1 Problem Definition

Documents (D) consist of words (W). For each word w in W, k options are given by outputs of k OCRs.

Each word has n features extracted from all its options. Denote n as the number of features. $x = \{x_1, x_2, \ldots, x_n\}$ is a set of features, which is the input of the classifier.

The classifier outputs a label $y \in [1, k+1]$ for each word. Class $y = i$ ($i \in [1, k]$) means that the ith option is the correct output for this word, and $y = k + 1$ means that none of the OCR outputs are correct.