Multilingual Text Compression via Subword Tokenization

Zhengze Yu
University of Chinese Academy of Sciences

1. Dataset

We use the *training-monolingual-news-commentary* corpus as the source text. The dataset covers five languages: **cs, de, en, es, fr**, with the same news commentary samples across languages. For testing, we randomly select several paragraphs from different positions in the news (the same positions for each language), and the remaining text is used as the training set.

The dataset used in this project is available at my **GitHub Repository (training-monolingual-news-commentary)**.

2. Methodology

We apply subword-based tokenization (**SubwordTokenization**, abbreviated as Subword) to split words into smaller subword units. The main idea is to use a finite vocabulary to cover all words, while minimizing the total number of tokens.

We further use **Byte Pair Encoding (BPE)**, also known as digram coding, a data compression algorithm that merges the most frequent pairs of characters iteratively to generate a fixed-size subword vocabulary. Initially, each word is split into characters; BPE merges the most frequent character pairs until the predefined vocabulary size is reached.

We train the subword model on the training set with vocabulary sizes of **500, 1000, 1500, 2000, 2500, 3000**. The trained model is then applied to the test set to segment the text, and the compressed size is computed as:

$$\text{Compression Rate} = \frac{\text{Encoded text length after segmentation}}{\text{Original text byte size}}$$

Language	500	1000	1500	2000	2500	3000
en	0.441	0.352	0.307	0.281	0.257	0.244
cs	0.467	0.407	0.359	0.338	0.297	0.287
de	0.430	0.356	0.320	0.278	0.264	0.257
es	0.417	0.337	0.305	0.285	0.253	0.235
fr	0.436	0.358	0.317	0.290	0.274	0.268

Table 1. Compression rates for different languages and vocabulary sizes.

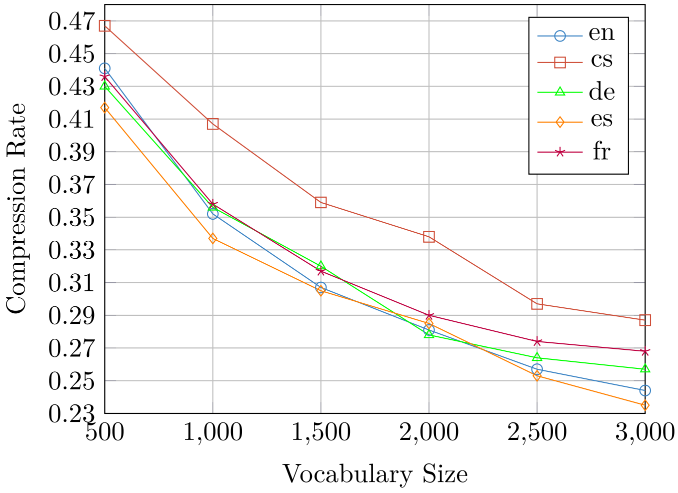

Figure 1. Compression rate vs. vocabulary size for five languages.

As the vocabulary size increases, the compression rate decreases. With a small vocabulary, words are over-segmented, leading to poor compression. As the vocabulary increases, compression improves. However, a very large vocabulary may ignore word roots or repetitive words, producing fewer tokens but less reasonable segmentation. Therefore, a medium-sized vocabulary is preferable.

The trend is similar across all five languages. At the same vocabulary size, the compression rate is lowest for **cs** and highest for **es**.