

Automatic Generation of Benchmarks for Entity Recognition and Linking – Appendix

Axel-Cyrille Ngonga Ngomo^{1,2}, Michael Röder¹, Diego Moussallem², Ricardo Usbeck¹, and René Speck²

¹Data Science Department, University of Paderborn, Germany
axel.ngonga|michael.roeder|ricardo.usbeck@upb.de

²AKSW Research Group, University of Leipzig, Germany
moussallem|speck@informatik.uni-leipzig.de

This document is the appendix of the paper "*Automatic Generation of Benchmarks for Entity Recognition and Linking*" and contains additional data that has been left out of the paper itself due to length restrictions.

In section 4.1 of the paper, different features of the manually created datasets and the BENGAL datasets are compared. Table 1 contains the complete list of datasets and their features while Table 2 shows the correlations of POS tag distributions of the single datasets.

In Section 4.2 of the paper, the performance of annotation systems on several datasets is discussed. Table 3 and 5 show the Micro F1-scores of the single annotation systems for a Named Entity Recognition (NER) experiment and an Annotation 2 Knowledge Base (A2KB) experiment, respectively. Also, Table 4 and Table 6 depict the performance of multilingual systems for NER and Disambiguation 2 Knowledge Base (D2KB) experiments on Portuguese datasets. Table 7 and 8 contain the person correlations of the datasets based on the results of the annotation systems.

Table 1: Features of the datasets used in our evaluation.

ID	Name	Doc. $ C $	Tokens $ T $	Entities $ E $	$ T / C $	$ E / C $	$ E / T $
D1	ACE2004	57	21312	306	373.9	5.4	0.01
D2	AIDA/CoNLL-Complete	1393	245008	34929	175.9	25.1	0.14
D3	AIDA/CoNLL-Test A	216	41757	5917	193.3	27.4	0.14
D4	AIDA/CoNLL-Test B	231	37687	5616	163.1	24.3	0.15
D5	AIDA/CoNLL-Training	946	165564	23396	175.0	24.7	0.14
D6	AQUAINT	50	11024	727	220.5	14.5	0.07
D7	DBpediaSpotlight	58	1661	330	28.6	5.7	0.20
D8	IITB	104	66531	18308	639.7	176.0	0.28
D9	KORE50	50	640	144	12.8	2.9	0.23
D10	Microposts2014-Test	1055	20648	1256	19.6	1.2	0.06
D11	Microposts2014-Train	2340	40684	3822	17.4	1.6	0.09
D12	MSNBC	20	10877	747	543.9	37.4	0.07
D13	N3-Reuters-128	128	15842	880	123.8	6.9	0.06
D14	N3-RSS-500	500	15504	1000	31.0	2.0	0.06
D15	OKE 2015 Task 1 evaluation	101	3064	664	30.3	6.6	0.22
D16	OKE 2015 Task 1 train	95	1946	341	20.5	3.6	0.18
B1	BENGAL Path 100	100	1202	362	12.02	3.6	0.30
B2	BENGAL Path Para 100	100	1250	362	12.5	3.6	0.29
B3	BENGAL Star 100	100	3039	880	30.39	8.8	0.29
B4	BENGAL Star Para 100	100	2772	543	27.72	5.43	0.19
B5	BENGAL Sym 100	100	2718	725	27.18	7.25	0.26
B6	BENGAL Sym Para 100	100	2621	612	26.21	6.12	0.23
B7	BENGAL Hybrid 100	100	1811	528	18.11	5.28	0.29
B8	BENGAL Hybrid Para 100	100	1732	420	17.32	4.2	0.24
B9	BENGAL Summary 100	100	2033	637	20.33	6.37	0.31
B10	BENGAL Summary Para 100	100	2025	646	20.25	6.46	0.31
B11	BENGAL Hybrid 10000	10000	556483	165254	55.6	16.5	0.30
B12	BENGAL Hybrid Long 10	10	9162	2417	241.7	916.2	0.26
B13	BENGAL Star Long 10	10	7369	316	31.6	736.9	0.04
P1	BENGAL PT-BR Path 100	100	1624	352	16.24	3.52	0.22
P2	BENGAL PT-BR Star 100	100	3765	793	37.65	7.93	0.21
P3	BENGAL PT-BR Sym 100	100	3091	685	30.91	6.85	0.22
P4	BENGAL PT-BR Hybrid 100	100	2606	543	26.06	5.43	0.21

Table 4: Micro F1-scores of the multilingual annotators for the NER experiments on chosen Portuguese datasets.

	Experiment Dataset ID	Babely Spotlight	
NER	HAREM-793-04789	0.23	0.17
	HAREM-561-02948	0.27	0.37
	HAREM-554-05073	0.67	0.67
	HAREM-192-03897	0.32	0.45
	HAREM-184-06038	0.42	0.07
	HAREM-88H-09142	0.44	0.31
	HAREM-28H-00303	0.50	0.30
	HAREM-27H-04908	0.67	0.67
	HAREM-27H-00966	0.25	0.19
	HAREM-09H-01758	0.42	0.48
	P1 BENGAL PT-BR Path 100	0.59	0.71
	P2 BENGAL PT-BR Star 100	0.53	0.70
	P3 BENGAL PT-BR Sym 100	0.47	0.69
	P4 BENGAL PT-BR Hybrid 100	0.51	0.69

Table 6: Micro F1-scores of the multilingual annotators for the D2KB experiments on chosen Portuguese datasets.

Experiment	Dataset ID	MAG	Babelfy	Spotlight	PBOH
D2KB	P1 BENGAL PT-BR Path 100	0.87	0.56	0.19	0.65
	P2 BENGAL PT-BR Star 100	0.80	0.85	0.56	0.83
	P3 BENGAL PT-BR Sym 100	0.76	0.68	0.44	0.68
	P4 BENGAL PT-BR Hybrid 100	0.84	0.77	0.54	0.82

Table 8: Pearson correlation of annotation performance result distributions of the single dataset for the A2KB experiment.

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	
D1																													
D2	0.60																												
D3	0.60	1.00																											
D4	0.60	1.00	1.00																										
D5	0.60	1.00	1.00	0.99																									
D6	0.82	0.53	0.54	0.56	0.52																								
D7	0.42	0.25	0.26	0.30	0.24	0.70																							
D8	0.35	0.51	0.52	0.55	0.50	0.70	0.88																						
D9	0.30	0.64	0.65	0.66	0.63	0.29	0.06	0.20																					
D10	0.59	0.76	0.76	0.77	0.75	0.78	0.63	0.85	0.30																				
D11	0.64	0.88	0.88	0.89	0.87	0.73	0.53	0.76	0.42	0.97																			
D12	0.66	0.81	0.82	0.82	0.80	0.76	0.36	0.53	0.73	0.77	0.81																		
D13	0.46	0.85	0.85	0.83	0.86	0.18	0.05	0.24	0.41	0.52	0.67	0.47																	
D14	0.66	0.84	0.84	0.83	0.85	0.39	0.26	0.39	0.39	0.59	0.74	0.48	0.92																
D15	0.56	0.68	0.67	0.69	0.67	0.30	0.20	0.27	0.54	0.40	0.51	0.44	0.72	0.76															
D16	0.53	0.57	0.57	0.60	0.55	0.46	0.31	0.39	0.51	0.41	0.46	0.44	0.49	0.57	0.90														
B1	0.35	0.53	0.53	0.57	0.52	0.57	0.57	0.72	0.24	0.60	0.55	0.45	0.29	0.35	0.53	0.71													
B2	0.37	0.63	0.63	0.66	0.62	0.54	0.56	0.74	0.32	0.63	0.61	0.48	0.42	0.49	0.62	0.75	0.98												
B3	0.25	0.65	0.65	0.67	0.64	0.38	0.25	0.54	0.43	0.54	0.54	0.47	0.44	0.45	0.63	0.75	0.89	0.92											
B4	0.47	0.73	0.73	0.76	0.72	0.59	0.41	0.66	0.39	0.69	0.70	0.57	0.48	0.56	0.64	0.76	0.92	0.94	0.95										
B5	0.06	0.45	0.45	0.48	0.43	0.24	0.19	0.46	0.38	0.33	0.33	0.27	0.23	0.28	0.49	0.69	0.82	0.84	0.94	0.86									
B6	0.04	0.44	0.44	0.48	0.43	0.23	0.20	0.46	0.37	0.33	0.33	0.26	0.22	0.27	0.49	0.69	0.82	0.84	0.94	0.86	1.00								
B7	0.41	0.67	0.67	0.70	0.66	0.66	0.52	0.75	0.40	0.76	0.72	0.67	0.36	0.38	0.53	0.69	0.94	0.93	0.89	0.93	0.76	0.77							
B8	0.50	0.67	0.67	0.70	0.66	0.73	0.56	0.75	0.36	0.76	0.72	0.66	0.35	0.43	0.55	0.72	0.95	0.94	0.88	0.95	0.77	0.77	0.99						
B9	0.39	0.42	0.42	0.46	0.40	0.48	0.54	0.59	0.31	0.46	0.45	0.34	0.25	0.37	0.74	0.89	0.86	0.86	0.77	0.80	0.73	0.74	0.80	0.82					
B10	0.40	0.45	0.46	0.50	0.44	0.51	0.53	0.60	0.32	0.48	0.47	0.36	0.28	0.39	0.73	0.89	0.88	0.88	0.80	0.83	0.76	0.77	0.82	0.85	1.00				
B11	0.24	0.57	0.58	0.61	0.56	0.41	0.20	0.48	0.50	0.46	0.46	0.46	0.32	0.37	0.56	0.77	0.82	0.84	0.94	0.90	0.96	0.96	0.82	0.84	0.73	0.77			
B12	0.32	0.74	0.73	0.72	0.74	0.31	0.24	0.48	0.10	0.63	0.67	0.44	0.70	0.60	0.48	0.38	0.69	0.72	0.69	0.73	0.46	0.47	0.71	0.69	0.45	0.48	0.47		
B13	0.81	0.59	0.61	0.63	0.58	0.85	0.56	0.57	0.48	0.63	0.64	0.67	0.38	0.58	0.59	0.74	0.56	0.58	0.49	0.64	0.42	0.40	0.60	0.68	0.58	0.62	0.59	0.21	