015; Kocmi et al., 2023), one critical but often overlooked component that has seen little change since 2015 is **tokenization**. MT models, like other NLP models, rely on predefined model vocabularies often created through data-driven segmentation to support an "open vocabulary." The open-vocabulary challenge refers to effective handling of any input words or phrases, even if they are not present in model's training data or fixed vocabulary. Current solutions, while generally effective, are primarily designed around how computers handle text storage (using bytes and Unicode), rather than how languages and scripts are linguistically structured, which can limit their effectiveness across different languages and settings.

In this thesis, we focus on overcoming three vocabulary challenges in machine translation; namely, optimizing vocabulary size, models' robustness to noise, and the vocabulary bottleneck in multilingual models.

Constructing a model vocabulary involves a delicate balance between enabling sufficient expressivity and robustness, and managing computational costs. Optimizing vocabulary sizes is a challenging and expensive task (Ding et al., 2019), particularly in the absence of reliable proxy metrics that correlate strongly with downstream tasks—necessitating full training runs to extrinsically compare vocabulary effectiveness for a downstream performance, which is computationally intensive and so under-explored. We demonstrate the importance of optimizing model vocabularies, and present the first work on expanding neural model vocabularies to address this challenge, enabling efficient online vocabulary optimization (Salesky et al, 2018).

Data-driven techniques to create open vocabularies potentially allow for an open vocabulary, where any previously unobserved word can be represented as a combination of more frequently observed substrings or characters. However, common training corpora can lack sufficient coverage to create a generalizeable vocabulary, particularly evident in morphologically-rich languages where only a few possible linguistic variations may be observed, or languages with non-standardized orthographies where variations in spelling and encoding further contribute to data sparsity. No matter the size of the vocabulary, there will be unseen input at inference time due to common phenomena in the wild such as spelling errors or unexpected codepoints, to which machine translation models notoriously lack robustness (Belinkov and Bisk, 2018). Previous methodologies do not adequately address these challenges, resulting in disparities in performance across languages and settings. We introduce **visual representations of text** for truly open-vocabulary tokenization (Salesky et al, 2021). We demonstrate that translation directly from pixels substantially improves robustness to a variety of forms of noise and results in more compositional text representations. Moreover, they eliminate the need to optimize a discrete vocabulary.

In multilingual settings, balancing capacity and coverage across diverse languages within a fixed parameter budget creates a **vocabulary bottleneck**. As we expand the number of languages to be covered by a particular model, with a finite vocabulary we inherently must reduce the capacity allocated to each language and script; on the other hand, if we maintain capacity per language, the number of parameters allocated to the vocabulary quickly explode and the model becomes intractable to train. A wide body of work has shown that multilingual models are affected by a language's script (Muller et al., 2021; Rust et al., 2021; Pfeiffer et al., 2021), necessitating vocabulary expansion and adaptation techniques. Recent work which expands the LLaMA-2 (Touvron et al., 2023) model vocabulary from English-only to 534 languages found that increasing the vocabulary size from 32k to 260k alone resulted in a 30% parameter increase (Lin et al., 2024). Another recent work found that a 4× increase in vocabulary size caused masked language model pretraining to take 2.5× longer (Liang et al., 2023). We identify the architectural requirements to apply visual representations of text to multilingual settings and at scale. We demonstrate that our approach enables **native cross-lingual transfer between scripts** with significantly greater positive transfer, particularly for low-resource languages (Salesky et al, 2023), and comes without the computational cost of byte-level or larger subword vocabularies.

Finally, we turn to a real-world visually-situated task, translation of text in natural images. While this task is typically approached with a cascade of multiple separate models including OCR and MT, we create a taxonomy of typical OCR errors and their impact on downstream translation performance. We present a novel benchmark that allows the evaluation of direct architectures such as those developed in the previous two chapters on this task, and show that multimodal models translating directly from pixels can reduce error propagation and improve robustness for this capability (Salesky et al, 2024).

This thesis advances robust open-vocabulary translation by introducing approaches such as vocabulary expansion and vocabulary-free, pixel-based text representations that bypass traditional tokenization bottlenecks. Furthermore, it presents a novel benchmark for visually-situated translation of text in natural images and demonstrates that end-to-end multimodal models can perform more robustly on this task. Together, these contributions provide highly adaptable, scalable models that improve translation quality and cross-lingual transfer across low-resource languages, diverse scripts, and real-world tasks.