
Handwriting Recognition – Solution

by Tom Payne

A very bad handwriting recognition program recently interpreted a high school student's note providing a 10 word excuse for being late to class. The output is below.

lie	charm	code	soil	rout	wake	he	us	this	moving
my	solemn	circle	did	hot	make	I'll	is	taxis	having
guy	clam	shute	raid	riot	sale	me	ugh	thai	running
bye	beam	clock	risk	not	wane	be	up	tear	morning
	alarm	visit	must			see			loving
	dream								

Task 1: Determine what the student's excuse was.

My alarm clock did not wake me up this morning.

Task 2: Explain how you could make the handwriting recognition program produce an output that is closer to what the student actually meant to write.

(Many possible answers here - this is one possibility.) You could compare the combinations of words to a huge corpus of English, say blogs on the internet, and see which words occur together and which don't. For example, you would get lots of hits on "my alarm clock" but few, if any, on "lie charm code," "my solemn circle," "my charm chute," "bye beam clock," etc. The program would first take each word in the first column and combine it with each word in the second column and search the corpus to determine which combinations are most likely. Then it would take the most likely two-word combinations (probably all those starting with "my"), and combine them with each word in the third column to form three-word combinations, and search the corpus for those. Then the most likely three-word combinations would be combined with each possible fourth word to form four word combinations, etc. until the one most likely 10 word combination is determined.