Contents

1 BPE Tokenizer

1.1 A: Counterexamples

In code.

1.2 B: Why is it generally impossible to build non-trivial tokenizers that preserve concatenation

Because preserving concatenation implies that concatenation of letter by letter tokenization is the same as tokenization of concatenation of the letters. Thus, tokenise("Then") = tokenise("T") + tokenise("h") + tokenise("e") + tokenise("n"). Thus, tokenisation has to be some representation of the alphabet.

1.3 Q 1.3

1.3.1 Longest token

The longest token contains the word References. This is possibly a corpus of research papers.

1.3.2 How can BPE compromise privacy

E.g., if corpus is medical history of a few patients, it may include patient names as part of tokenization.

1.4 Q 1.4 English vs Thai

1.5 A: Number of Tokens

Number of tokens used for English is 119 and number of tokens used for Thai is 636.

1.6 B: Effect of Small Corpus

If BPE is trained on a bigger corpus, it is likely to find useful compression of data based on language structure such as frequent words and freqeuent n-grams. However, a smaller corpus may lead to a tokenization that is not representative of the content at large. This is problematic because tokens will not correspond to language structure and this will make 1. Training more expensive (more tokens to represent the same information) 2. Will lose out on long range relationships (each batch will contain less information).