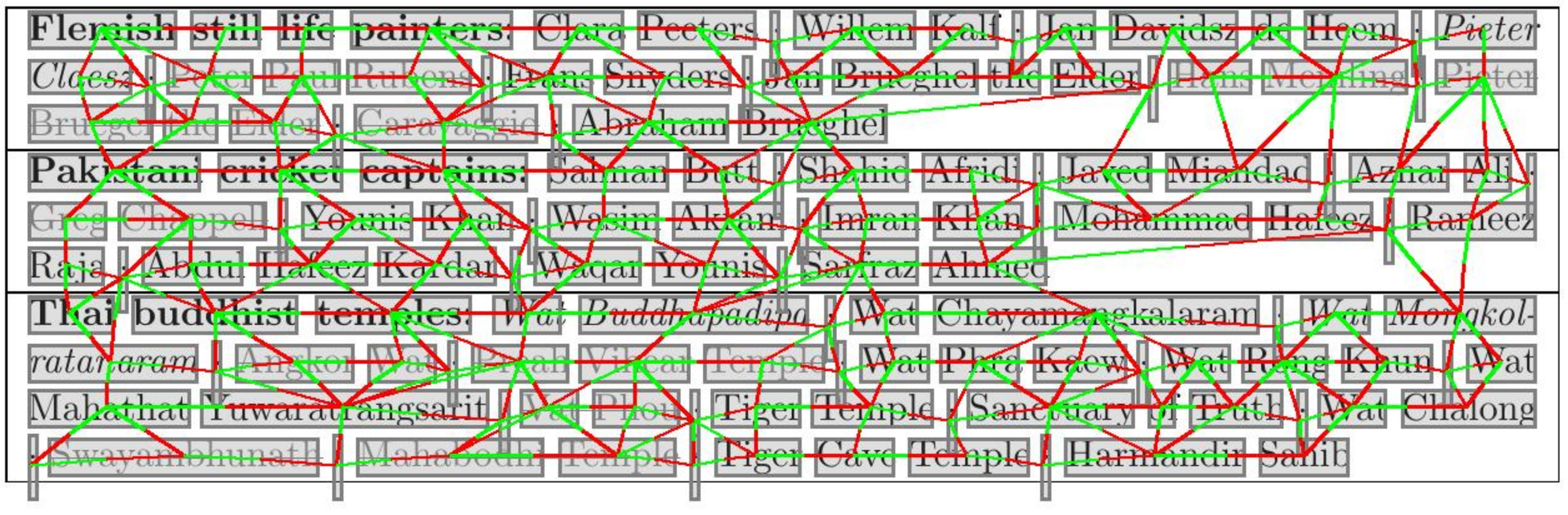

Table 4: Instances extracted for several fine-grained classes from Wikipedia. Lists shown are from Mods_I. Instances in italics were also returned by Hearst∩. Strikethrough denotes incorrect.

	UniformSet		WeightedSet	
	Coverage	MAP	Coverage	MAP
Baseline	95 / 70	0.01	98 / 74	0.01
Hearst	9 / 9	0.63	8 / 8	0.80
Hearst∩	13 / 12	0.62	9 / 9	0.80
Mods_H raw	56 / 32	0.23	50 / 30	0.16
Mods_H RR	56 / 32	0.29	50 / 30	0.25
Mods_I raw	62 / 36	0.18	59 / 38	0.20
Mods_I RR	62 / 36	0.24	59 / 38	0.23

Table 5: Coverage and precision for populating Wikipedia category pages with instances. "Coverage" is the number of class labels (out of 100) for which at least one instance was returned, followed by the number for which at least one correct instance was returned. "MAP" is mean average precision. MAP does not punish methods for returning empty lists, thus favoring the baseline (see Figure 2).

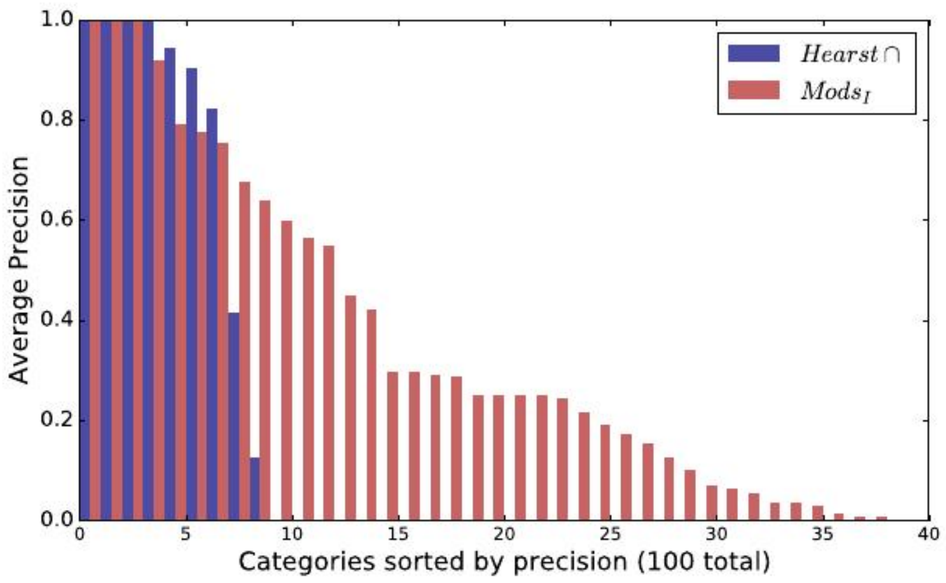

Figure 2: Distribution of AP over 100 class labels in WeightedSet. The proposed method (red) and the baseline method (blue) achieve high AP for the same number of classes, but Mods_I additionally finds instances for classes for which the baseline returns nothing.

estimates. For each of these labels, we manually check the top 10 instances proposed by each method to determine whether each belongs to the class. Table 6 shows the precision scores for each method computed against the original Wikipedia list of instances and against our manually-augmented list of gold instances. The overall ordering of the systems does not change, but the precision scores increase notably after re-annotation. We continue to evaluate against the Wikipedia lists, but acknowledge that reported precision is likely an underestimate of true precision.

	Wikipedia	Gold
Hearst	0.56	0.79
Hearst∩	0.53	0.78
Mods_H	0.23	0.39
Mods_I	0.24	0.42
Hearst+Mods_H	0.43	0.63
Hearst+Mods_I	0.43	0.63

Table 6: P@10 before vs. after re-annotation; Wikipedia underestimates true precision.

	UniformSet		WeightedSet	
	AUC	Recall	AUC	Recall
Baseline	0.55	0.23	0.53	0.28
Hearst	0.56	0.03	0.52	0.02
Hearst∩	0.57	0.04	0.53	0.02
Mods_H	0.68	0.08	0.60	0.06
Mods_I	0.71	0.09	0.65	0.09
Hearst∩+Mods_H	0.70	0.09	0.61	0.08
Hearst∩+Mods_I	0.73	0.10	0.66	0.10

Table 7: Recall of instances on Wikipedia category pages, measured against the full set of instances from all pages in sample. AUC captures tradeoff between true and false positives.