

Results

Joseph Boyd

June 22, 2015

Contents

1	Introduction	2
1.1	Motivation	4
1.2	Aims	4
1.3	Research Questions	4
1.4	Main Results	4
1.5	Thesis outline	4
2	Previous/related Work	4
2.1	Background Theory	4
2.2	Grobid	4
3	Implementation and Design	4
4	Results and Analysis	4
5	Conclusion	4
5.1	Summary	4
5.2	Research Answers	4
5.3	Future Work	4
6	References	4
7	Appendices	4
8	Baseline	4
8.1	Header model - Cora dataset	4
8.2	Header model - Cora dataset appending HEP dataset	6
8.3	Header model - Cora and HEP combined datasets	8
8.4	Header model - HEP dataset	10
8.5	Header model - HEP dataset appending CORA dataset	12
8.6	Header model - HEP dataset appending 1/3 CORA dataset	14
8.7	Header model - HEP dataset appending 2/3 CORA dataset	16
8.8	Segmentation model - Cora dataset	18
8.9	Segmentation model - Cora dataset appending HEP dataset	20
8.10	Segmentation model - Cora and HEP combined datasets	22
8.11	Segmentation model - HEP dataset	24
8.12	Segmentation model - HEP dataset appending CORA dataset	26

9 Regularisation	28
9.1 Header model - $L2 = 0$	28
9.2 Header model - $L2 = 1e^{-6}$	30
9.3 Header model - $L2 = 1e^{-5}$	32
9.4 Header model - $L2 = 1e^{-4}$	34
9.5 Header model - $L2 = 1e^{-3}$	36
10 Dictionaries	38
10.1 Header model - HEP dataset	38
10.2 Header model - HEP dataset appending CORA dataset	40
10.3 Segmentation model - HEP dataset	42
10.4 Segmentation model - HEP dataset appending CORA dataset	44
11 Dictionaries + stop words	46
11.1 Header model - HEP dataset	46
11.2 Header model - HEP dataset appending CORA dataset	48
11.3 Segmentation model - HEP dataset	50
11.4 Segmentation model - HEP dataset appending CORA dataset	52

1 Introduction

$$\max\{a, b\}$$

1.1 Motivation

1.2 Aims

1.3 Research Questions

1.4 Main Results

1.5 Thesis outline

2 Previous/related Work

2.1 Background Theory

2.2 Grobid

3 Implementation and Design

4 Results and Analysis

5 Conclusion

5.1 Summary

5.2 Research Answers

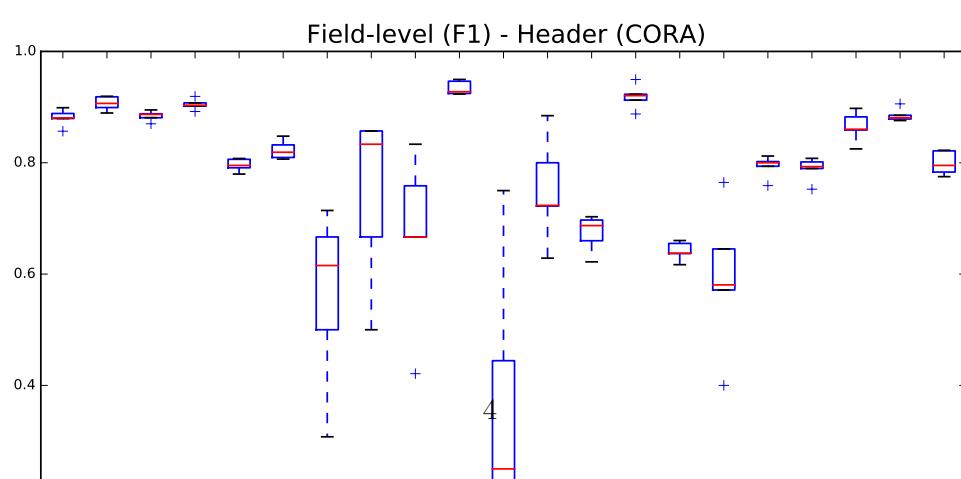
5.3 Future Work

6 References

7 Appendices

8 Baseline

8.1 Header model - Cora dataset

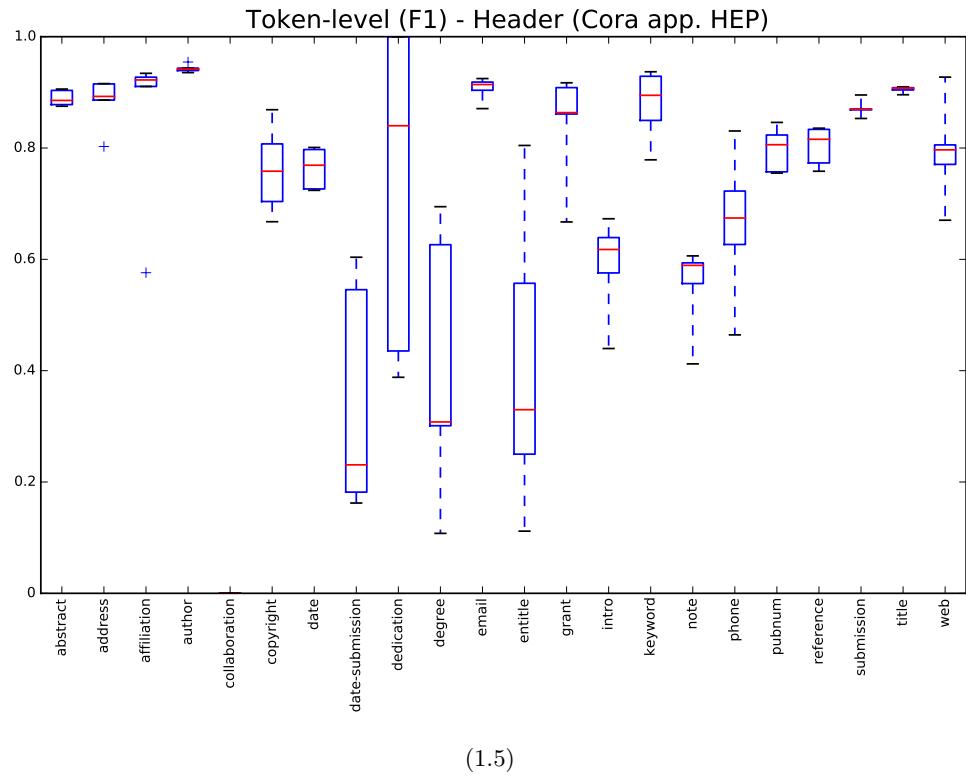
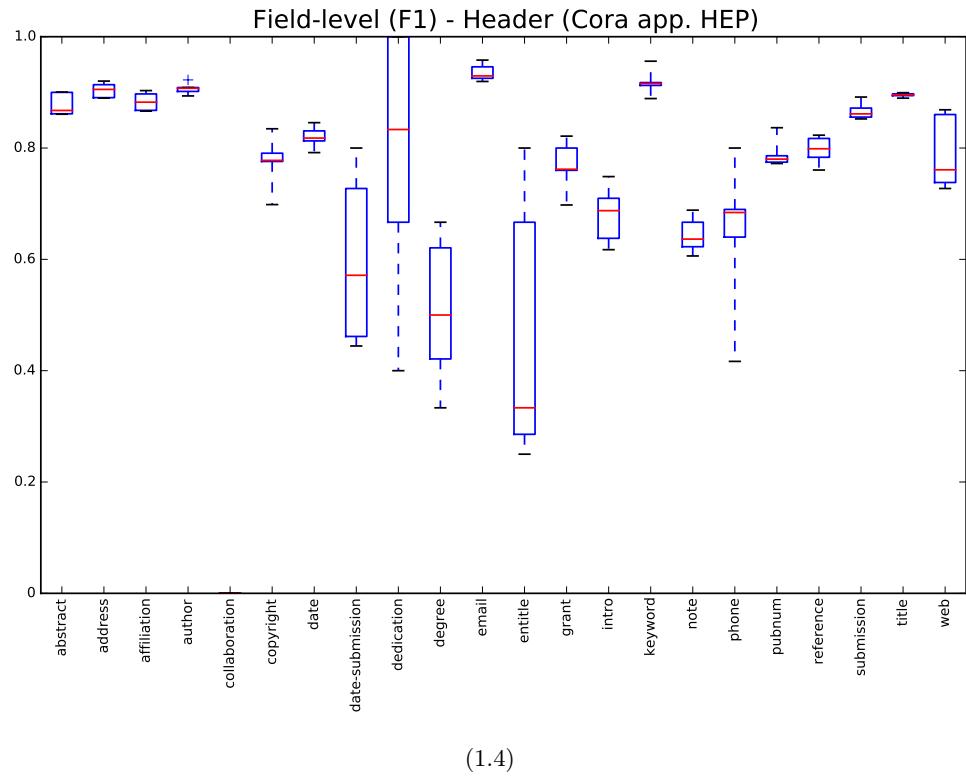


Confusion matrix - Header (CORA)

abstract	4377	76	4	78	36			181	4	17	18869	291	1092	9		184	8	388	33		
address	24	13769	580	94	10	11		5	6	76		30	175	13	341		3	173	10	33	
affiliation	50	611	22397	213	5	3		11	63	21		54	56	7	728		1	54	138	5	
author	69	144	183	2099	1	1		32	12	62		82	35	521		11	72	5	100	21	
copyright	77		10	11	3491	13			27	7		182	296		642		22	97	10	22	
date	21	14		2	13	209		11		2	13		56		224	10	52	61	175	21	12
date-submission					25	42						18		5		8	97				
dedication	15			2				202			4			42							
degree	39	6	59	136	46	10			698	2			296			22			69		
email	46	59	35	106	1				75			27	6	239	209	23	40			30	
entitle	208						10			94						1		132			
grant	71	16	27	2	26				6		3660	116	13	386		25			1	1	
intro	33711	1469	21	45	245	25		3	88		53285	104	7223		46	467		409	243		
keyword	265									206	7330	98			32	118	3	25			
note	2337	135	568	817	460	163	55	15	91	129	89	330	3558	556	14324		290	950	119	1339	220
phone		23							200			14		24	487	28	5				
pubnum	72	8	14	7	18	33				3	1	80	27	369	20	3086	52	5	82	7	
reference	131	103	73	146	23	77		14		16		39	88	42	769		86	8250	63	186	18
submission	14	8	20	5	9	182	15		11	11					184		5	32	3466	19	4
title	482	32	89	68	19	10	1			29	8	1	360		1167		43	139	18	33842	14
web	18	5			12					90			80		315	17		35		3	2665
	abstract	address	affiliation	author	copyright	date	date-submission	dedication	degree	email	entitle	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web

(1.3)

8.2 Header model - Cora dataset appending HEP dataset

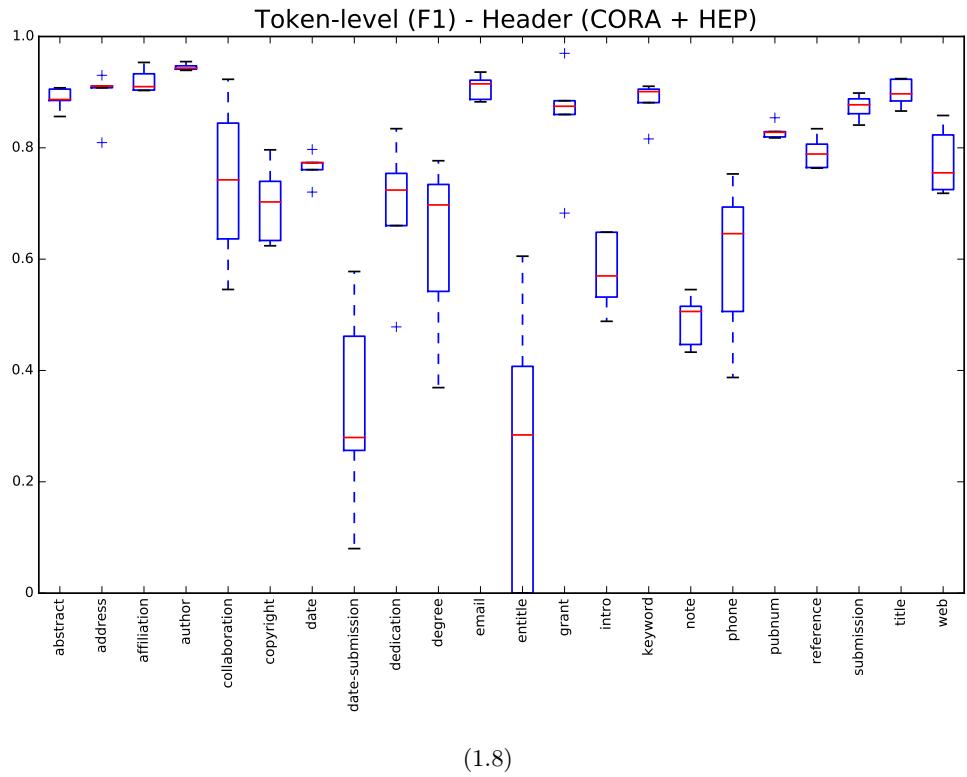
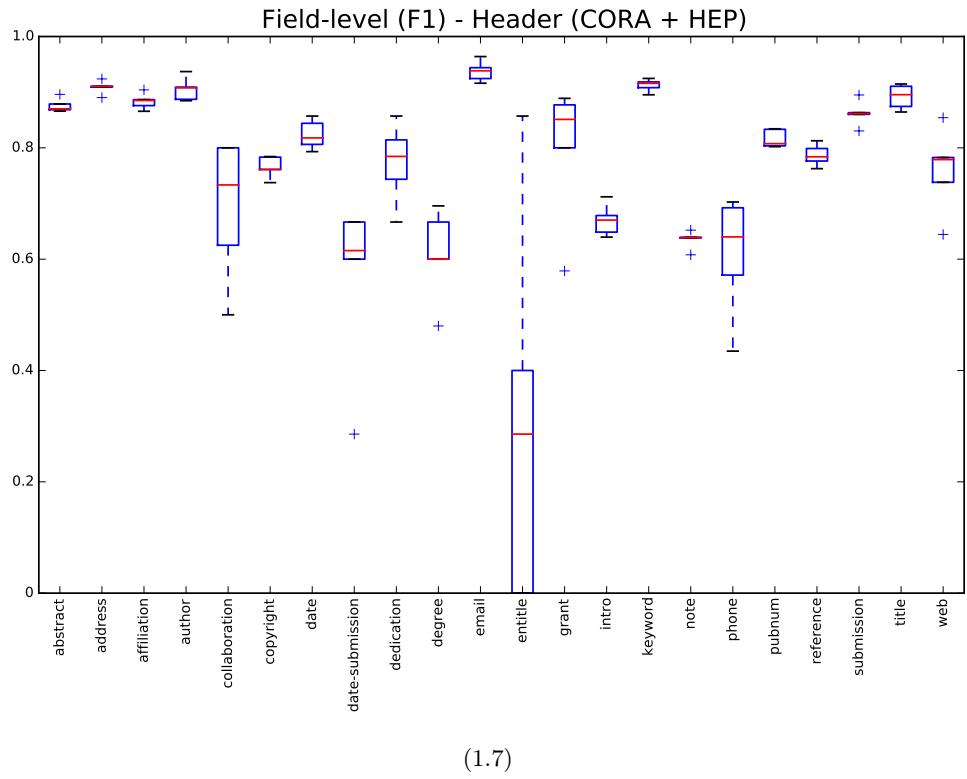


Confusion matrix - Header (Cora app. HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	dedication	degree	email	entitle	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web
abstract	4457	51	49	91		36	2		108		4		2	18104	520	915		3	151	2	354	84
address	34	13711	629	104		13	11			12	70	7	26	188	6	349	2	143		21	26	
affiliation	84	488	2532	241		3	1			44	14	19	55	63	17	667		7	55	1	121	5
author	89	121	228	2406				2		13	54		4	47	38	432	6	84	5	97	1	
collaboration																						
copyright	305	3	73	5		3406	7				8		28	228	667		34	115		3	25	
date	37	22	1	7		12	2054			1	3		59	202	10	58	106	174	19	3		
date-submission		1				23	52							4	18			97				
dedication									191					14	28							
degree	8	9	163	138		46	7			670	8				232		22		80			
email	67	69	32	46		10	5			10	7464			47	51	323	187	17	72		15	
entitle	99			1					10		125		33	17	42		1		117			
grant	147	13	29			20				14		3590	152	13	324		25			12	2	
intro	35306	1066	5806	27		55	33		700			49878	36	3557		42	352		301	225		
keyword	325											230	7326	77		34	82	3				
note	3050	185	384	859	4	351	119	55	12	678	143	85	369	2678	481	14028	3	259	977	159	1374	
phone	14	18				23	57			160			4		20	525	28	12				
pubnum	40	23	4	12		23		14			11		131	13	341	21	3020	109	5	72	2	
reference	176	79	88	86		89	58		14		7		98	31	662		68	6445	16	203	6	
submission	27	8	11	5		1	202	22		11	8		21	154			44	3449	20	2		
title	422	3	93	78		2	8	4		3	15		1	370	4	1025		57	146	11	2407	
web	128					29		35			63		62	238	32		22	2	5	2624		

(1.6)

8.3 Header model - Cora and HEP combined datasets



Confusion matrix - Header (CORA + HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	dedication	degree	email	entitle	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	0.984	0.122	0.016	0.21		0.094	0.028	0.002	0.068		0.042		0.05	0.862	0.256	0.44		0.036	0.356	0.2	0.43	0.028	
address	0.548	0.952	0.402	0.622		0.336	0.512	0.05		0.4	0.296		0.028	0.808	0.152	0.696		0.1	0.634		0.4	0.1	
affiliation	0.536	0.216	0.964	0.39		0.184	0.346			0.688	0.248		0.12	0.672	0.358	0.678		0.166	0.71	0.016	0.854	0.038	
author	0.45	0.46	0.302	0.974		0.2	0.066			0.164	0.262		0.688	0.2	0.67	0.06	0.136	0.79	0.644	0.536	0.33		
collaboration		0.066	0.366		0.776									0.2						0			
copyright	0.774		0.439	0.142		0.946	0.34			0.2	0.206		0.764	0.2	0.692		0.28	0.804	0.2	0.306	0.21		
date	0.6	0.822	0.114			0.172	0.968	0.2			0.326		1.0		0.808		0.77	0.924	0.782	0.526	0.184		
date-submission						0.258	1.0						0.106		0.072				0.98			0.062	
dedication	0.4							0.8					0.4		0.2								
degree		0.03	0.188	0.208		0.03	0.004			0.838	0.062			0.692			0.098		0.506	0.018			
email	0.544	0.298	0.132	0.634		0.286				0.968		0.034	0.502	0.1	0.592	0.414	0.044	0.336		0.2	0.506		
entitle	0.8								0.2			0.6		0.4					0.2		0.8		
grant	0.206	0.06	0.264							0.24		0.976	0.35	0.614	0.032	0.328							
intro	0.876	0.04	0.01	0.108		0.054	0.018			0.028		0.016	0.96	0.088	0.324		0.006	0.126		0.262	0.054		
keyword	0.85	0.132	0.012			0.01				0.034		0.718	0.982	0.796		0.732	0.2	0.036					
note	0.678	0.272	0.43	0.37		0.08	0.582	0.35	0.2	0.2	0.316	0.414	0.654	0.718	0.508	0.622		0.5	0.684	0.74	0.56	0.378	
phone	0.2	0.406									0.908				0.474	0.972	0.216	0.2			0.4		
pubnum	0.49	0.56	0.218	0.592		0.556	0.344						0.88	0.326	0.822	0.348	0.972	0.84		0.92	0.54		
reference	0.788	0.458	0.428	0.352		0.548	0.22		0.104		0.184		0.518	0.352	0.682		0.466	0.96	0.514	0.7	0.496		
submission	0.444	0.014	0.15	0.266		0.1	0.62	0.492		0.4	0.41		0.112	0.2	0.746		0.126	0.748	0.982	0.46	0.23		
title	0.844	0.094	0.62	0.34		0.356	0.27	0.086		0.22	0.47		0.632	0.2	0.784		0.608	0.864	0.6	0.982	0.42		
web	0.8	0.2				0.4		0.066			0.536		0.744		0.794	0.252		0.352	0.4	0.004	0.966		

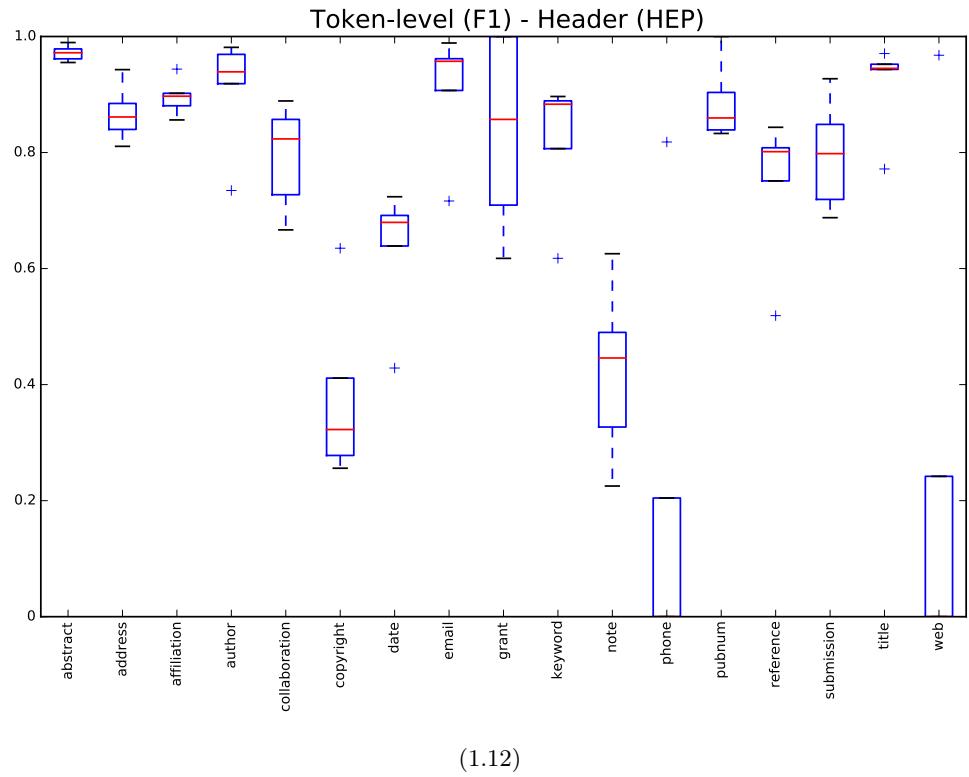
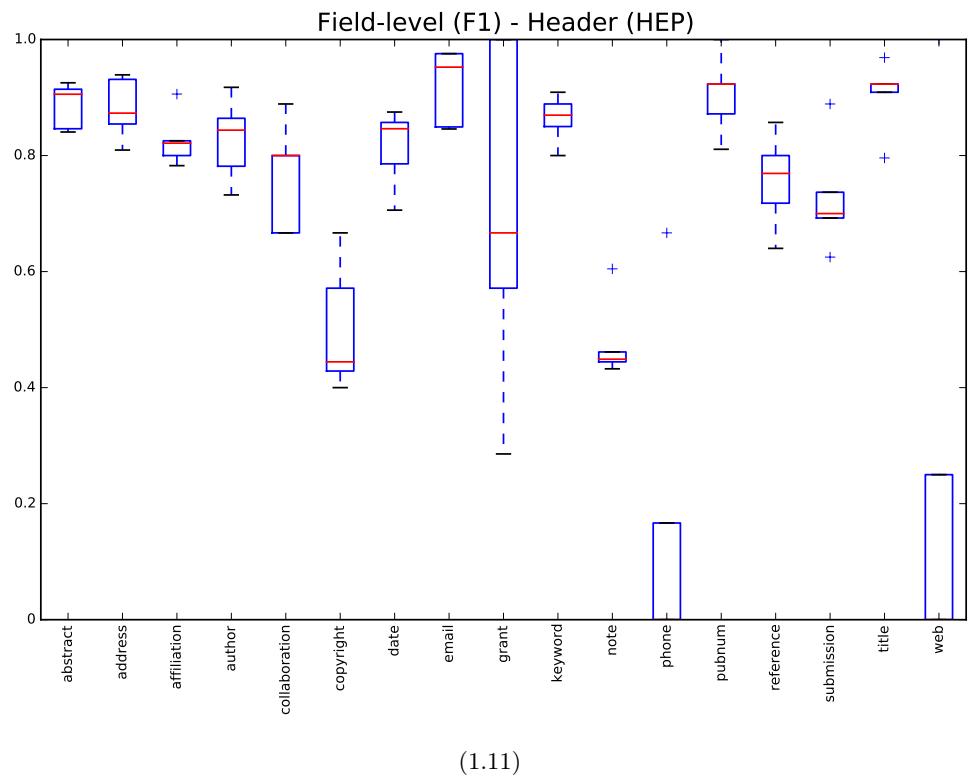
(1.9)

Confusion matrix - Header (CORA + HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	dedication	degree	email	entitle	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	6039	59	12	116		140	4	8	108		101		109	24985	311	1448	19	281	2	468	12		
address	61	5570	703	135		31	38	2		6	86		10	134	26	372		9	222		21	13	
affiliation	43	717	22994	317	2	97	9			52	35		23	93	28	687		9	61	1	150	5	
author	107	132	288	27397		4	1			9	44		86	1	545	6	40	54	10	127	21		
collaboration		1	7		26															2			
copyright	267		69	12		3596	12			27	44		448	4	717		39	134	4	41	21		
date	18	28	9			14	2263	11			8		78		266		41	99	160	10	7		
date-submission						30	53		201				18		4			86			4		
dedication	19												30		15								
degree		9	59	67		46	7			667	5			450			22		45	6			
email	43	70	23	48		37					9460		1	122	3	385	195	7	56		3	27	
entitle	175						10				94		50					1		115			
grant	47	26	36							16		4155	99	401	26	15							
intro	33347	645	29	26		159	37			71		49	53296	21	8440		34	405		525	300		
keyword	354	34	3			1				17		302	3466	214			252	3	9				
note	2257	205	701	783	6	635	155	55	12	79	142		402	4612	530	14472		256	820	219	1416	221	
phone	14	14													34	503	16	5			25		
pubnum	26	13	13	11		28	59						152	28	327	38	3839	111		65	33		
reference	189	120	84	103		95	77	14			6		167	94	913		117	9505	32	147	20		
submission	17	2	16	12		13	208	32			11	9		14	1	238		5	48	4152	32	13	
title	466	7	94	126		17	8	3		3	19		462	2	1121		62	201	11	5710	18		
web	15	6				28		10			60		142		381	55		21	23	3	2598		

(1.10)

8.4 Header model - HEP dataset



Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web
abstract	0.988	0.004	0.002	0.014		0.008	0.006	0.006	0.026	0.07	0.008		0.024	0.036	0.01	0.032	
address	0.4	0.906	0.38	0.428	0.02	0.008		0.066		0.102	0.072			0.016		0.014	
affiliation	0.29	0.218	0.928	0.236		0.006		0.016		0.04	0.158			0.008		0.004	
author	0.202	0.208	0.074	0.966	0.012	0.076		0.034		0.032	0.012		0.094	0.006	0.044	0.512	
collaboration			0.132	0.2	0.858	0.05											0.15
copyright	0.97	0.008	0.02	0.03		0.938					0.32		0.056	0.4	0.4		
date	0.2	0.2	0.054				0.92				0.14	0.286		0.408	0.518		
email	0.188	0.1	0.002	0.018		0.002		0.98		0.004	0.258	0.028		0.004	0.2	0.01	
grant	0.6								1.0		0.4				0.2		
keyword	0.376	0.118	0.282							0.972	0.08			0.122	0.2		
note	0.506	0.194	0.268	0.154		0.072	0.2	0.522		0.32	0.764		0.188	0.604	0.124	0.56	0.074
phone											0.6	0.2					
pubnum	0.018			0.082							0.06		0.984	0.2	0.2		0.2
reference	0.54	0.13	0.018	0.084		0.152	0.156			0.22	0.612		0.088	0.908		0.316	
submission	0.324	0.2					0.4	0.042	0.2		0.456			0.952			0.058
title	0.288			0.066							0.028			0.032		0.992	
web	0.2										0.412		0.2	0.2			0.188
	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web

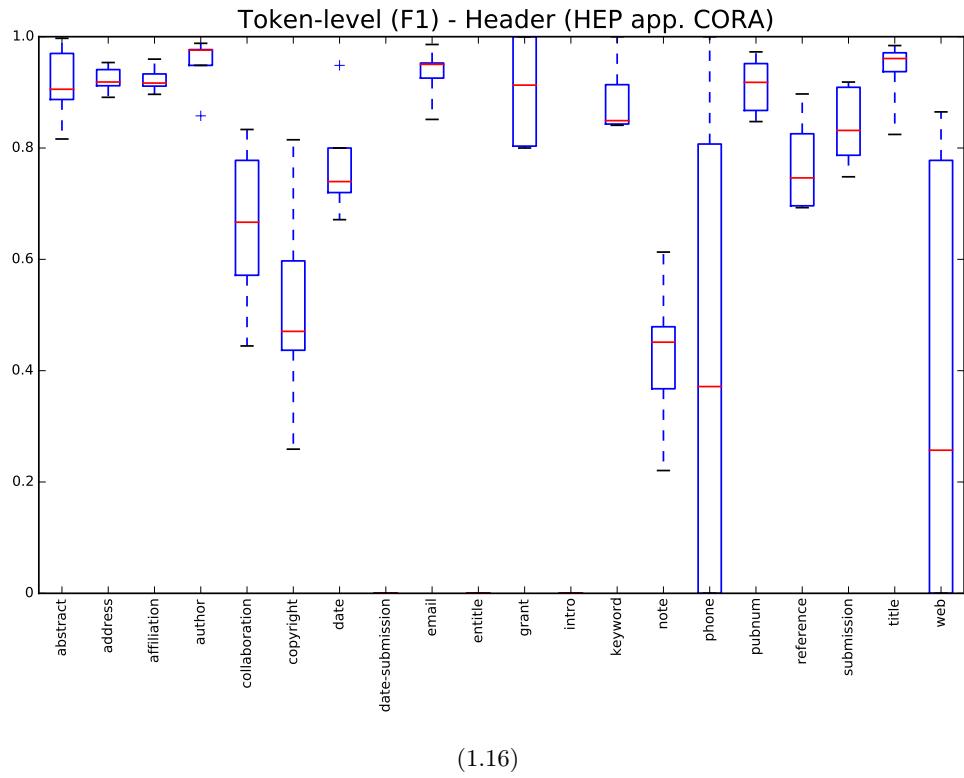
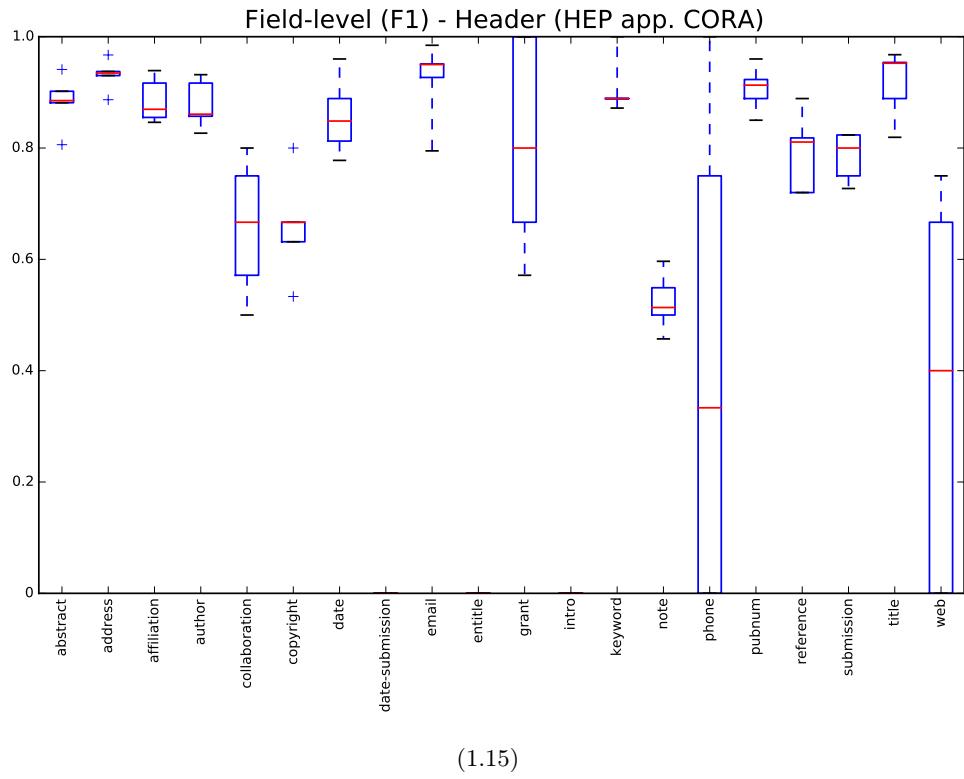
(1.13)

Confusion matrix - Header (HEP)

	abstract	11	3	84		11	21	14	17	64	19		42	120	17	59	
abstract	23027																
address	10	2604	199	151	1	33		8		40	9			68		63	
affiliation	38	198	3415	128	2	20		4		36	34			24		12	
author	45	22	77	3742	1	18		2		83	1			18	12	5	24
collaboration			2	3	28	1										4	
copyright	275	6	15	5		147					43		10	15	12		
date	3	5	6				151			7	15			18	29		
email	35	15	3	43		6		1866		9	66	4		11	6	21	
grant	31								290		45				10		
keyword	87	13	12							1359	30			38	41		
note	212	27	38	14		40	16	202		78	517		9	137	30	99	14
phone										47	9						
pubnum	1			21						7		776	17	4		24	
reference	114	21	6	75		24	14			56	82		15	1137		13	
submission	16	7				35	27	15		69				660		9	
title	75			9						9			20		1900		
web	9									36		23	19			15	
	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web

(1.14)

8.5 Header model - HEP dataset appending CORA dataset



Confusion matrix - Header (HEP app. CORA)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	email	entitle	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web
abstract	0.97																			
address		0.952	0.246	0.248																
affiliation	-0.052	0.192	0.94	0.288		0.032	0.046		0.042	0.04			0.004	0.48						
author		0.17	0.16	0.974		0.074							0.008	0.294		0.094			0.044	
collaboration		0.066	0.334		0.916								0.2	0.45					0.1	
copyright	-0.428		0.014	0.03		0.906							0.176	0.222	0.05	0.016	0.272			0.006
date							0.976								0.054		0.18	0.2	0.054	
date-submission																				
email	0.128	0.034	0.112	0.006					0.962		0.2	0.2		0.294						
entitle																				
grant											1.0			0.4						
intro																				
keyword		0.054	0.1										0.246	0.97	0.11		0.006	0.372		
note	0.556	0.034	0.264	0.208	0.014	0.14	0.04		0.482		0.2	0.276	0.264	0.846		0.22	0.444	0.2	0.518	0.104
phone									0.32					0.08	0.4					
pubnum													0.046		0.006	0.018			0.2	
reference		0.118	0.018	0.096		0.03	0.12					0.2	0.222	0.502		0.21	0.906		0.2	
submission						0.288	0.232		0.242				0.422				0.926		0.146	
title											0.094		0.626			0.016		0.986		0.6
web													0.2		0.2					

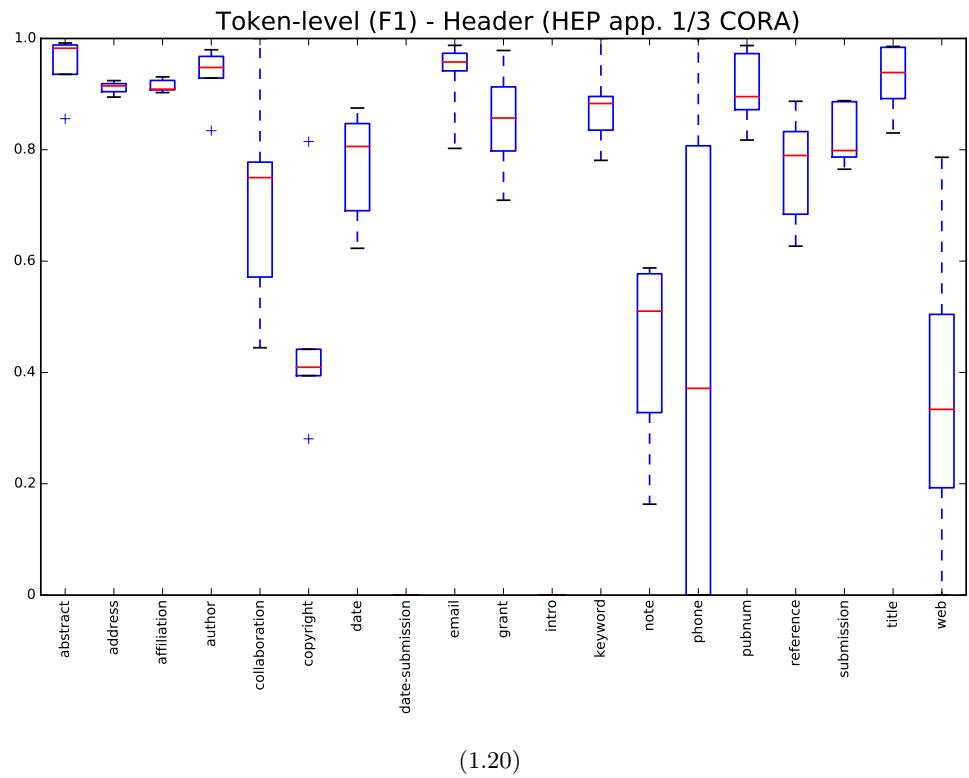
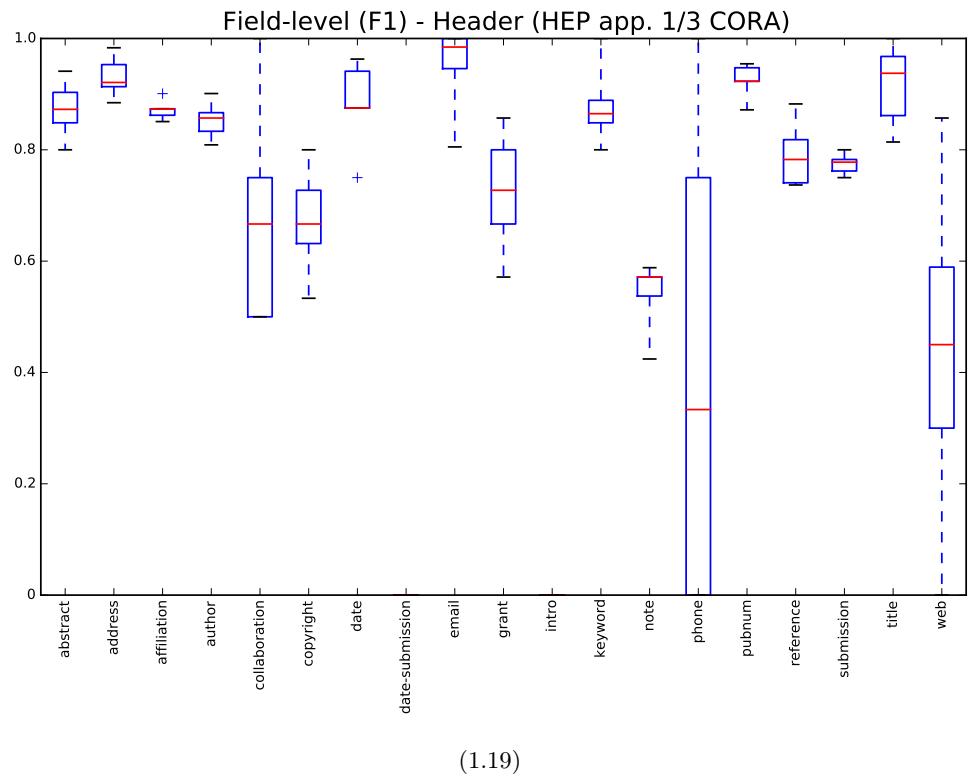
(1.17)

Confusion matrix - Header (HEP app. CORA)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	email	entitle	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	1969							8	13		17	2957	63	731		19	2	6			
address		2926	96	18				5	18			93		6	24						
affiliation	-14	183	3488	69	2	6	6		24	17		2	96		4						
author		33	35	3014		14			1		14		115		19		5				
collaboration		1	4		22						2		7				2				
copyright	59		10	5		201					39		184	9	1	15		5			
date							202						3		16	7	6				
date-submission																					
email	7	6	9	8					15.5		9	17		90							
entitle																					
grant											330		46								
intro												8	1468	22		1	124				
keyword		6	11									16	190	63	823		33	70	16	34	19
note	55	5	21	16	2	39	10		21			18	106	187							
phone									30				4	22							
pubnum												3		322	1				24		
reference		17	6	32		13	14					18	106	187		25	1133	6			
submission						26	57		31				74				630		20		
title												64		118		12		1790			
web													23		19		60				

(1.18)

8.6 Header model - HEP dataset appending 1/3 CORA dataset



Confusion matrix - Header (HEP app. 1/3 CORA)

	abstract	address	affiliation	author	collaboration	copyright	date	date submission	email	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web
abstract	0.982																		
address	0.5	0.944	0.258	0.116				0.2	0.074							0.032		0.04	
affiliation	0.332	0.202	0.93	0.082				0.032	0.038	0.042	0.028				0.022	0.284		0.022	
author	0.008	0.376	0.098	0.972				0.074		0.046				0.142	0.416	0.094	0.018	0.044	0.002
collaboration	0.2	0.066	0.334		0.9										0.05				0.15
copyright	0.34		0.014	0.04		0.93								0.064	0.17	0.468	0.05		
date	0.14		0.054				0.968										0.2	0.17	0.064
date-submission																			
email	0.04	0.006		0.01		0.1			0.97		0.2		0.236						
grant	0.2									1.0			0.2			0.2			
intro																			
keyword	0.376	0.028									0.372	0.972	0.114				0.162		
note	0.616	0.034	0.114	0.218		0.134	0.2		0.51	0.374	0.27	0.314	0.836		0.152	0.616	0.262	0.516	0.094
phone									0.32				0.08	0.4					
pubnum											0.018			0.998	0.2			0.2	
reference	0.13	0.018	0.096		0.03	0.126			0.042		0.2	0.158	0.65		0.042	0.914		0.206	
submission	0.2				0.4	0.2			0.042			0.382				0.944			0.058
title	0.034		0.078			0.022					0.062	0.356			0.016		0.974		0.6
web													0.2		0.2				0.6

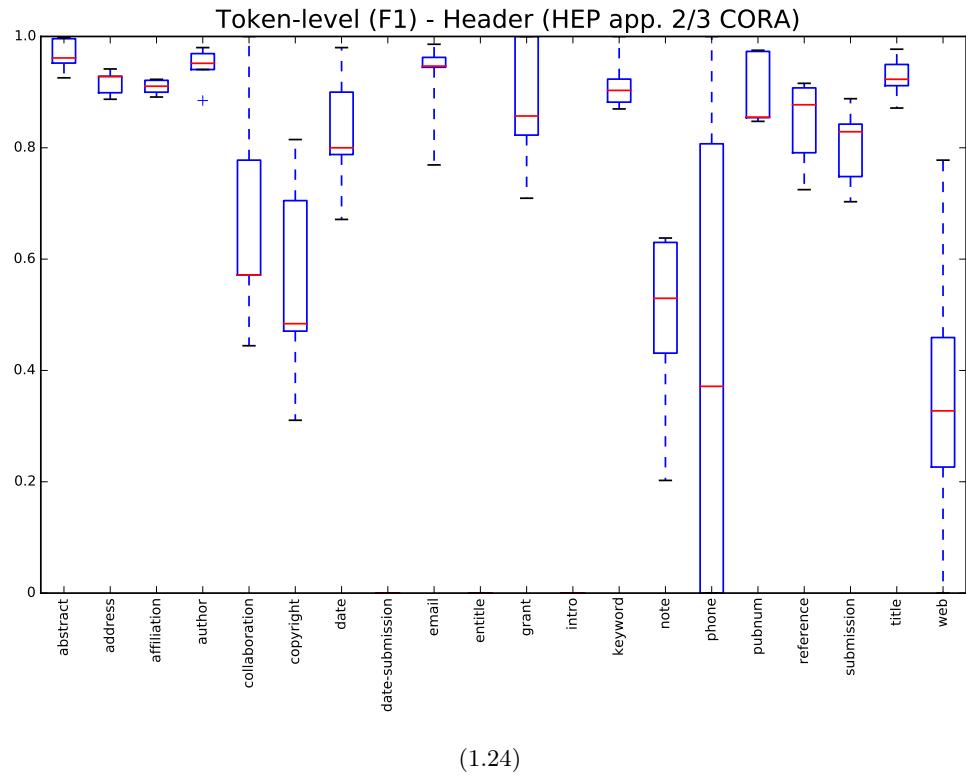
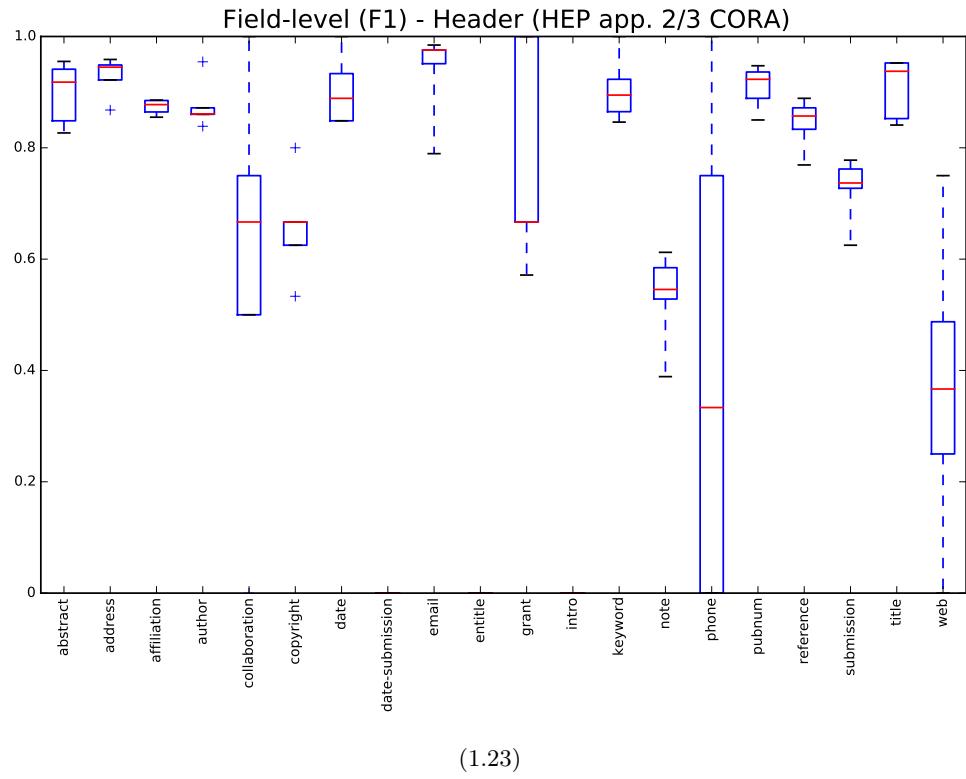
(1.21)

Confusion matrix - Header (HEP app. 1/3 CORA)

	abstract	address	affiliation	author	collaboration	copyright	date	date submission	email	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web
abstract	21387			32			12	28	91	17	1271	63	532	49	18		14		
address	13	2897	138	44			2	8				73			11				
affiliation	48	191	3469	78	2	6	5		22	5		6	65			14			
author	14	38	42	3731		14			4			5	125		19	51	5	2	
collaboration	2	1	5		24							2					4		
copyright	134		10	12		196				15	28		75	9			3	46	
date	7		6				198								9	14			
date-submission																			
email	3	2		15		9			1963		5		88						
grant	8									325			42		1				
intro																			
keyword	92	3								22	1374	23			66				
note	23	5	12	19		34	8		22	36	179	62	761		23	154	46	31	18
phone												4	22						
pubnum												1		621	4			24	
reference		21	6	32		13	15				18	54	234		7	1149		8	
submission	4					40	48		27			66				644			9
title	26			10			3				38		165		12		1759		
web														23		19		60	

(1.22)

8.7 Header model - HEP dataset appending 2/3 CORA dataset



Confusion matrix - Header (HEP app. 2/3 CORA)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	email	entitle	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	0.986	0.004				0.01	0.002	0.034			0.026	0.282	0.068	0.16		0.026	0.01	0.004	0.022		
address	0.2	0.946	0.284	0.352			0.2		0.102				0.07		0.1						
affiliation	0.272	0.192	0.934	0.254		0.032	0.046		0.042	0.064		0.2	0.002	0.31							
author	0.014	0.282	0.066	0.98		0.074		0.034			0.2	0.288		0.094		0.044					
collaboration	0.2	0.066	0.334		0.866						0.2	0.2		0.1					0.1		
copyright	0.228		0.014	0.03		0.904					0.064	0.184		0.128	0.05	0.016	0.272	0.2		0.082	
date			0.054				0.062									0.2	0.17				
date-submission																					
email	0.24	0.034	0.134	0.004					0.966			0.2	0.252								
entitle																					
grant	0.2										1.0		0.4								
intro												0.2	0.114		0.122						
keyword	0.176	0.054	0.1								0.2	0.982									
note	0.6	0.034	0.152	0.208		0.052			0.532			0.47	0.282	0.806		0.244	0.432	0.256	0.64	0.026	
phone									0.32					0.08	0.4						
pubnum												0.2		0.008	0.018				0.2		
reference	0.074	0.118	0.018	0.096		0.11	0.12				0.322	0.076	0.348		0.21	0.92		0.2			
submission	0.512					0.288	0.232	0.2	0.042				0.23				0.924		0.058		
title	0.102		0.006	0.094								0.106	0.588		0.03		0.986				
web													0.2		0.2			0.6			

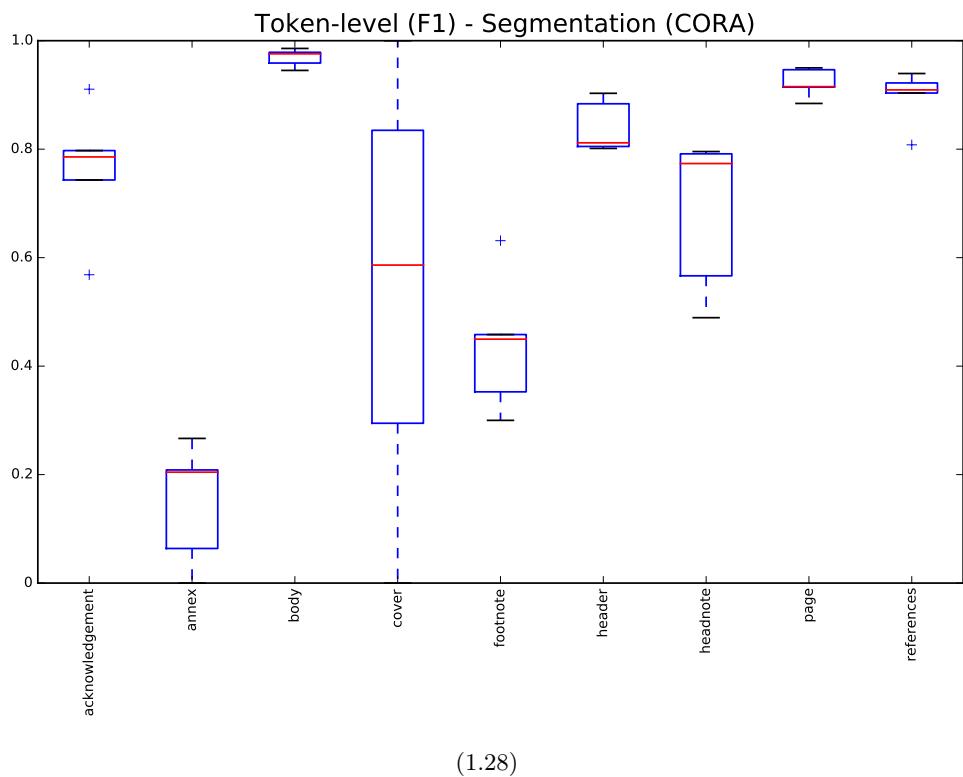
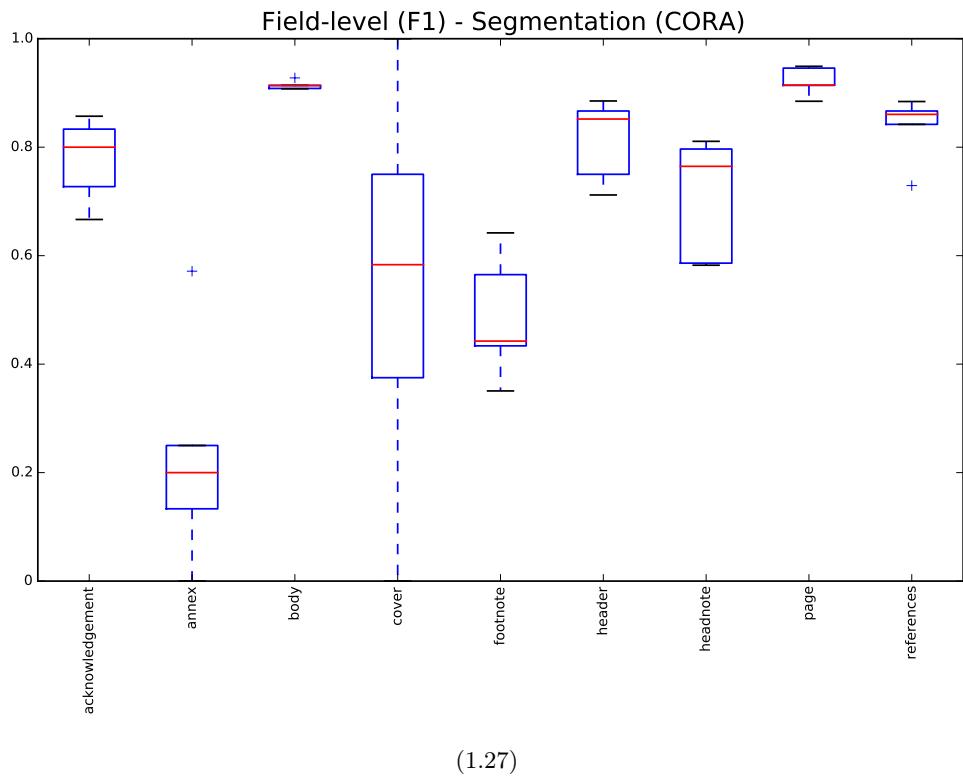
(1.25)

Confusion matrix - Header (HEP app. 2/3 CORA)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	email	entitle	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web
abstract	2211	2			105	8	94		17	442	63	573		37	19	7	31			
address	3	2951	139	24			2		5			53		9						
affiliation	25	182	3461	75	2	6	6		25	27		20	1	78					3	
author	26	35	30	378		14			2			4		136		19		5	1	
collaboration	2	1	5		22						2		4					2		
copyright	48		10	5		248					15	11		89	9	1	15	3		74
date			6				212									2	14			
date-submission																				
email	21	6	12	2					195			17		82						
entitle																				
grant	8										313			55						
intro												3	1456	23			38			
keyword	43	6	11									181	79	845		30	89	35	64	6
note	3	5	23	16		31			26				4	22						
phone									30											
pubnum												5		1				19		
reference	38	17	6	32		21	14				29	37	93		25	1239		6		
submission	32					26	57	22	27				57				608		9	
title	16		4	17							9	130				23		1814		
web												23		19		19		60		

(1.26)

8.8 Segmentation model - Cora dataset

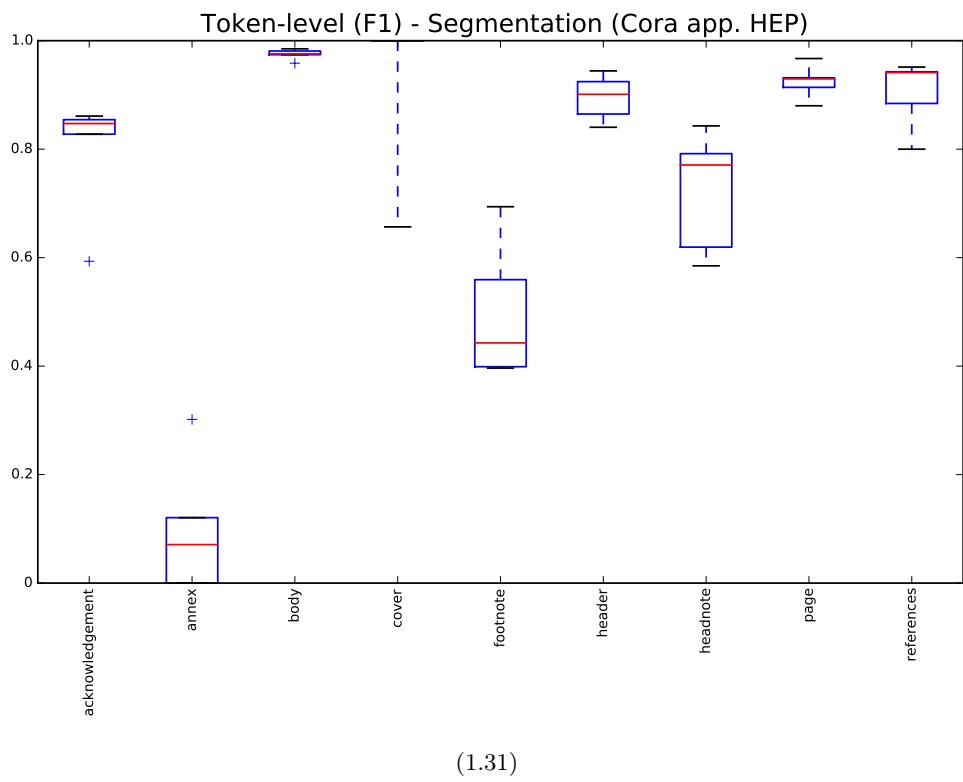
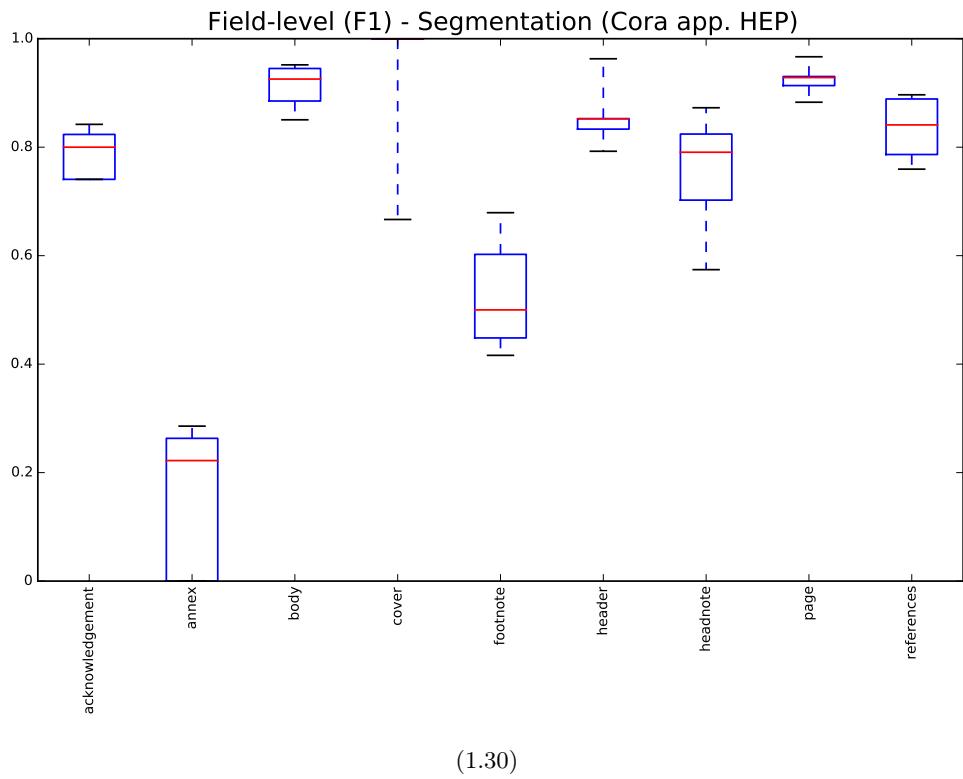


Confusion matrix - Segmentation (CORA)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	294	17	80		5	2			39
annex	31	109	1776		13		1	1	60
body	16	143	82791	45	113	374	11	26	430
cover				59		46			
footnote	4	62	467		355	28	51	2	80
header		6	491		19	2676	3	3	11
headnote		8	223		40	23	472	4	44
page		5	104		5	5		1050	15
references	6	77	1157		10		1	4	8317

(1.29)

8.9 Segmentation model - Cora dataset appending HEP dataset

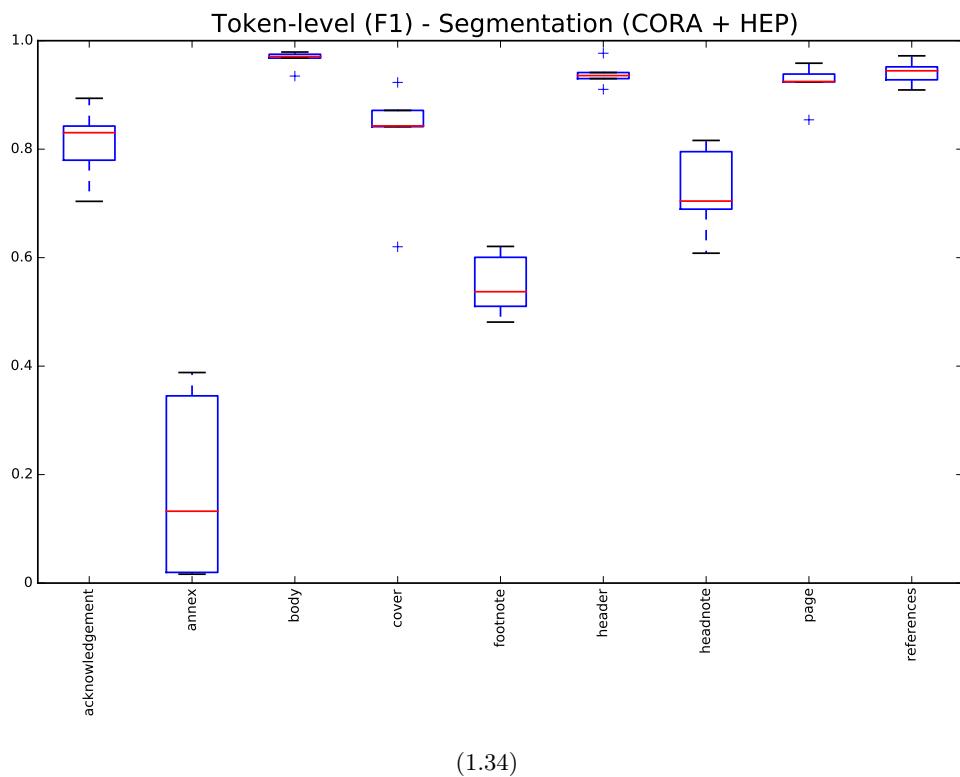
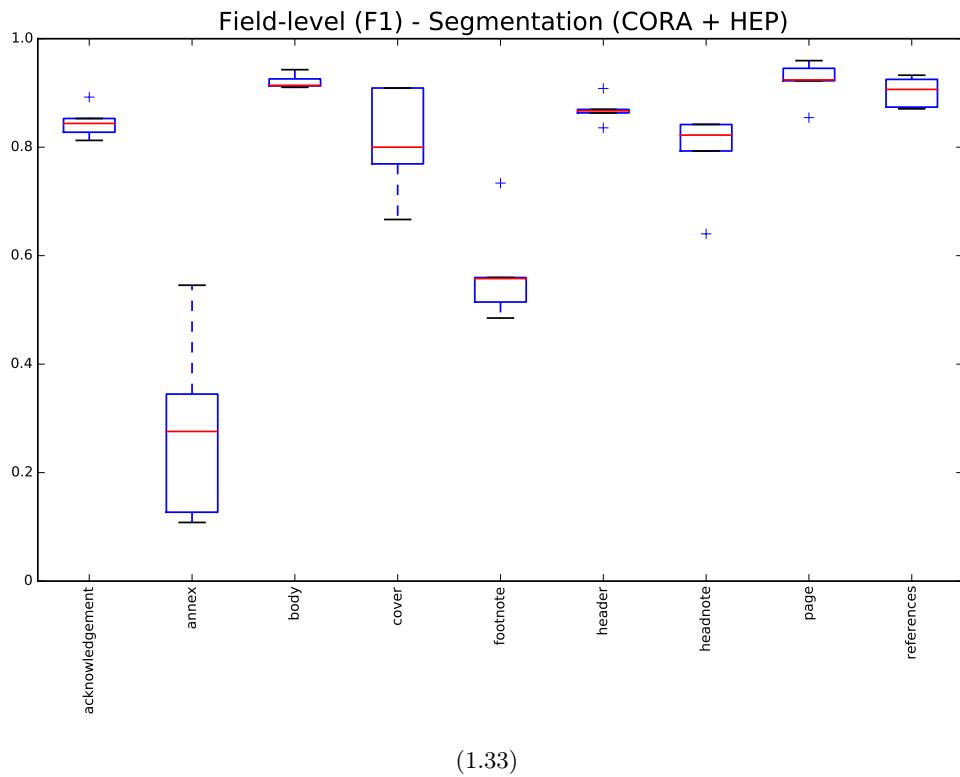


Confusion matrix - Segmentation (Cora app. HEP)

	acknowledgement	14	73	13				14
acknowledgement	323	14	73	13				14
annex	10	202	1482	2			1	294
body	54	427	82613	113	231	32	35	444
cover			82		23			
footnote		40	483	382	31	42	4	67
header			322	15	2856	7	2	7
headnote		9	208	30	20	506	8	33
page		2	83	3	5	4	1072	15
references		472	369	10	17	13	11	8680

(1.32)

8.10 Segmentation model - Cora and HEP combined datasets

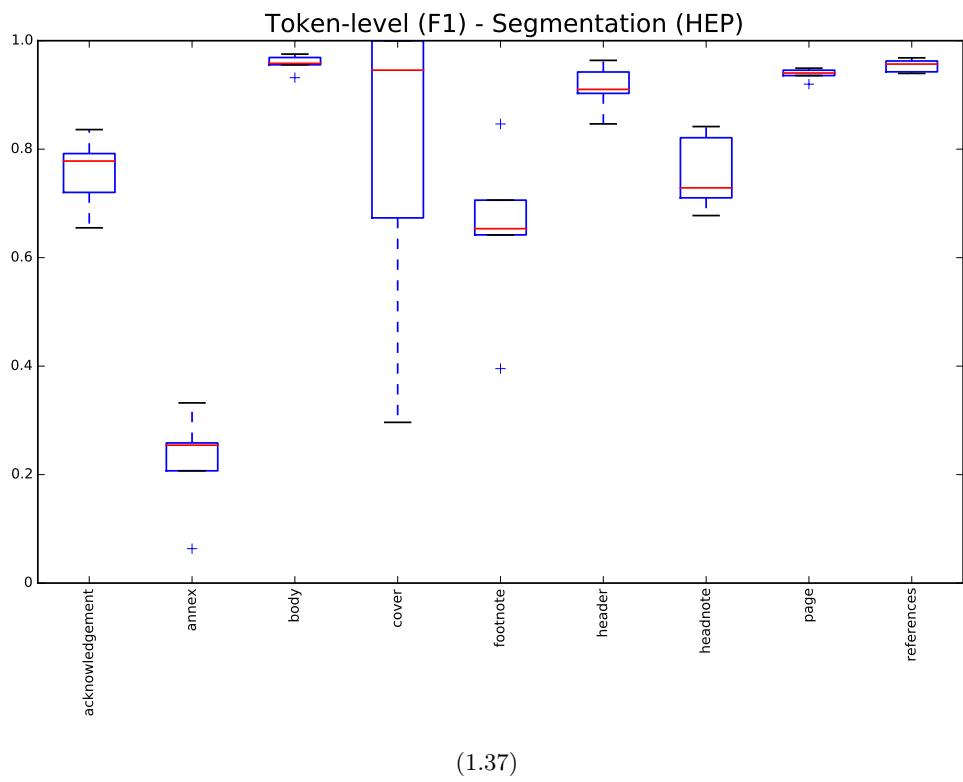
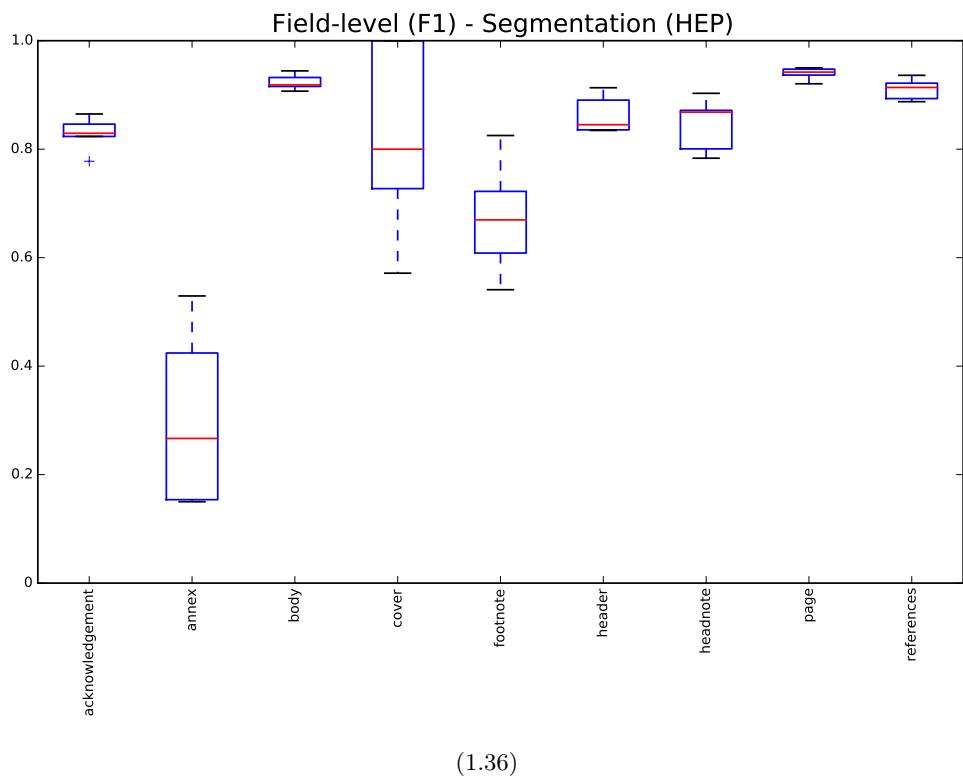


Confusion matrix - Segmentation (CORA + HEP)

	acknowledgement	23	283		5	19	1		24
acknowledgement	900								
annex	27	1455	11391		4	33	4	4	86
body	30	3478	268883	65	321	475	72	70	925
cover				319	5	71	1		
footnote	4	44	1021	3	1261	69	340	45	80
header		7	731	18	55	15569	23	10	100
headnote		16	644		57	90	1859	62	83
page	2	8	244	3	22	23	24	3300	30
references		348	752		16	36	35	13	20031
	acknowledgement	annex	body	cover	footnote	header	headnote	page	references

(1.35)

8.11 Segmentation model - HEP dataset

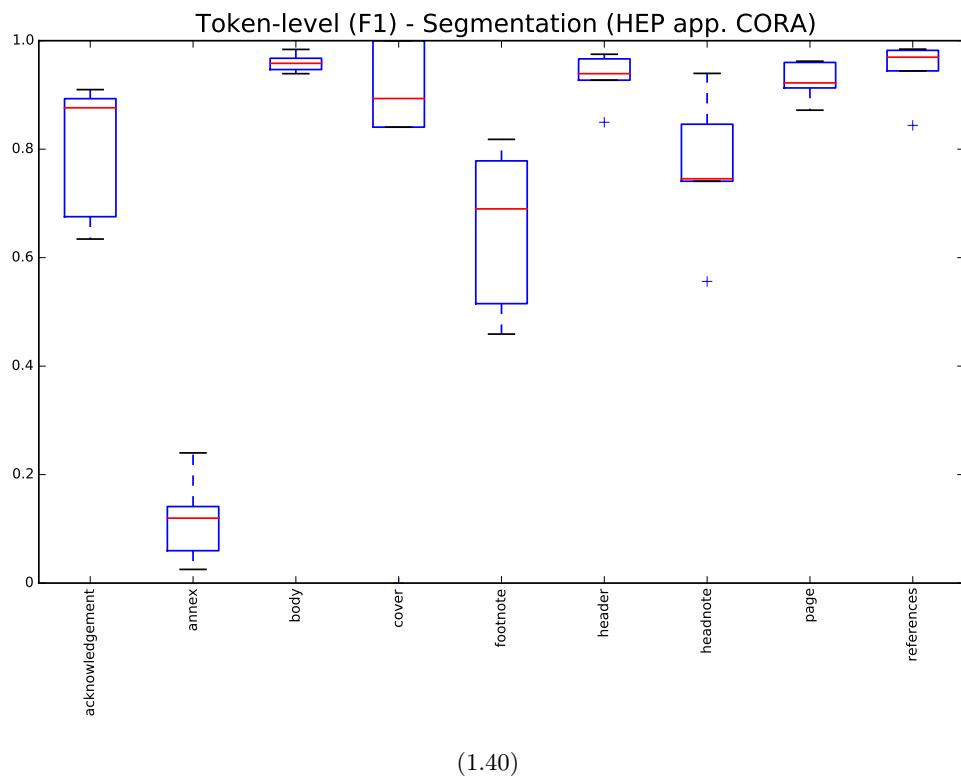
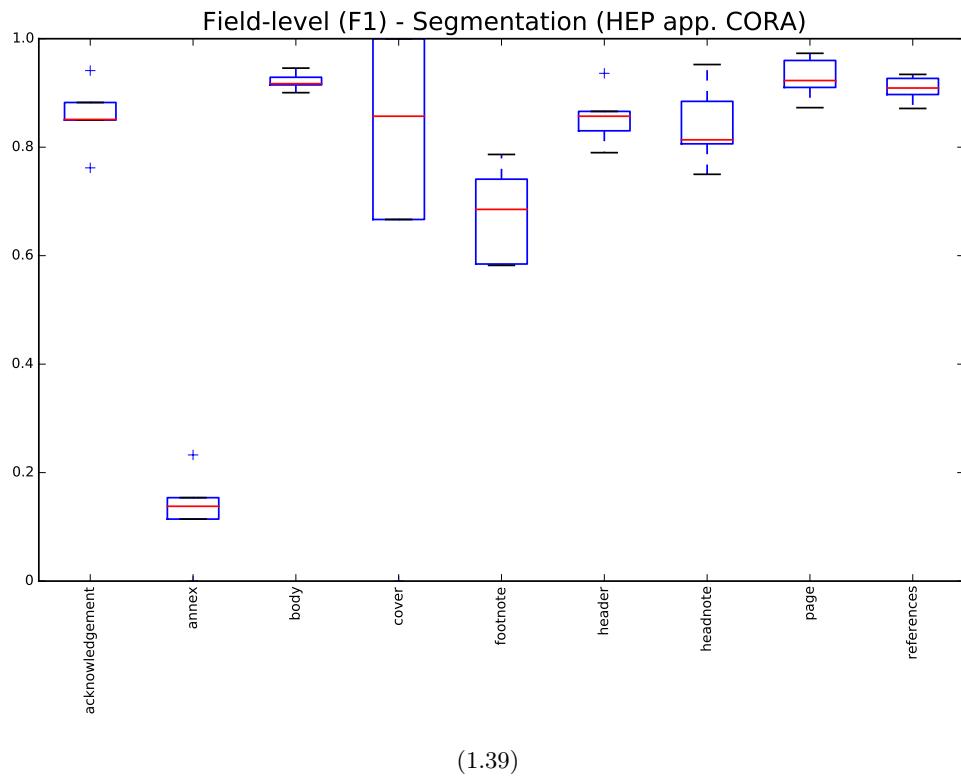


Confusion matrix - Segmentation (HEP)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references	
acknowledgement	533	14	260			10			1	
annex	18	1481	9443		1	17	2	2	49	
body	11	2727	186687	63	140	373	57	38	274	
cover				269	5	16	1			
footnote	4	1	497		1078	32	160	25	21	
header		5	1582	140	84	11408	9	5	71	
headnote		13	404		75	76	1348	40	41	
page	2	6	116		20	14	11	2292	11	
references	45	7	479		17	39	24	6	11042	

(1.38)

8.12 Segmentation model - HEP dataset appending CORA dataset



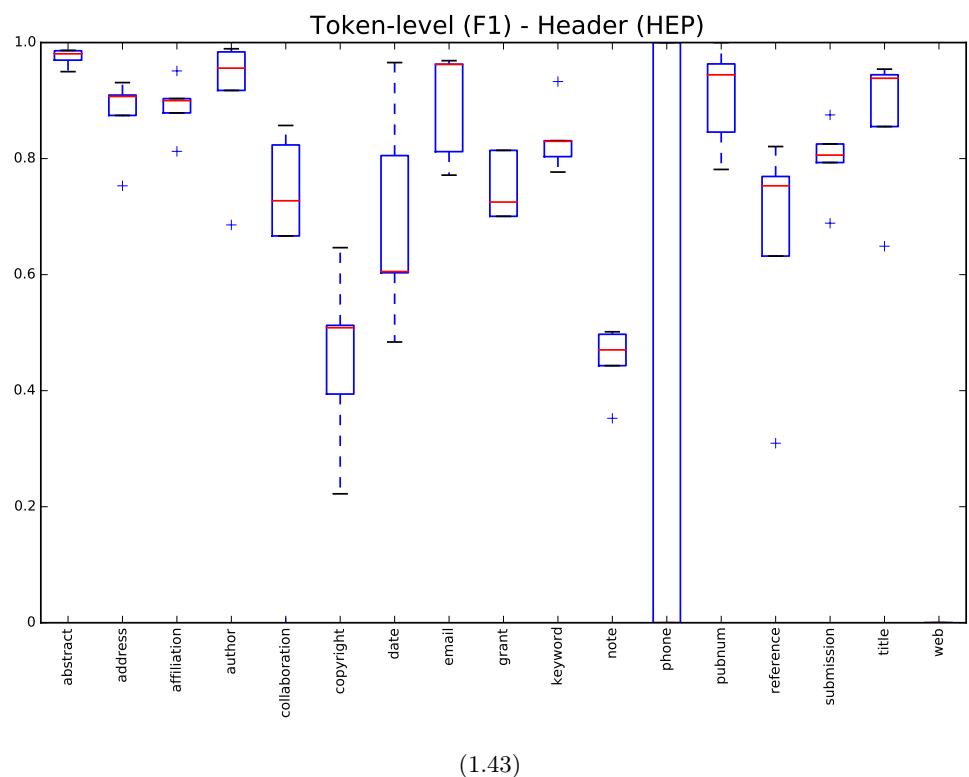
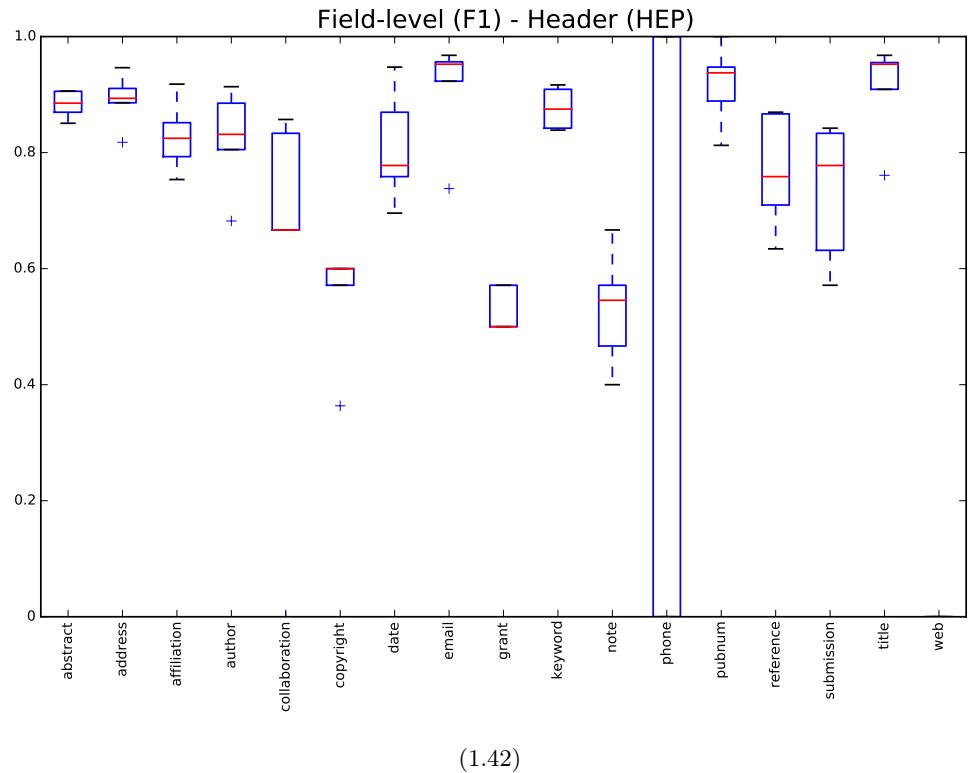
Confusion matrix - Segmentation (HEP app. CORA)

	acknowledgement	8	229			27	1		1
acknowledgement	552								
annex	21	1214	9662			32	6	2	76
body	31	3798	185852		268	121	43	35	222
cover				259	5	26	1		
footnote	4	1	524		1152	33	45	32	27
header		36	547	19	36	12390	14	7	255
headnote	1	12	454		49	71	1302	43	65
page	2	6	130	2	14	15	18	2261	24
references		30	541		16	4	14	3	11051
	acknowledgement	annex	body	cover	footnote	header	headnote	page	references

(1.41)

9 Regularisation

9.1 Header model - $L2 = 0$



Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web
abstract	0.988	0.024	0.004		0.008	0.014	0.006	0.026	0.092	0.04		0.01	0.012	0.014	0.036		
address	0.246	0.922	0.37	0.348	0.02			0.13		0.1	0.038			0.054		0.21	
affiliation	0.23	0.19	0.93	0.494				0.076			0.262			0.048		0.164	
author	0.11	0.154	0.114	0.96		0.074		0.072		0.464		0.094	0.038	0.044	0.508		
collaboration			0.266	0.2	0.714					0.2					0.3		
copyright	0.676	0.008	0.022			0.848			0.2		0.33		0.056	0.384	0.2		
date		0.29	0.108	0.114			0.916				0.14		0.68	0.32	0.2		
email	0.84	0.058		0.028		0.2		0.986					0.048		0.008		
grant	0.4					0.2			0.8		0.2				0.2		
keyword	0.26	0.054	0.346						0.944	0.046	0.156		0.418				
note	0.488	0.048	0.186	0.138	0.2	0.268	0.2	0.342	0.546	0.79		0.4	0.4	0.124	0.49	0.074	
phone							0.2			0.2	0.4						
pubnum	0.2									0.072		0.988	0.142			0.2	
reference	0.126	0.13	0.044	0.15		0.08	0.078		0.2	0.262	0.552		0.088	0.918		0.316	
submission	0.448					0.2	0.6	0.042	0.2		0.11			0.968	0.2	0.058	
title	0.282			0.068						0.63			0.2		0.98		
web	0.2									0.6		0.2		0.4			

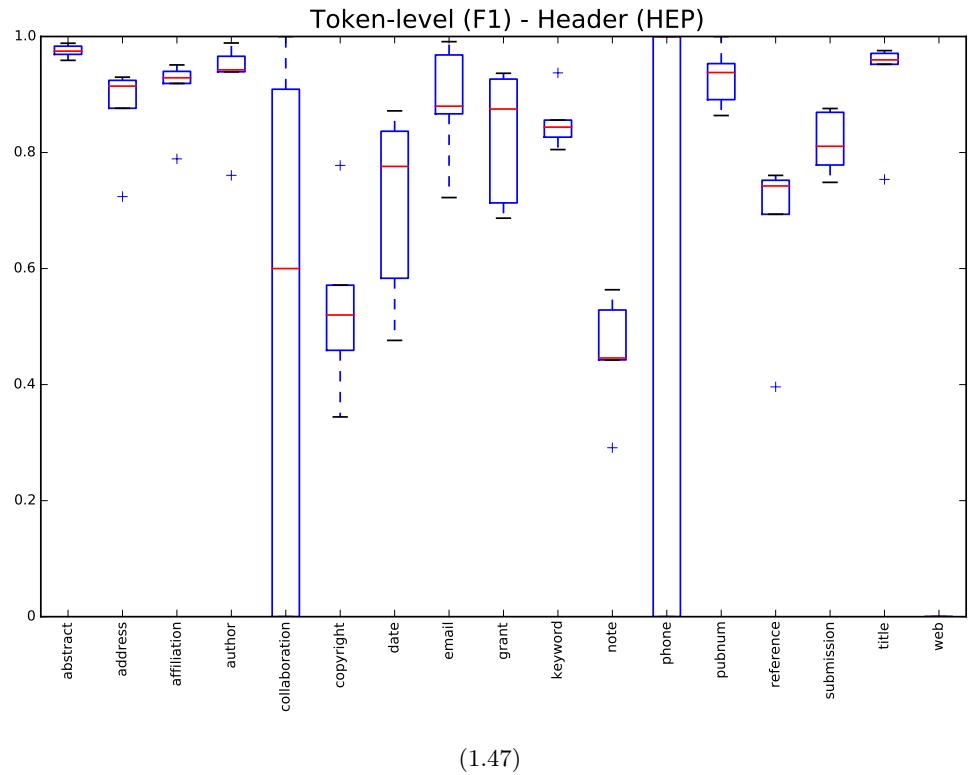
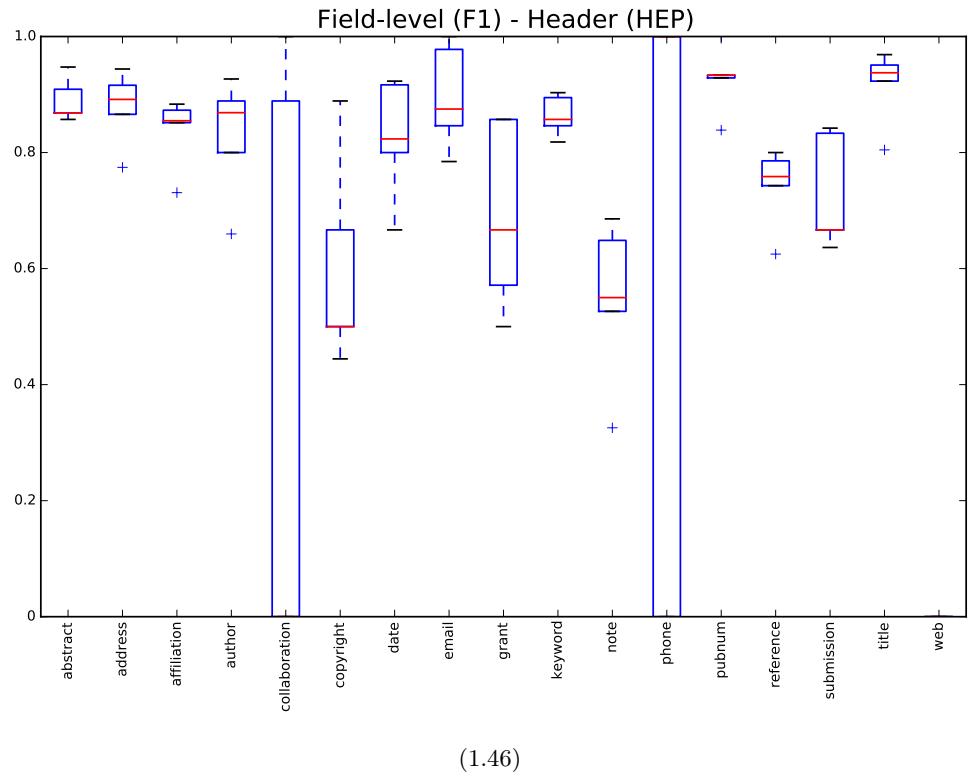
(1.44)

Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web
abstract	22897		43	77		11	14	3	17	77	100		21	128	25	96	
address	12	2548	174	130	1			8		9	10			241		53	
affiliation	16	192	3384	89				16			30			157		27	
author	42	32	83	3728		14		8		18		18		73	5	29	
collaboration			3	3	25					2					5		
copyright	190	6	17			222			12		25		10	37	9		
date		10	9	4			150				7			14	37	3	
email	74	22		62		9		1789						109		20	
grant	20					22			293		21				20		
keyword	47	6	19						1287	7	80			134			
note	139	15	41	10	1	46	8	178		85	573		12	169	30	112	
phone							10			13	33					14	
pubnum	11									4		784	27			24	
reference	80	21	15	66		10	7		18	77	121		15	1114		13	
submission	16					25	49	9	15		30			671	14	9	
title	79			17						96			66		1755		
web	9									35		23		35			

(1.45)

9.2 Header model - $L2 = 1e^{-6}$



Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	0.092			0.01				0.006	0.026	0.056	0.008			0.008	0.01	0.008		
address	0.4	0.918	0.35	0.426				0.13		0.1	0.036			0.08		0.002		
affiliation	0.316	0.182	0.944	0.228				0.016			0.284			0.068				
author	0.424	0.176	0.112	0.978		0.074		0.034		0.318		0.094	0.056	0.044	0.22			
collaboration		0.2	0.2	0.4	0.526	0.05										0.3		
copyright	0.512	0.008	0.022			0.888					0.412			0.4	0.2			
date	0.37	0.2	0.054	0.114			0.912							0.48	0.688			
email	0.344	0.072	0.016	0.014		0.034		0.974			0.06			0.056		0.002		
grant	0.4								1.0		0.6							
keyword	0.564	0.054	0.346						0.962	0.16				0.158				
note	0.638	0.138	0.182	0.288		0.13	0.2	0.48		0.546	0.8		0.222	0.498	0.24	0.348		
phone								0.2			0.2	0.6						
pubnum	0.2										0.354		0.988	0.046	0.2		0.2	
reference	0.412	0.13	0.018	0.09		0.144				0.42	0.42			0.914		0.354		
submission	0.6	0.2	0.026			0.2	0.2	0.042	0.174		0.198				0.97			
title	0.31			0.048							0.262			0.124		0.986		
web	0.2			0.026							0.6		0.2		0.376			

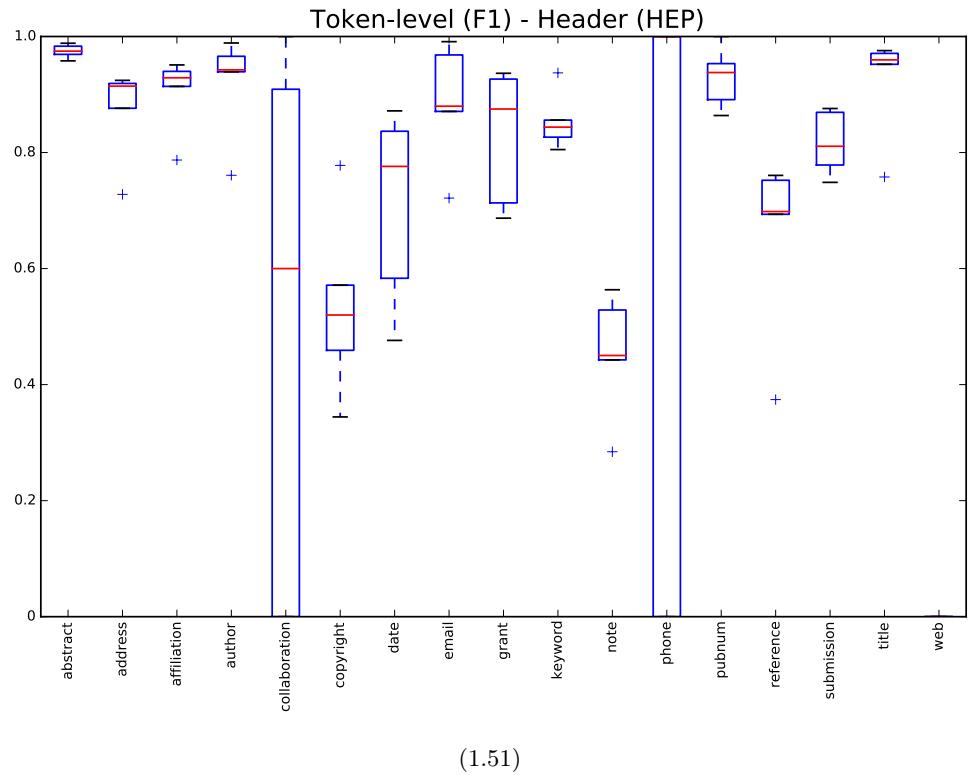
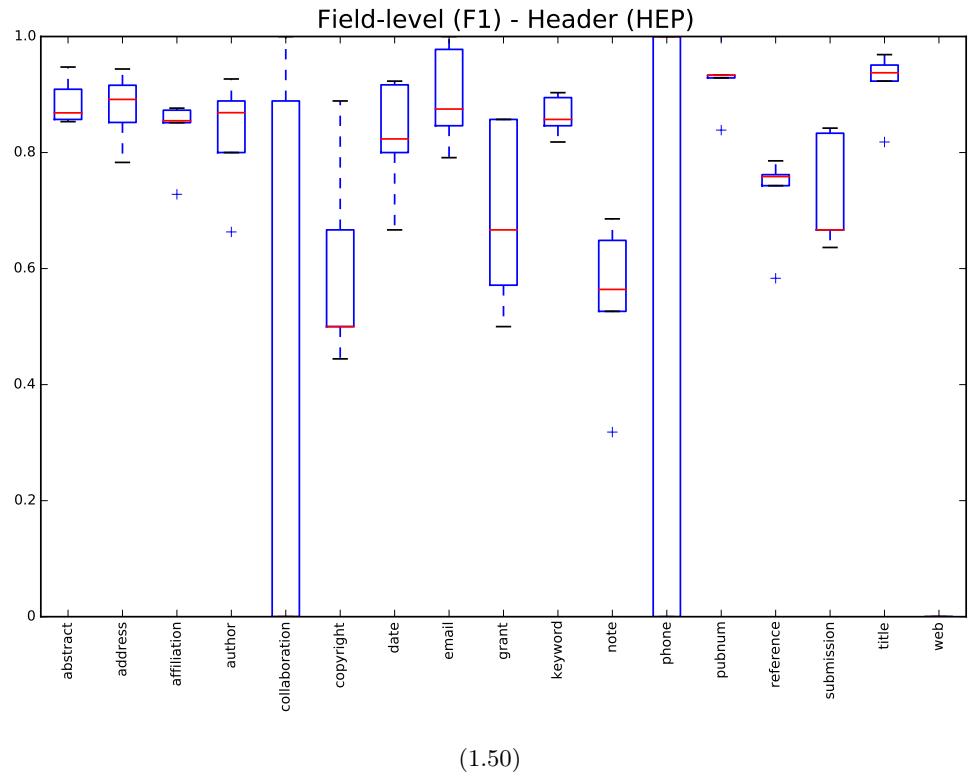
(1.48)

Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	23210			80			3	17	64	17			83	12	23			
address	10	2527	141	109				8		9	18			355		9		
affiliation	24	149	3377	87				4		43				226		1		
author	60	14	78	3715		14		3		24			18	105	5	14		
collaboration		2	3	6	21	1									5			
copyright	185	6	17			215					87			15	3			
date	12	5	6	4			142							11	54			
email	91	19	4	33		3		1798			7			127		3		
grant	42								300		34							
keyword	108	6	19						139	30				78				
note	209	16	34	12		29	8	186		78	550		7	206	47	51		
phone								10		13	33							
pubnum	11									30		783	3	4		19		
reference	98	21	6	37		25	16	9	13	77	111			1166		16		
submission	25	23	2			25	16			37				688				
title	103			4							21			74		1811		
web	9		2								35		23		33			

(1.49)

9.3 Header model - $L2 = 1e^{-5}$



Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	0.092			0.01				0.006	0.026	0.056	0.024			0.008	0.01	0.008		
address	0.4	0.912	0.35	0.426				0.134		0.1	0.036			0.246		0.002		
affiliation	0.316	0.182	0.942	0.23				0.016			0.238			0.174				
author	0.424	0.176	0.112	0.978		0.074		0.034		0.318		0.094	0.042	0.044	0.22			
collaboration		0.2	0.2	0.4	0.526	0.05									0.3			
copyright	0.512	0.008	0.022			0.888					0.412			0.4	0.2			
date	0.37	0.2	0.054	0.114			0.912							0.48	0.688			
email	0.344	0.072	0.016	0.014		0.034		0.974			0.06			0.054		0.002		
grant	0.4								1.0		0.6							
keyword	0.564	0.054	0.346						0.962	0.16				0.158				
note	0.638	0.138	0.182	0.288		0.13	0.2	0.48		0.546	0.8		0.222	0.498	0.24	0.348		
phone								0.2			0.2	0.6						
pubnum	0.2									0.354		0.988	0.046	0.2		0.2		
reference	0.412	0.13	0.018	0.164		0.144				0.42	0.42			0.908		0.354		
submission	0.6	0.2	0.026			0.2	0.2	0.042	0.174		0.198			0.97				
title	0.31			0.048						0.262			0.122		0.986			
web	0.2			0.026						0.6		0.2		0.376		title		

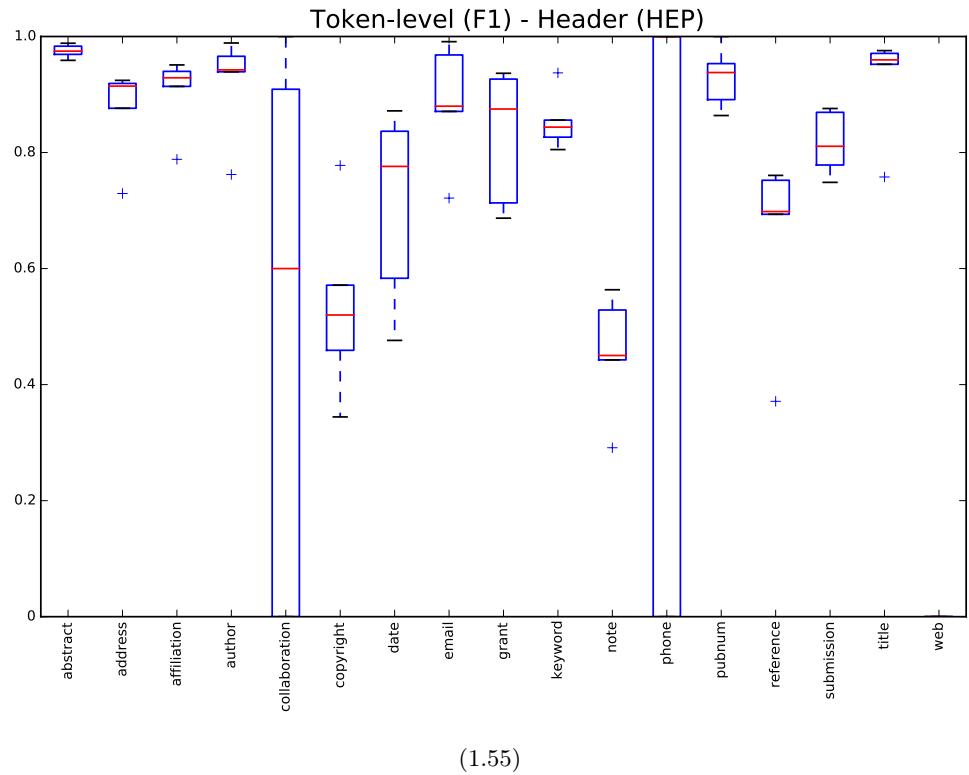
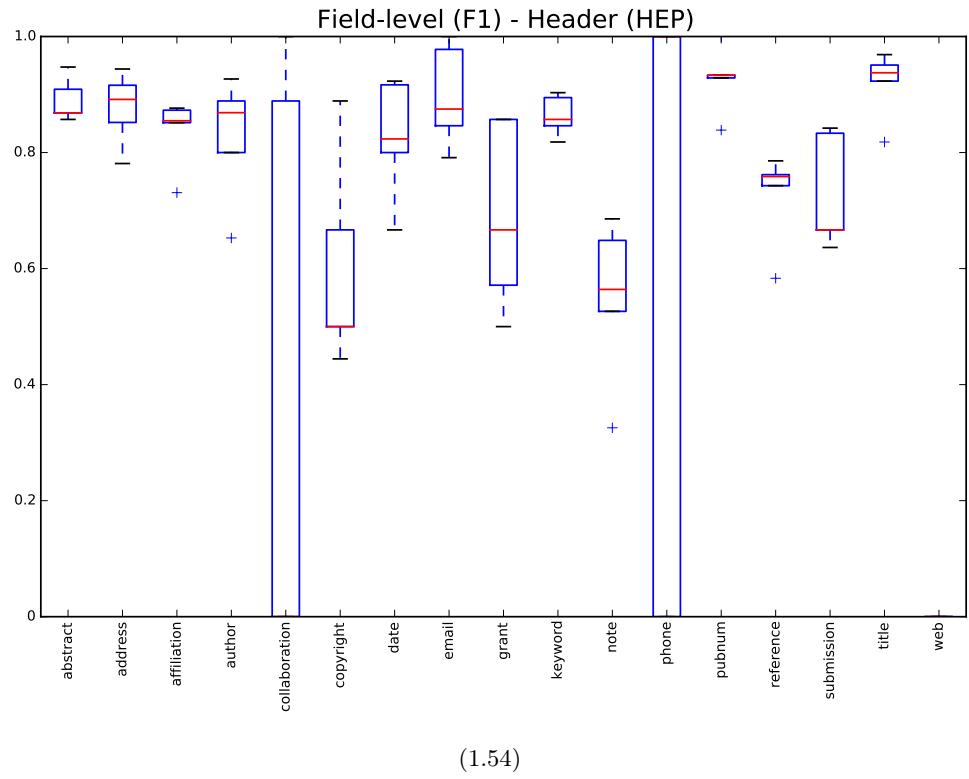
(1.52)

Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	23199			80			3	17	64	31			80	12	23			
address	10	2523	147	110				9		9	18			351		9		
affiliation	24	152	3369	90				4			37			234		1		
author	60	14	78	3741		14		3			24			18	79	5	14	
collaboration		2	3	6	21	1									5			
copyright	185	6	17			215					87			15	3			
date	12	5	6	4			142							11	54			
email	91	19	4	33		3		1803			7			122		3		
grant	42								300		34							
keyword	108	6	19						1339	30				78				
note	209	16	34	12		29	8	186		78	550		7	206	47	51		
phone								10		13	33							
pubnum	11									30		783	3	4		19		
reference	98	21	6	75		25	16	9	13	77	111			1128		16		
submission	25	23	2			25	16			37				688				
title	103			4						21			71		1814			
web	9		2							35		23		33				

(1.53)

9.4 Header model - $L2 = 1e^{-4}$



Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	0.092			0.01				0.006	0.026	0.056	0.008			0.008	0.01	0.008		
address	0.4	0.912	0.348	0.426				0.134		0.1	0.036			0.248		0.002		
affiliation	0.316	0.186	0.942	0.23				0.016			0.238			0.174				
author	0.424	0.176	0.112	0.978		0.074		0.034		0.318		0.094	0.044	0.044	0.22			
collaboration		0.2	0.2	0.4	0.526	0.05										0.3		
copyright	0.512	0.008	0.022			0.888					0.412			0.4	0.2			
date	0.37	0.2	0.054	0.114			0.912							0.48	0.688			
email	0.344	0.072	0.016	0.014		0.034		0.974			0.06			0.054		0.002		
grant	0.4								1.0		0.6							
keyword	0.564	0.054	0.346						0.962	0.16				0.158				
note	0.638	0.138	0.182	0.288		0.13	0.2	0.48		0.546	0.8		0.222	0.498	0.24	0.348		
phone								0.2			0.2	0.6						
pubnum	0.2										0.354		0.988	0.046	0.2		0.2	
reference	0.412	0.13	0.018	0.164		0.144				0.42	0.42			0.908		0.354		
submission	0.6	0.2	0.026			0.2	0.2	0.042	0.174		0.198				0.97			
title	0.31			0.048							0.262			0.122		0.986		
web	0.2			0.026							0.6		0.2		0.376			

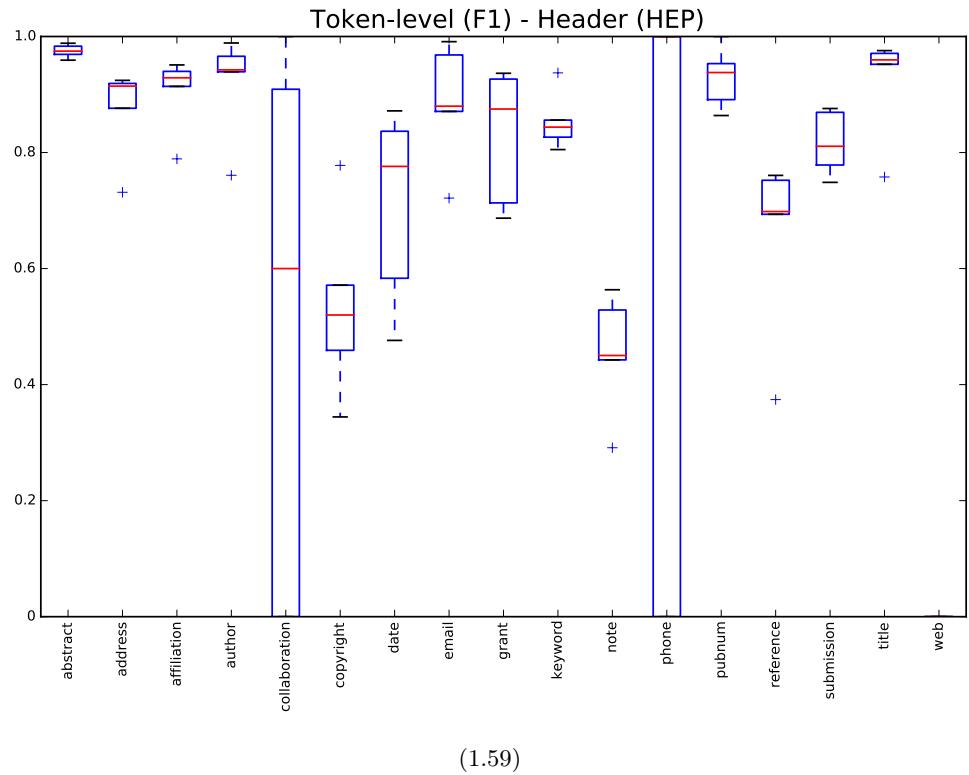
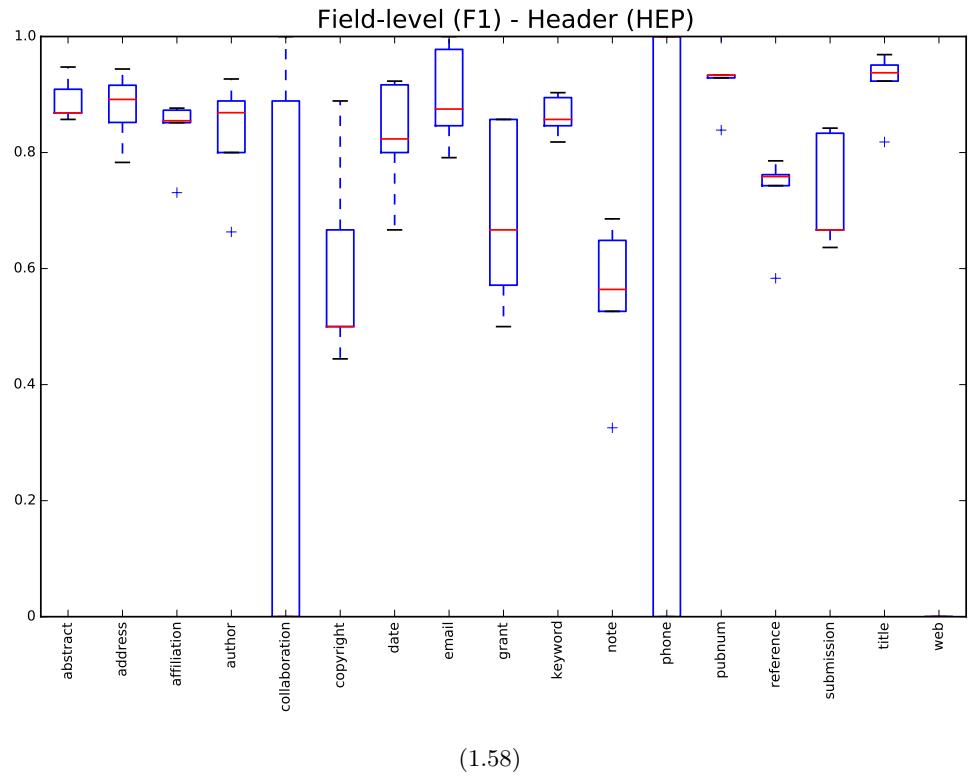
(1.56)

Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	23210			80			3	17	64	17			83	12	23			
address	10	2525	145	105				9		9	18			356		9		
affiliation	24	151	3370	90				4			37			234		1		
author	60	14	78	3738		14		3			24			18	82	5	14	
collaboration		2	3	6	21	1										5		
copyright	185	6	17			215					87			15	3			
date	12	5	6	4			142							11	54			
email	91	19	4	30		3		1803			7			125		3		
grant	42								300		34							
keyword	108	6	19						1339	30				78				
note	209	16	34	12		29	8	186		78	550		7	206	47	51		
phone								10		13	33							
pubnum	11									30		783	3	4		19		
reference	98	21	6	75		25	16	9	13		111			1128		16		
submission	25	23	2			25	16				37				688			
title	103			4							21			71		1814		
web	9		2							35		23		33				

(1.57)

9.5 Header model - $L2 = 1e^{-3}$



Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	0.092			0.01				0.006	0.026	0.056	0.008			0.008	0.01	0.008		
address	0.4	0.912	0.35	0.426				0.134		0.1	0.036			0.246		0.002		
affiliation	0.316	0.182	0.942	0.23				0.016			0.238			0.174				
author	0.424	0.176	0.112	0.978		0.074		0.034		0.318		0.094	0.042	0.044	0.22			
collaboration		0.2	0.2	0.4	0.526	0.05										0.3		
copyright	0.512	0.008	0.022			0.888					0.412			0.4	0.2			
date	0.37	0.2	0.054	0.114			0.912							0.48	0.688			
email	0.344	0.072	0.016	0.014		0.034		0.974			0.06			0.054		0.002		
grant	0.4								1.0		0.6							
keyword	0.564	0.054	0.346						0.962	0.16				0.158				
note	0.638	0.138	0.182	0.288		0.13	0.2	0.48		0.546	0.8		0.222	0.498	0.24	0.348		
phone								0.2			0.2	0.6						
pubnum	0.2									0.354		0.988	0.046	0.2		0.2		
reference	0.412	0.13	0.018	0.164		0.144				0.42	0.42			0.908		0.354		
submission	0.6	0.2	0.026			0.2	0.2	0.042	0.174		0.198				0.97			
title	0.31			0.048						0.262				0.122		0.986		
web	0.2			0.026						0.6		0.2		0.376				

(1.60)

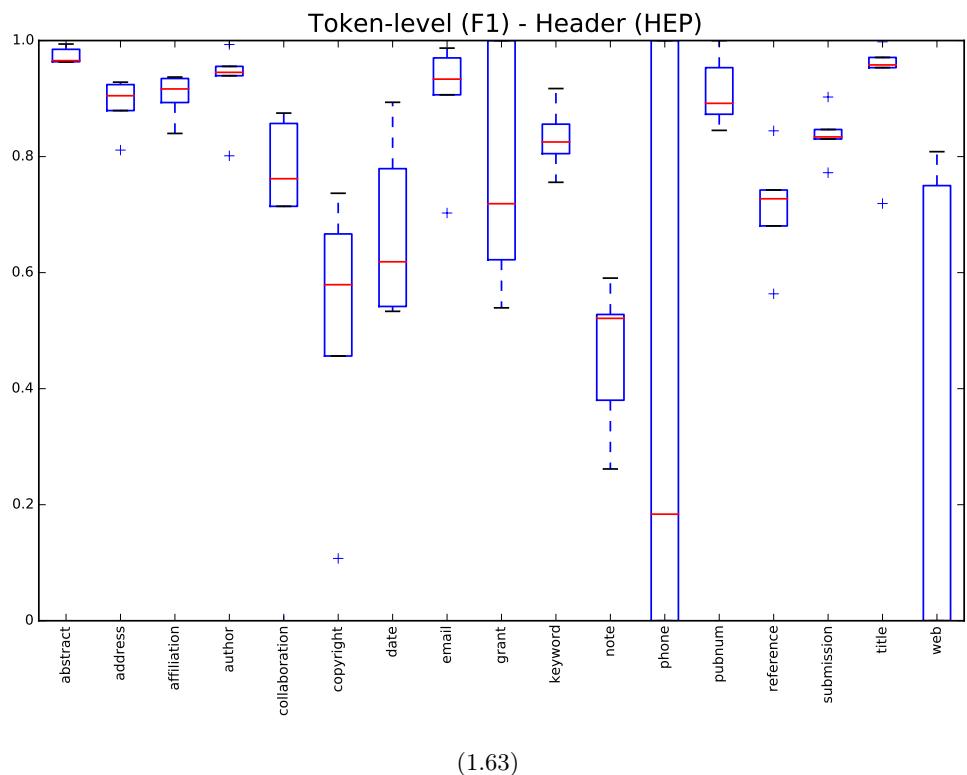
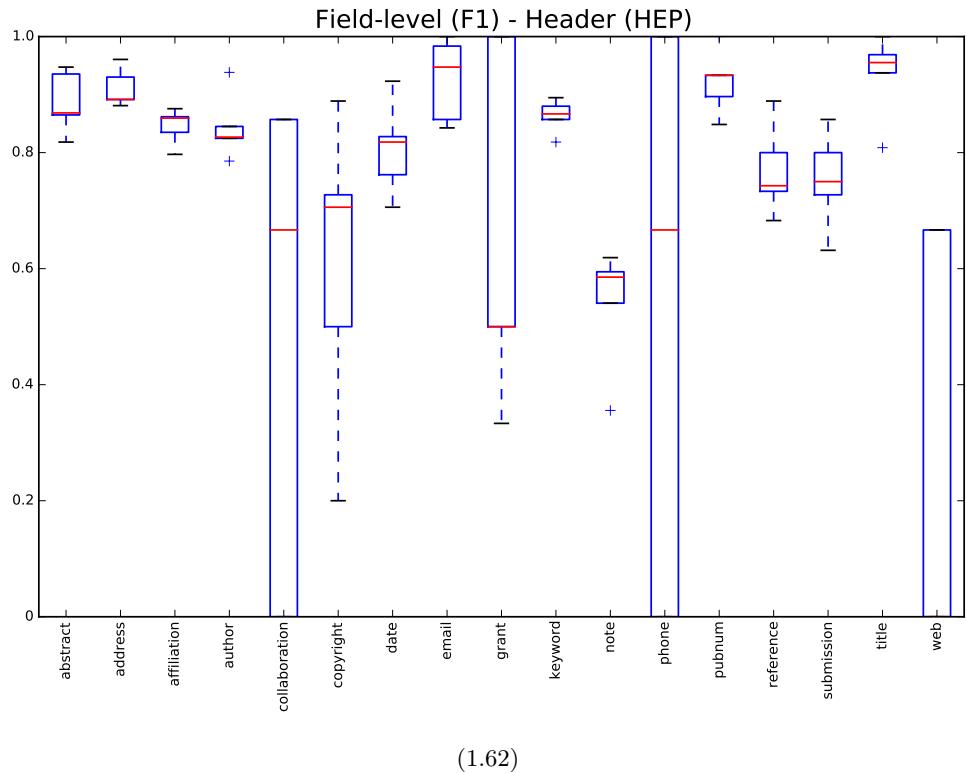
Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	23213			80			3	17	64	17			80	12	23			
address	10	2529	141	110				9		9	18			351		9		
affiliation	24	152	3369	90				4			37			234		1		
author	60	14	78	3741		14		3			24			18	79	5	14	
collaboration		2	3	6	21	1										5		
copyright	185	6	17			215					87			15	3			
date	12	5	6	4			142							11	54			
email	91	19	4	33		3		1803			7			122		3		
grant	42								300		34							
keyword	108	6	19						1339	30				78				
note	209	16	34	12		29	8	186		78	550		7	206	47	51		
phone								10		13	33							
pubnum	11									30		783	3	4		19		
reference	98	21	6	75		25	16	9	13		111			1128		16		
submission	25	23	2			25	16				37				688			
title	103			4							21			71		1814		
web	9		2							35		23		33				

(1.61)

10 Dictionaries

10.1 Header model - HEP dataset



Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web
abstract	0.992		0.03	0.016		0.104	0.01	0.006	0.026	0.056	0.006		0.01	0.002	0.008	0.026	
address	0.532	0.92	0.32	0.22	0.02		0.134	0.13		0.1	0.016			0.024		0.21	
affiliation	0.51	0.262	0.936	0.12			0.046	0.036			0.312			0.02		0.084	
author	0.126	0.216	0.1	0.968		0.074		0.064	0.004		0.174		0.094	0.01	0.044	0.334	
collaboration			0.332	0.2	0.686	0.05					0.2					0.1	
copyright	0.312	0.008	0.02			0.888				0.072	0.53			0.512	0.4		
date	0.37	0.2	0.054				0.93				0.134			0.48	0.36		
email	0.556	0.044	0.026	0.014				0.978			0.058			0.012		0.006	
grant	0.2					0.2			1.0		0.2				0.2	0.2	
keyword	0.53		0.2			0.15				0.954	0.16	0.156		0.362			
note	0.672	0.138	0.186	0.248	0.014	0.214	0.2	0.48		0.546	0.836		0.138	0.466	0.062	0.66	
phone							0.2				0.2	0.6					
pubnum	0.2								0.082		0.272		0.988	0.046		0.2	
reference	0.176	0.33	0.018	0.164		0.136	0.078			0.316	0.394		0.088	0.932		0.354	
submission	0.4	0.2				0.2	0.448				0.244			0.966	0.2		
title	0.046			0.088							0.116			0.048	0.994		
web	0.2										0.612		0.2			0.388	

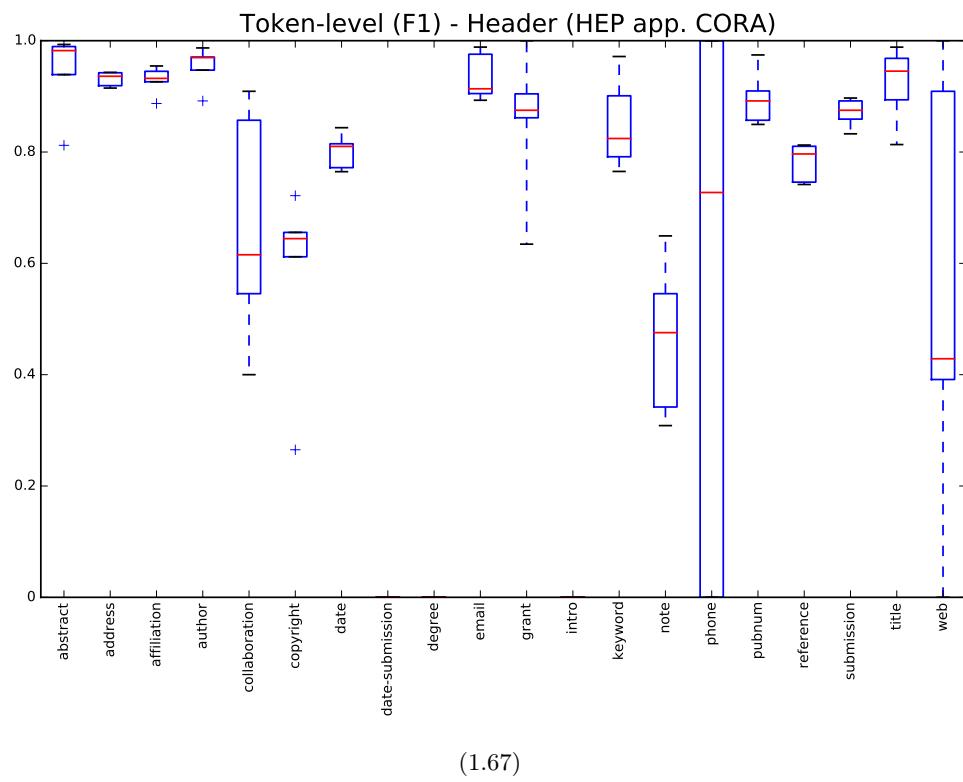
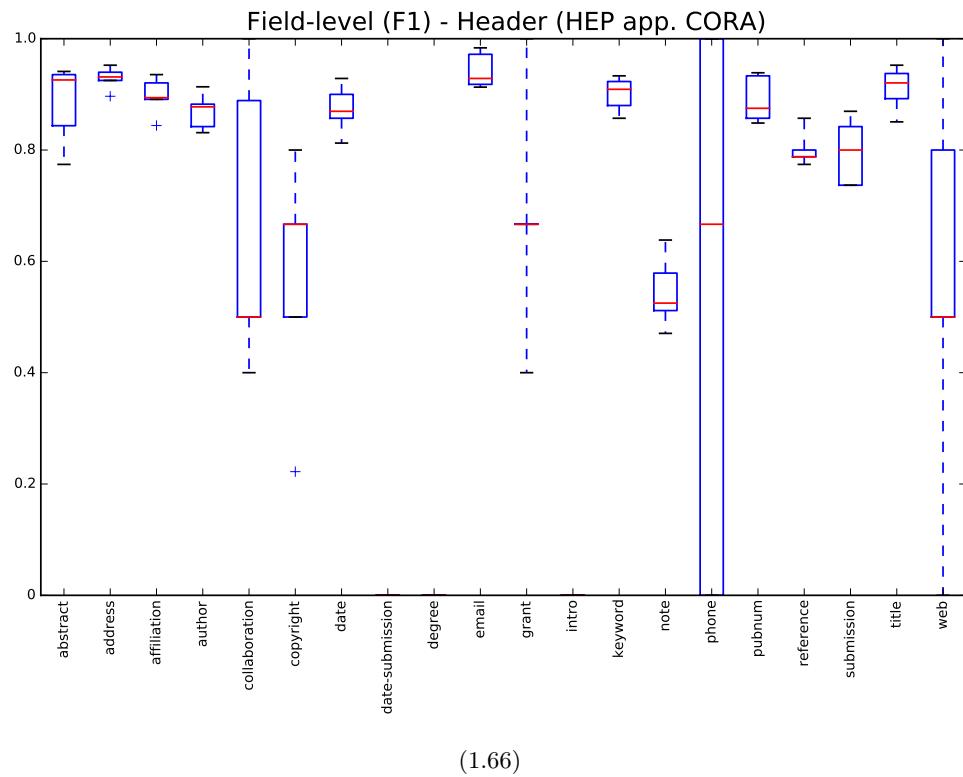
(1.64)

Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web
abstract	23088		58	91		55	12	3	17	64	13		21	19	14	44	
address	32	2714	156	89	1		12	8		9	10			104		51	
affiliation	70	162	3446	70	2		6	15			41			65		34	
author	34	28	60	3801		14		17	6		28		18	19	5	20	
collaboration			4	3	26	1				2					2		
copyright	197	6	15			200			12		61			25	12		
date	12	5	6				161				14			11	25		
email	112	13	3	31				1867			18			26		15	
grant	9					20				303	22			10	12		
keyword	99		3			6				1261	30	80		101			
note	182	16	54	25	1	54	8	179		78	570		4	185	30	47	
phone								10			13	33					
pubnum	11								21		9		787	3		19	
reference	85	67	6	75		14	14			54	81		15	1130		16	
submission	11	16				25	35				69			676	6		
title	36			8							42			37		1890	
web	9										36		23			34	

(1.65)

10.2 Header model - HEP dataset appending CORA dataset



Confusion matrix - Header (HEP app. CORA)

abstract	0.99					0.008		0.002		0.026	0.282	0.054	0.222		0.032		0.01	0.02	
address		0.942	0.43	0.264			0.2			0.198				0.104		0.1			
affiliation	0.01	0.16	0.948	0.1		0.032	0.046			0.054				0.416					
author		0.214	0.094	0.976	0.012	0.074				0.034				0.242		0.094		0.044	
collaboration		0.2	0.066	0.534	0.2	0.976								0.2				0.1	
copyright	0.328			0.014	0.03		0.826				0.08		0.354		0.33	0.05	0.016	0.272	0.2
date							0.974							0.054		0.4	0.34	0.054	
date-submission																			
degree																			
email	0.328	0.012	0.094	0.02						0.968	0.2	0.308		0.16					
grant	0.2							0.04		0.986				0.8					
intro																			
keyword	0.4	0.054	0.064						0.034		0.282	0.95	0.11		0.006	0.408			
note	0.538	0.034	0.276	0.232		0.128	0.008			0.456	0.2	0.234	0.146	0.844		0.35	0.08	0.4	0.626
phone										0.32				0.08	0.6				
pubnum														0.2		0.996	0.064		
reference		0.118	0.018	0.096		0.114	0.138				0.2	0.354	0.548		0.188	0.91		0.2	
submission	0.2					0.088	0.296				0.018		0.304			0.954		0.146	
title			0.2	0.16			0.042				0.062		0.432		0.016		0.984		
web														0.2		0.2		0.8	

(1.68)

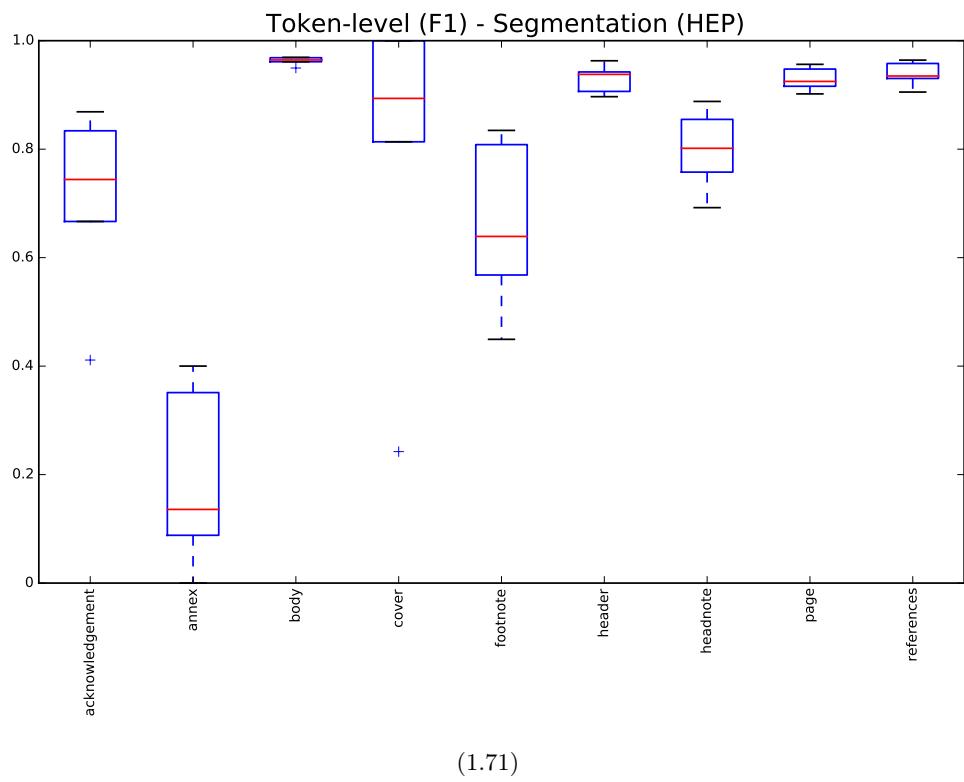
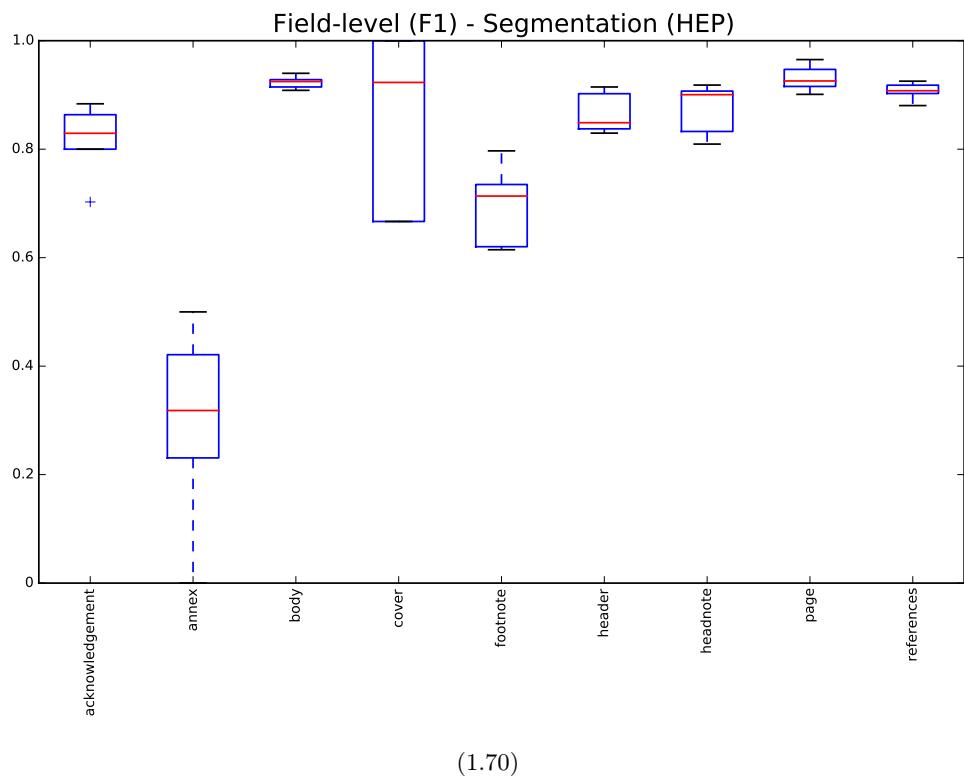
Confusion matrix - Header (HEP app. CORA)

This treemap visualization illustrates the distribution of document types across different categories. The categories are listed on the left, and the document types are represented by colored rectangles within each category. The size of each rectangle corresponds to the count of documents in that specific category.

Category	Document Type	Count
abstract	abstract	20863
	address	2937
affiliation	affiliation	118
	author	12
collaboration	author	2
	collaboration	130
copyright	affiliation	58
	copyright	3551
date	affiliation	2
	date	1
degree	date	6
	degree	1
email	degree	29
	email	17
grant	email	5
	grant	8
intro	grant	7
	intro	1
keyword	intro	6
	keyword	40
note	keyword	9
	note	6
phone	note	7
	phone	45
pubnum	phone	25
	pubnum	64
reference	pubnum	6
	reference	17
submission	reference	32
	submission	4
title	submission	21
	title	11
web	title	18
	web	3
abstract	11	
address	2	
affiliation	6	
author	35	
collaboration	2	
copyright	258	
date	202	
degree	17	
email	18	
grant	1946	
intro	12	
keyword	1320	
note	22	
phone	42	
pubnum	1	
reference	155	
submission	34	
title	72	
web	1	
abstract	8	
address	20	
affiliation	8	
author	148	
collaboration	125	
copyright	19	
date	4	
degree	3	
email	9	
grant	1	
intro	9	
keyword	15	
note	3	
phone	14	
pubnum	5	
reference	6	
submission	20	
title	12	
web	60	

(1.69)

10.3 Segmentation model - HEP dataset



Confusion matrix - Segmentation (HEP)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	0.97	0.17	0.918				0.034		0.204
annex	0.406	0.55	0.686			0.2	0.002		0.306
body	0.008	0.074	0.984		0.012	0.014			0.026
cover				0.98	0.072	0.11	0.014		
footnote	0.05		0.604		0.792	0.454	0.452	0.25	0.324
header		0.054	0.292	0.274	0.084	0.92	0.02	0.024	0.192
headnote		0.054	0.384		0.394	0.22	0.826	0.126	0.108
page	0.068	0.008	0.266		0.28	0.076	0.182	0.928	0.102
references	0.196		0.304		0.022	0.136	0.02	0.01	0.974

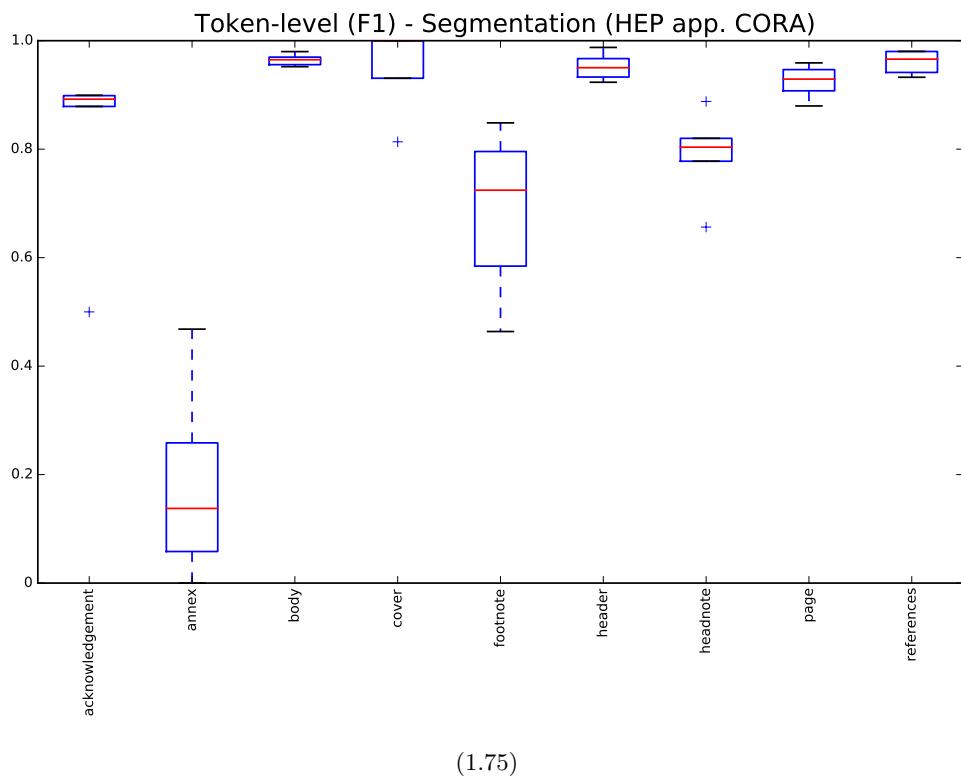
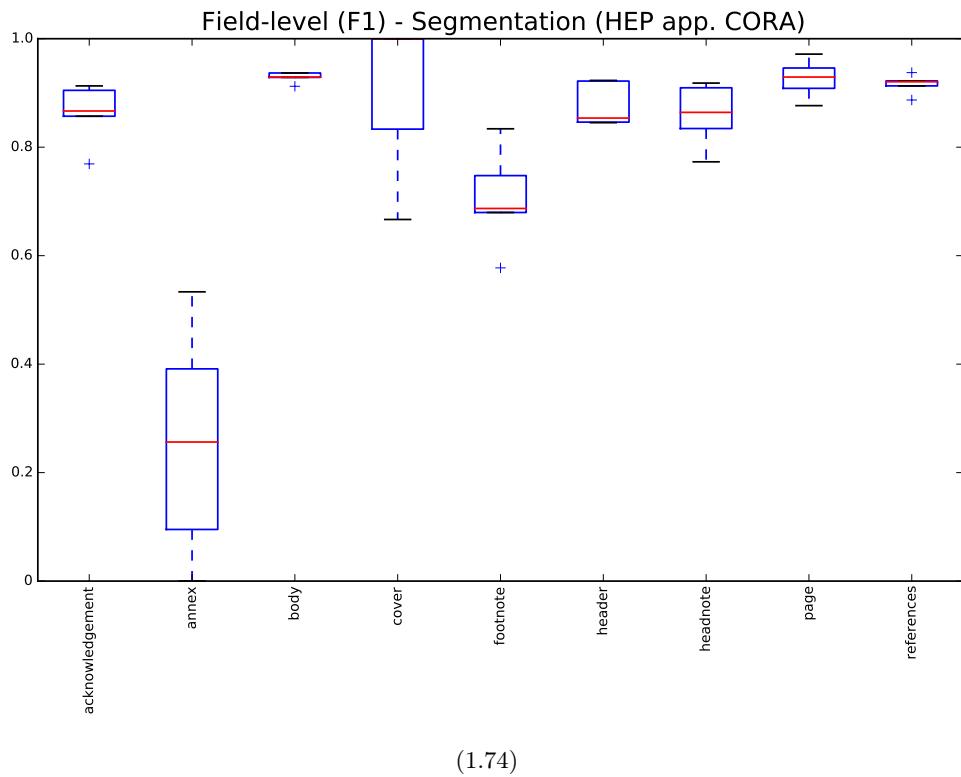
(1.72)

Confusion matrix - Segmentation (HEP)

	acknowledgement	12	258				1		18
acknowledgement	529								
annex	13	1631	9236				3	2	2
body	65	1771	186441		132	356	45	37	503
cover				269	5	16	1		
footnote	6		511		1159	39	49	32	22
header		3	975	70	78	12078	10	6	28
headnote		9	378		62	72	1386	47	35
page	2	2	136		21	12	17	2260	14
references	71		522		10	73	19	4	10837

(1.73)

10.4 Segmentation model - HEP dataset appending CORA dataset



Confusion matrix - Segmentation (HEP app. CORA)

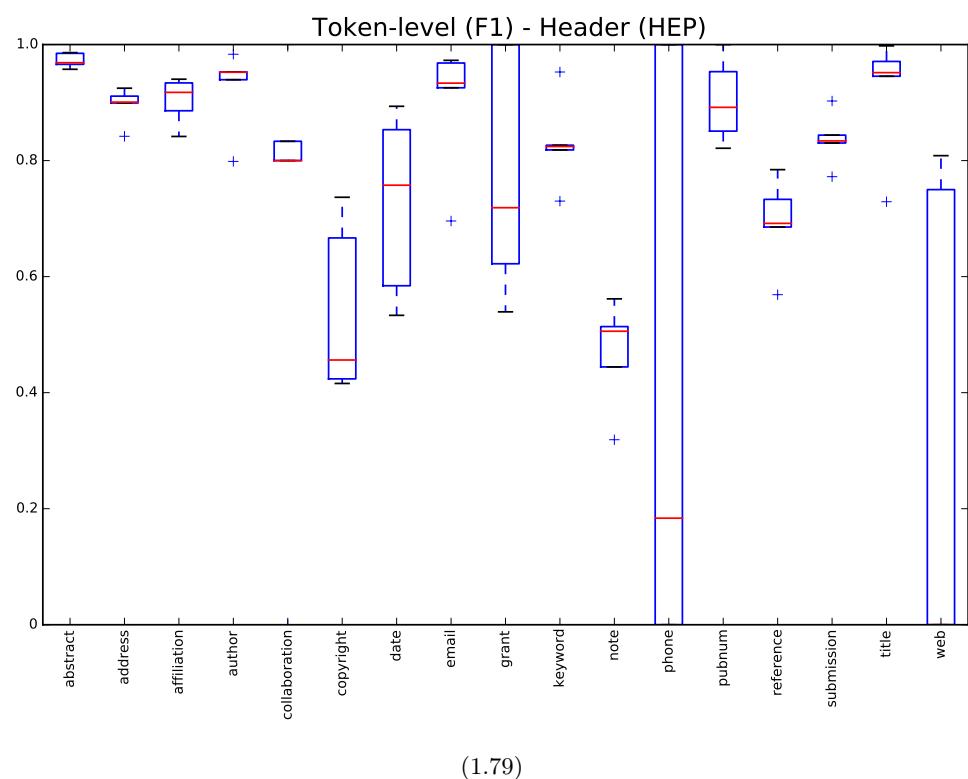
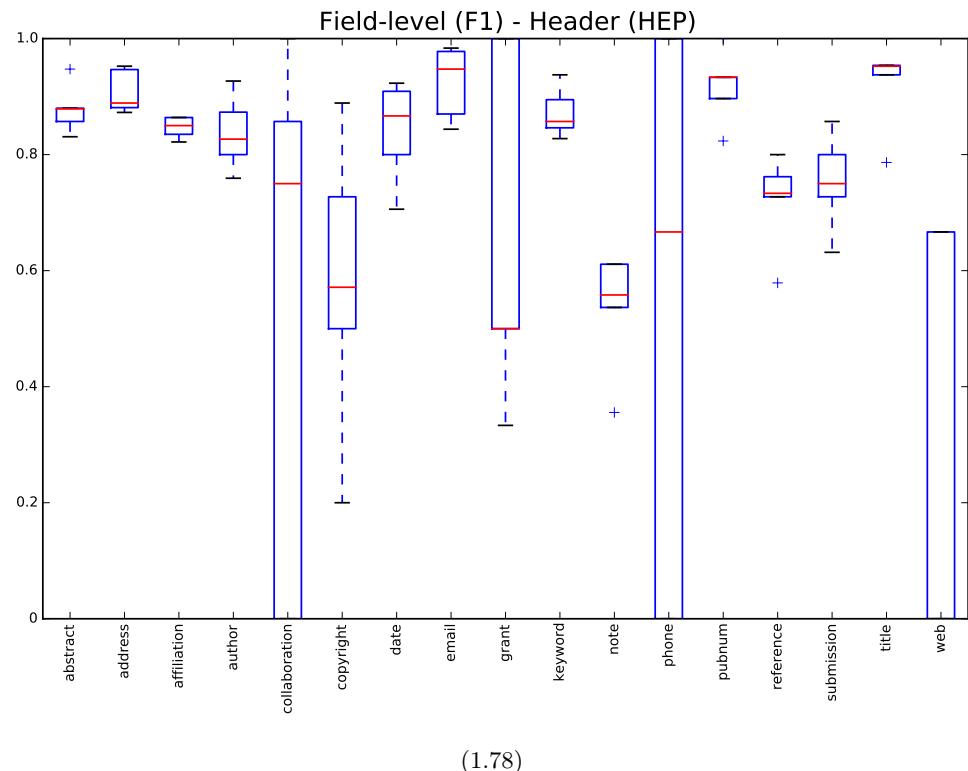
	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	0.972	0.342	0.904			0.17			
annex	0.406	0.368	0.644			0.4	0.004		0.258
body	0.004	0.086	0.986		0.016	0.01			0.024
cover				0.98	0.072	0.31	0.014		
footnote	0.032	0.038	0.552		0.814	0.272	0.394	0.18	0.296
header		0.068	0.316	0.018	0.07	0.932	0.048	0.036	0.164
headnote		0.032	0.358		0.292	0.242	0.828	0.104	0.156
page	0.068	0.066	0.28		0.194	0.096	0.094	0.934	0.34
references		0.228	0.25		0.026	0.048	0.012	0.014	0.986
	(1.76)								

Confusion matrix - Segmentation (HEP app. CORA)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	580	22	199			17			
annex	13	1765	9179			36	2	2	16
body	20	2354	186193		209	191	40	41	302
cover				259	5	26	1		
footnote	4	10	447		1228	37	24	37	31
header		13	523	1	48	12571	12	8	72
headnote		4	428		58	62	1334	46	57
page	2	5	129		29	13	9	2259	18
references		15	338		15	7	16	4	11141
	(1.77)								

11 Dictionaries + stop words

11.1 Header model - HEP dataset



Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web
abstract	0.992		0.03	0.016		0.008	0.014	0.006	0.026	0.048	0.006		0.01	0.002	0.008	0.026	
address	0.532	0.916	0.336	0.218	0.02			0.142		0.1	0.016			0.122		0.208	
affiliation	0.522	0.244	0.934	0.124				0.038			0.222			0.058		0.08	
author	0.136	0.242	0.1	0.968		0.074		0.064	0.004		0.248		0.094	0.006	0.044	0.4	
collaboration			0.4	0.2	0.664	0.05										0.1	
copyright	0.512	0.116	0.112	0.2		0.888			0.072		0.314			0.328	0.4		
date	0.34	0.2	0.054				0.942				0.14			0.48	0.36		
email	0.556	0.054	0.026	0.01				0.978			0.05			0.004		0.006	
grant	0.2					0.2			1.0		0.2				0.2	0.2	
keyword	0.33	0.054	0.346				0.004		0.954	0.16	0.156			0.156			
note	0.61	0.138	0.184	0.236		0.214	0.2	0.48		0.546	0.81		0.148	0.418	0.062	0.63	
phone							0.2				0.2	0.6					
pubnum	0.2							0.082		0.272		0.988	0.046			0.2	
reference	0.458	0.234	0.018	0.37		0.014	0.078			0.516	0.716		0.288	0.912		0.238	
submission	0.6	0.2				0.2	0.448	0.042			0.198			0.966	0.2		
title	0.092			0.062						0.226			0.048		0.994		
web	0.2									0.612		0.2			0.388		

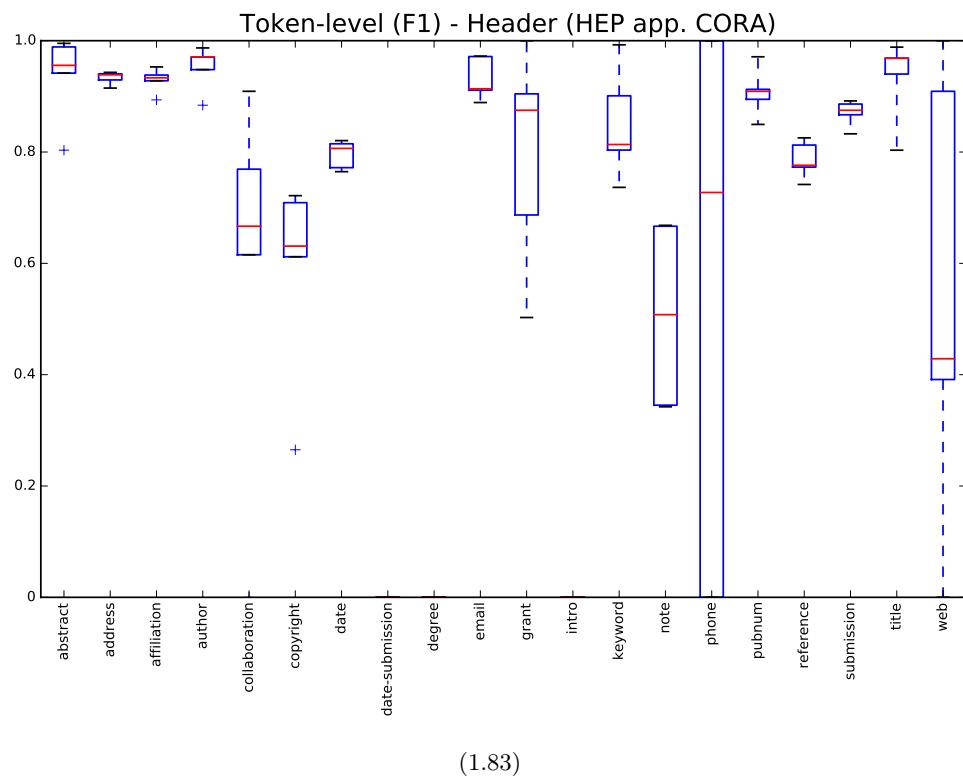
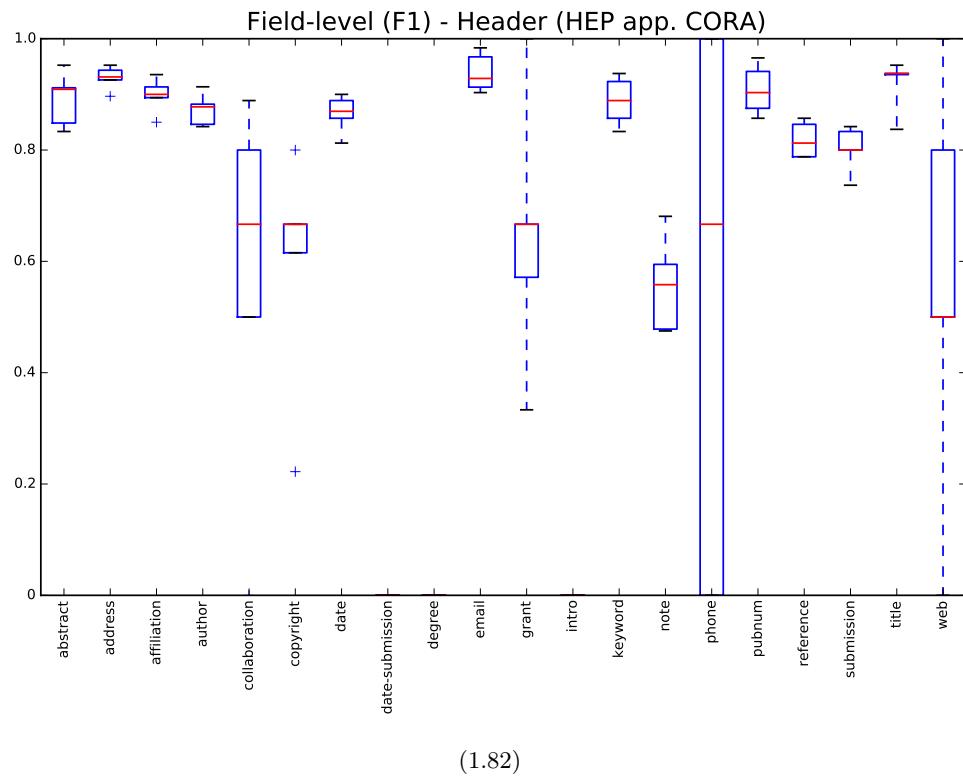
(1.80)

Confusion matrix - Header (HEP)

	abstract	address	affiliation	author	collaboration	copyright	date	email	grant	keyword	note	phone	pubnum	reference	submission	title	web
abstract	23168		51	91		11	14	3	17	53	13		21	14	15	38	
address	32	2773	163	75	1			10		9	10			71		42	
affiliation	82	169	3469	73				16		30				46		26	
author	55	32	74	3782		14		14	6	21		18	13	5	16		
collaboration			5	3	27	1									2		
copyright	175	13	22	3		213			12	31			47	12			
date	10	5	6				170			7			11	25			
email	112	15	3	23				1893			18			9		12	
grant	9					20				22				10	12		
keyword	87	6	19					1		1280	30	80		77			
note	230	16	53	27		54	8	186		78	547		12	141	30	51	
phone								10		13	33						
pubnum	11								21	9		787	3			19	
reference	145	45	6	83		3	7			99	101		24	1035		9	
submission	32	11				25	33	9			46			676	6		
title	72			12						26			37		1866		
web	9									36		23				34	

(1.81)

11.2 Header model - HEP dataset appending CORA dataset



Confusion matrix - Header (HEP app. CORA)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	degree	email	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web	
abstract	0.98																				
address		0.944	0.422	0.264			0.008	0.002			0.168			0.026	0.456	0.062	0.136	0.024		0.01	
affiliation	-0.076	0.162	0.946	0.1		0.032	0.038			0.054			0.402								
author		0.216	0.092	0.978		0.074				0.034		0.008	0.242		0.094		0.044				
collaboration		0.066	0.534	0.2	0.776								0.2						0.1		
copyright	-0.314		0.014	0.03		0.836					0.12	0.162	0.434		0.318	0.05	0.016	0.272	0.2	0.006	
date							0.968							0.054		0.4	0.412	0.054			
date-submission																					
degree																					
email	0.328	0.012	0.14	0.024							0.966		0.6		0.218						
grant	0.2								0.04		0.986		0.8								
intro																					
keyword	0.4	0.054	0.064							0.034		0.282	0.95	0.114		0.398					
note	0.36	0.234	0.276	0.16		0.128	0.008				0.456	0.2	0.192	0.15	0.854		0.306	0.178	0.4	0.666	0.104
phone											0.32			0.08	0.6						
pubnum												0.2		0.112	0.254			0.008	0.018		0.2
reference		0.118	0.018	0.096		0.094	0.138				0.002		0.2	0.354	0.594		0.21	0.928		0.104	
submission						0.288	0.296					0.2		0.094	0.428			0.016	0.956		0.058
title							0.042											0.016	0.984		
web																		0.2		0.8	

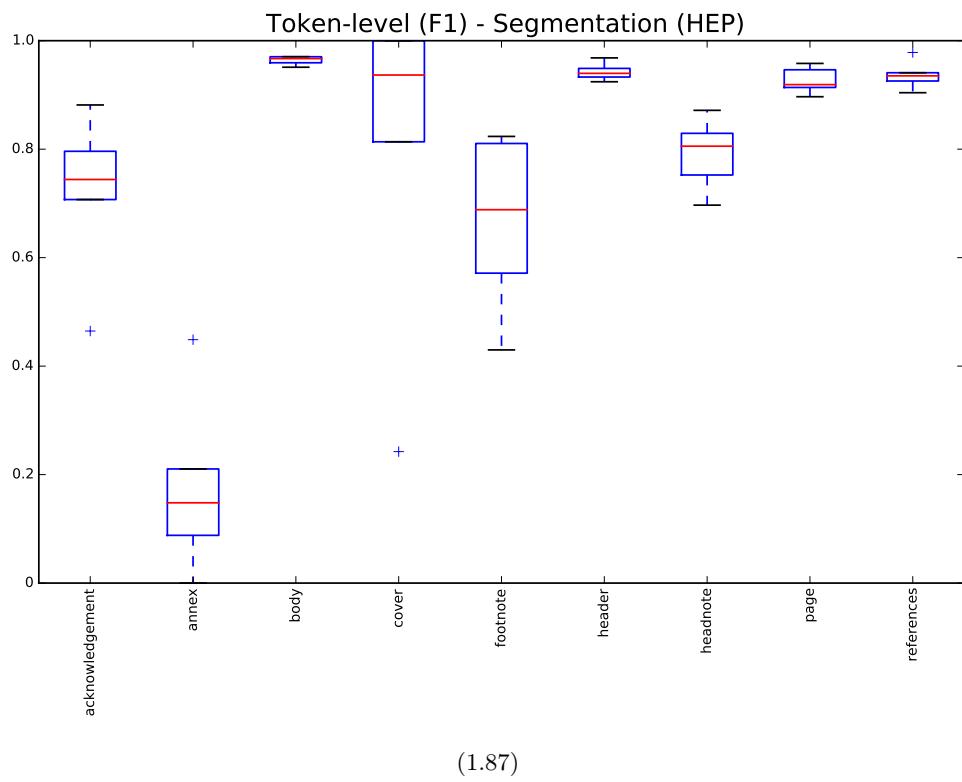
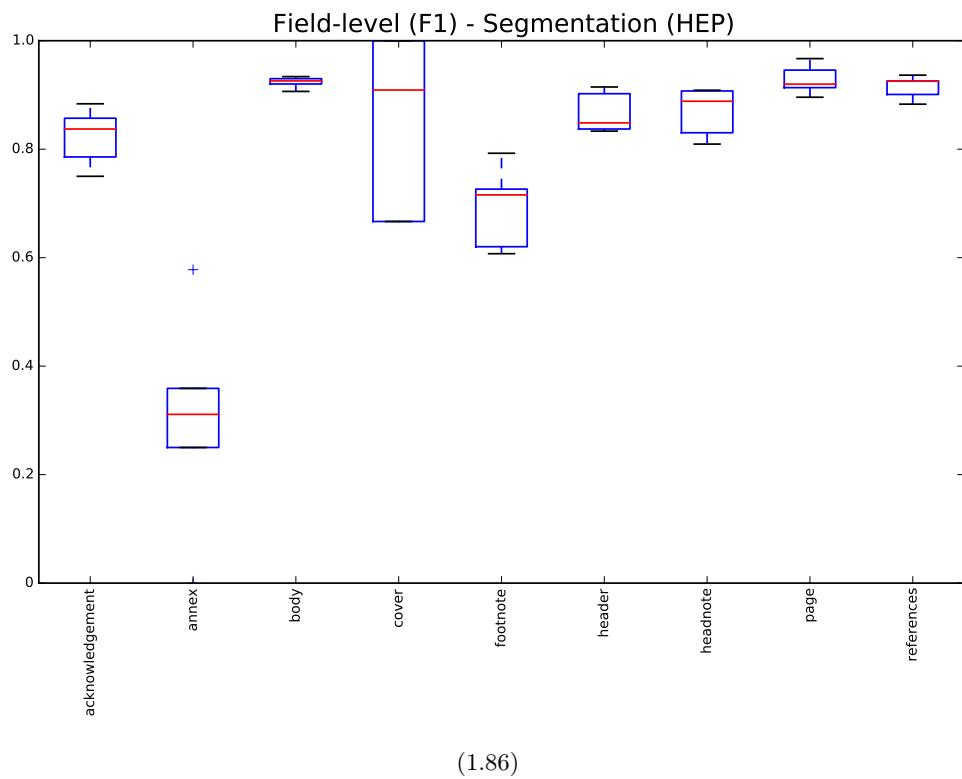
(1.84)

Confusion matrix - Header (HEP app. CORA)

	abstract	address	affiliation	author	collaboration	copyright	date	date-submission	degree	email	grant	intro	keyword	note	phone	pubnum	reference	submission	title	web
abstract	2051					11	7			17	2296	52	581		12	6	12	3		
address		2967	103	12			2			18			75		9					
affiliation	-21	130	3546	58	2	6	5			8			135							
author	32	28	3818		14					2	14		118		19		5			
collaboration		1	7	3	23								2					2		
copyright	59		10	5		276				3	38	41		63	9	1	15	3		5
date							198						3			9	18	6		
date-submission																				
degree																				
email	17	11	16	8						193		37		63						
grant	8								6		294			68						
intro																				
keyword	36	6	7							17		12	1309	23			170			
note	17	6	45	21		39	6			18	16	216	43	852		39	36	24	36	19
phone										18			4	34						
pubnum														325	1			24		
reference		17	6	32		13	18			1	18	83	169		25	1164		11		
submission						26	35			4	14		38			712			9	
title				17		3					64		117			12		1800		
web														23		19		60		

(1.85)

11.3 Segmentation model - HEP dataset



Confusion matrix - Segmentation (HEP)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	0.956	0.37	0.892				0.034		0.1
annex	0.406	0.576	0.67			0.2	0.002		0.266
body	0.01	0.06	0.984		0.012	0.014			0.026
cover				0.98	0.072	0.31	0.014		
footnote	0.04	0.03	0.63		0.784	0.406	0.452	0.244	0.334
header		0.098	0.308	0.084	0.072	0.922	0.02	0.024	0.05
headnote		0.092	0.386		0.374	0.218	0.828	0.108	0.14
page	0.068	0.066	0.284		0.228	0.088	0.182	0.934	0.11
references	0.136	0.232	0.302		0.022	0.136	0.02	0.01	0.964

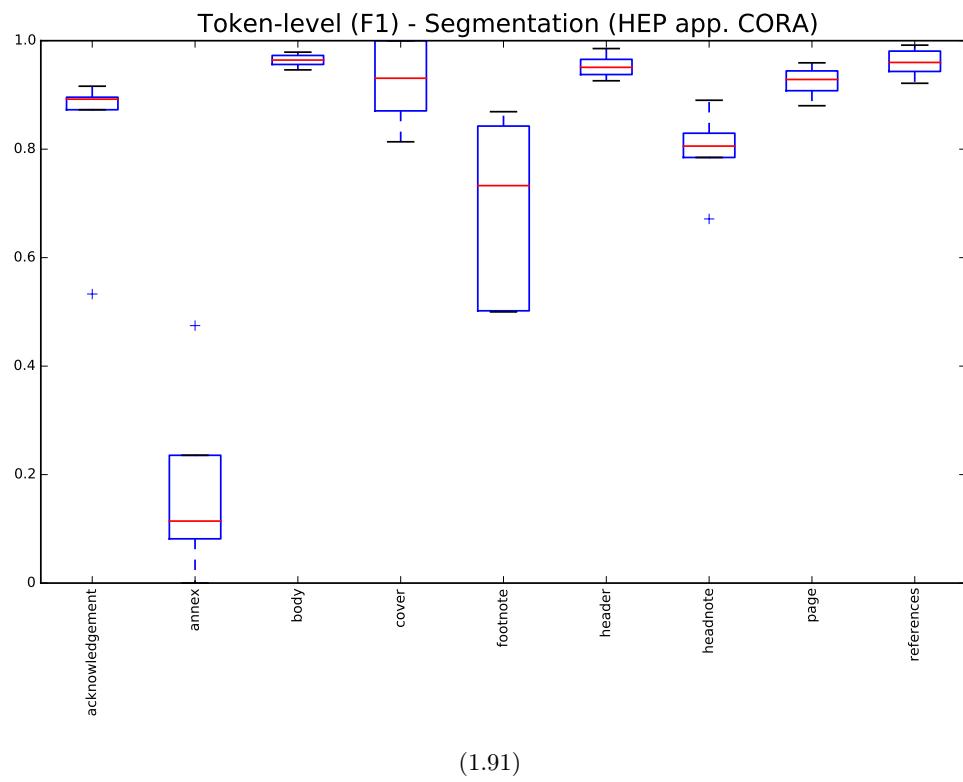
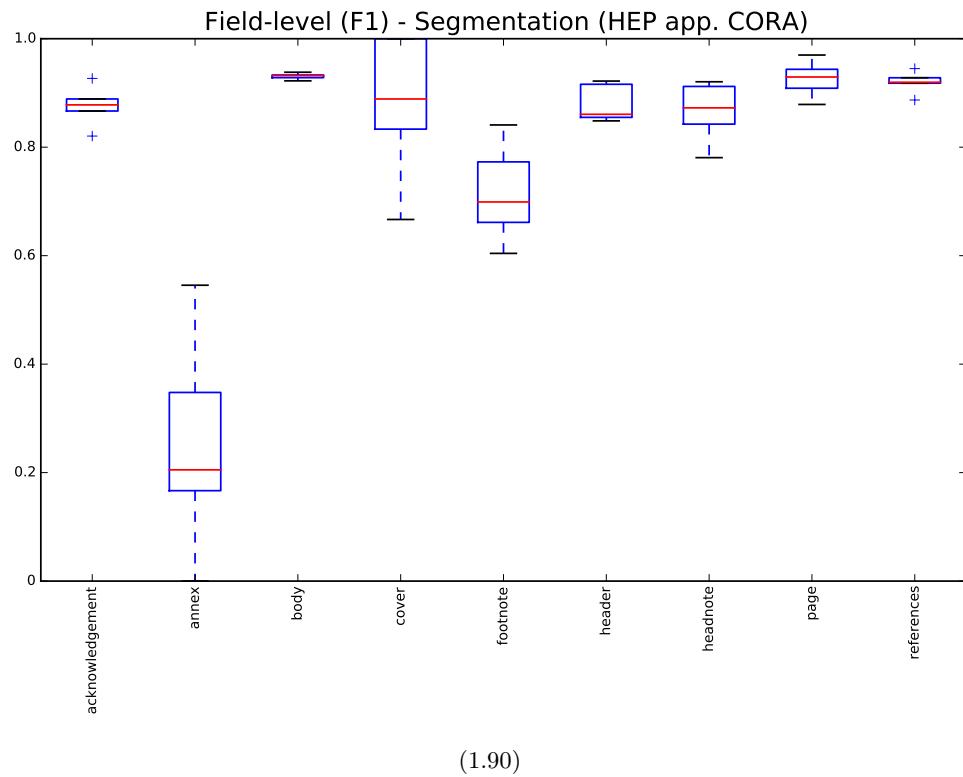
(1.88)

Confusion matrix - Segmentation (HEP)

	acknowledgement	18	264				1		3
acknowledgement	532								
annex	13	1430	9315				3	2	248
body	71	1474	186744		147	313	51	40	510
cover				259	5	26	1		
footnote	5	3	511		1165	32	49	33	20
header		8	848	50	37	12259	17	6	23
headnote		12	391		58	69	1380	46	33
page	2	5	141		21	13	17	2253	12
references	54	65	463		10	73	19	4	10848

(1.89)

11.4 Segmentation model - HEP dataset appending CORA dataset



Confusion matrix - Segmentation (HEP app. CORA)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	0.972	0.028	0.924			0.312			
annex	0.206	0.4	0.622			0.4	0.004		0.314
body	0.002	0.096	0.986		0.014	0.008			0.024
cover				0.98	0.072	0.51	0.014		
footnote	0.032		0.596		0.816	0.272	0.386	0.23	0.312
header			0.324	0.018	0.084	0.932	0.048	0.036	0.152
headnote		0.04	0.352		0.27	0.242	0.828	0.108	0.132
page	0.068	0.048	0.3		0.204	0.108	0.142	0.934	0.118
references		0.028	0.304		0.022	0.048	0.014	0.014	0.99

(1.92)

Confusion matrix - Segmentation (HEP app. CORA)

	acknowledgement	annex	body	cover	footnote	header	headnote	page	references
acknowledgement	578	2	211			27			
annex	8	1591	9298			36	2	2	76
body	11	2746	185961		172	135	37	40	248
cover				237	5	48	1		
footnote	4		446		1245	37	21	39	26
header			535	1	46	12563	12	9	82
headnote		6	418		47	63	1351	48	56
page	2	3	131		27	14	10	2260	17
references		7	348		10	7	17	4	11143

(1.93)