Natural Language Processing

Homework #4

Due on January 8, 2027

Ada Lovelace

Problem 1

Explain why byte-pair encoding (BPE) tokenization can represent out-of-vocabulary words that a fixed word-level vocabulary cannot.

Solution BPE 从字符级别开始，通过统计训练语料中相邻符号对的出现频率，反复合并最高频的符号对，逐步构建一个包含子词单元的词表。由于任意单词都可以被拆分为若干个已知的子词片段（在最坏情况下可以退化到单个字符），因此即使某个完整单词从未在训练语料中出现过，BPE 分词器仍然可以用已知的子词单元将其表示出来，而不会产生“未登录词”（OOV）问题。

Problem 2

Given the bigram counts $C(\text{the}, \text{cat}) = 12$ and $C(\text{the}) = 400$, compute the maximum-likelihood estimate of $P(\text{cat} \mid \text{the})$, and explain one drawback of using raw MLE counts for language modeling.

Solution The maximum-likelihood bigram probability is

$$P(\text{cat} \mid \text{the}) = \frac{C(\text{the}, \text{cat})}{C(\text{the})} = \frac{12}{400} = 0.03. \tag{1}$$

一个主要缺陷是：任何在训练语料中从未共现过的二元组都会被赋予概率 0，这会导致模型在遇到测试集中真实存在但训练时未见过的词对时，将整句话的概率计算为 0 （数据稀疏问题）。常见的缓解方法包括加一平滑（Laplace smoothing）和 Kneser-Ney 平滑等。

Problem 3

Try citation. For example, thanks Josh Davis for his original work on the latex-homework-template project [1].

Solution Citations use plain `bibtex` via `\bibliographystyle{abbrv}` and `\bibliography{references.bib}` at the end of the document.

参考文献

[1] J. Davis. latex-homework-template. https://github.com/jdavis/latex-homework-template, 2020.