Category Performance

Code	Name	Size	STY		CUI	
			Pred.	Gold	Pred.	Gold
T204	Eukaryote	780,146	**75.58**	**86.01**	47.79	53.50
T103	Chemical	644,616	**72.96**	**85.77**	37.76	44.49
T007[†]	Bacterium	350,134	66.12	78.35	46.12	52.00
T017	Anatomical Structure	183,300	53.68	72.51	34.65	46.77
T038	Biologic Function	182,802	63.73	82.59	44.85	56.36
T033	Finding	123,495	37.67	65.80	35.11	56.08
T058	Health Care Activity	121,800	55.14	74.80	35.31	46.60
T037[†]	Injury or Poisoning	92,384	60.95	83.43	38.17	52.96
T082	Spatial Concept	40,259	49.34	80.00	39.08	62.15
T074[†]	Medical Device	20,364	42.32	65.20	21.94	33.54
T170	Intellectual Product	20,075	41.36	72.54	30.60	49.16
T005[†]	Virus	17,794	**77.99**	**88.05**	37.74	41.51
T201[†]	Clinical Attribute	9,675	57.79	73.05	**59.42**	**75.65**
T097[†]	Professional or Occupational Group	5,422	**73.41**	**90.75**	**54.05**	**61.85**
T168[†]	Food	3,829	48.65	65.88	40.88	54.39
T092[†]	Organization	1,896	57.82	78.51	39.26	54.64
T098	Population Group	1,758	**69.14**	**88.45**	**62.82**	**79.09**
T031[†]	Body Substance	1,700	67.33	82.18	**58.91**	**71.29**
T062	Research Activity	1,294	62.73	79.26	**54.69**	**70.32**
T091[†]	Biomedical Occupation or Discipline	816	43.62	65.43	32.98	48.94
T022[†]	Body System	478	36.36	44.32	29.55	37.50

Table 1. Accuracy results for each semantic type covered in the test split of MedMentions st21pv. STY and CUI results correspond to performance on linking mentions to the semantic type and concepts. Showing results using predicted and gold spans. Types marked with [†] have fewer than 1,000 test instances. Bold results correspond to the top 5 performing types.

Analysing accuracy results for specific categories, shown in Table 1, provides some interesting insights. Notably, it becomes evident that there's a large variation in performance depending on the category. As expected, both semantic types (STY) and concepts (CUI) corresponding to larger categories (i.e. more candidates) tend to be harder to predict than smaller categories, and categories that perform well on STY linking also tend to perform well on CUI linking. However, there are some clear exceptions to this tendency (e.g. T204, T103) where it's seems that it's possible to accurately predict STYs, even for large categories, but remain unreliable at predicting CUIs. Additionally, we can also see some dramatic improvements when using gold spans for some categories (e.g. T082), suggesting that our mention recognition is holding back performance on those categories.