Results

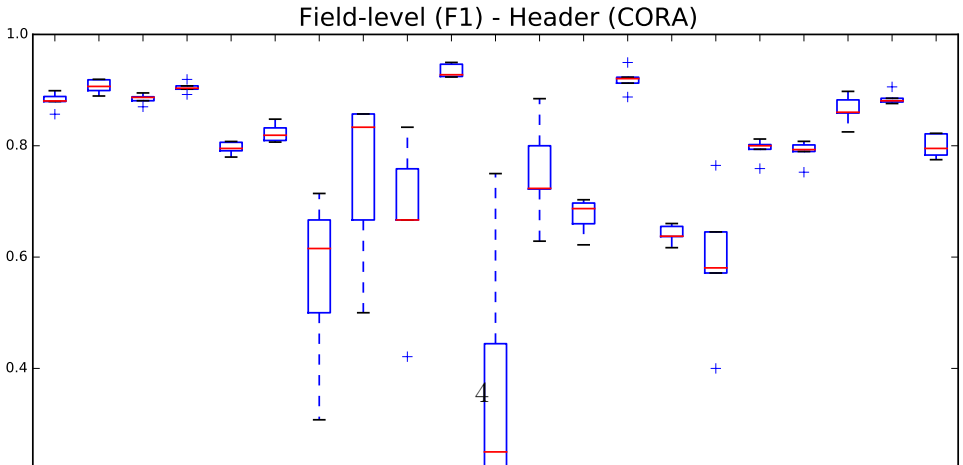
Joseph Boyd

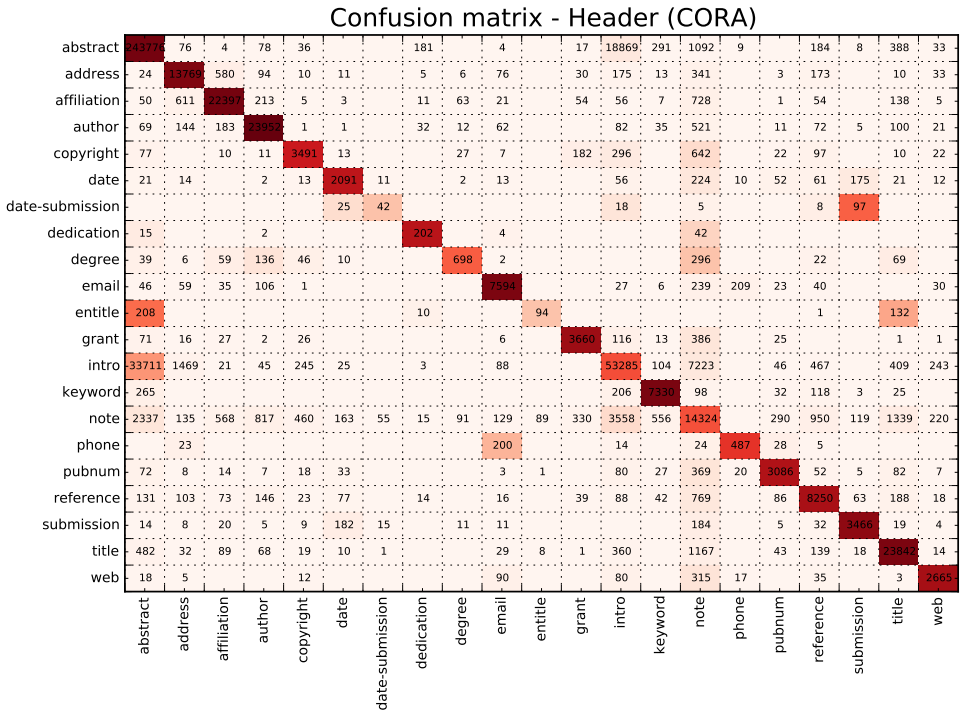
June 22, 2015

Contents

1 Introduction

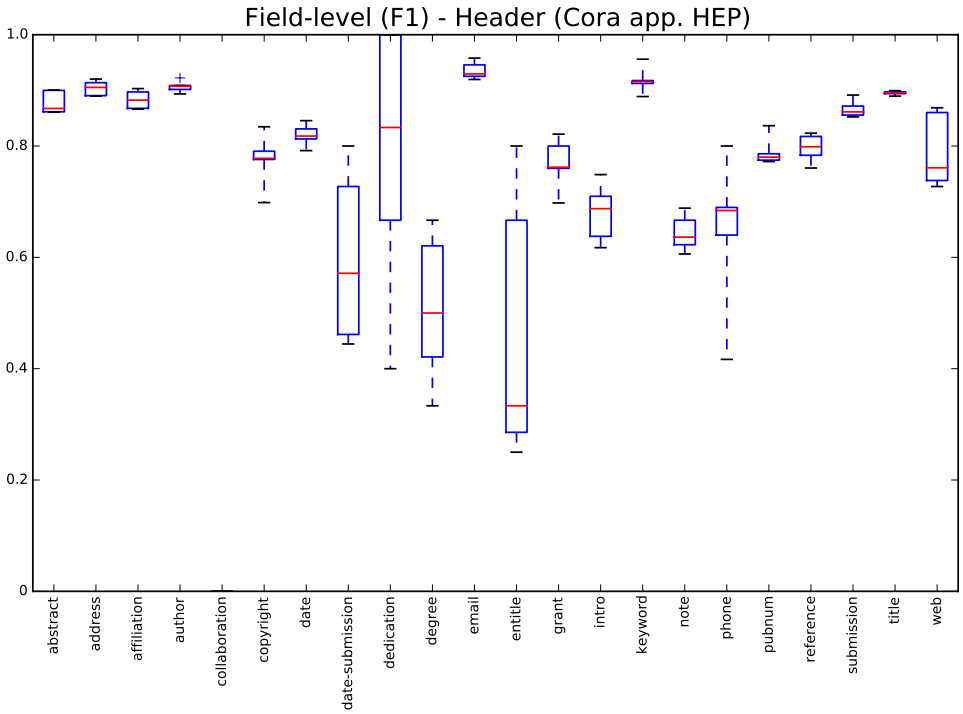
$$\max\{a, b\}$$



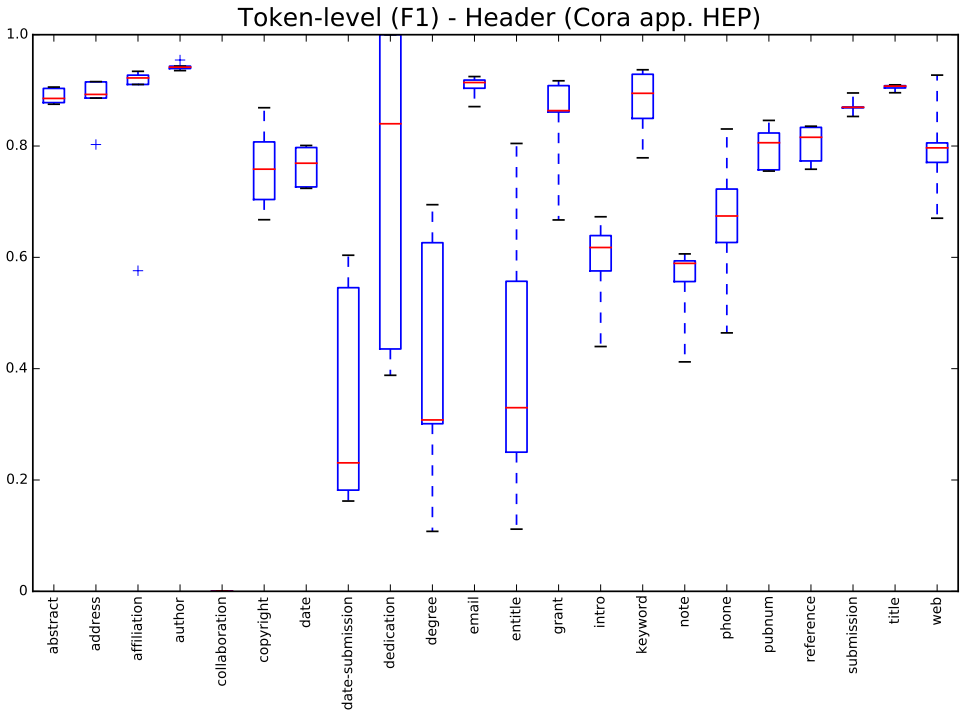

Confusion matrix - Header (CORA)
abstract 24377 76 4 78 36 181 4 17 18869 291 1092 9 184 8 388 33
address 24 13769 580 94 10 11 5 6 76 30 175 13 341 3 173 10 33
affiliation 50 611 22397 213 5 3 11 63 21 54 56 7 728 1 54 138 5
author 69 144 183 3992 1 1 32 12 62 82 35 521 11 72 5 100 21
copyright 77 10 11 3491 13 27 7 182 296 642 22 97 10 22
date 21 14 2 13 2091 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 7594 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 63 188 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 23842 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)

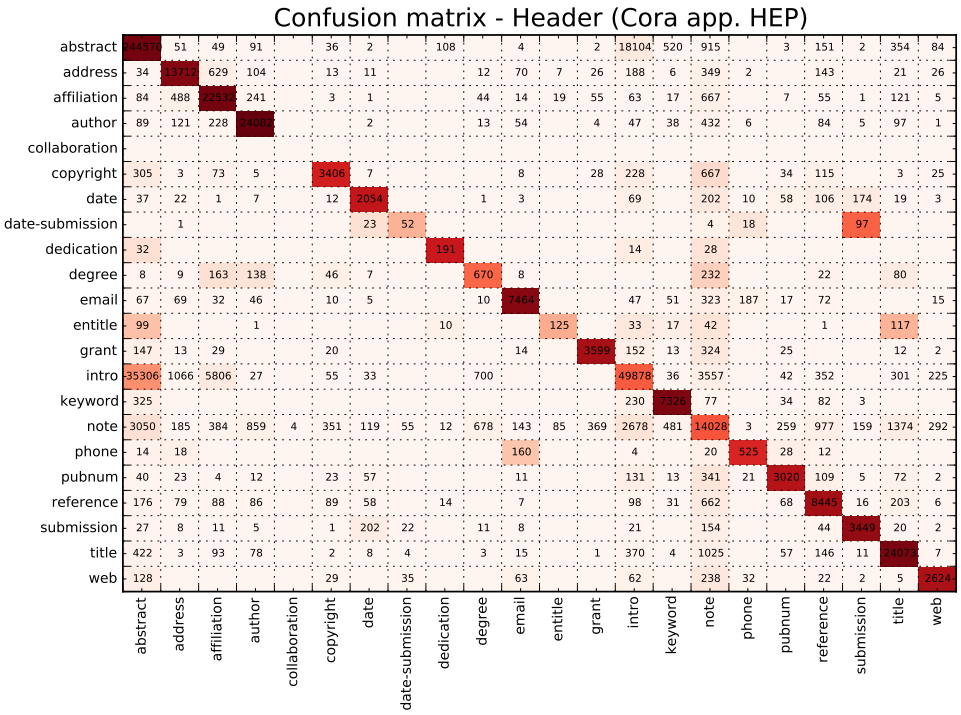
8.2 Header model - Cora dataset appending HEP dataset



(1.4)

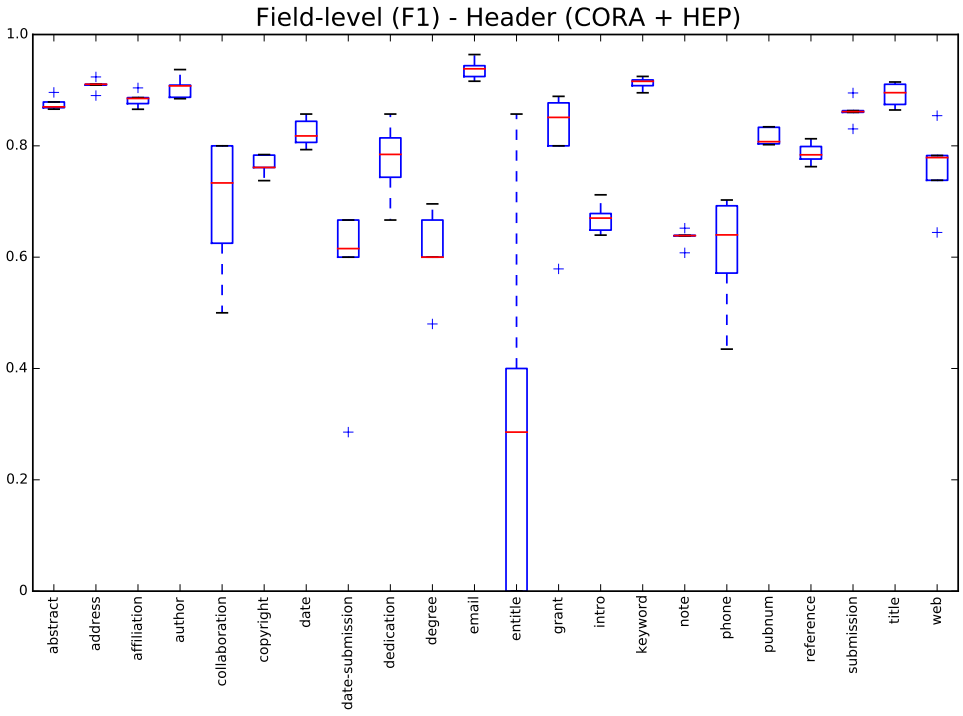


(1.5)

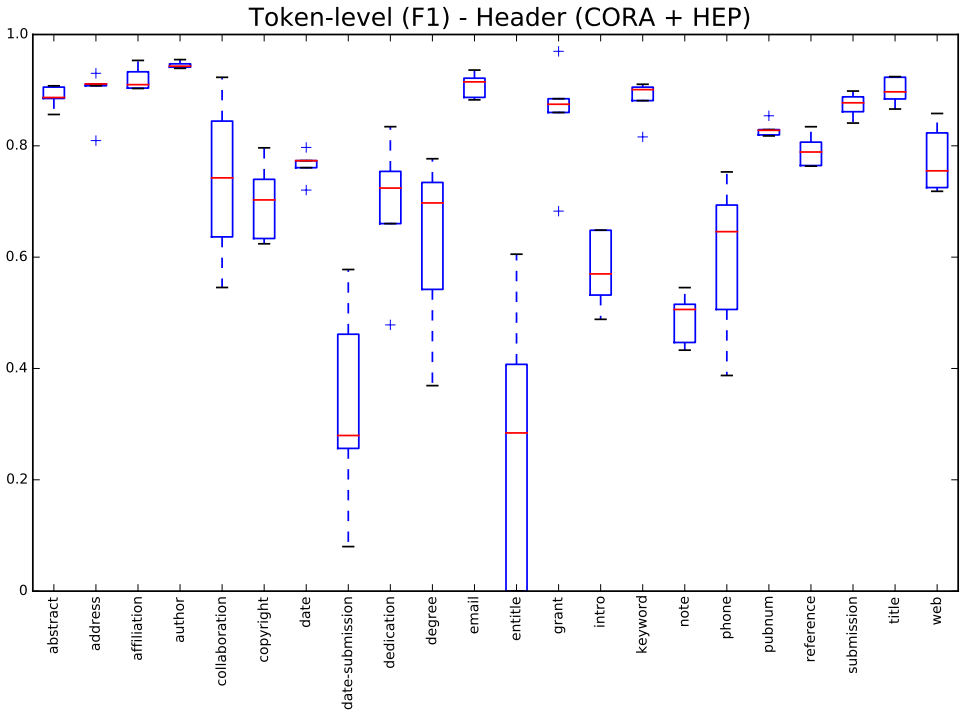

Confusion matrix - Header (Cora app. HEP)
abstract 24457 51 49 91 36 2 108 4 2 18104 520 915 3 151 2 354 84
address 34 13712 629 104 13 11 12 70 7 26 188 6 349 2 143 21 26
affiliation 84 488 22534 241 3 1 44 14 19 55 63 17 667 7 55 1 121 5
author 89 121 228 24082 2 13 54 4 47 38 432 6 84 5 97 1
collaboration
copyright 305 3 73 5 3406 7 8 28 228 667 34 115 3 25
date 37 22 1 7 12 2054 1 3 69 202 10 58 106 174 19 3
date-submission 1 23 52 4 18 97
dedication 32 191 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 3599 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 292
phone 14 18 160 4 20 525 28 12
pubnum 40 23 4 12 23 57 11 131 13 341 21 3020 109 5 72 2
reference 176 79 88 86 89 58 14 7 98 31 662 68 8445 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 15 370 4 1025 57 146 11 24073 7
web 128 29 35 63 62 238 32 22 2 5 2624
abstract address affiliation author collaboration copyright date date-submission dedication degree email entitle grant intro keyword note phone pubnum reference submission title web

(1.6)

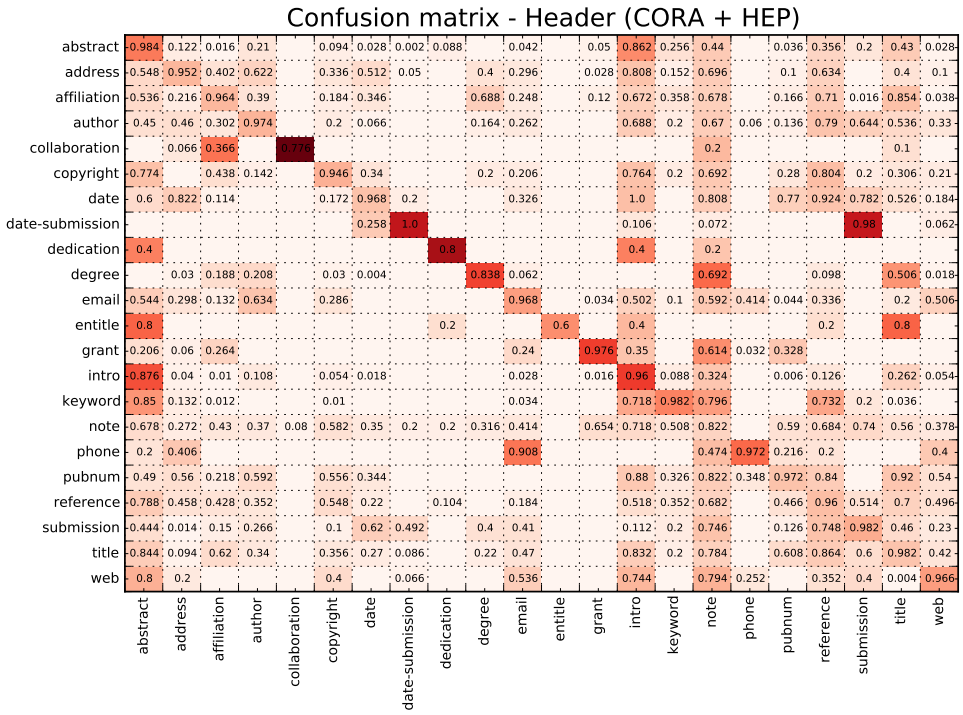
8.3 Header model - Cora and HEP combined datasets



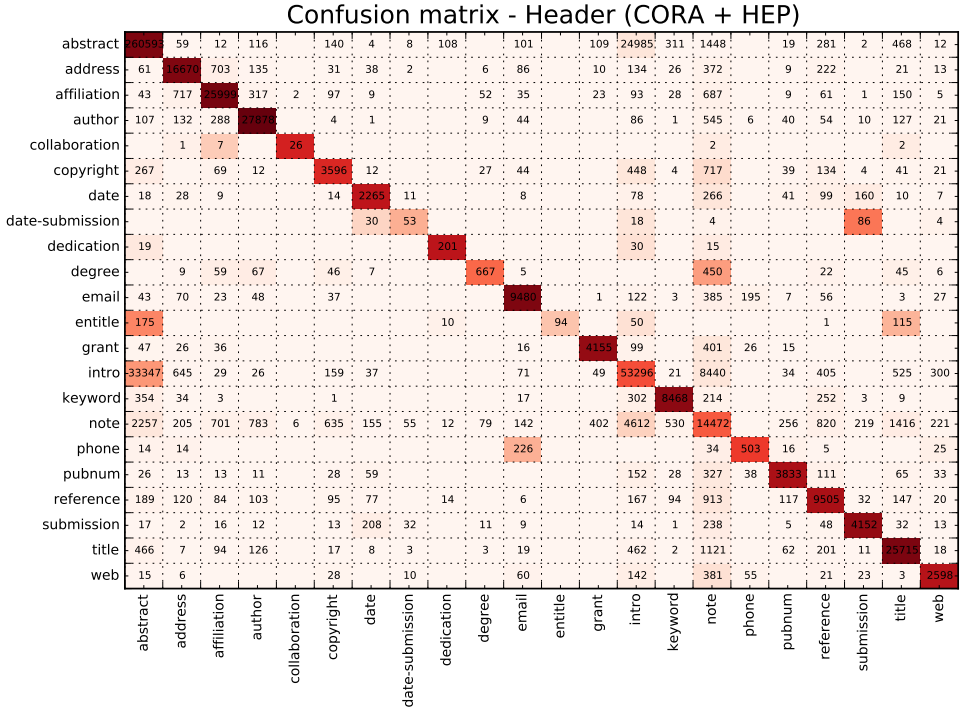
(1.7)



(1.8)

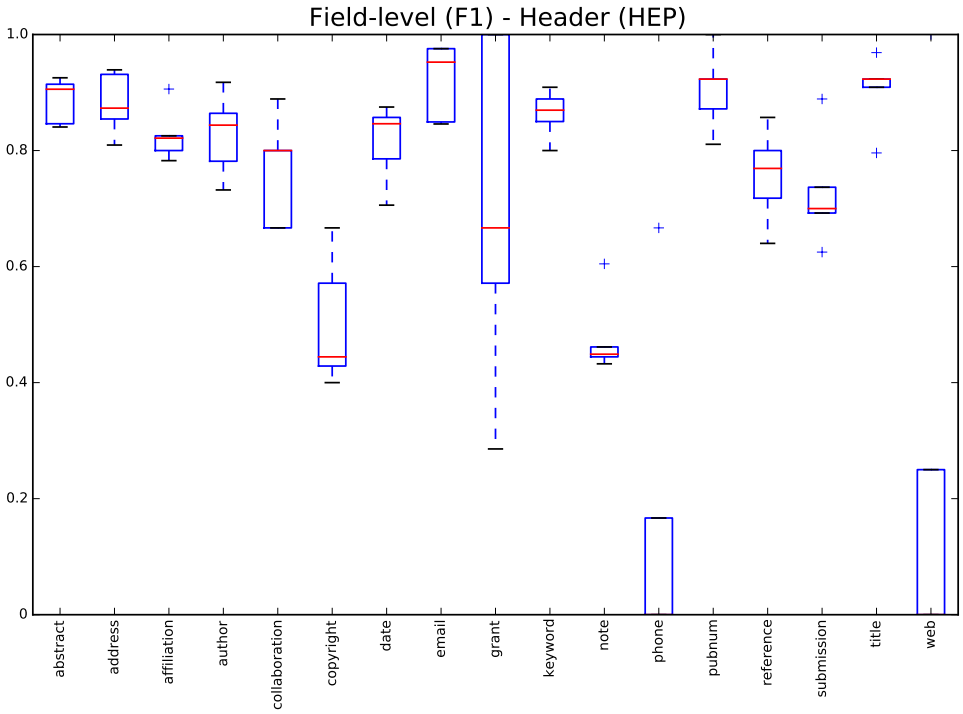

Confusion matrix - Header (CORA + HEP)
abstract address affiliation author collaboration copyright date date-submission dedication degree email entitle grant intro keyword note phone pubnum reference submission title web

(1.9)

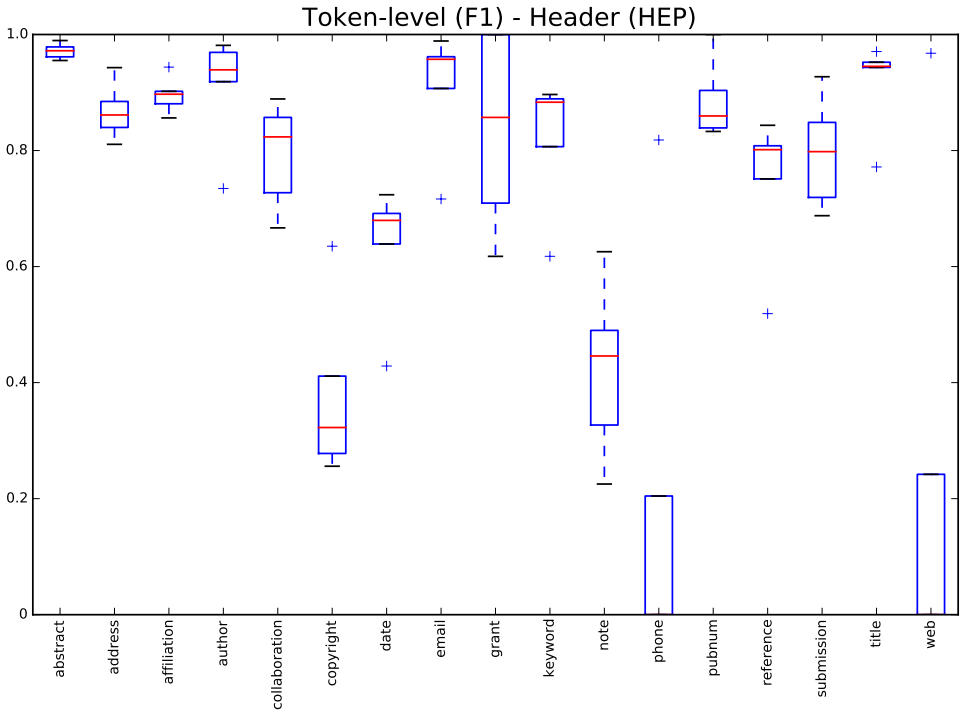

Confusion matrix - Header (CORA + HEP)
abstract address affiliation author collaboration copyright date date-submission dedication degree email entitle grant intro keyword note phone pubnum reference submission title web

(1.10)

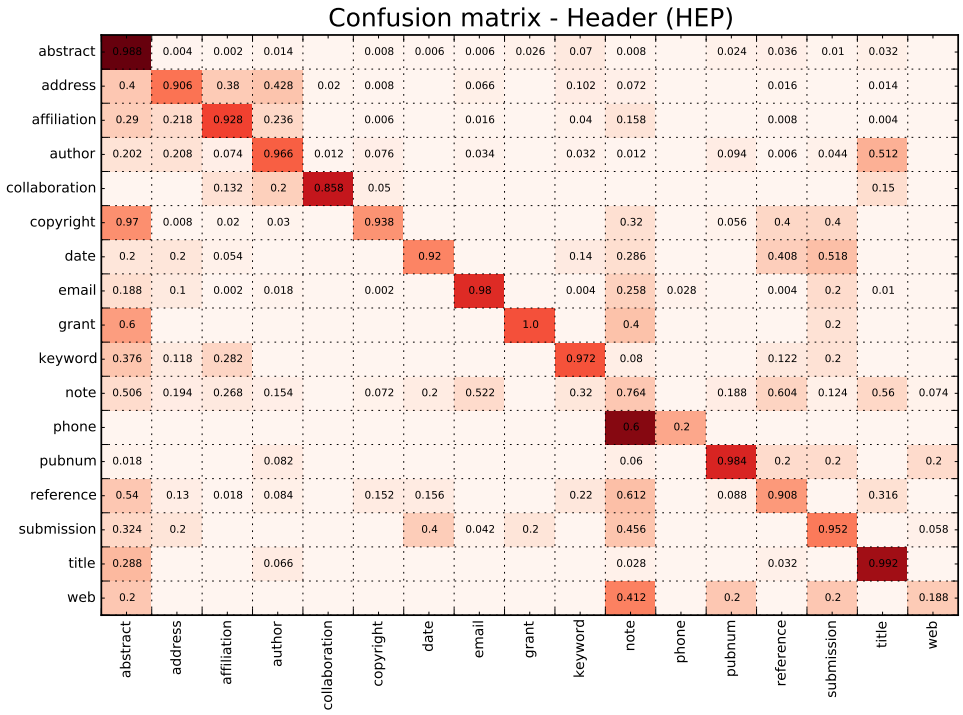
8.4 Header model - HEP dataset



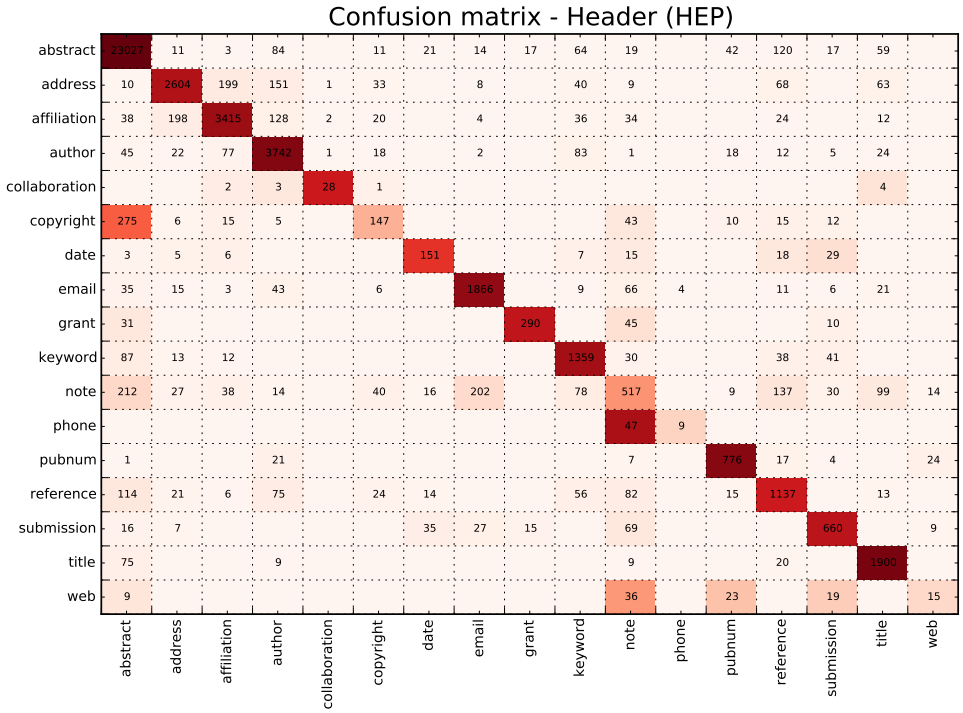
(1.11)



(1.12)

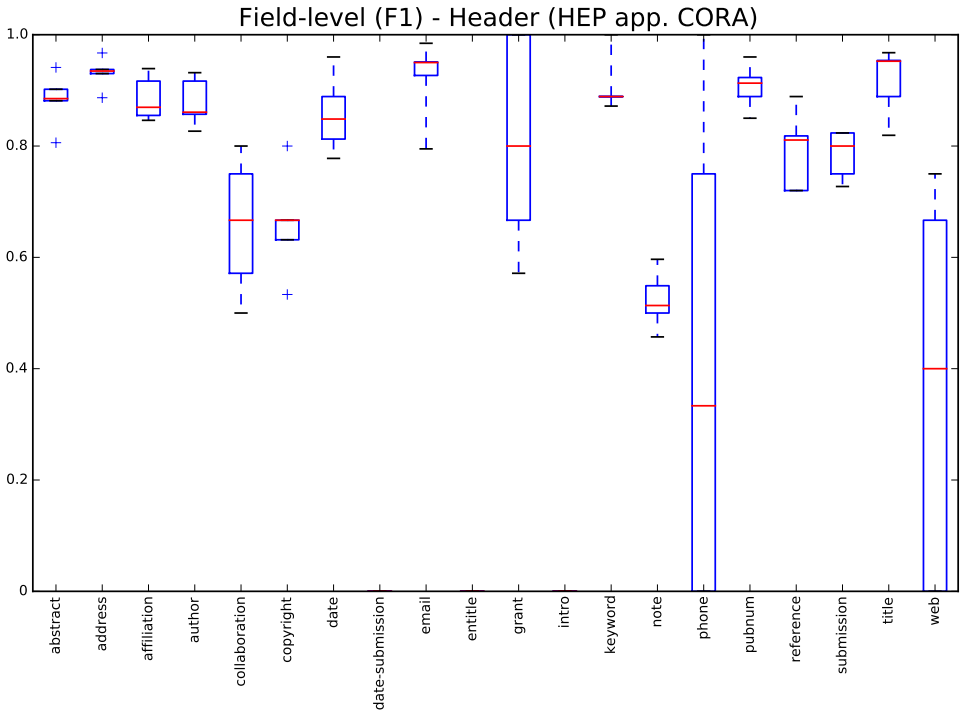

Confusion matrix - Header (HEP)
abstract 0.988 0.004 0.002 0.014 0.008 0.006 0.006 0.026 0.07 0.008 0.024 0.036 0.01 0.032
address 0.4 0.906 0.38 0.428 0.02 0.008 0.066 0.102 0.072 0.016 0.014
affiliation 0.29 0.218 0.928 0.236 0.006 0.016 0.04 0.158 0.008 0.004
author 0.202 0.208 0.074 0.966 0.012 0.076 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)

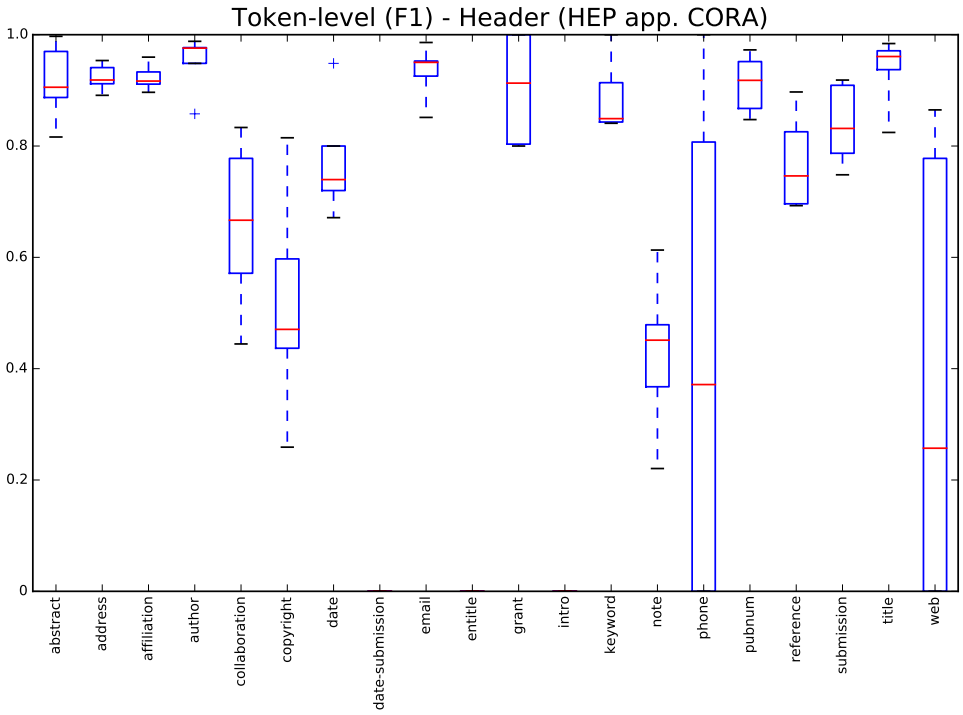

Confusion matrix - Header (HEP)
abstract 23027 11 3 84 11 21 14 17 64 19 42 120 17 59
address 10 2604 199 151 1 33 8 40 9 68 63
affiliation 38 198 3415 128 2 20 4 36 34 24 12
author 45 22 77 3742 1 18 2 83 1 18 12 5 24
collaboration 2 3 28 1 4
copyright 275 6 15 5 147 43 10 15 12
date 3 5 6 151 7 15 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)

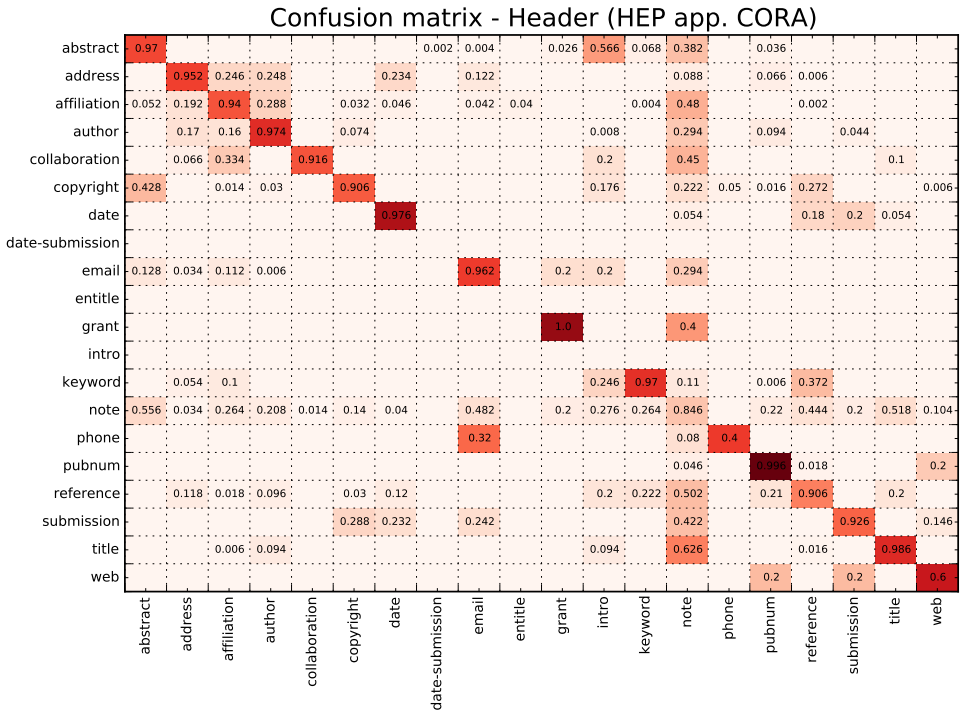
8.5 Header model - HEP dataset appending CORA dataset



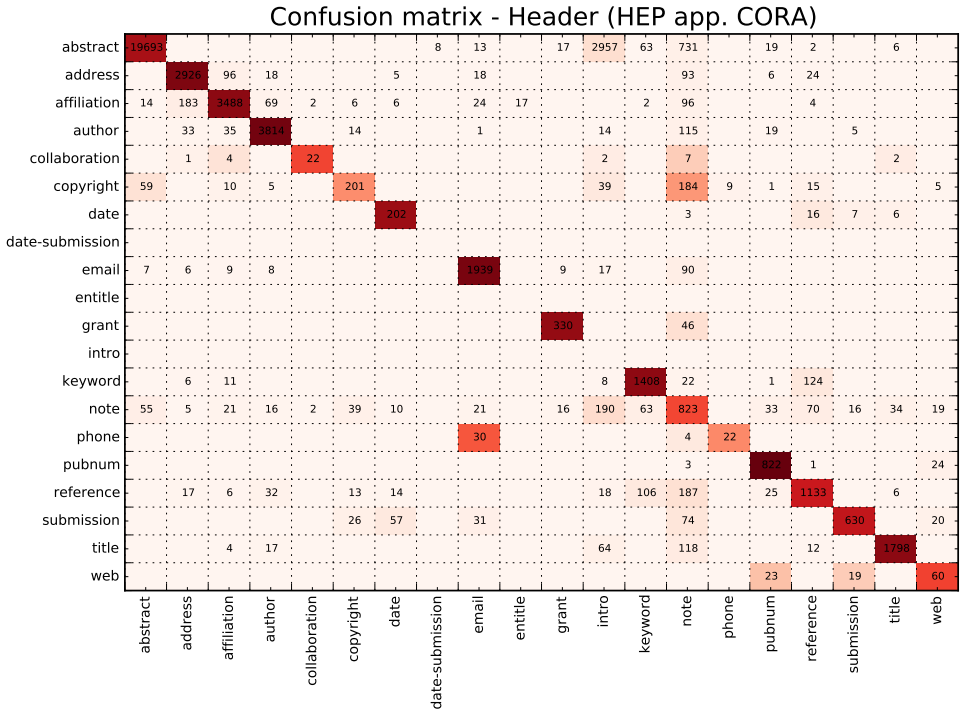
(1.15)



(1.16)

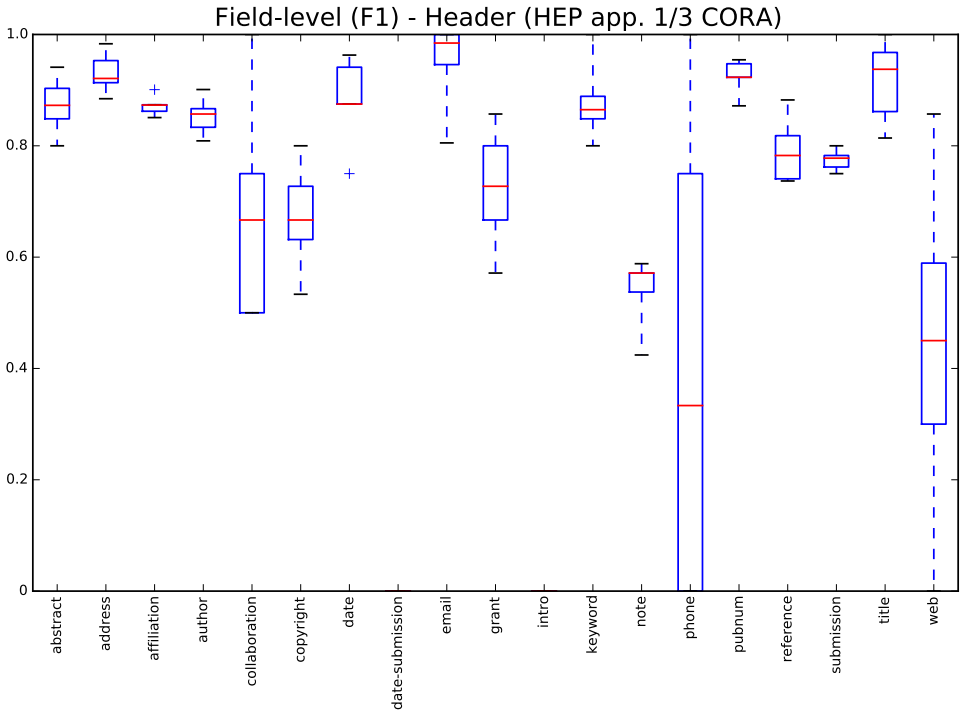

Confusion matrix - Header (HEP app. CORA)
abstract
address
affiliation
author
collaboration
copyright
date
date-submission
email
entitle
grant
intro
keyword
note
phone
pubnum
reference
submission
title
web

(1.17)

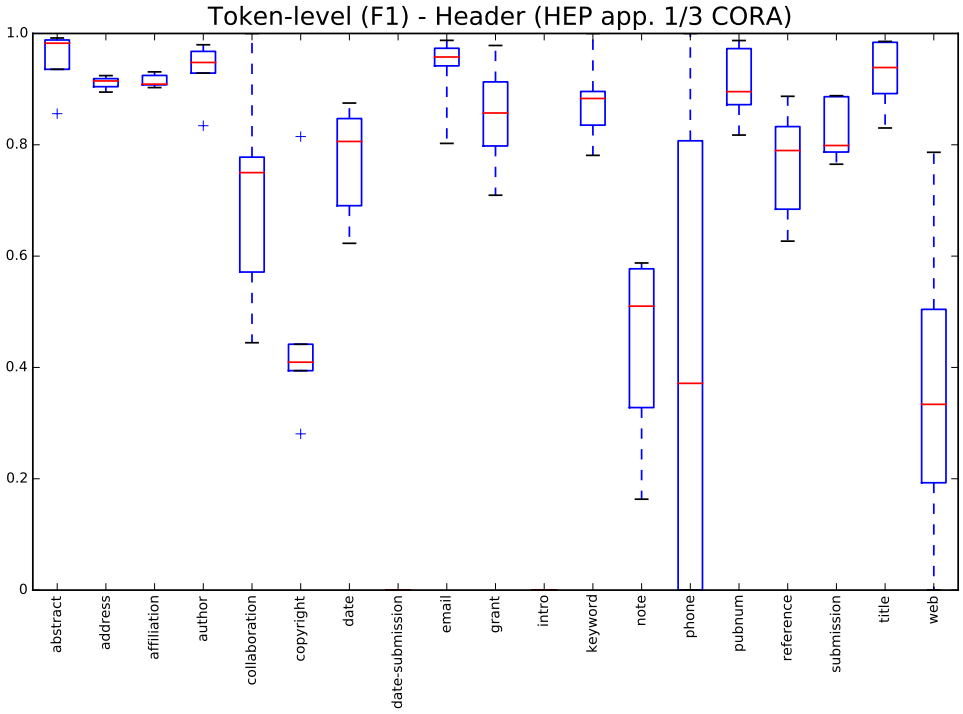

Confusion matrix - Header (HEP app. CORA)
abstract
address
affiliation
author
collaboration
copyright
date
date-submission
email
entitle
grant
intro
keyword
note
phone
pubnum
reference
submission
title
web

(1.18)

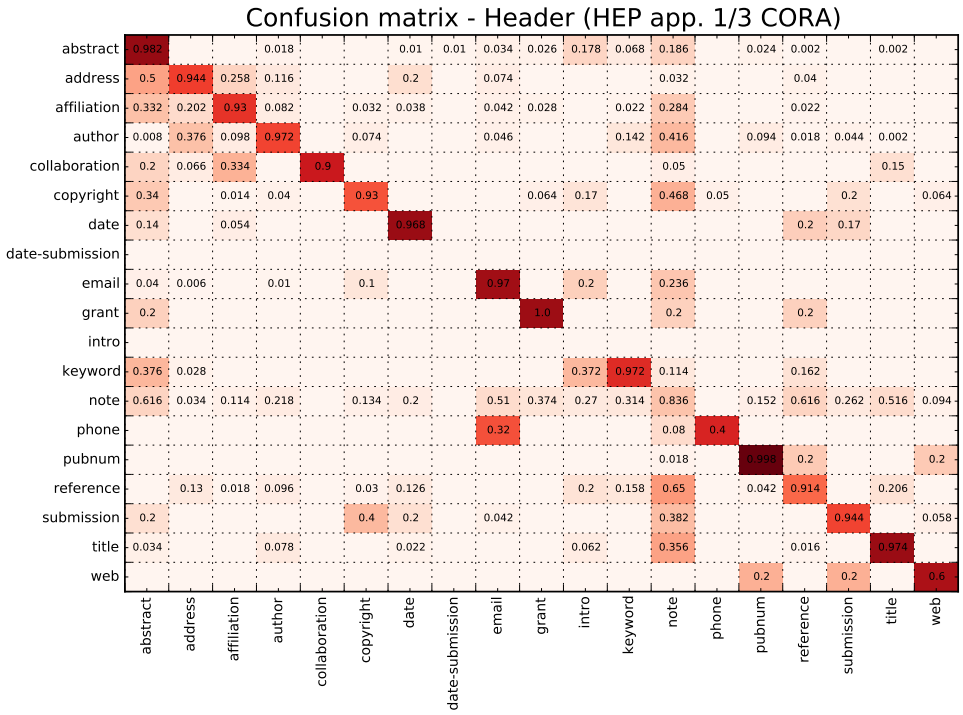
8.6 Header model - HEP dataset appending 1/3 CORA dataset



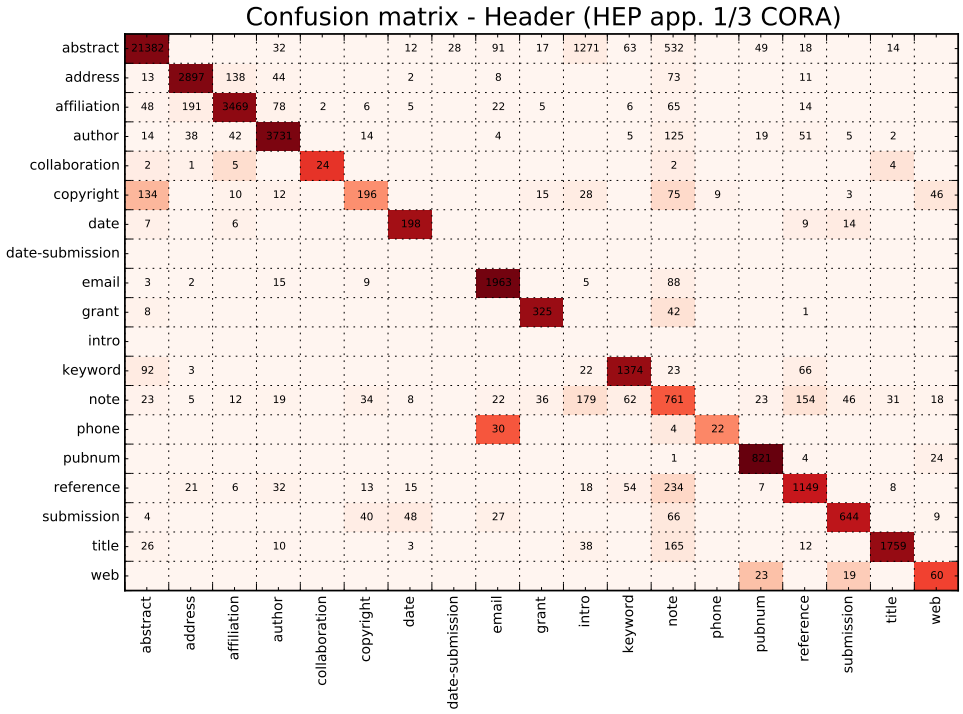
(1.19)



(1.20)

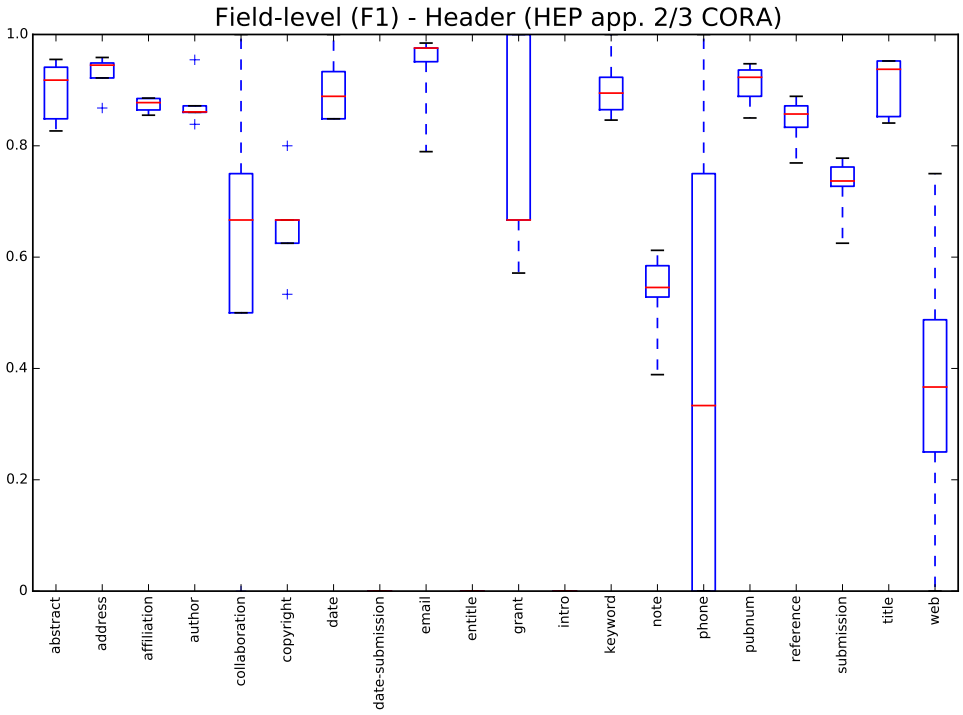

Confusion matrix - Header (HEP app. 1/3 CORA)
abstract 0.982 0.018 0.01 0.01 0.034 0.026 0.178 0.068 0.186 0.024 0.002 0.002
address 0.5 0.944 0.258 0.116 0.2 0.074 0.032 0.04
affiliation 0.332 0.202 0.93 0.082 0.032 0.038 0.042 0.028 0.022 0.284 0.022
author 0.008 0.376 0.098 0.972 0.074 0.046 0.142 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 0.2 0.064
date 0.14 0.054 0.968 0.2 0.17
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 0.2 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
web 0.2 0.2 0.6
abstract address affiliation author collaboration copyright date date-submission email grant intro keyword note phone pubnum reference submission title web

(1.21)

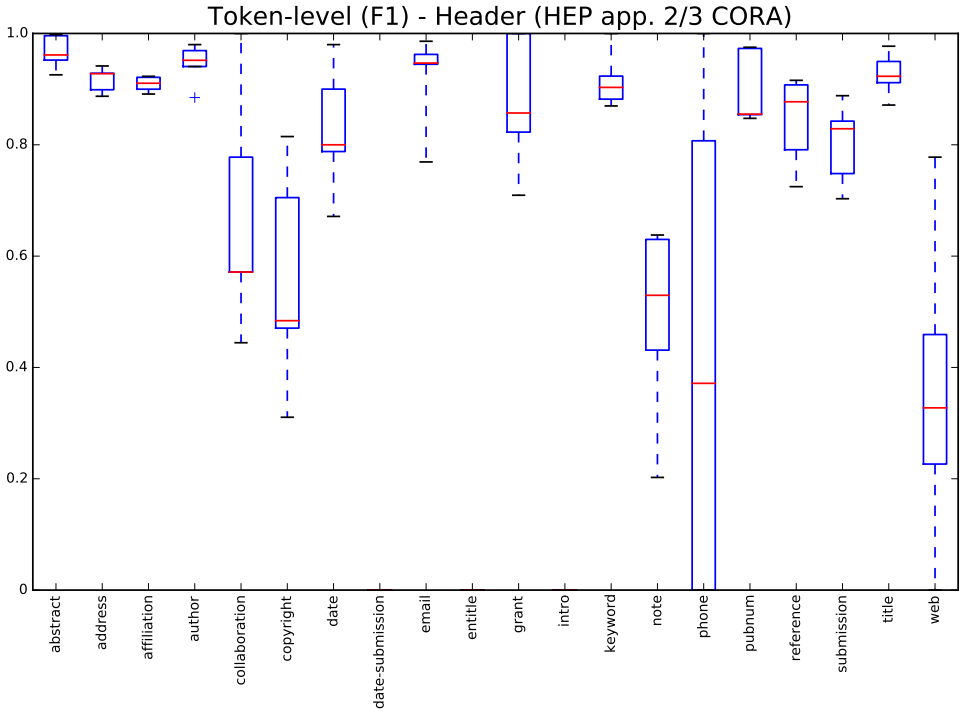

Confusion matrix - Header (HEP app. 1/3 CORA)
abstract 21382 32 12 28 91 17 1271 63 532 49 18 14
address 13 2897 138 44 2 8 73 11
affiliation 48 191 3469 78 2 6 5 22 5 6 65 14
author 14 38 42 3731 14 4 5 125 19 51 5 2
collaboration 2 1 5 24 2 4
copyright 134 10 12 196 15 28 75 9 3 46
date 7 6 198 9 14
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 30 4 22
pubnum 1 821 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
abstract address affiliation author collaboration copyright date date-submission email grant intro keyword note phone pubnum reference submission title web

(1.22)

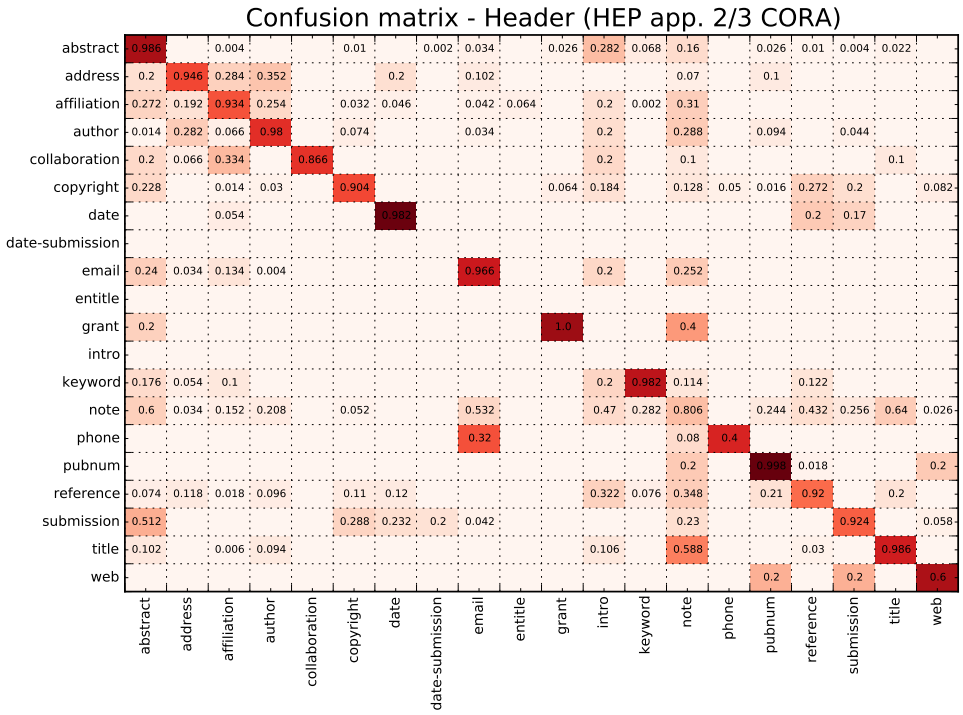
8.7 Header model - HEP dataset appending 2/3 CORA dataset



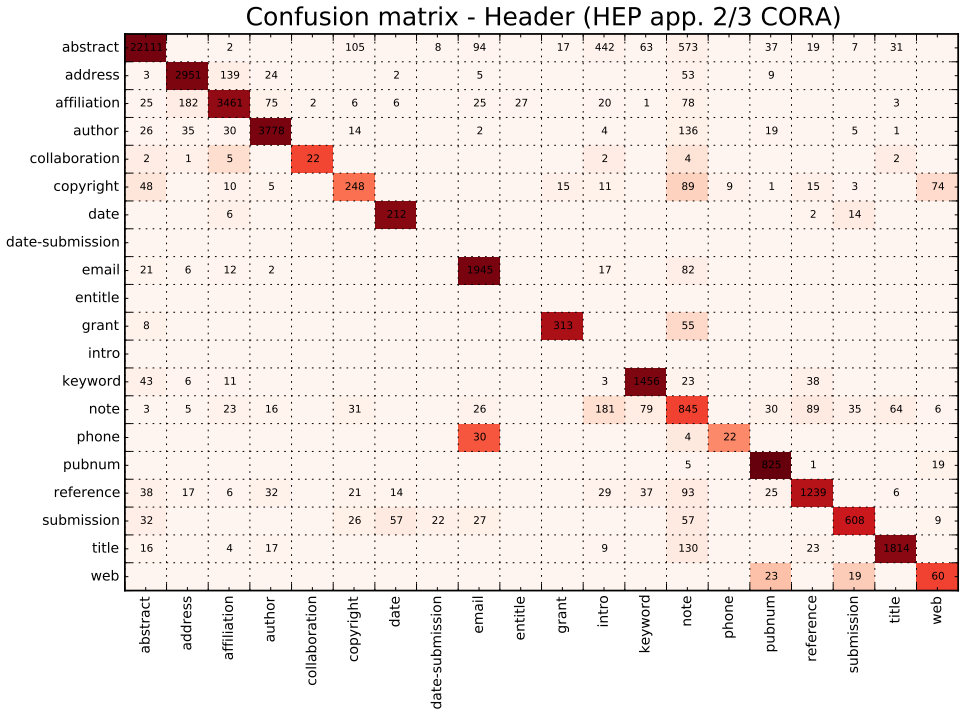
(1.23)



(1.24)

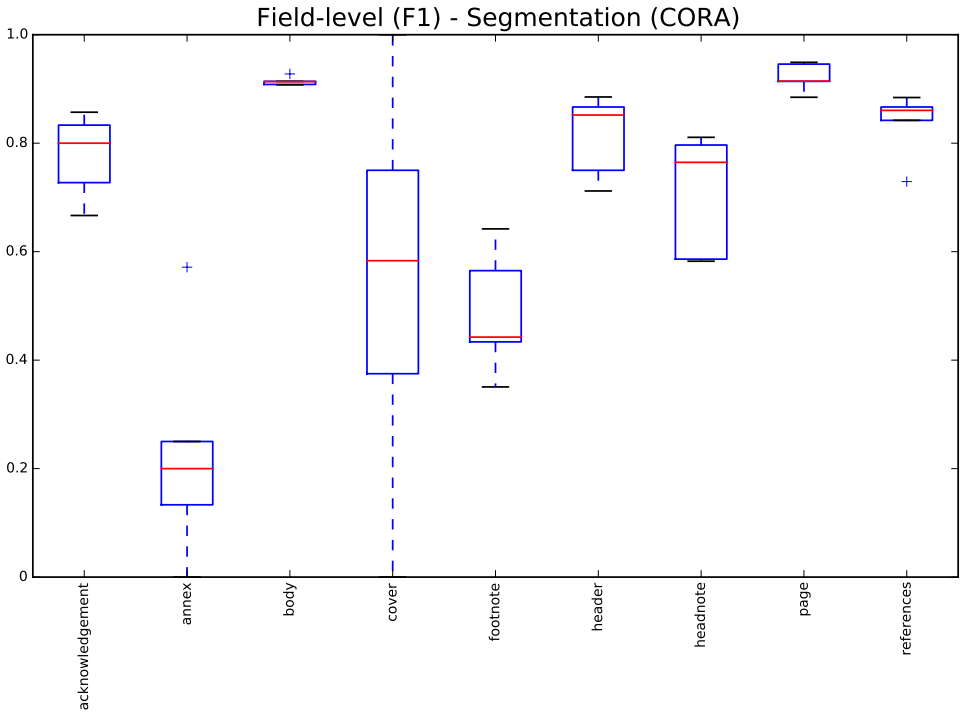

Confusion matrix - Header (HEP app. 2/3 CORA)
abstract
address
affiliation
author
collaboration
copyright
date
date-submission
email
entitle
grant
intro
keyword
note
phone
pubnum
reference
submission
title
web

(1.25)

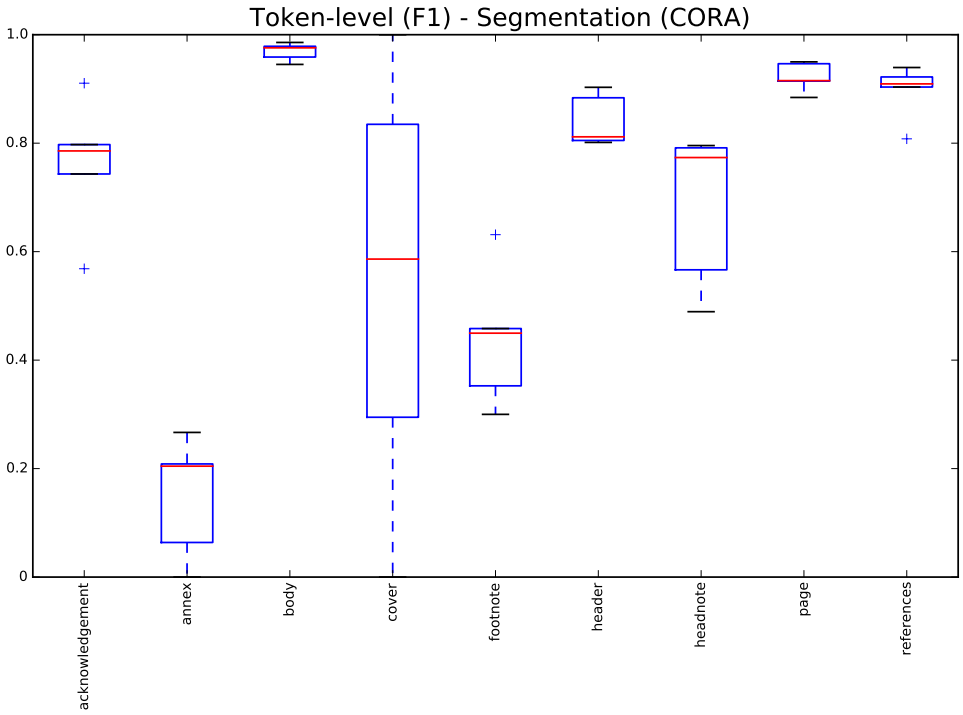

Confusion matrix - Header (HEP app. 2/3 CORA)
abstract
address
affiliation
author
collaboration
copyright
date
date-submission
email
entitle
grant
intro
keyword
note
phone
pubnum
reference
submission
title
web

(1.26)

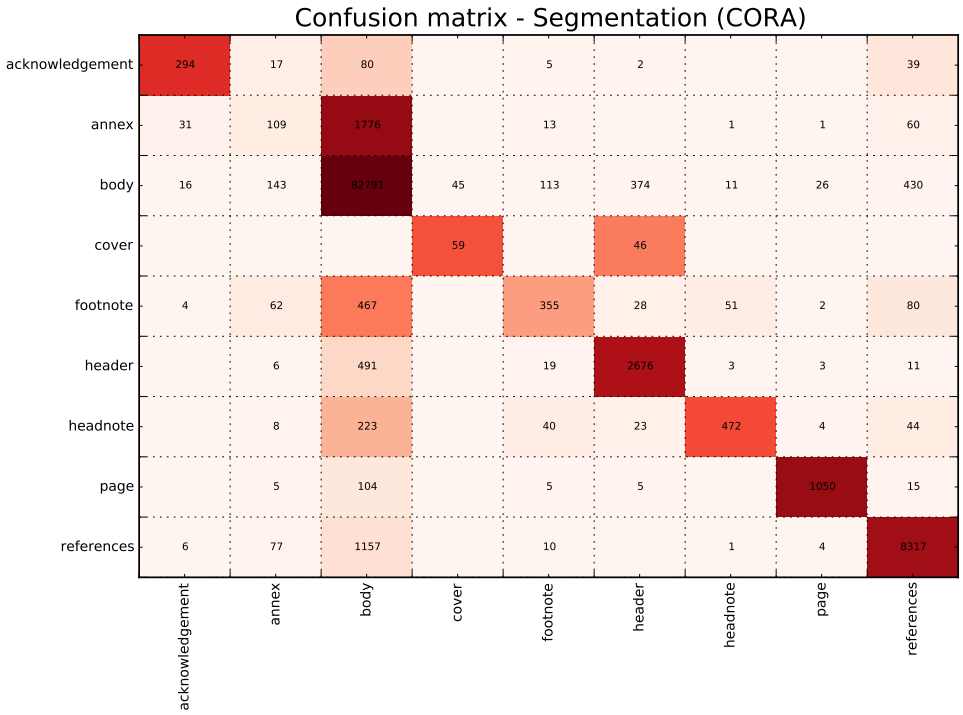
8.8 Segmentation model - Cora dataset



(1.27)

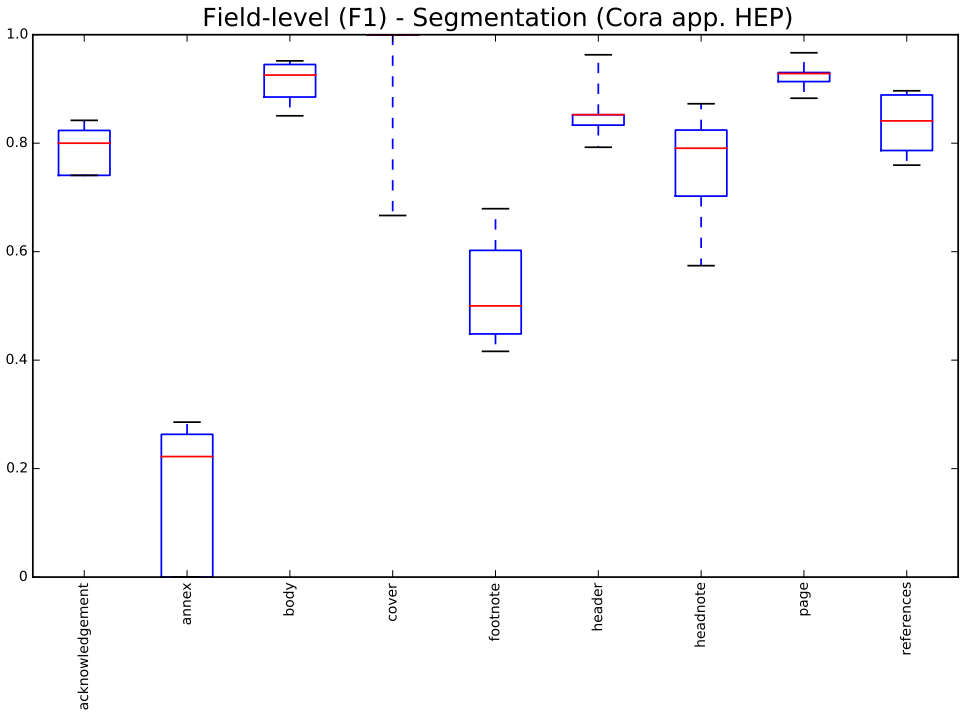


(1.28)

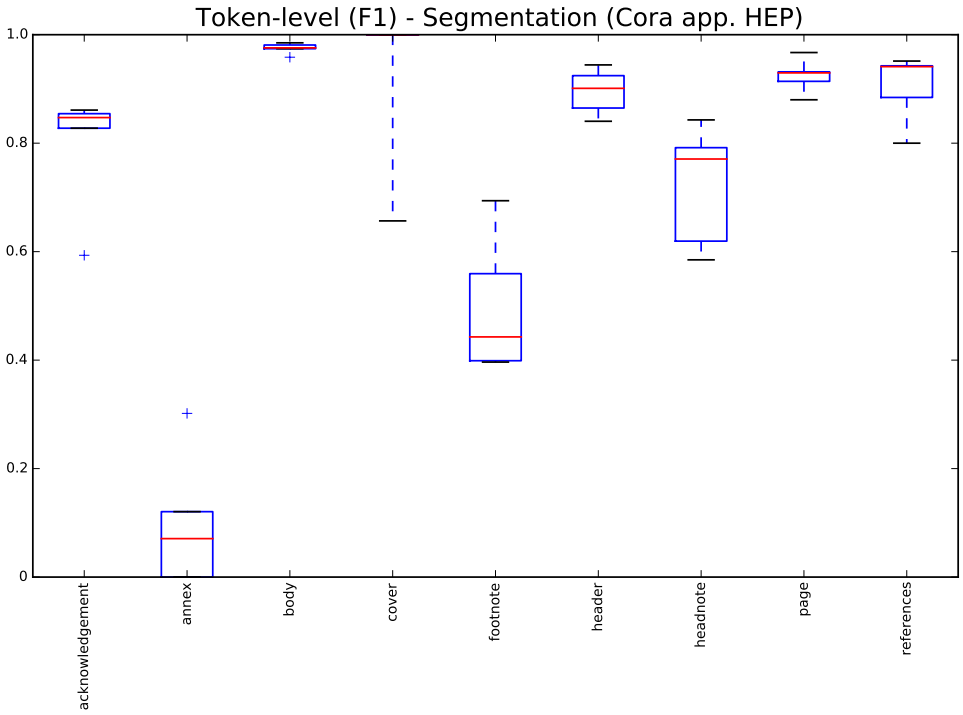

Confusion matrix - Segmentation (CORA)
acknowledgement
annex
body
cover
footnote
header
headnote
page
references
294 17 80 5 2 39
31 109 1776 13 1 1 60
16 143 82791 45 113 374 11 26 430
59 46
4 62 467 355 28 51 2 80
6 491 19 2676 3 3 11
8 223 40 23 472 4 44
5 104 5 5 1050 15
6 77 1157 10 1 4 8317

(1.29)

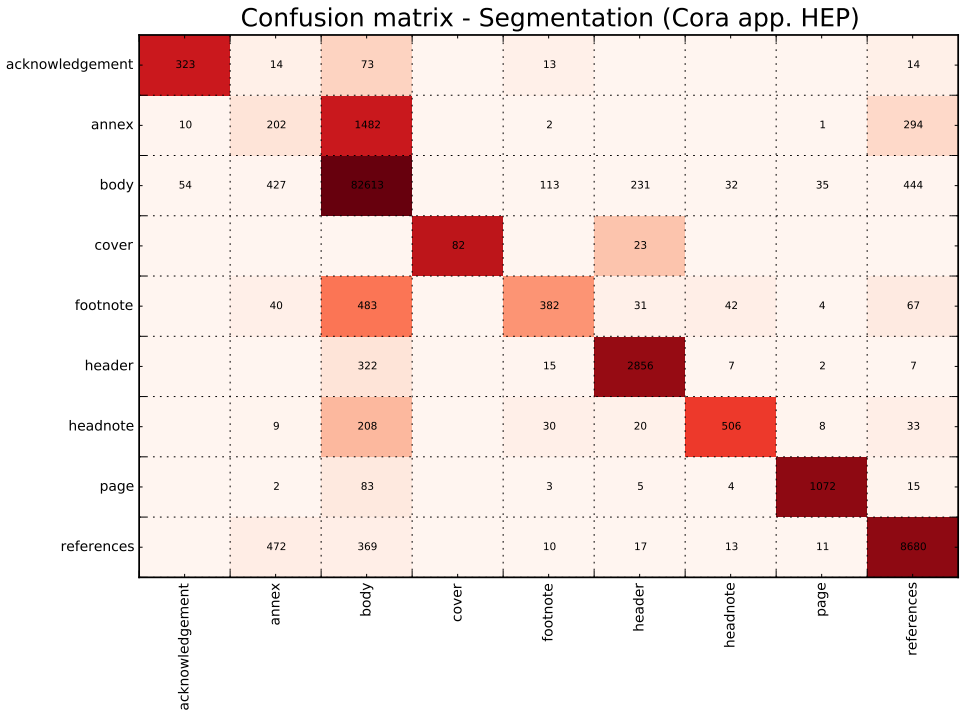
8.9 Segmentation model - Cora dataset appending HEP dataset



(1.30)

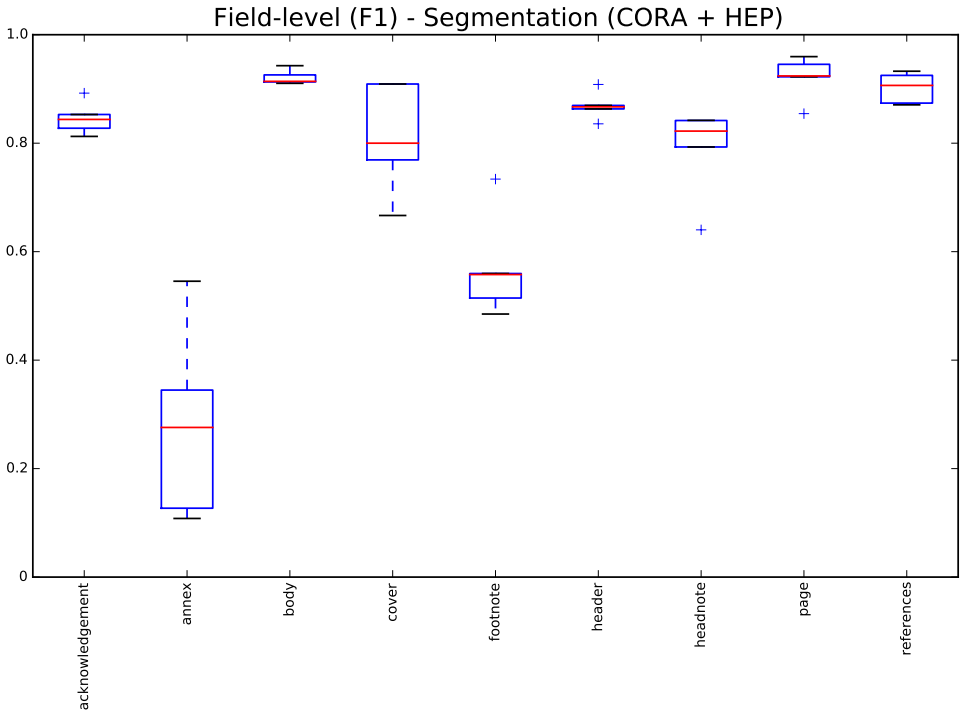


(1.31)

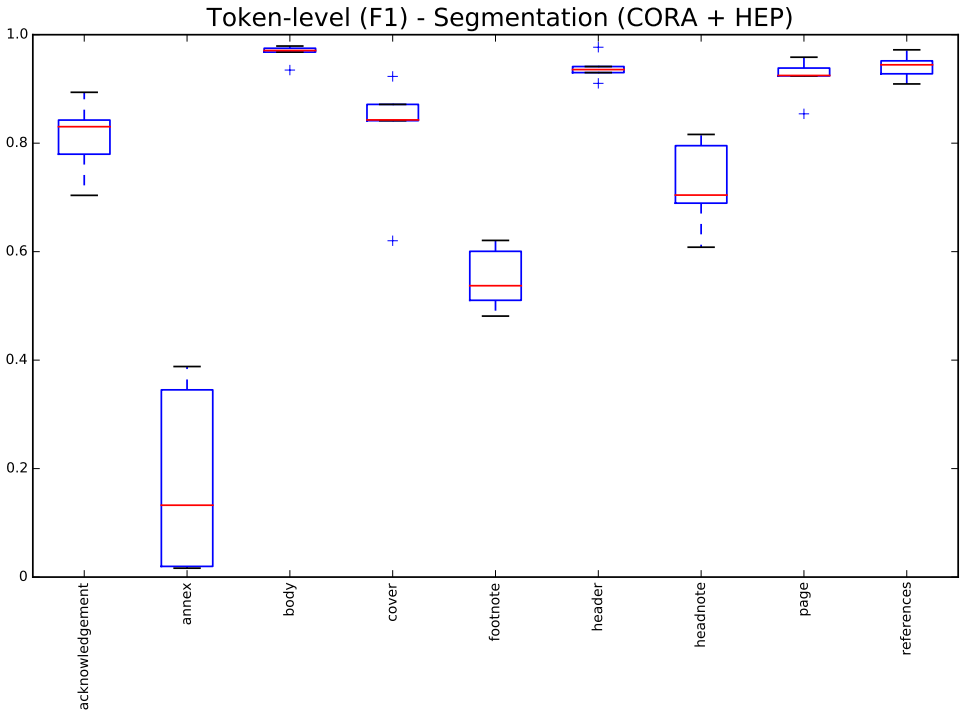

Confusion matrix - Segmentation (Cora app. HEP)
acknowledgement
annex
body
cover
footnote
header
headnote
page
references
323 14 73 13 14
10 202 1482 2 1 294
54 427 82613 113 231 32 35 444
82 23
40 483 382 31 42 4 67
322 15 2856 7 2 7
9 208 30 20 506 8 33
2 83 3 5 4 1072 15
472 369 10 17 13 11 8680

(1.32)

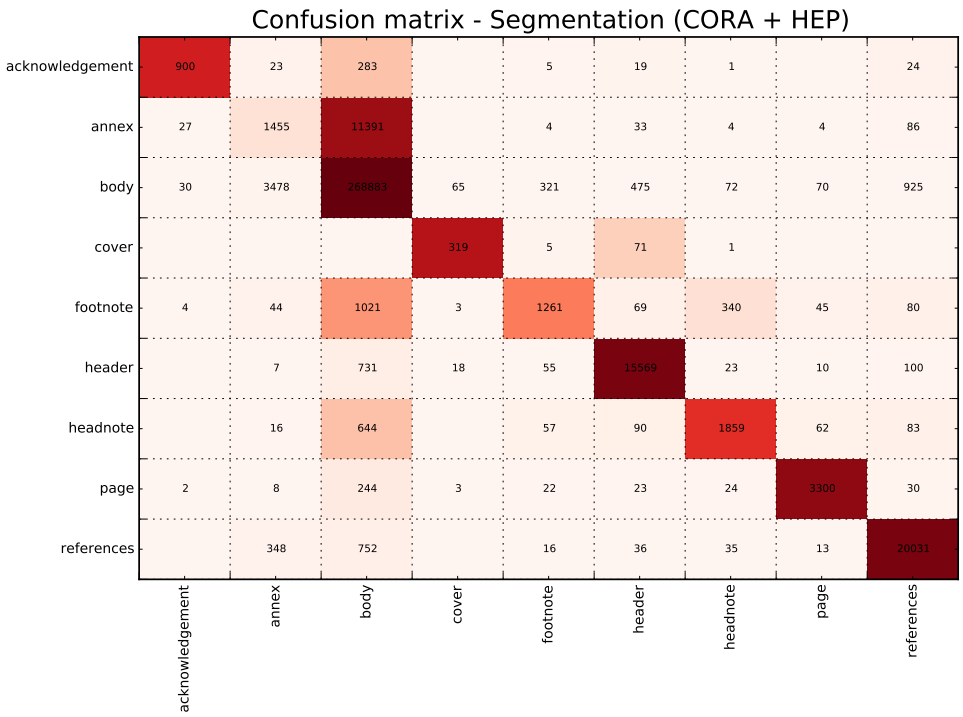
8.10 Segmentation model - Cora and HEP combined datasets



(1.33)

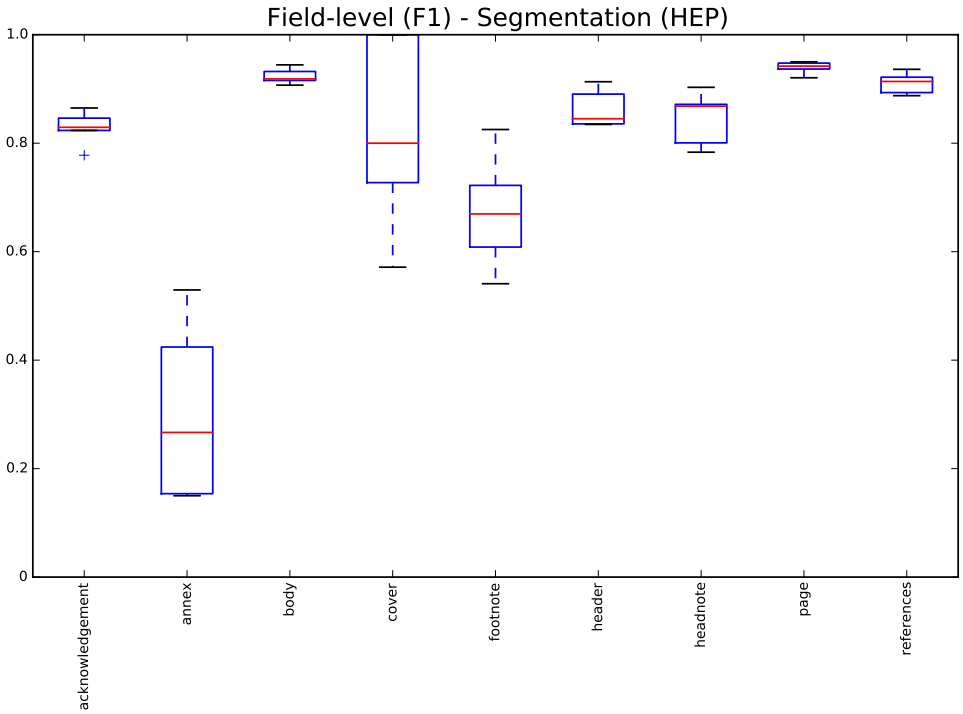


(1.34)

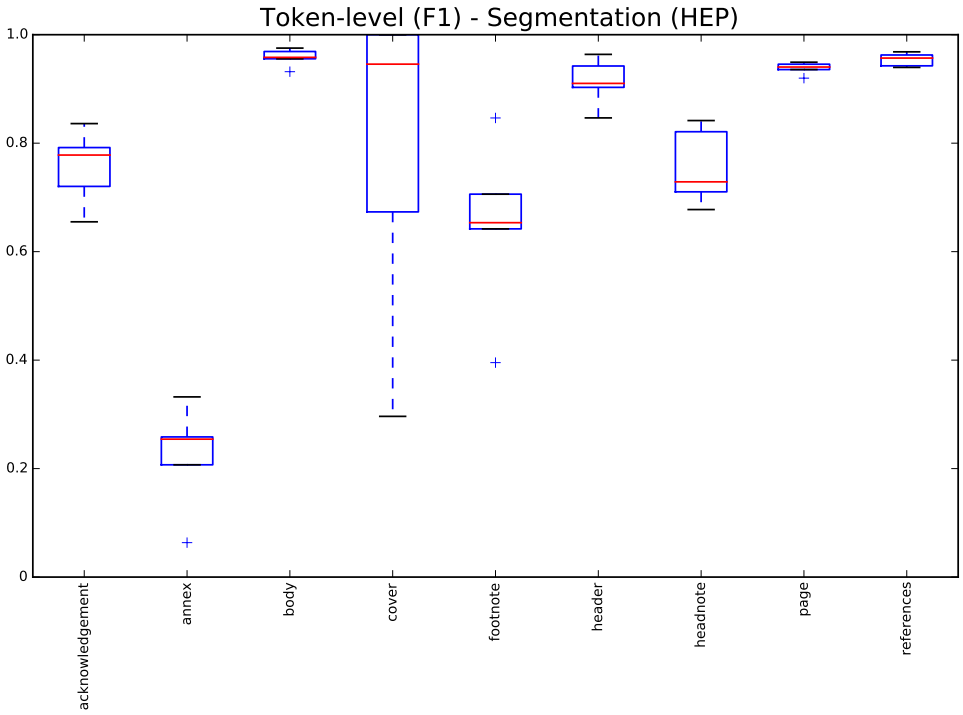

Confusion matrix - Segmentation (CORA + HEP)
acknowledgement
annex
body
cover
footnote
header
headnote
page
references
900 23 283 5 19 1 24
27 1455 11391 4 33 4 4 86
30 3478 268883 65 321 475 72 70 925
319 5 71 1
4 44 1021 3 1261 69 340 45 80
7 731 18 55 15569 23 10 100
16 644 57 90 1859 62 83
2 8 244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

(1.35)

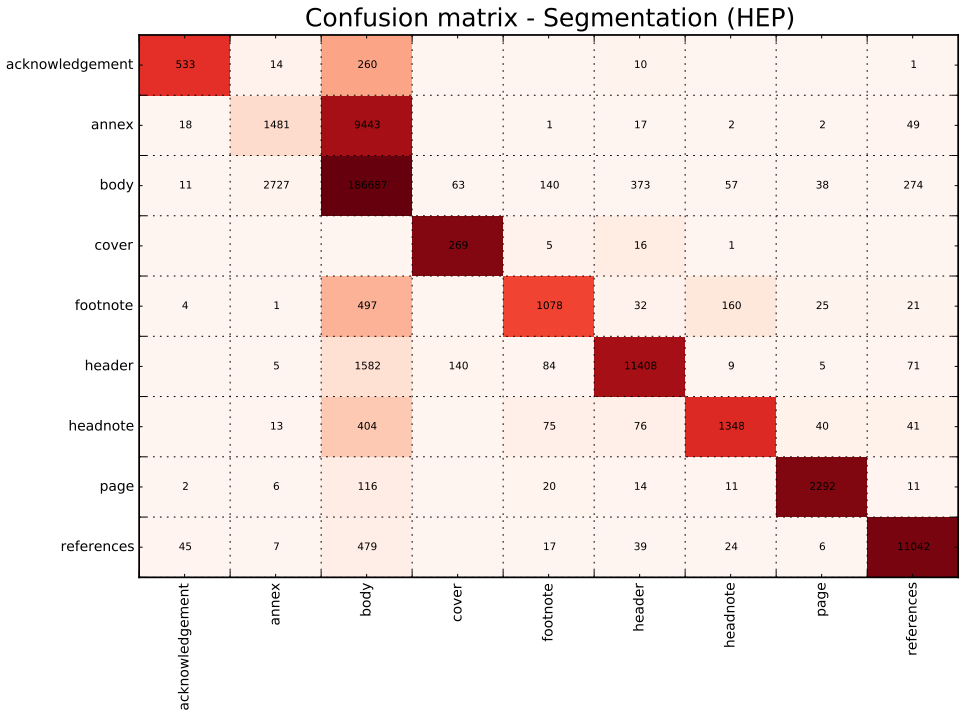
8.11 Segmentation model - HEP dataset



(1.36)

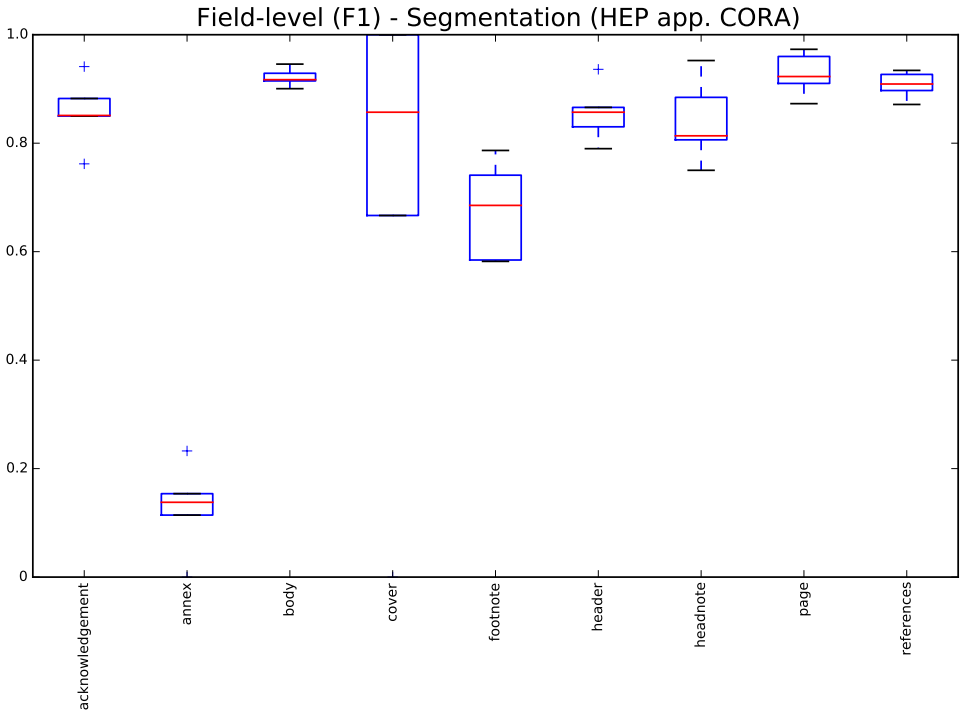


(1.37)

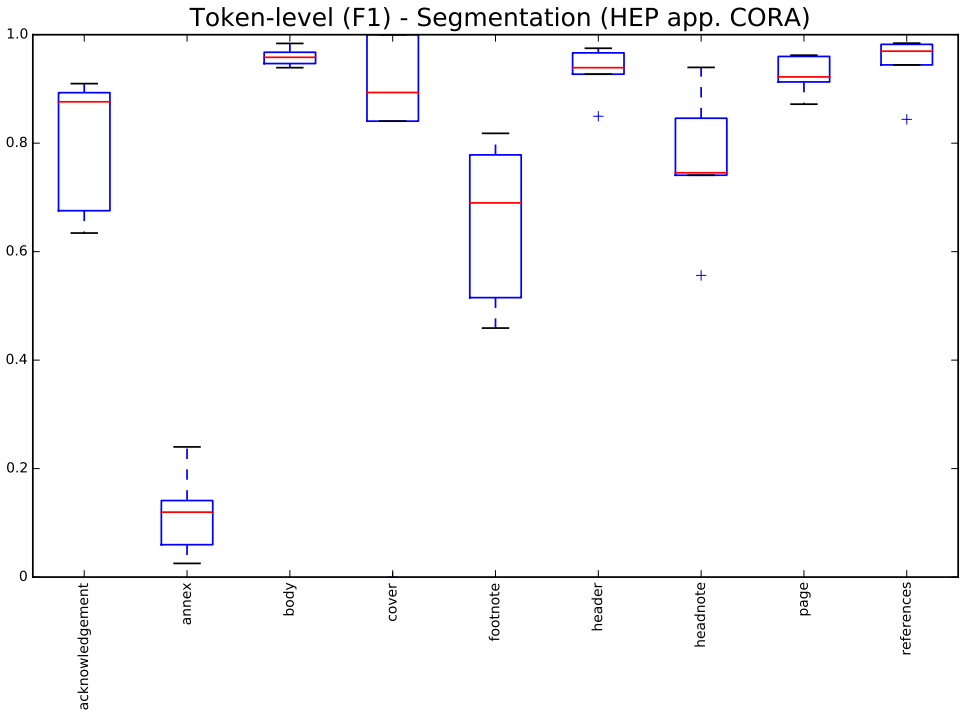

Confusion matrix - Segmentation (HEP)
acknowledgement
annex
body
cover
footnote
header
headnote
page
references
533 14 260 10 1
18 1481 9443 1 17 2 2 49
11 2727 186687 63 140 373 57 38 274
269 5 16 1
4 1 497 1078 32 160 25 21
5 1582 140 84 11408 9 5 71
13 404 75 76 1348 40 41
2 6 116 20 14 11 2292 11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

(1.38)

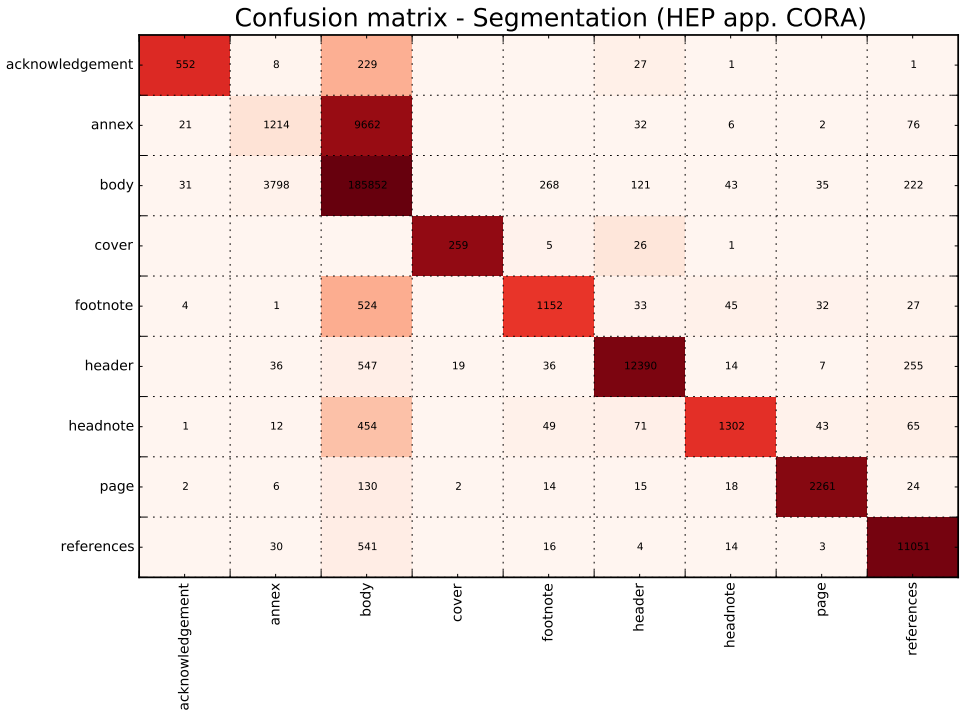
8.12 Segmentation model - HEP dataset appending CORA dataset



(1.39)



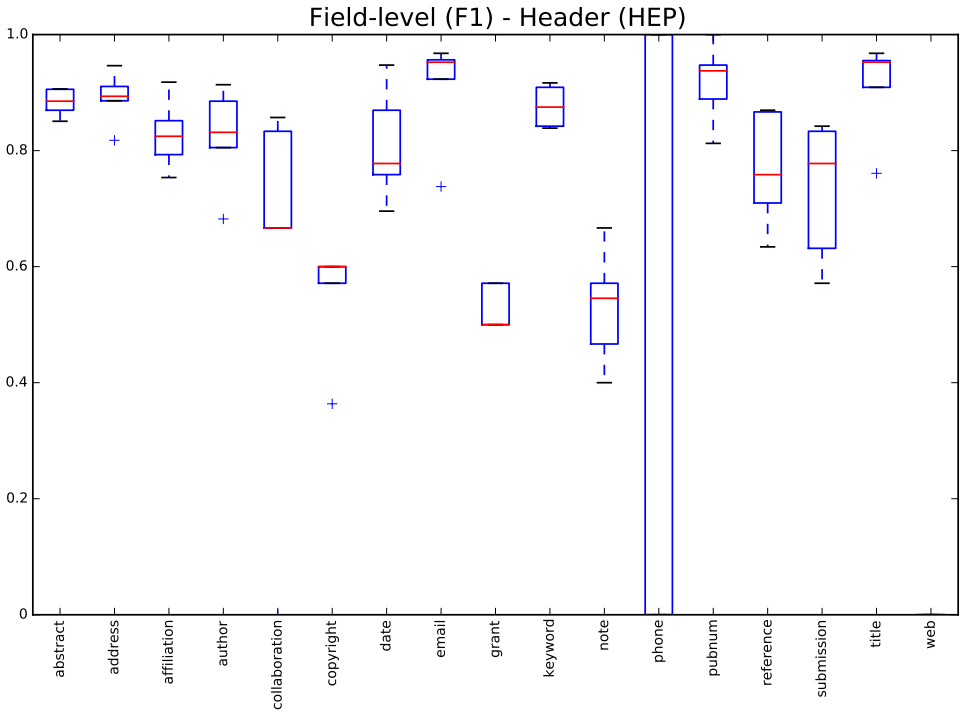
(1.40)


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

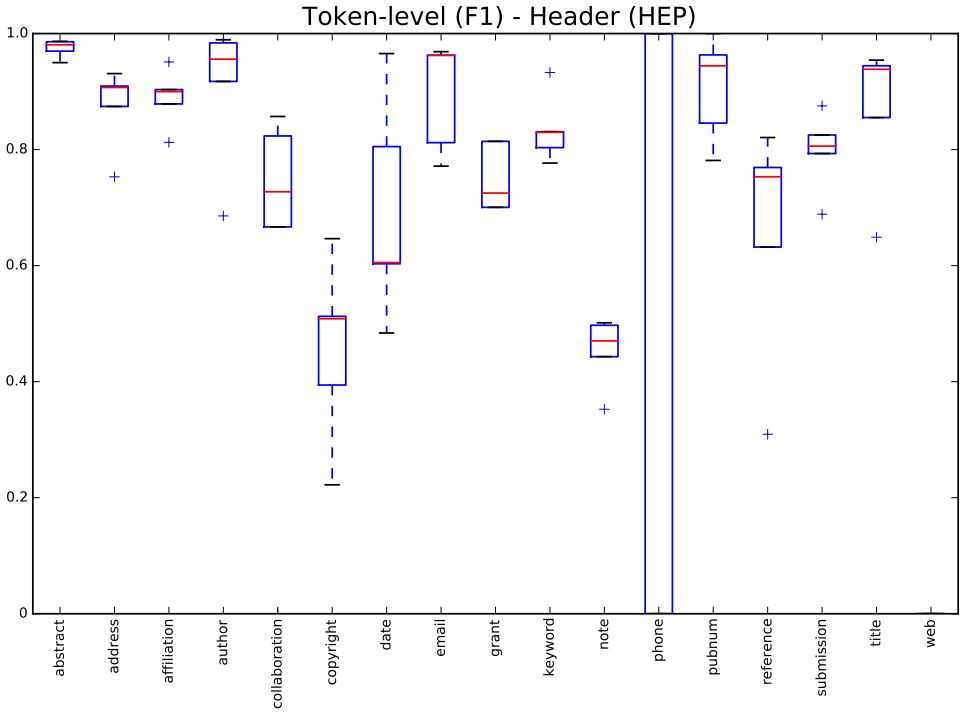
(1.41)

9 Regularisation

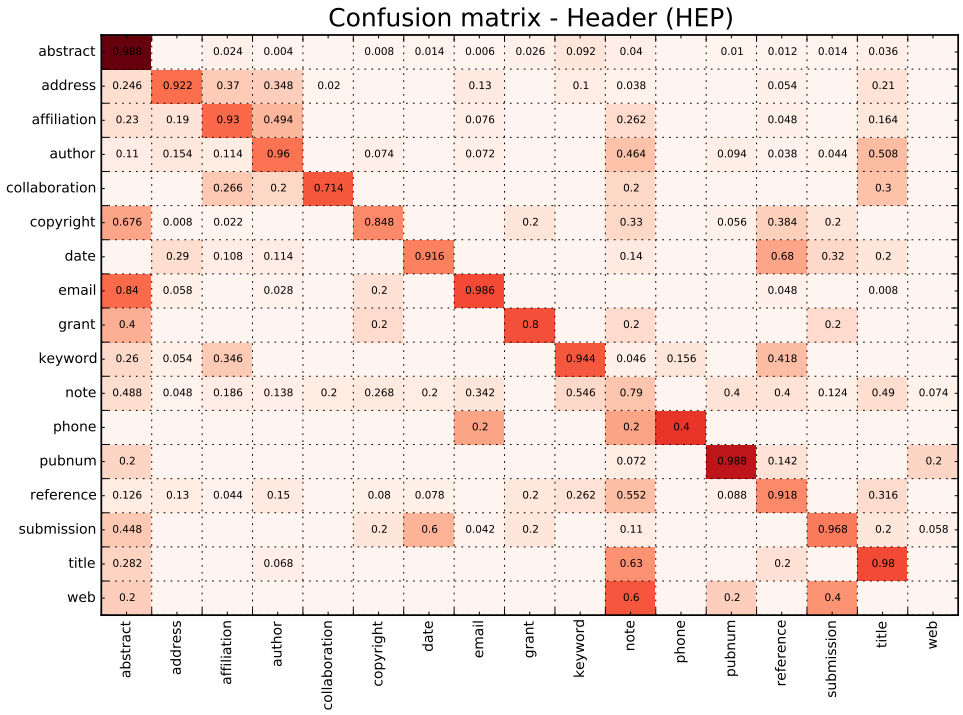
9.1 Header model - $L2 = 0$



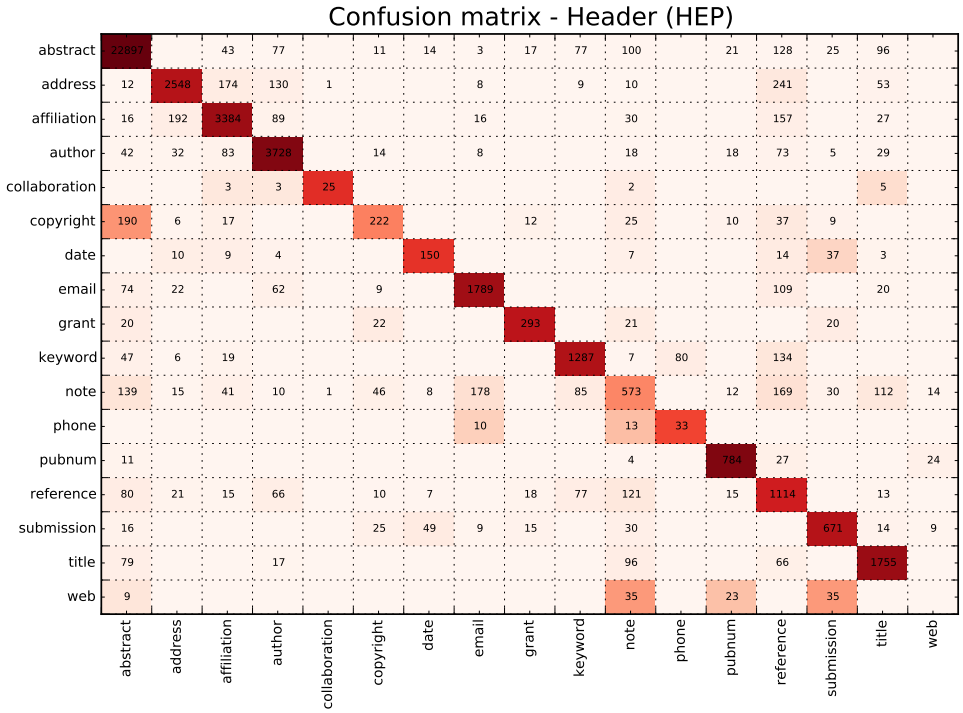
(1.42)



(1.43)

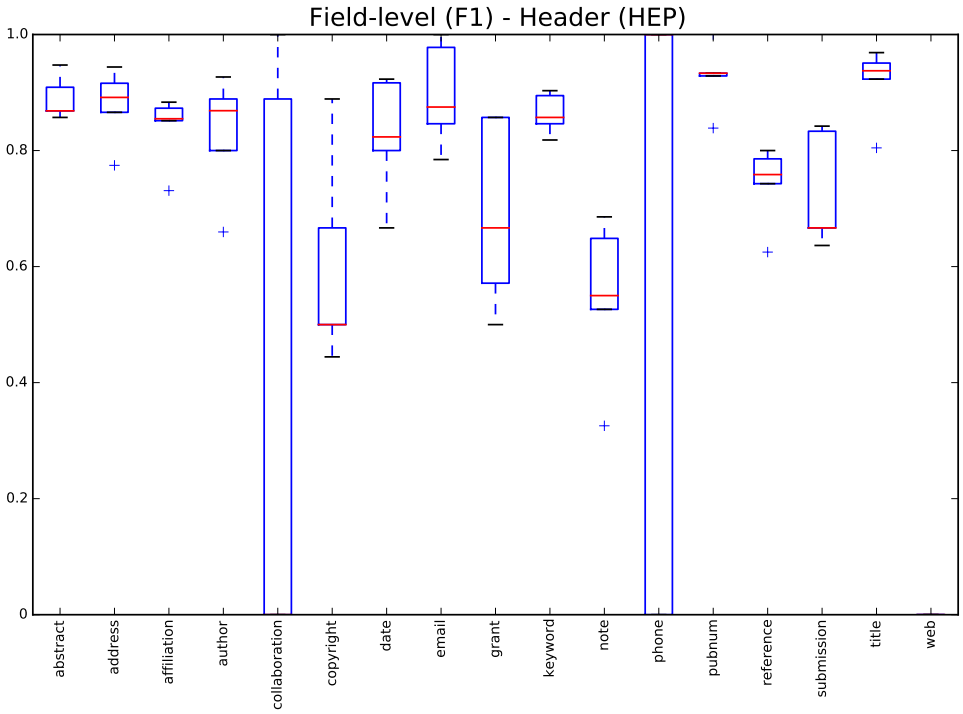

Confusion matrix - Header (HEP)
abstract 0.988 0.024 0.004 0.008 0.014 0.006 0.026 0.092 0.04 0.01 0.012 0.014 0.036
address 0.246 0.922 0.37 0.348 0.02 0.13 0.1 0.038 0.054 0.21
affiliation 0.23 0.19 0.93 0.494 0.076 0.262 0.048 0.164
author 0.11 0.154 0.114 0.96 0.074 0.072 0.464 0.094 0.038 0.044 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.44)

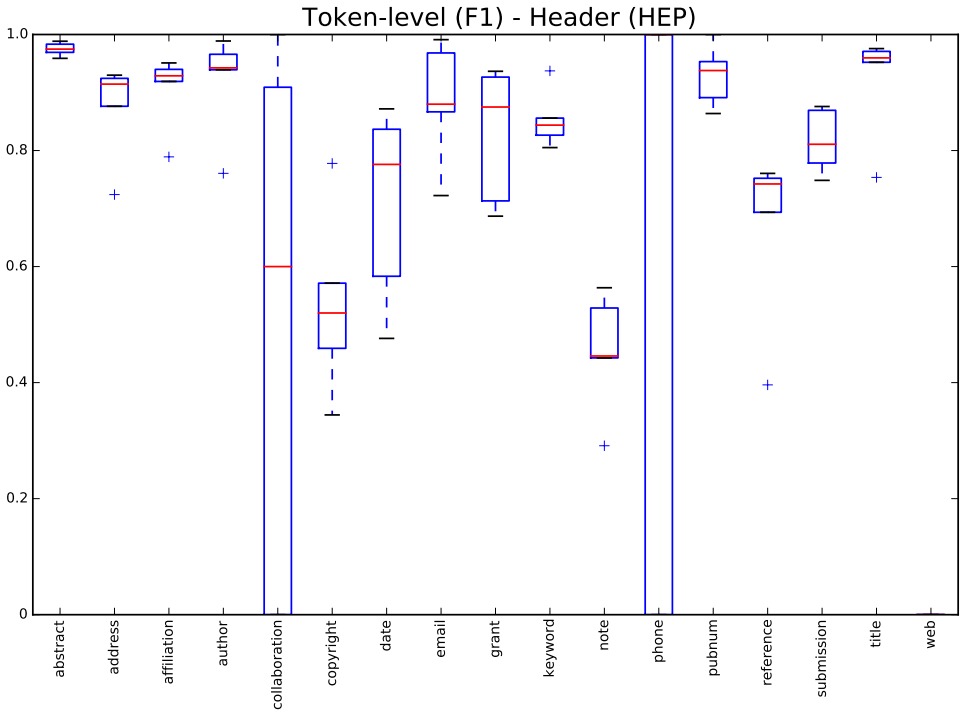

Confusion matrix - Header (HEP)
abstract 22897 43 77 11 14 3 17 77 100 21 128 25 96
address 12 2548 174 130 1 8 9 10 241 53
affiliation 16 192 3384 89 16 30 157 27
author 42 32 83 3728 14 8 18 18 73 5 29
collaboration 3 3 25 2 5
copyright 190 6 17 222 12 25 10 37 9
date 10 9 4 150 7 14 37 3
email 74 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 14
phone 10 13 33
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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.45)

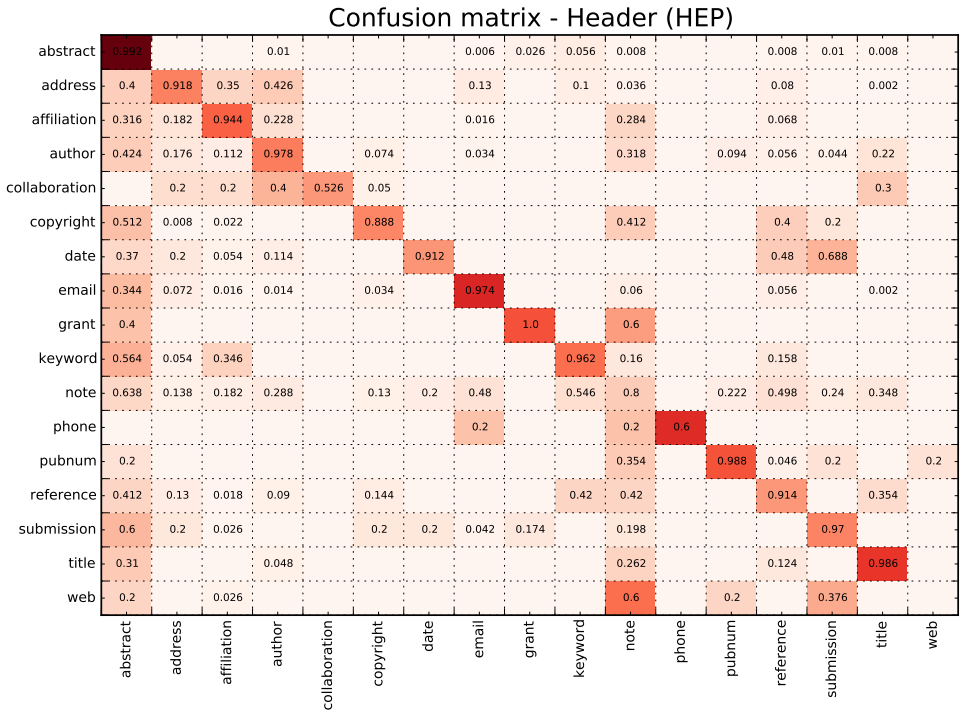
9.2 Header model - $L2 = 1e^{-6}$



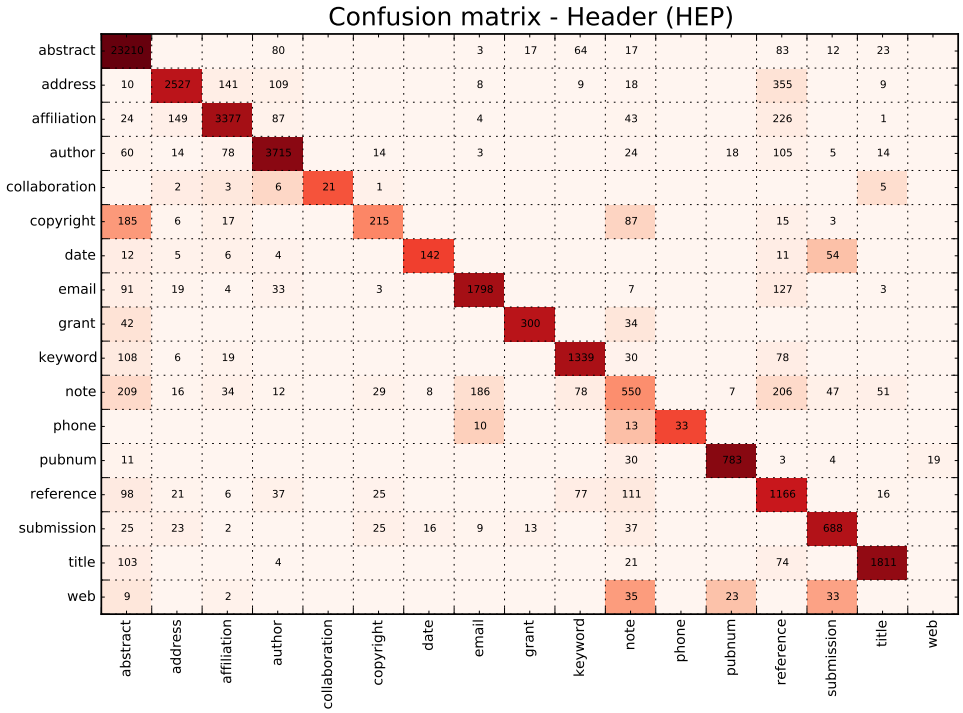
(1.46)



(1.47)

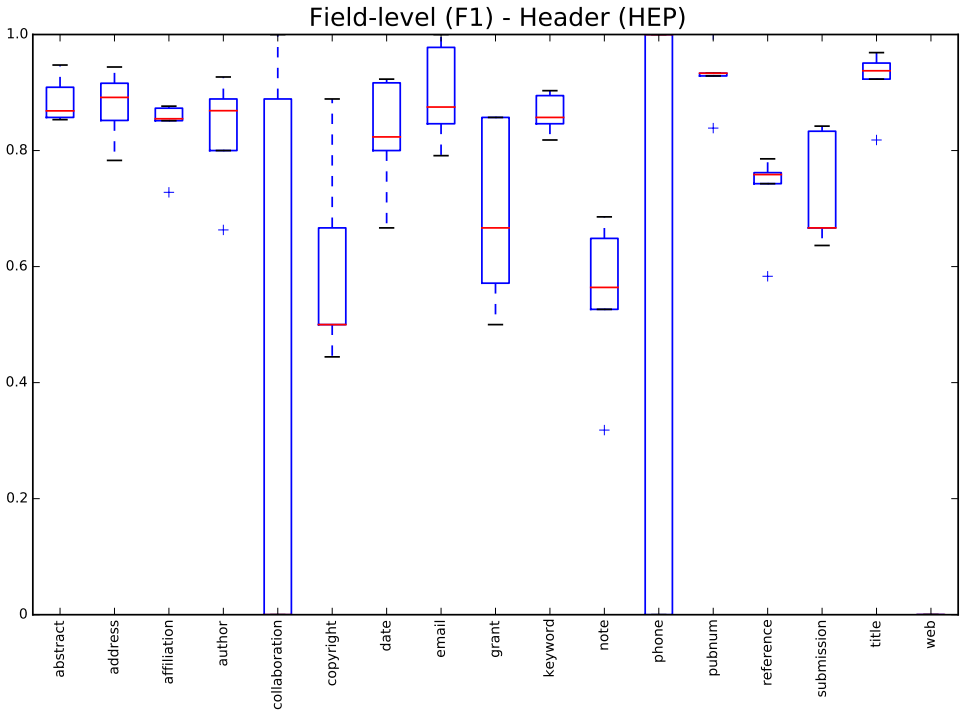

Confusion matrix - Header (HEP)
abstract 0.992 0.01 0.006 0.026 0.056 0.008 0.008 0.01 0.008
address 0.4 0.918 0.35 0.426 0.13 0.1 0.036 0.08 0.002
affiliation 0.316 0.182 0.944 0.228 0.016 0.284 0.068
author 0.424 0.176 0.112 0.978 0.074 0.034 0.318 0.094 0.056 0.044 0.22
collaboration 0.2 0.2 0.4 0.526 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.48)

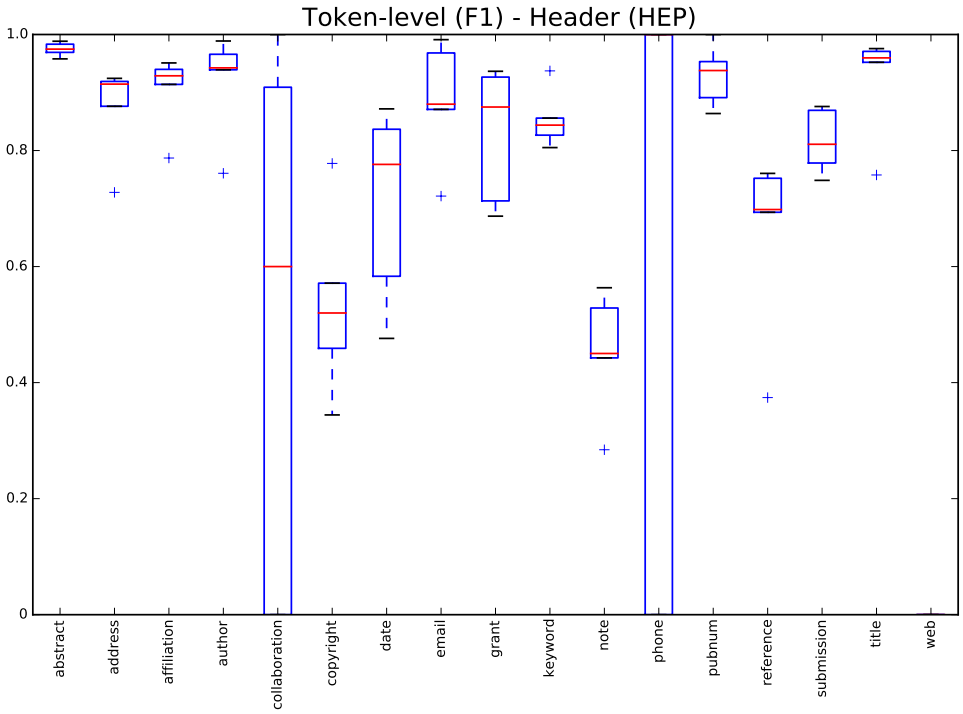

Confusion matrix - Header (HEP)
abstract 23210 80 3 17 64 17 83 12 23
address 10 2527 141 109 8 9 18 355 9
affiliation 24 149 3377 87 4 43 226 1
author 60 14 78 3715 14 3 24 18 105 5 14
collaboration 2 3 6 21 1 5
copyright 185 6 17 215 87 15 3
date 12 5 6 4 142 11 54
email 91 19 4 33 3 1798 7 127 3
grant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

(1.49)

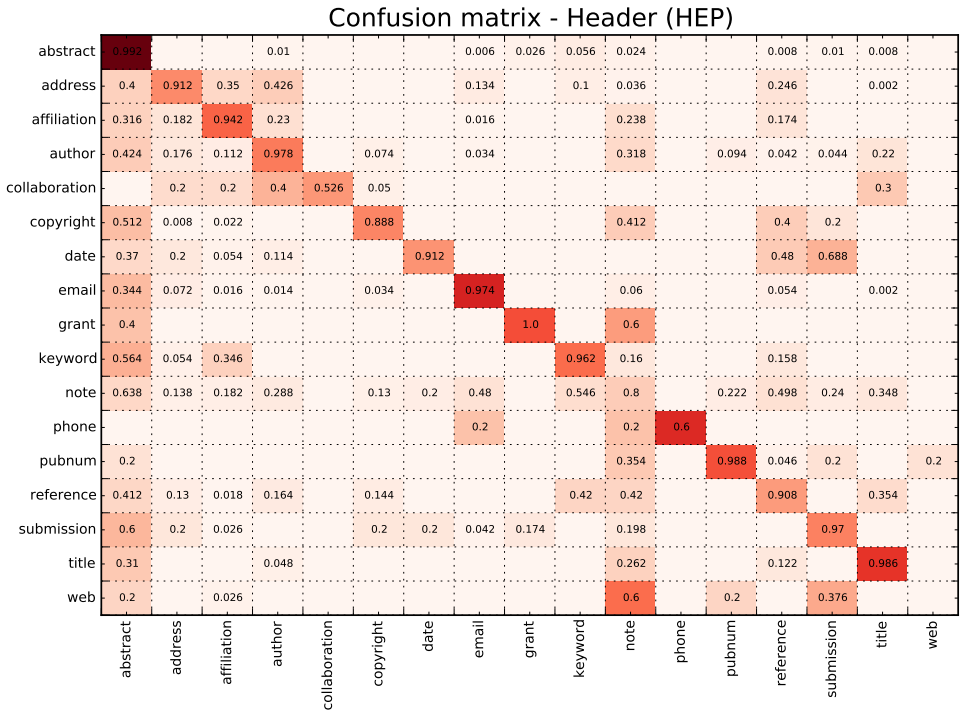
9.3 Header model - $L2 = 1e^{-5}$



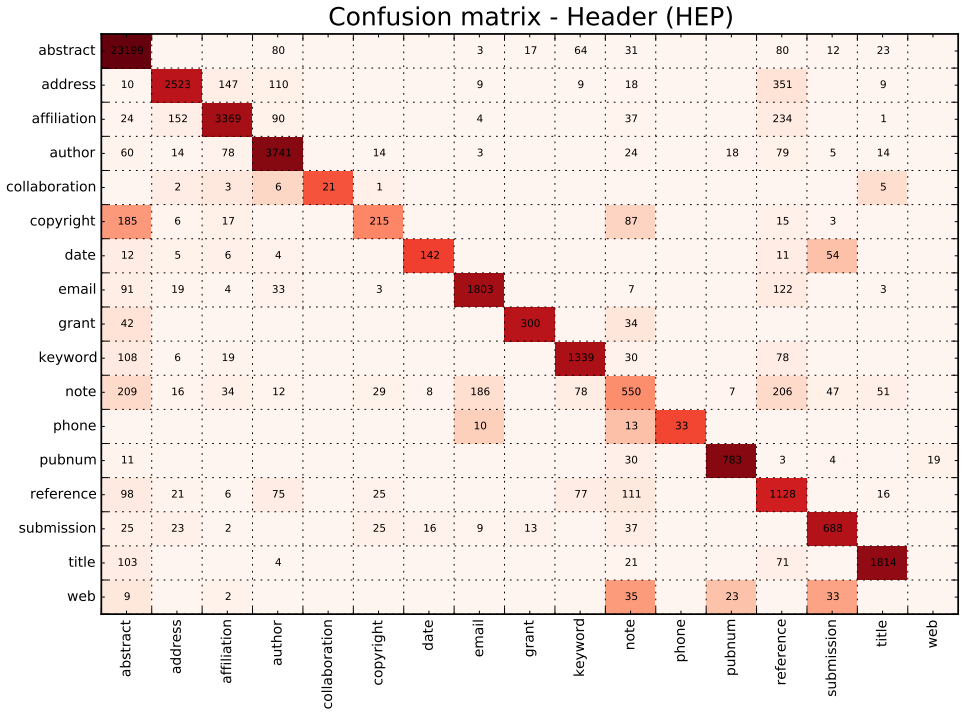
(1.50)



(1.51)

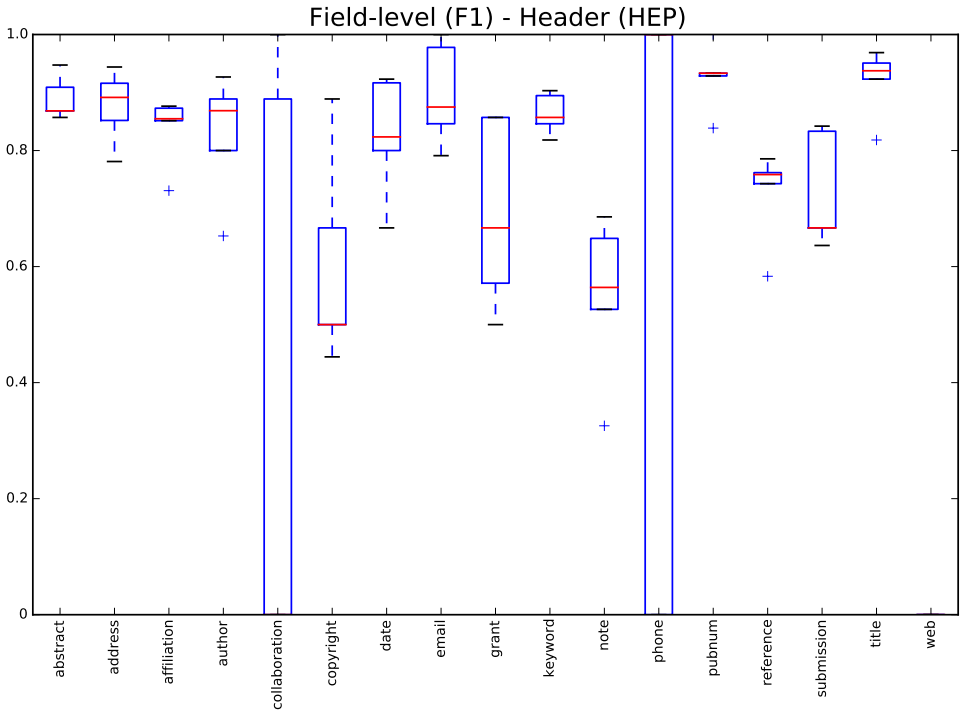

Confusion matrix - Header (HEP)
abstract 0.992 0.01 0.006 0.026 0.056 0.024 0.008 0.01 0.008
address 0.4 0.912 0.35 0.426 0.134 0.1 0.036 0.246 0.002
affiliation 0.316 0.182 0.942 0.23 0.016 0.238 0.174
author 0.424 0.176 0.112 0.978 0.074 0.034 0.318 0.094 0.042 0.044 0.22
collaboration 0.2 0.2 0.4 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.52)

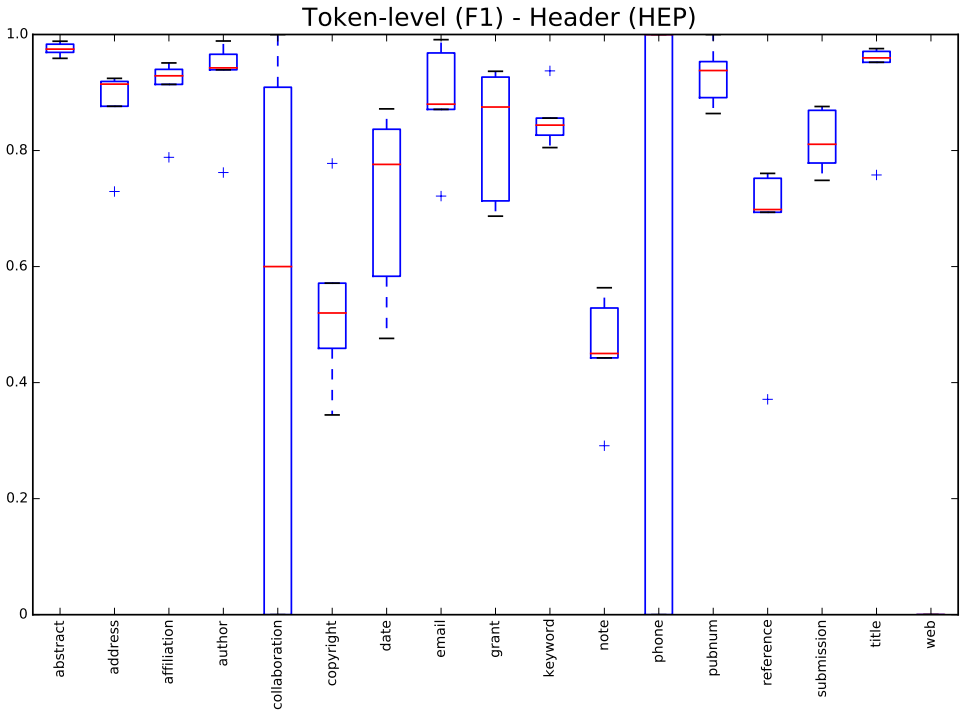

Confusion matrix - Header (HEP)
abstract 23199 80 3 17 64 31 80 12 23
address 10 2523 147 110 9 9 18 351 9
affiliation 24 152 3369 90 4 37 234 1
author 60 14 78 3741 14 3 24 18 79 5 14
collaboration 2 3 6 21 1 5
copyright 185 6 17 215 87 15 3
date 12 5 6 4 142 11 54
email 91 19 4 33 3 1803 7 122 3
grant 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 77 111 1128 16
submission 25 23 2 25 16 9 13 37 688
title 103 4 21 71 1814
web 9 2 35 23 33
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.53)

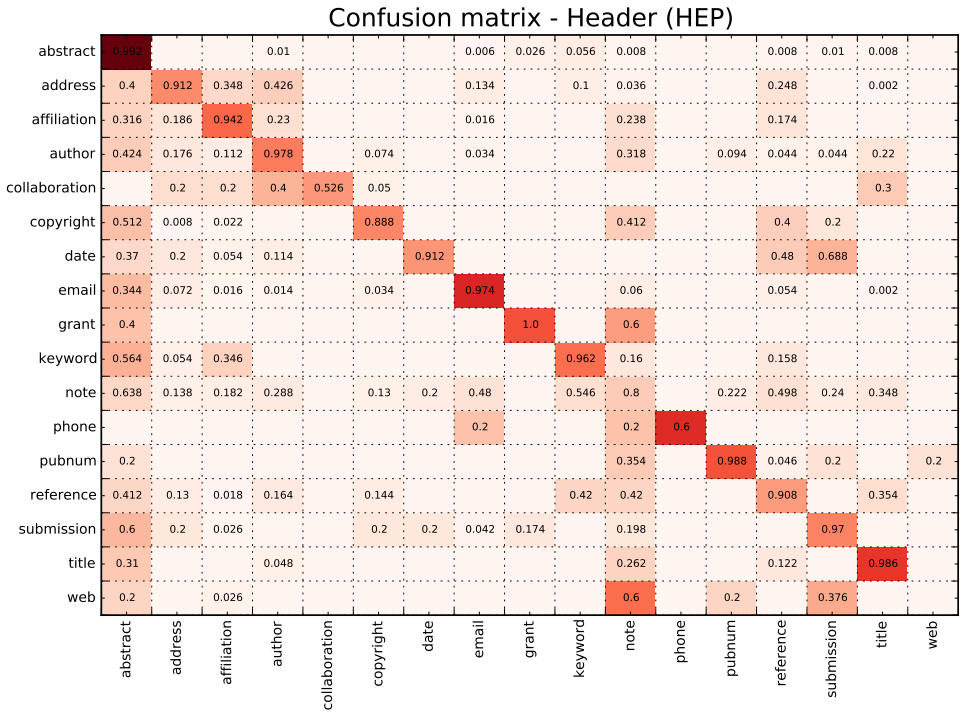
9.4 Header model - $L2 = 1e^{-4}$



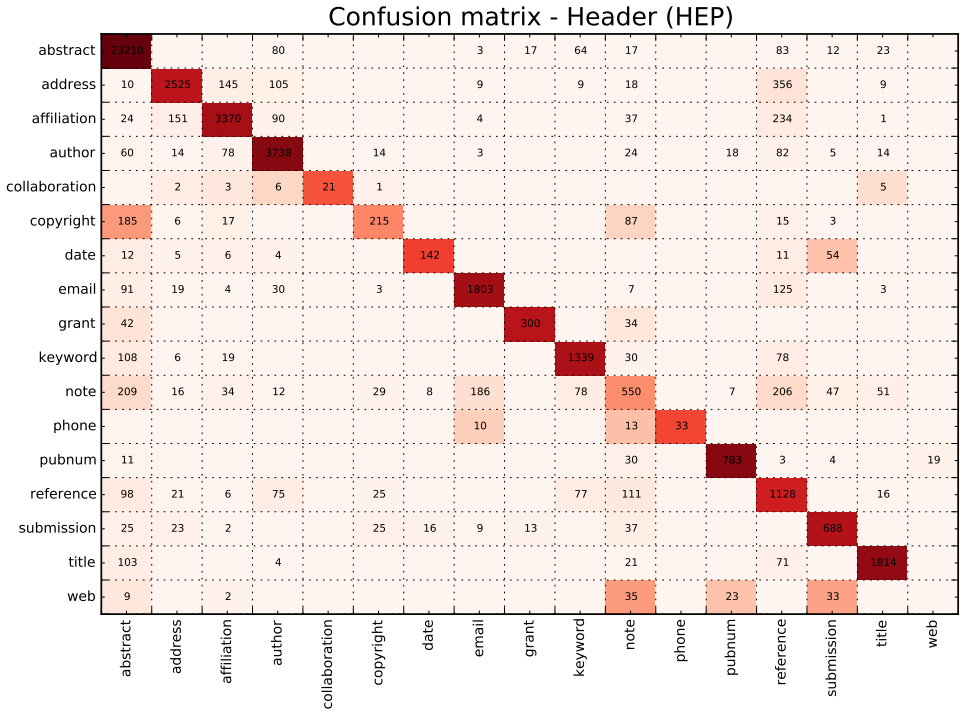
(1.54)



(1.55)

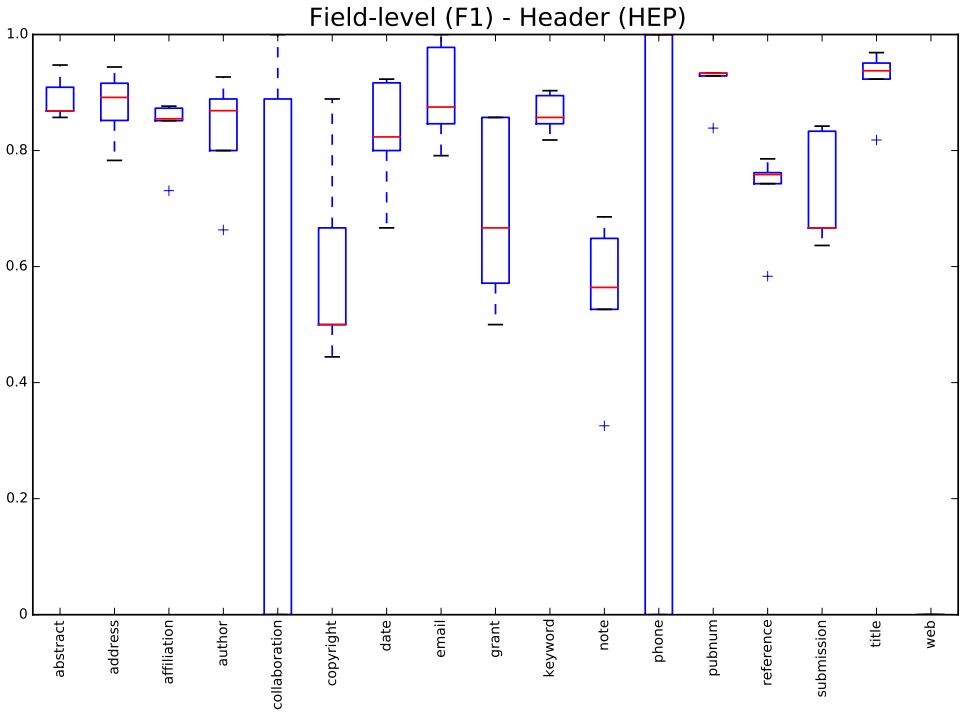

Confusion matrix - Header (HEP)
abstract 0.992 0.01 0.006 0.026 0.056 0.008 0.008 0.01 0.008
address 0.4 0.912 0.348 0.426 0.134 0.1 0.036 0.248 0.002
affiliation 0.316 0.186 0.942 0.23 0.016 0.238 0.174
author 0.424 0.176 0.112 0.978 0.074 0.034 0.318 0.094 0.044 0.044 0.22
collaboration 0.2 0.2 0.4 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.56)

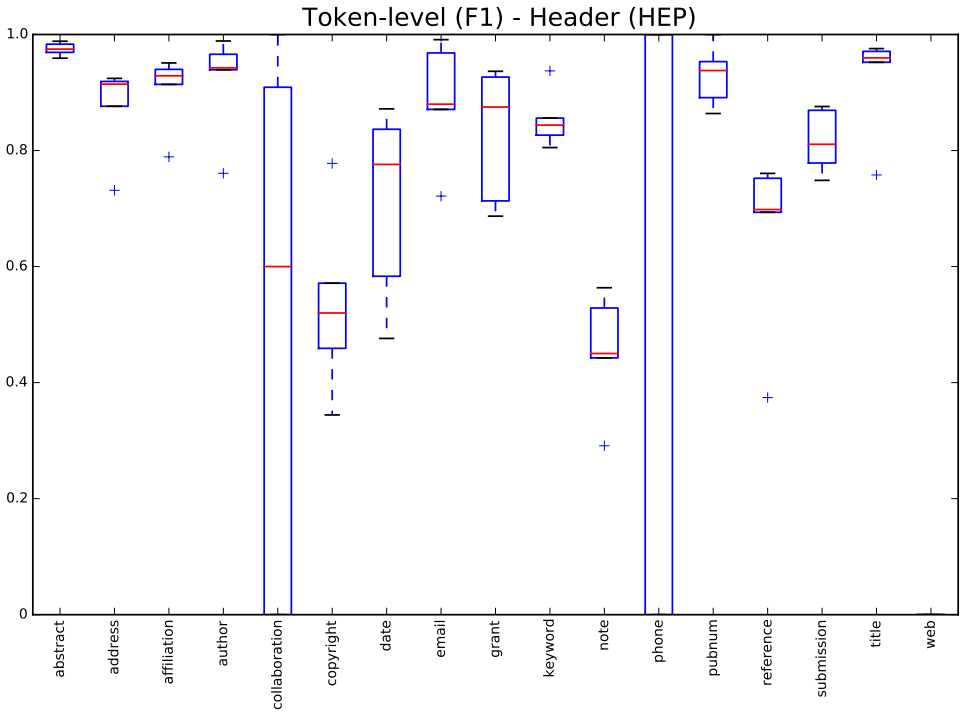

Confusion matrix - Header (HEP)
abstract 23210 80 3 17 64 17 83 12 23
address 10 2525 145 105 9 9 18 356 9
affiliation 24 151 3370 90 4 37 234 1
author 60 14 78 3738 14 3 24 18 82 5 14
collaboration 2 3 6 21 1 5
copyright 185 6 17 215 87 15 3
date 12 5 6 4 142 11 54
email 91 19 4 30 3 1803 7 125 3
grant 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 77 111 1128 16
submission 25 23 2 25 16 9 13 37 688
title 103 4 21 71 1814
web 9 2 35 23 33
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.57)

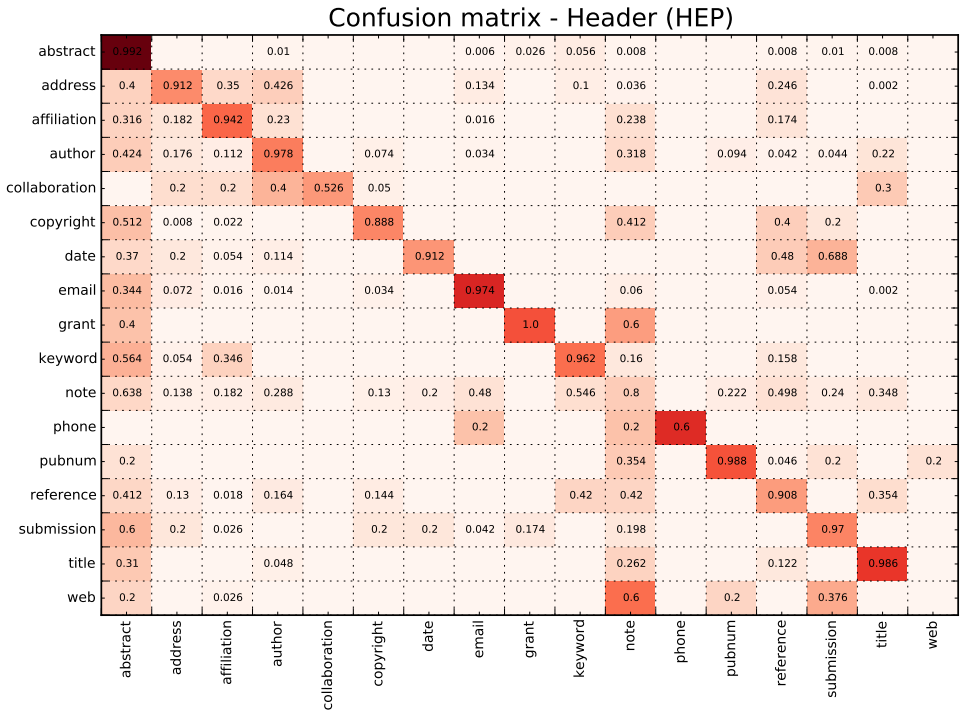
9.5 Header model - $L2 = 1e^{-3}$



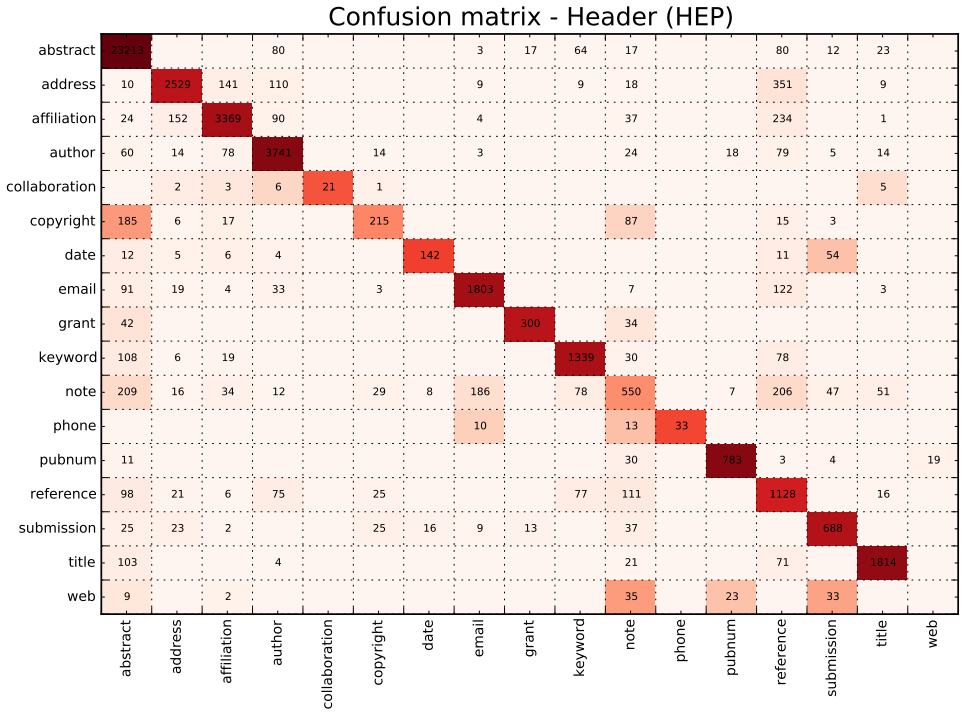
(1.58)



(1.59)


Confusion matrix - Header (HEP)
abstract 0.992 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

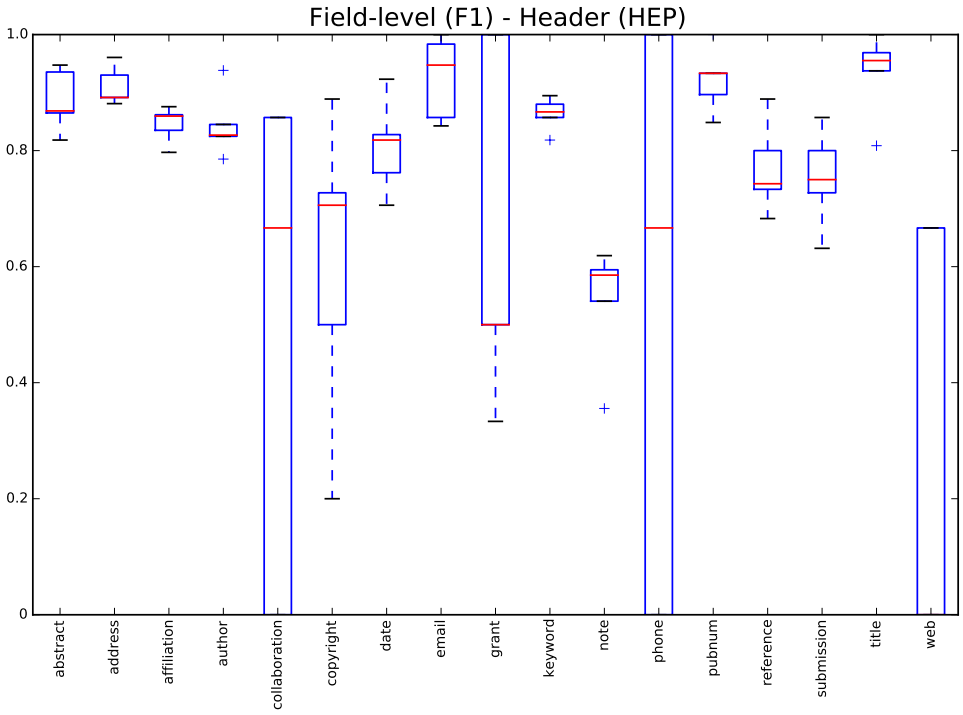
(1.60)


Confusion matrix - Header (HEP)
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 77 111 1128 16
submission 25 23 2 25 16 9 13 37 688
title 103 4 21 71 1814
web 9 2 35 23 33
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

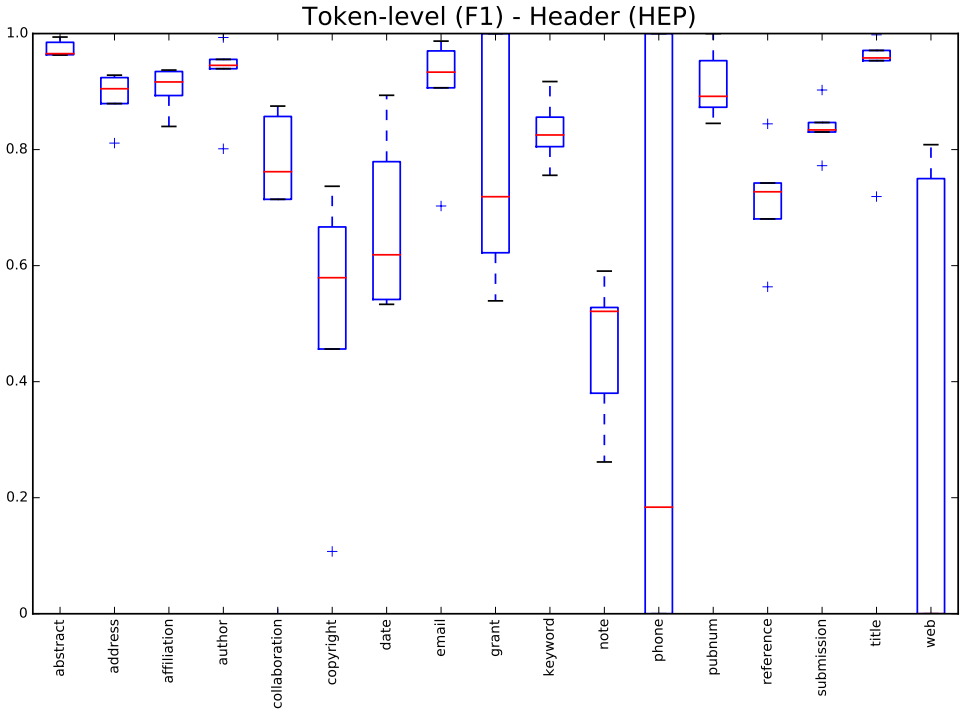
(1.61)

10 Dictionaries

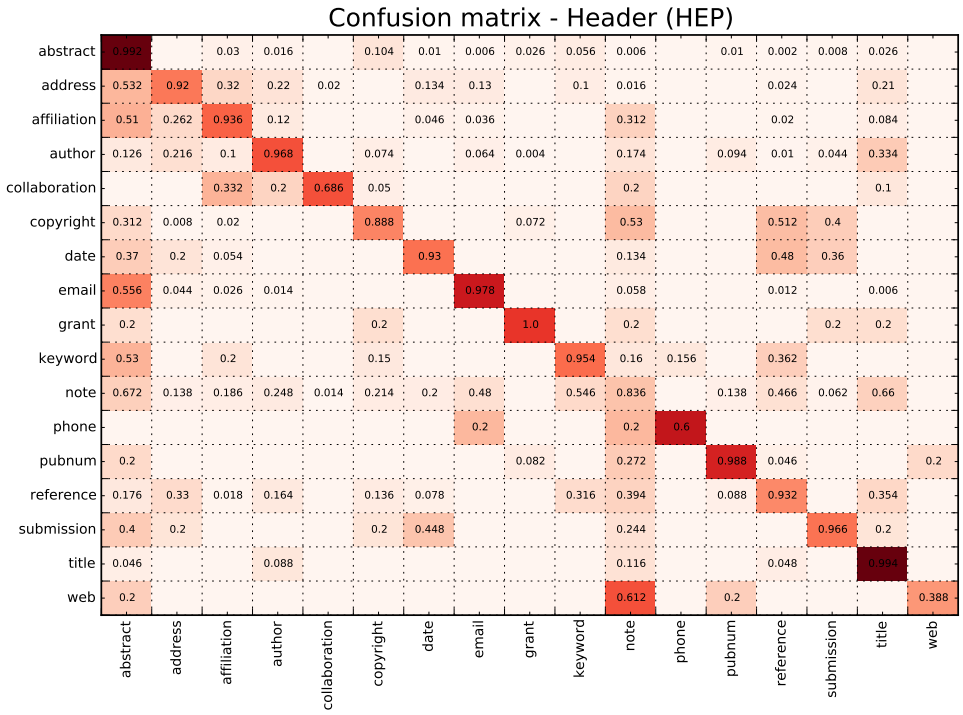
10.1 Header model - HEP dataset



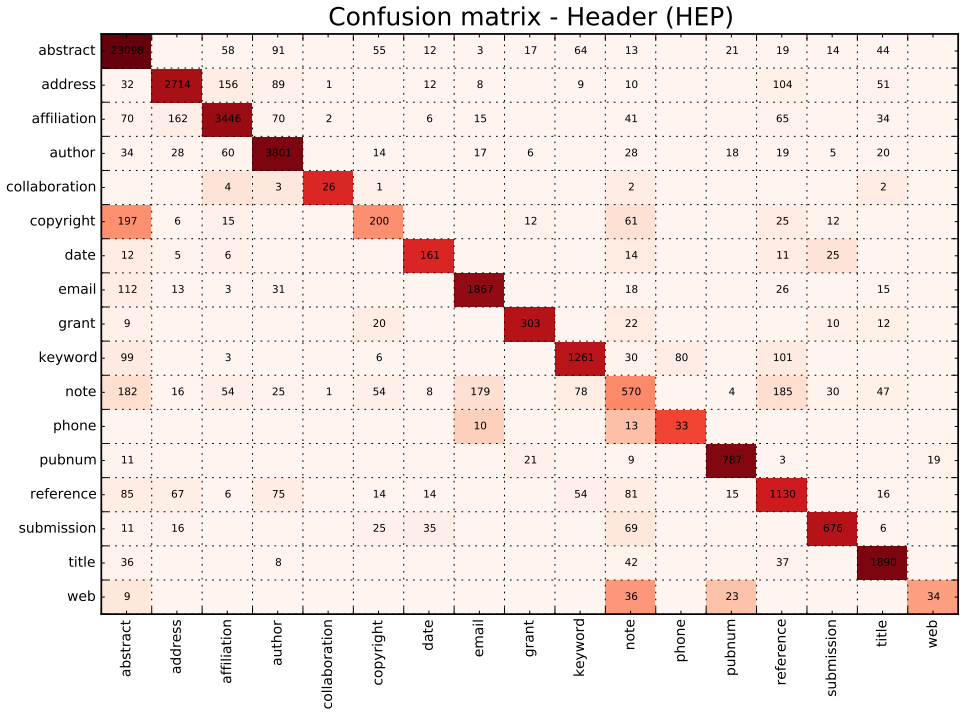
(1.62)



(1.63)

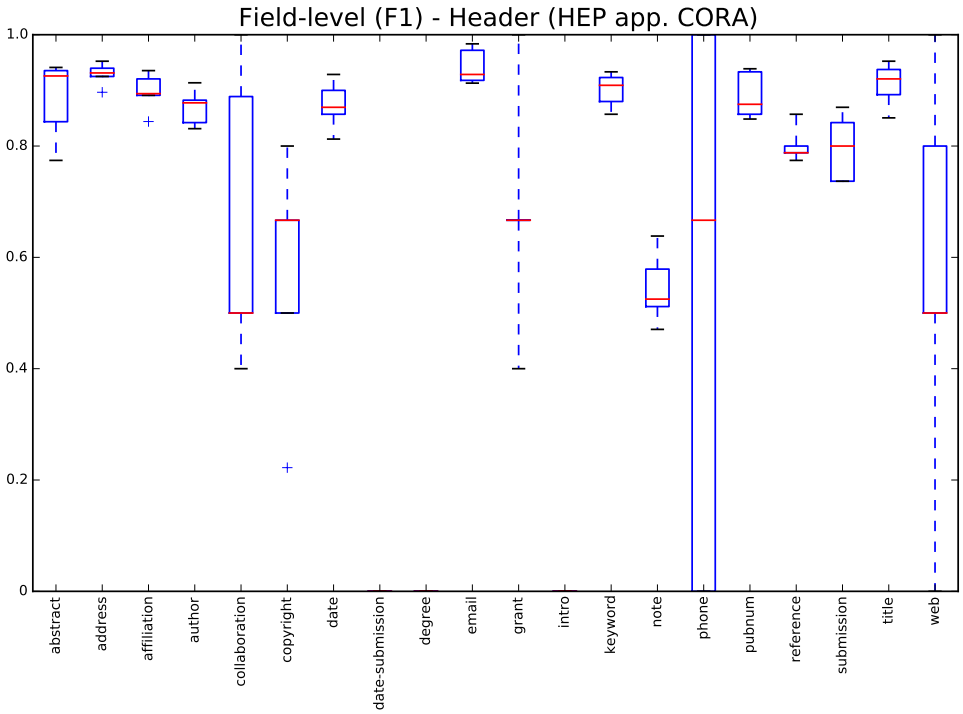

Confusion matrix - Header (HEP)
abstract 0.992 0.03 0.016 0.104 0.01 0.006 0.026 0.056 0.006 0.01 0.002 0.008 0.026
address 0.532 0.92 0.32 0.22 0.02 0.134 0.13 0.1 0.016 0.024 0.21
affiliation 0.51 0.262 0.936 0.12 0.046 0.036 0.312 0.02 0.084
author 0.126 0.216 0.1 0.968 0.074 0.064 0.004 0.174 0.094 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.64)

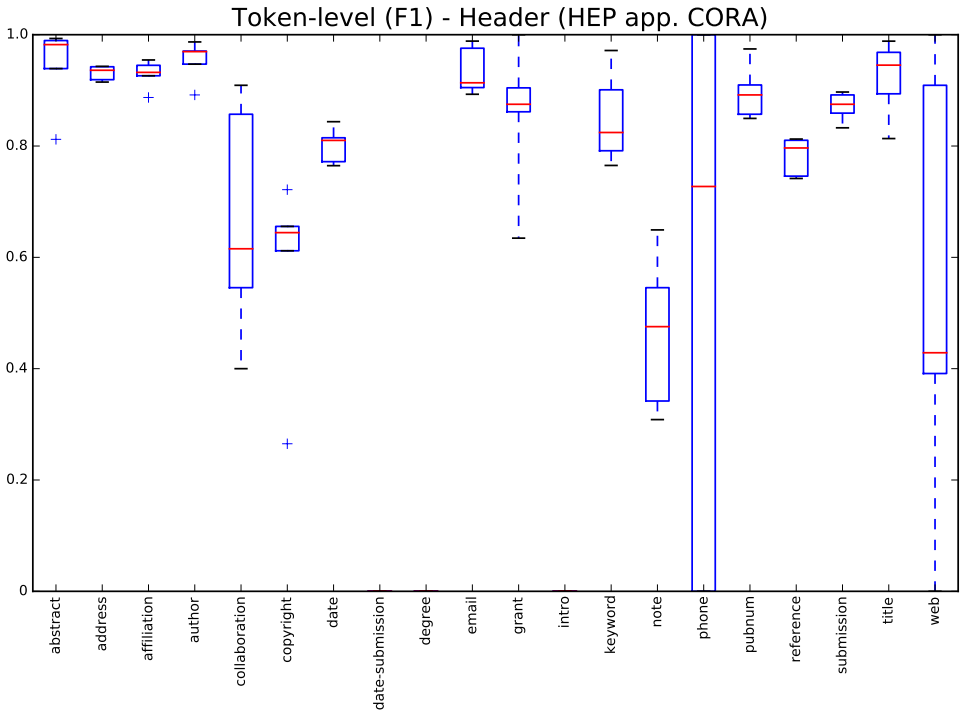

Confusion matrix - Header (HEP)
abstract 23098 58 91 55 12 3 17 64 13 21 19 14 44
address 32 2714 156 89 1 12 8 9 10 104 51
affiliation 70 162 3446 70 2 6 15 41 65 34
author 34 28 60 3801 14 17 6 28 18 19 5 20
collaboration 4 3 26 1 2 2
copyright 197 6 15 200 12 61 25 12
date 12 5 6 161 14 11 25
email 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.65)

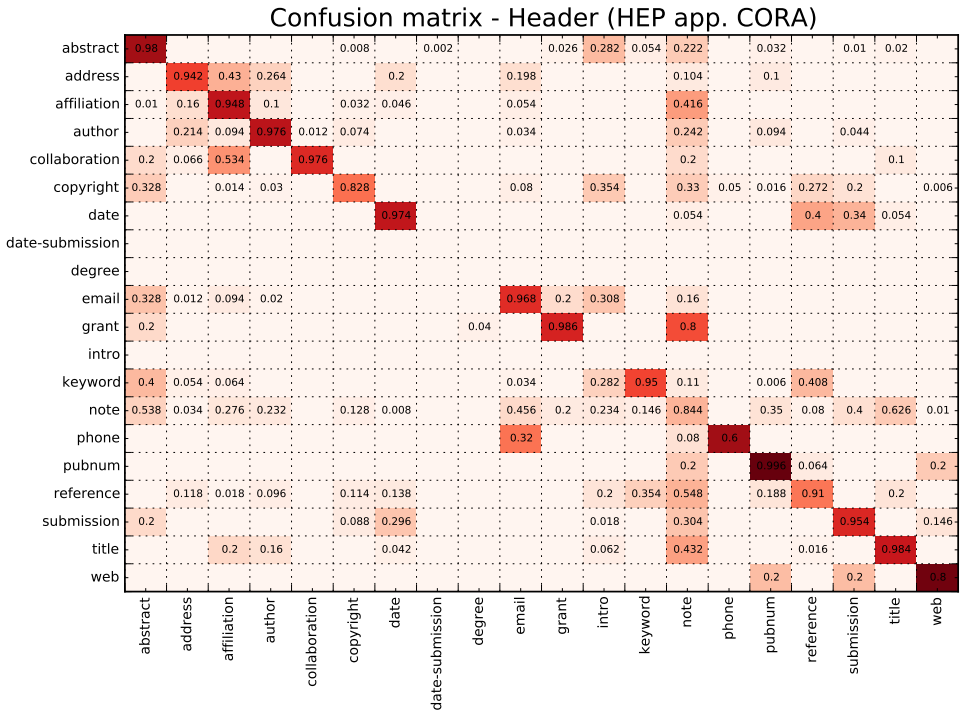
10.2 Header model - HEP dataset appending CORA dataset



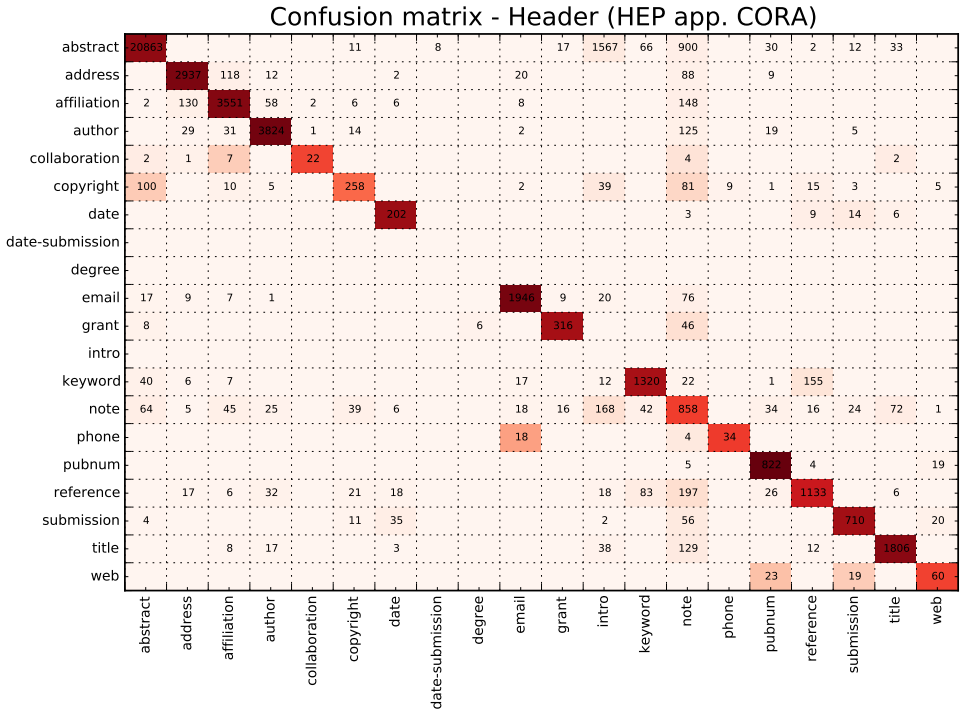
(1.66)



(1.67)

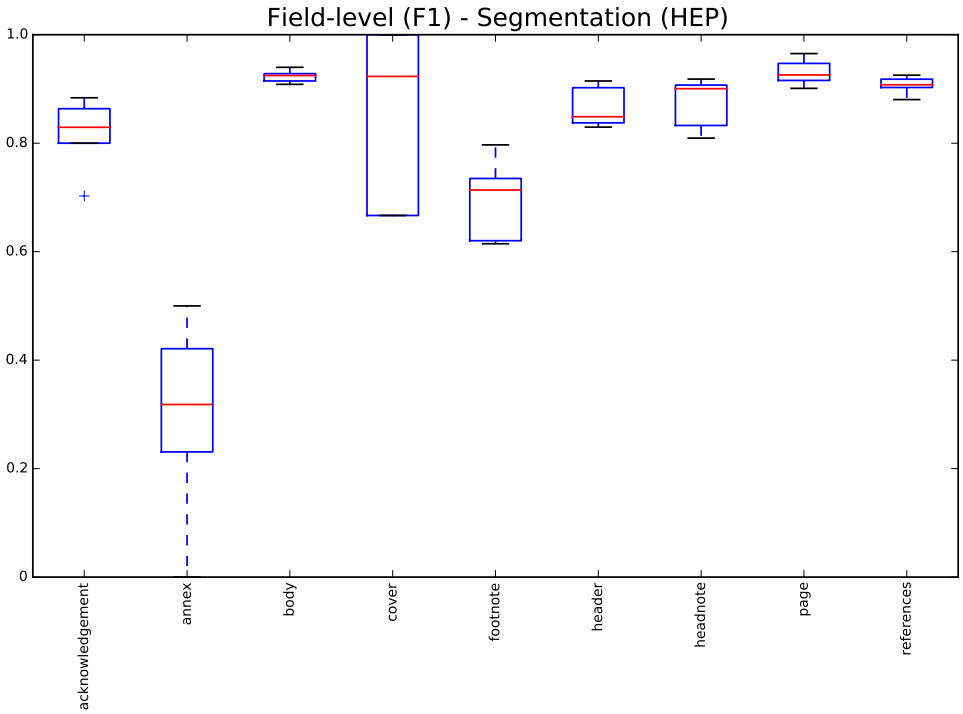

Confusion matrix - Header (HEP app. CORA)
abstract
address
affiliation
author
collaboration
copyright
date
date-submission
degree
email
grant
intro
keyword
note
phone
pubnum
reference
submission
title
web

(1.68)

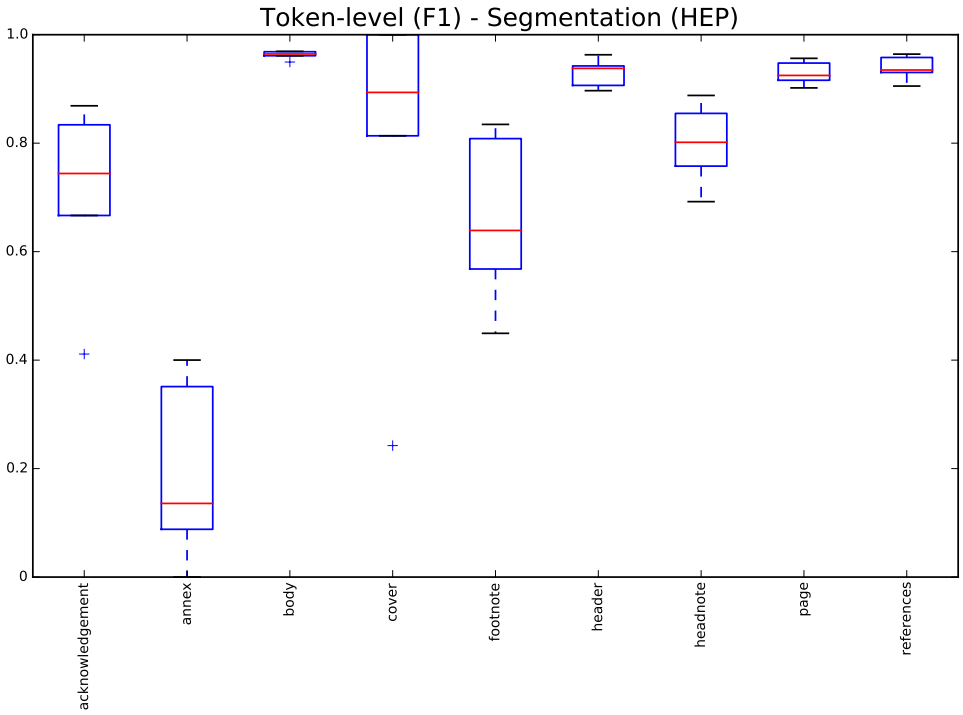

Confusion matrix - Header (HEP app. CORA)
abstract
address
affiliation
author
collaboration
copyright
date
date-submission
degree
email
grant
intro
keyword
note
phone
pubnum
reference
submission
title
web

(1.69)

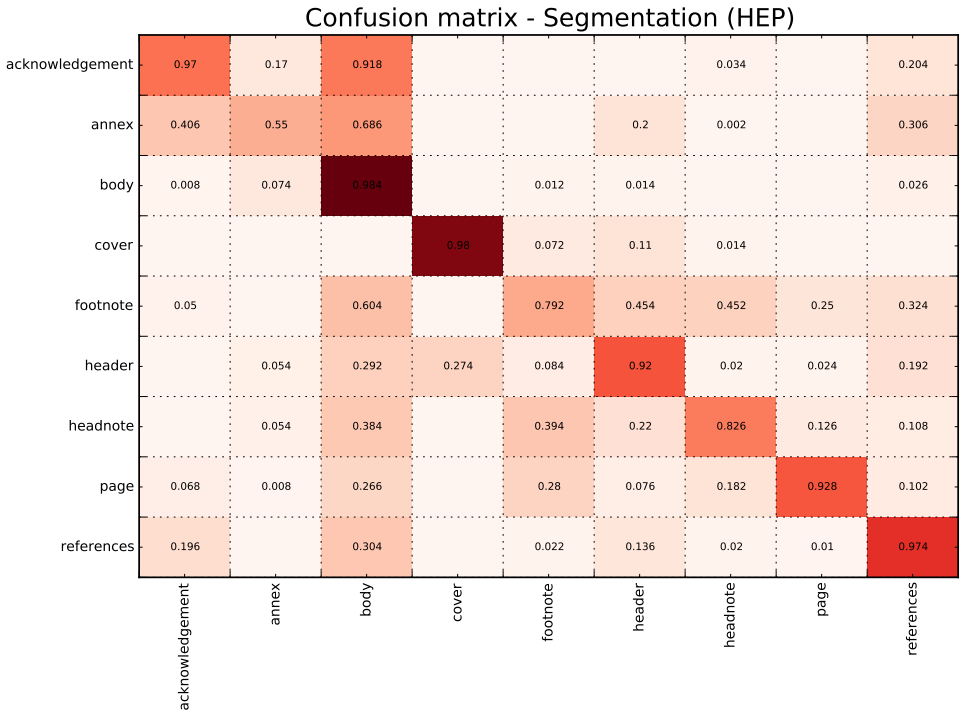
10.3 Segmentation model - HEP dataset



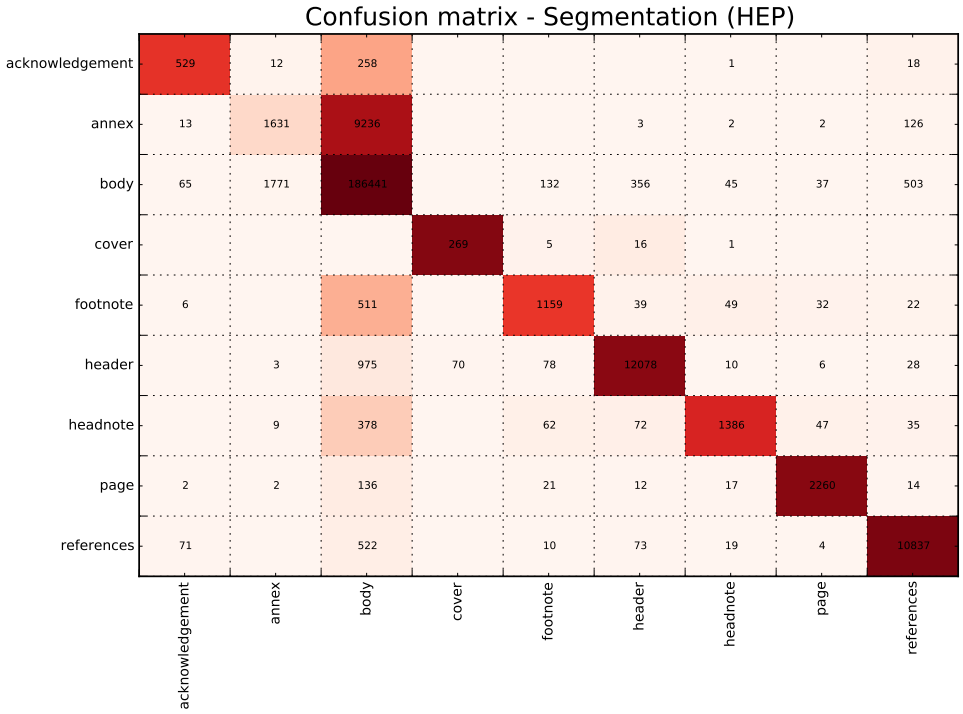
(1.70)



(1.71)

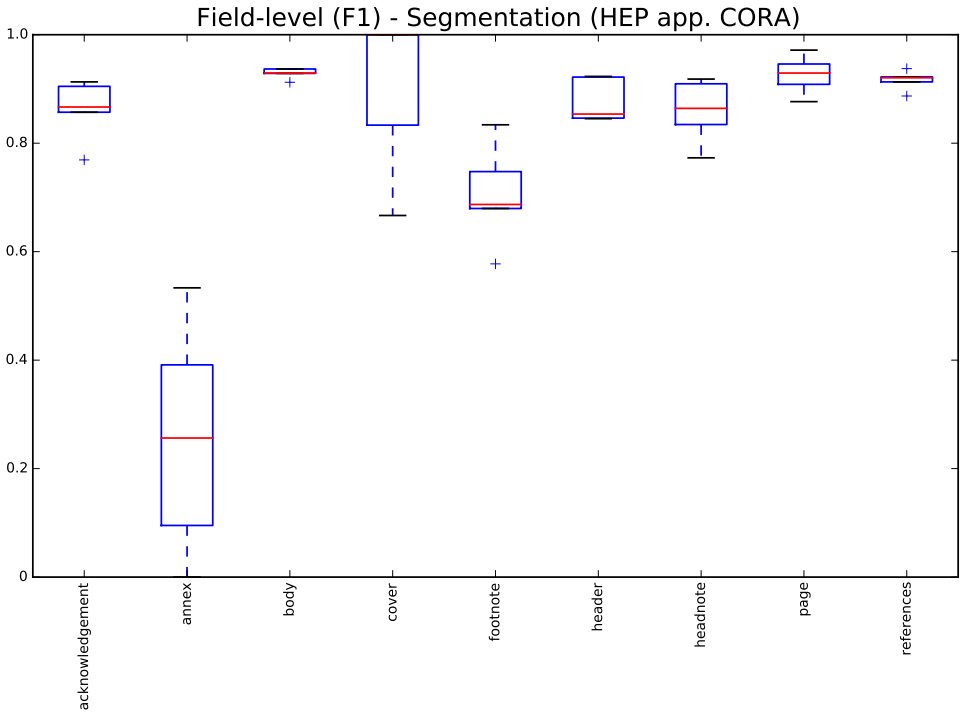

Confusion matrix - Segmentation (HEP)
acknowledgement 0.97 0.17 0.918 0.034 0.204
annex 0.406 0.55 0.686 0.2 0.002 0.306
body 0.008 0.074 0.984 0.012 0.014 0.026
cover 0.98 0.072 0.11 0.014
footnote 0.05 0.604 0.792 0.454 0.452 0.25 0.324
header 0.054 0.292 0.274 0.084 0.92 0.02 0.024 0.192
headnote 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
acknowledgement annex body cover footnote header headnote page references

(1.72)

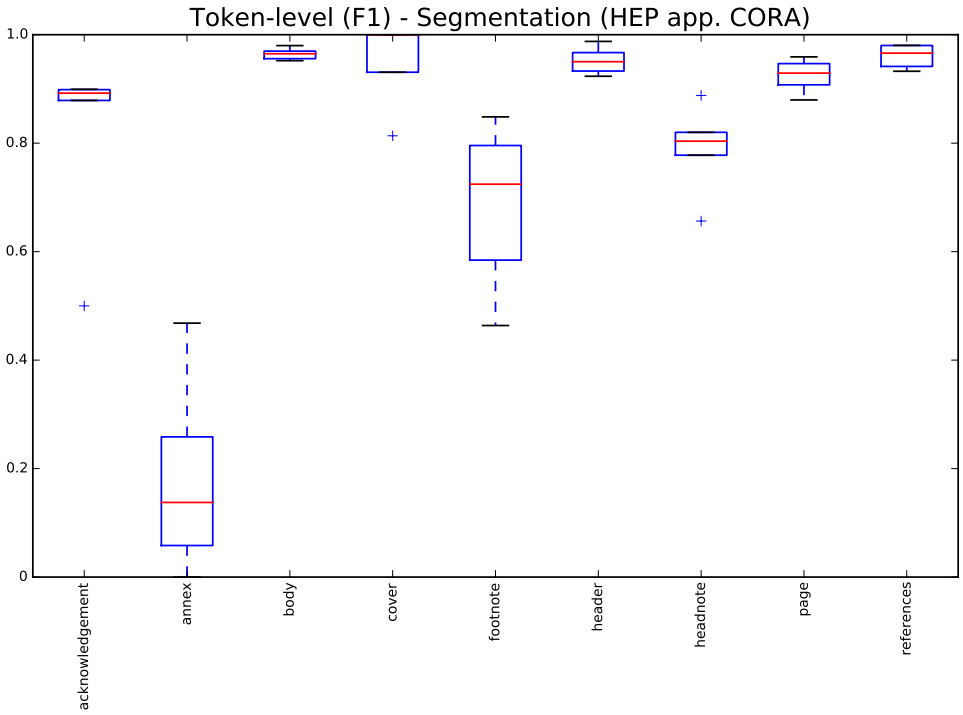

Confusion matrix - Segmentation (HEP)
acknowledgement 529 12 258 1 18
annex 13 1631 9236 3 2 2 126
body 65 1771 186441 132 356 45 37 503
cover 269 5 16 1
footnote 6 511 1159 39 49 32 22
header 3 975 70 78 12078 10 6 28
headnote 9 378 62 72 1386 47 35
page 2 2 136 21 12 17 2260 14
references 71 522 10 73 19 4 10837
acknowledgement annex body cover footnote header headnote page references

(1.73)

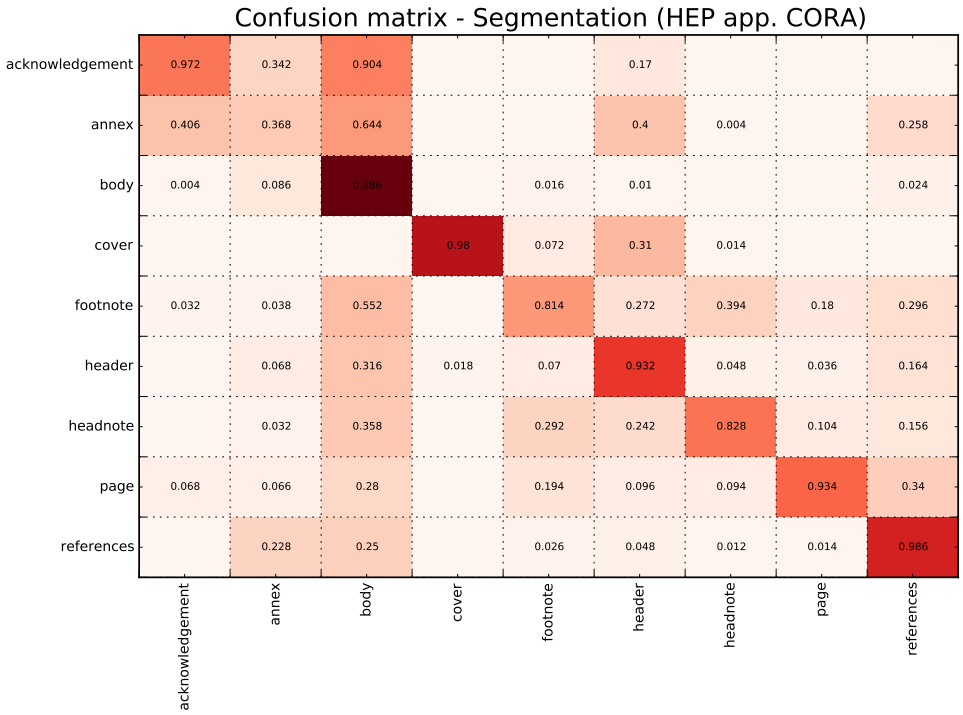
10.4 Segmentation model - HEP dataset appending CORA dataset



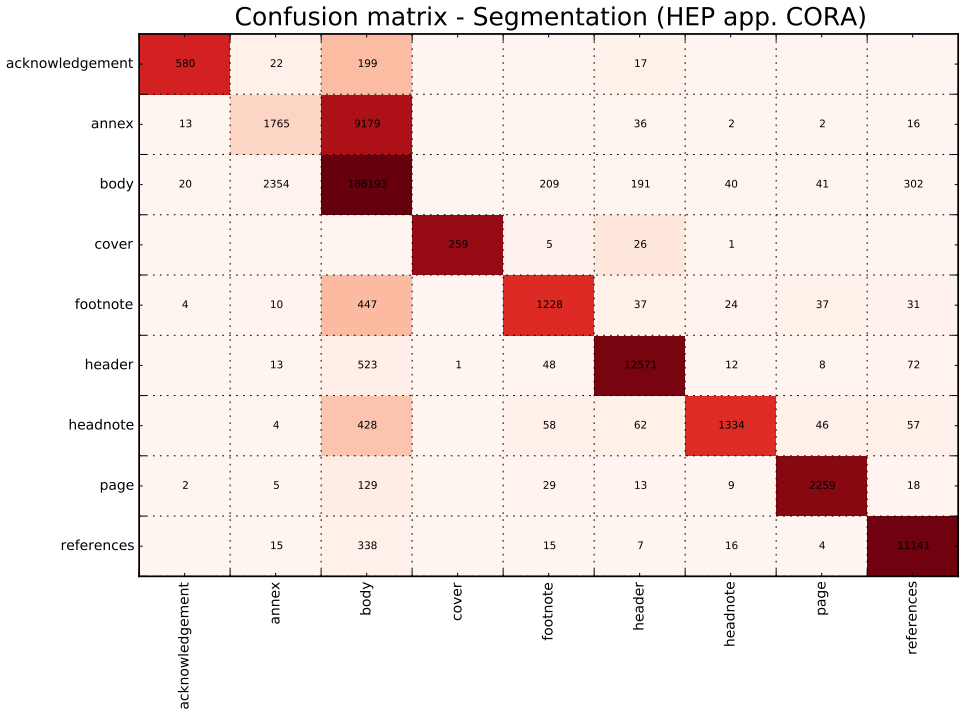
(1.74)



(1.75)


Confusion matrix - Segmentation (HEP app. CORA)
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
acknowledgement
annex
body
cover
footnote
header
headnote
page
references

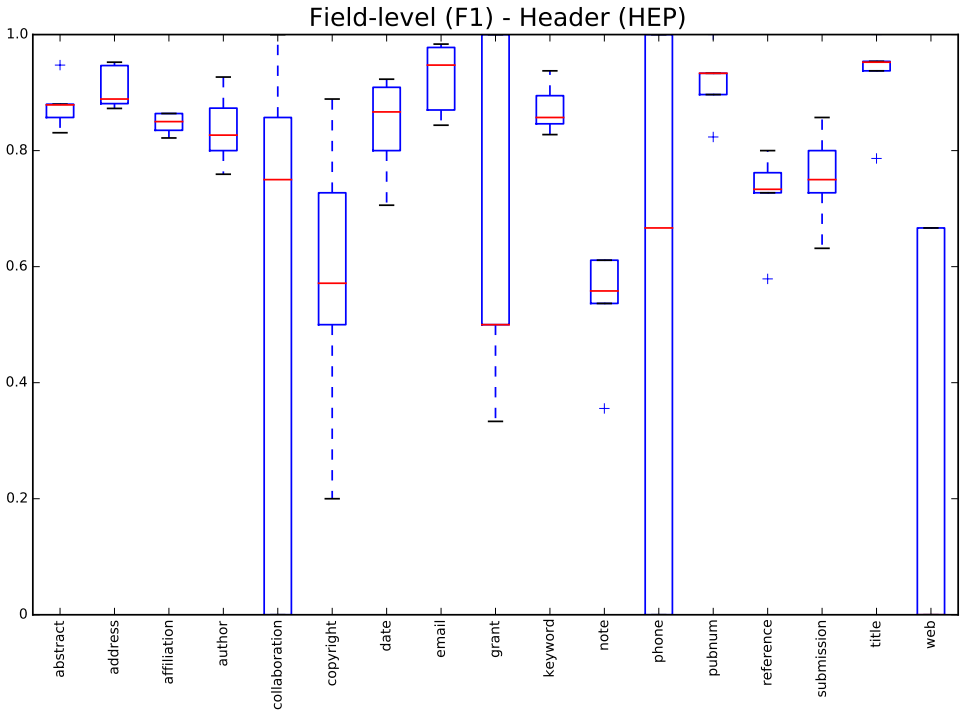
(1.76)


Confusion matrix - Segmentation (HEP app. CORA)
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
acknowledgement
annex
body
cover
footnote
header
headnote
page
references

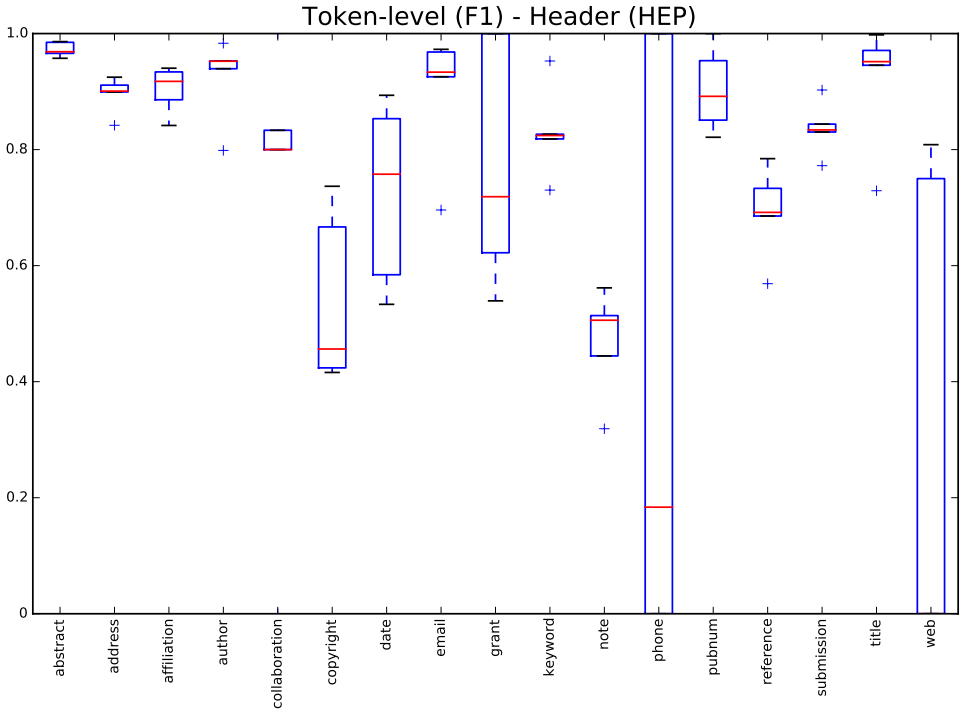
(1.77)

11 Dictionaries + stop words

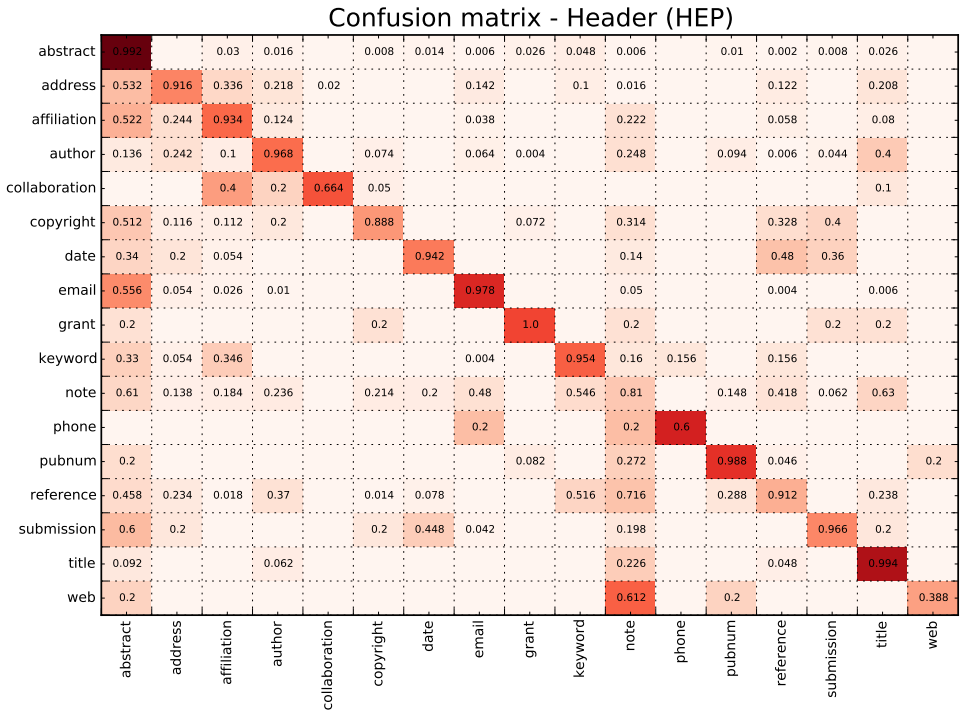
11.1 Header model - HEP dataset



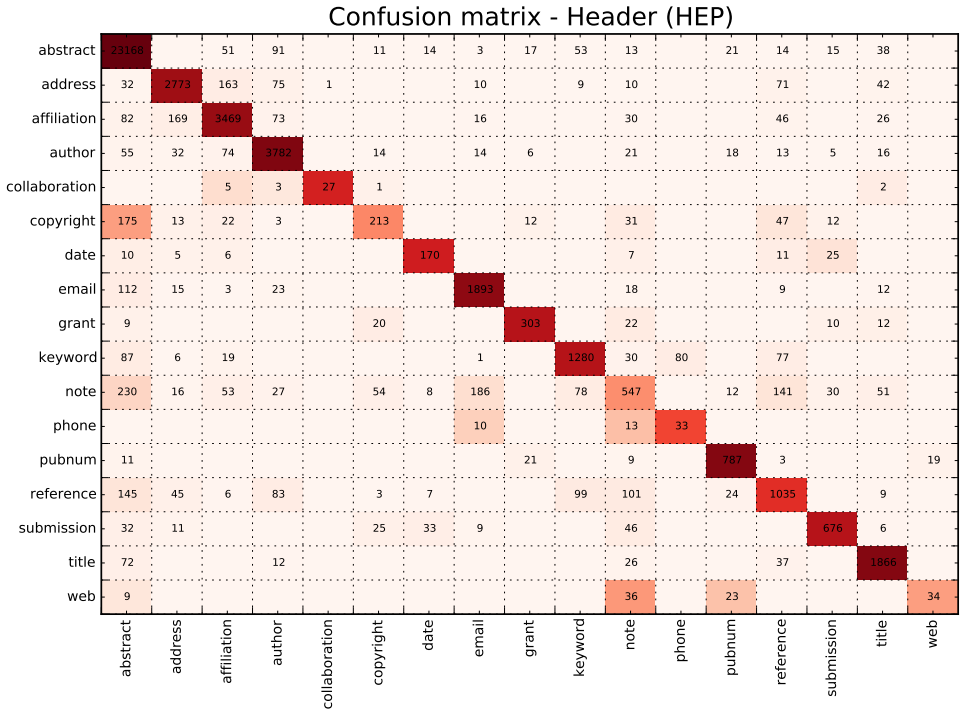
(1.78)



(1.79)

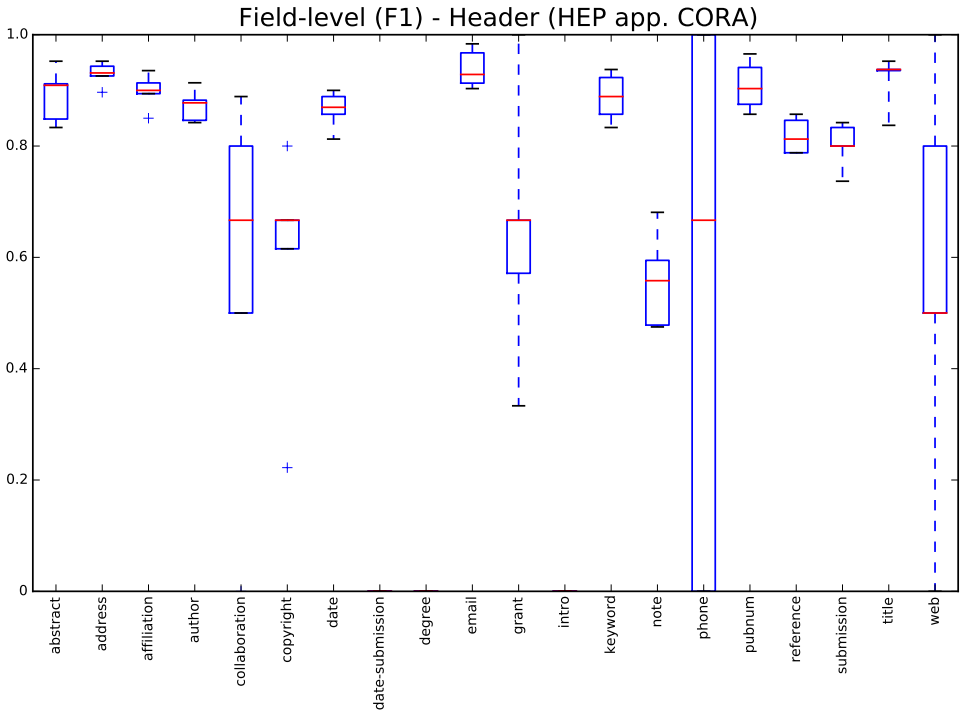

Confusion matrix - Header (HEP)
abstract 0.992 0.03 0.016 0.008 0.014 0.006 0.026 0.048 0.006 0.01 0.002 0.008 0.026
address 0.532 0.916 0.336 0.218 0.02 0.142 0.1 0.016 0.122 0.208
affiliation 0.522 0.244 0.934 0.124 0.038 0.222 0.058 0.08
author 0.136 0.242 0.1 0.968 0.074 0.064 0.004 0.248 0.094 0.006 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.80)

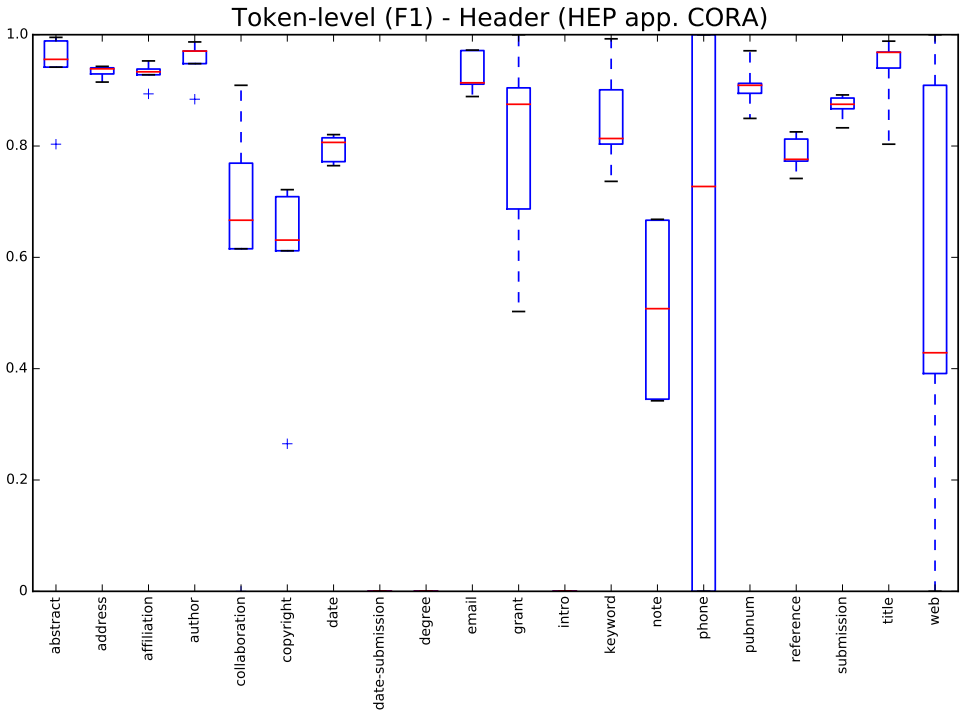

Confusion matrix - Header (HEP)
abstract 23168 51 91 11 14 3 17 53 13 21 14 15 38
address 32 2773 163 75 1 10 9 10 71 42
affiliation 82 169 3469 73 16 30 46 26
author 55 32 74 3782 14 14 6 21 18 13 5 16
collaboration 5 3 27 1 2
copyright 175 13 22 3 213 12 31 47 12
date 10 5 6 170 7 11 25
email 112 15 3 23 1893 18 9 12
grant 9 20 303 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
abstract address affiliation author collaboration copyright date email grant keyword note phone pubnum reference submission title web

(1.81)

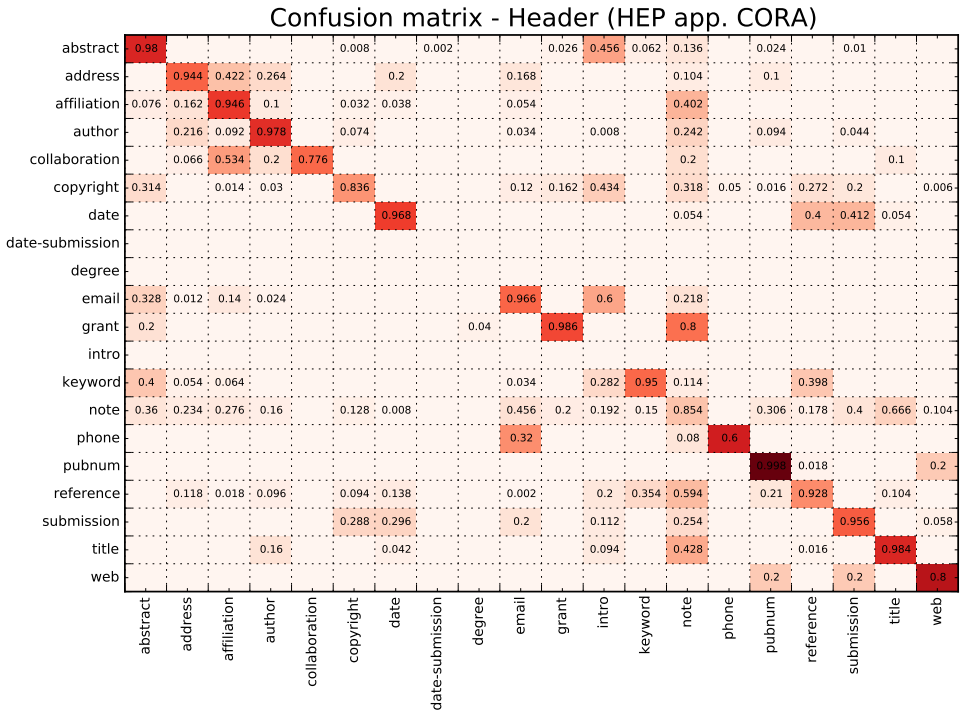
11.2 Header model - HEP dataset appending CORA dataset



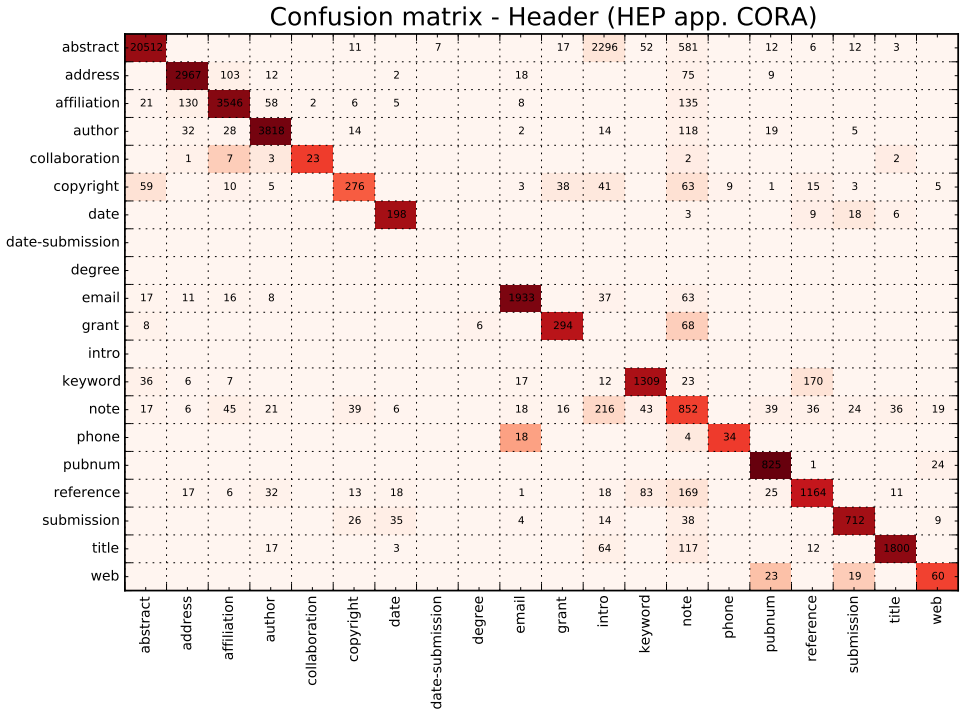
(1.82)



(1.83)

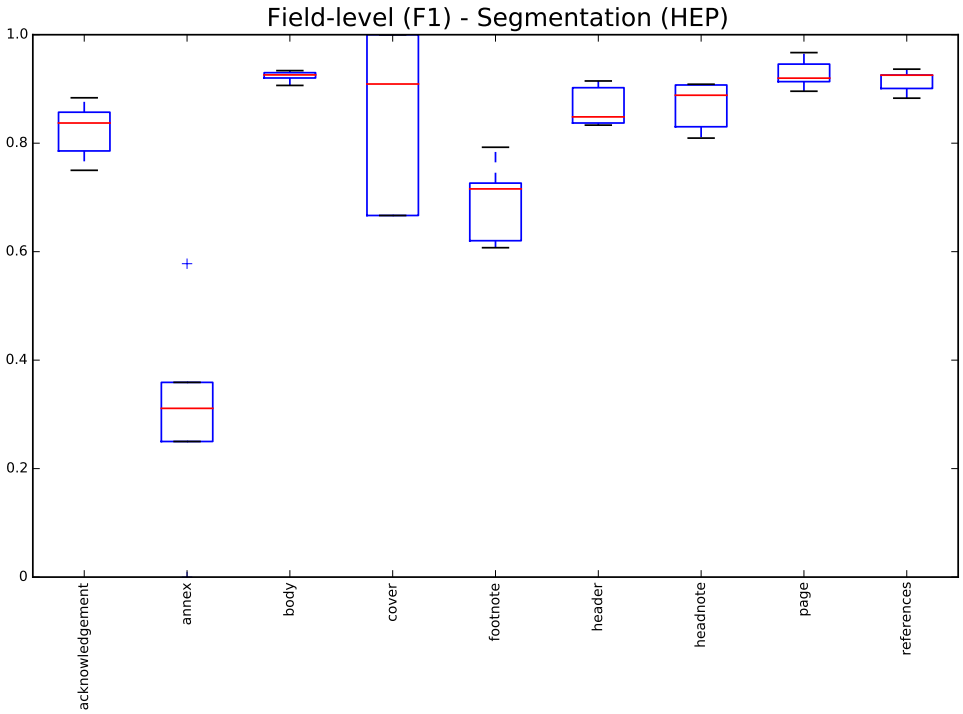

Confusion matrix - Header (HEP app. CORA)
abstract
address
affiliation
author
collaboration
copyright
date
date-submission
degree
email
grant
intro
keyword
note
phone
pubnum
reference
submission
title
web

(1.84)

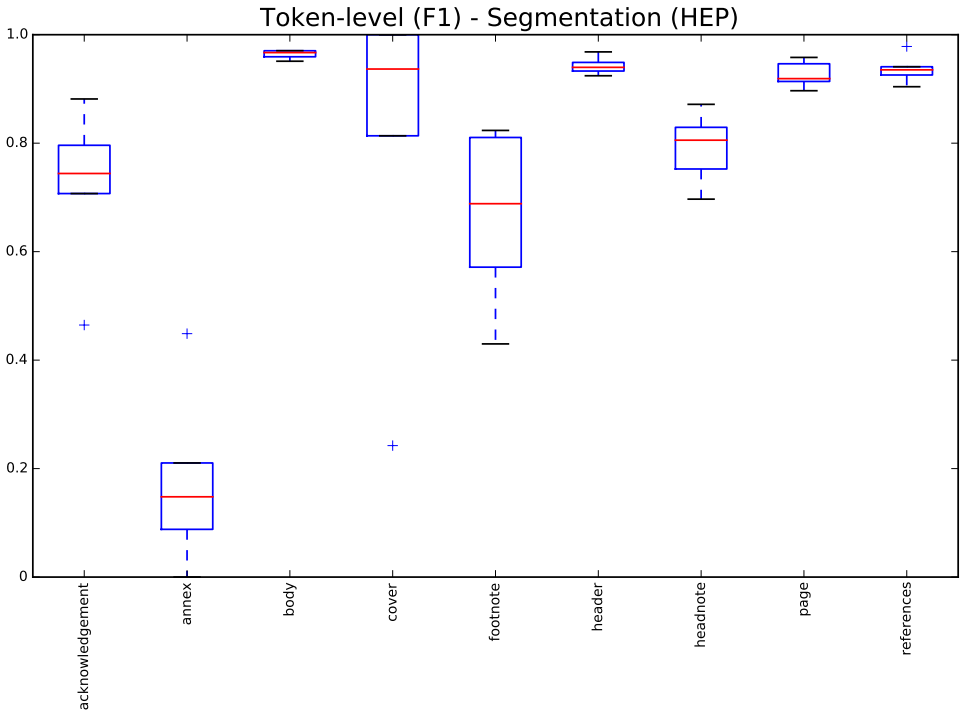

Confusion matrix - Header (HEP app. CORA)
abstract
address
affiliation
author
collaboration
copyright
date
date-submission
degree
email
grant
intro
keyword
note
phone
pubnum
reference
submission
title
web

(1.85)

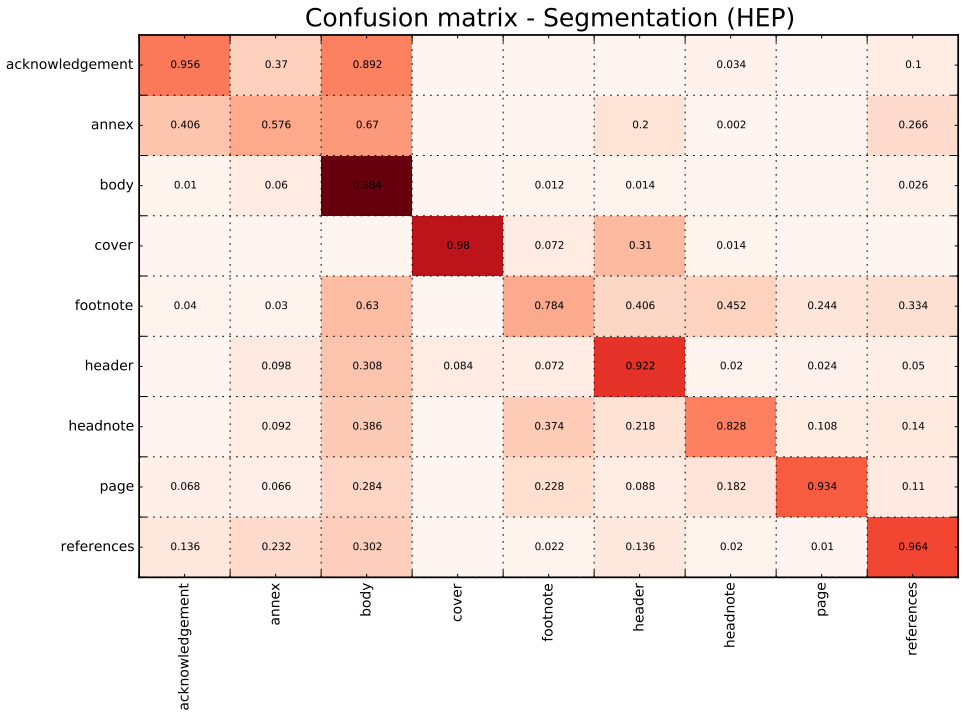
11.3 Segmentation model - HEP dataset



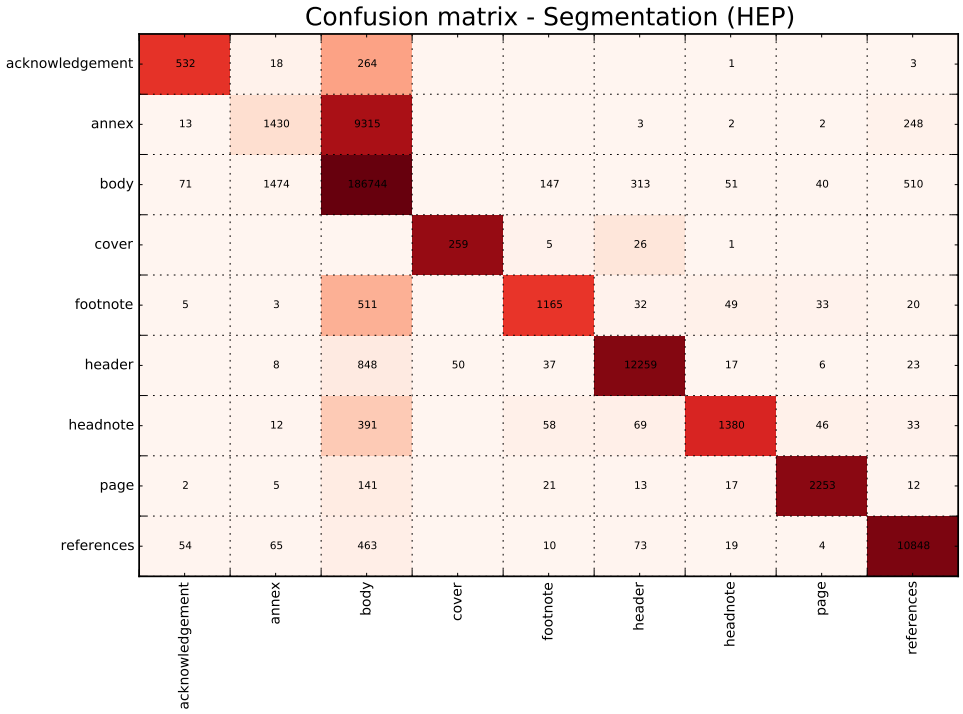
(1.86)



(1.87)

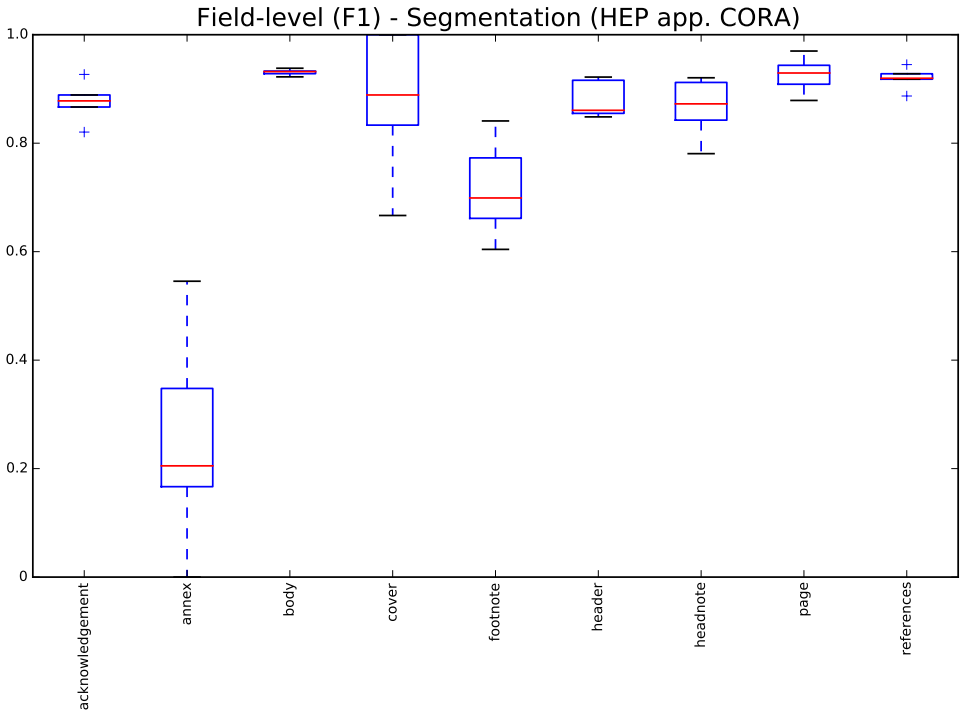

Confusion matrix - Segmentation (HEP)
acknowledgement 0.956 0.37 0.892 0.034 0.1
annex 0.406 0.576 0.67 0.2 0.002 0.266
body 0.01 0.06 0.984 0.012 0.014 0.026
cover 0.98 0.072 0.31 0.014
footnote 0.04 0.03 0.63 0.784 0.406 0.452 0.244 0.334
header 0.098 0.308 0.084 0.072 0.922 0.02 0.024 0.05
headnote 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
acknowledgement
annex
body
cover
footnote
header
headnote
page
references

(1.88)

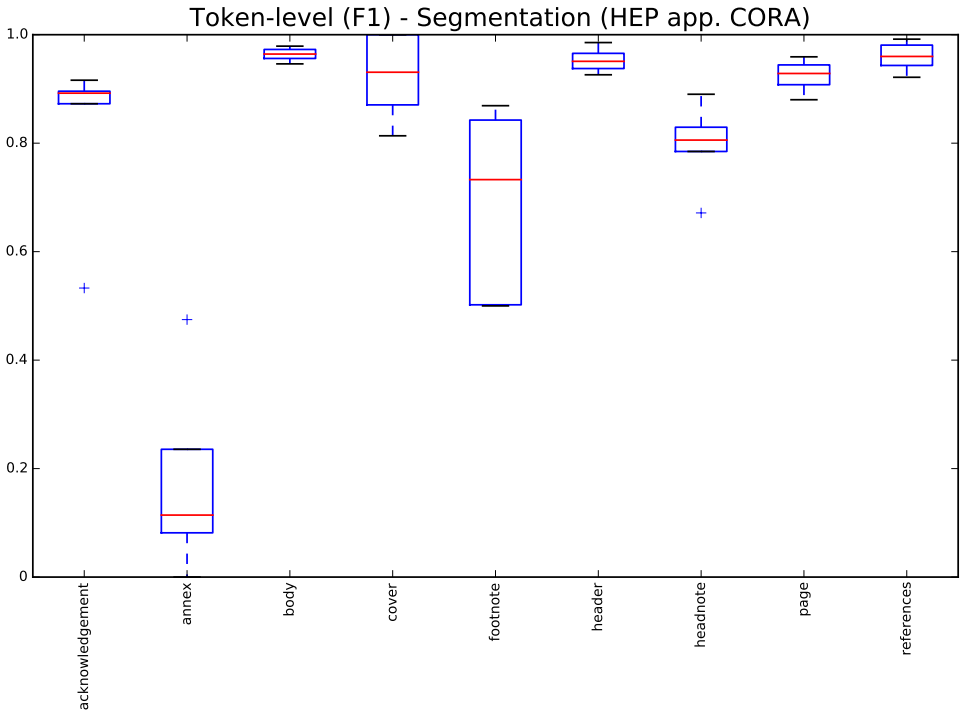

Confusion matrix - Segmentation (HEP)
acknowledgement 532 18 264 1 3
annex 13 1430 9315 3 2 2 248
body 71 1474 186744 147 313 51 40 510
cover 259 5 26 1
footnote 5 3 511 1165 32 49 33 20
header 8 848 50 37 12259 17 6 23
headnote 12 391 58 69 1380 46 33
page 2 5 141 21 13 17 2253 12
references 54 65 463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

(1.89)

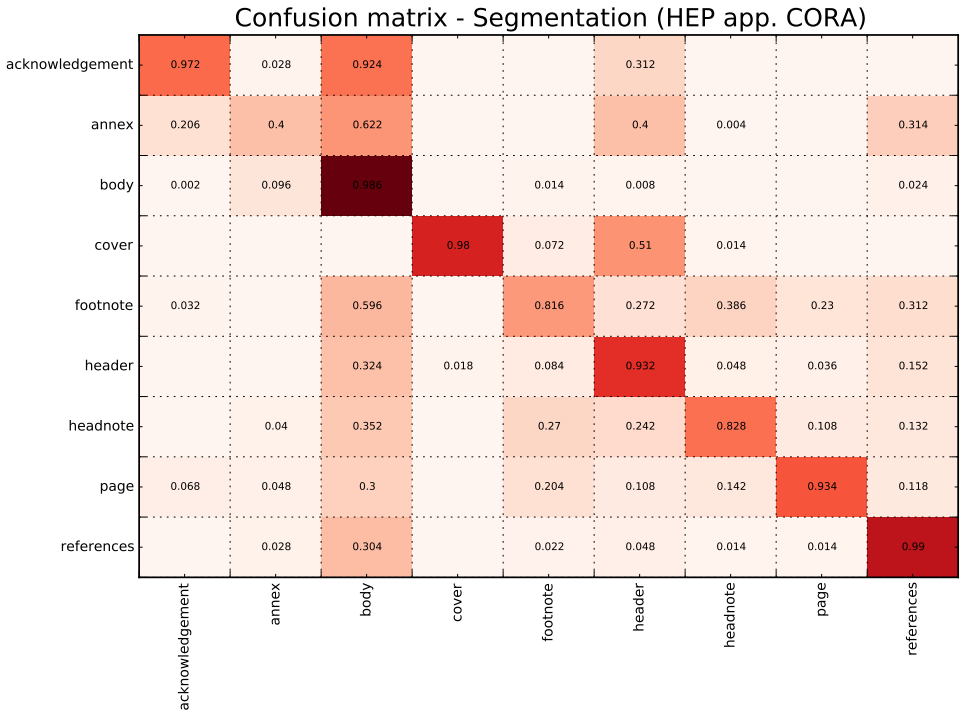
11.4 Segmentation model - HEP dataset appending CORA dataset



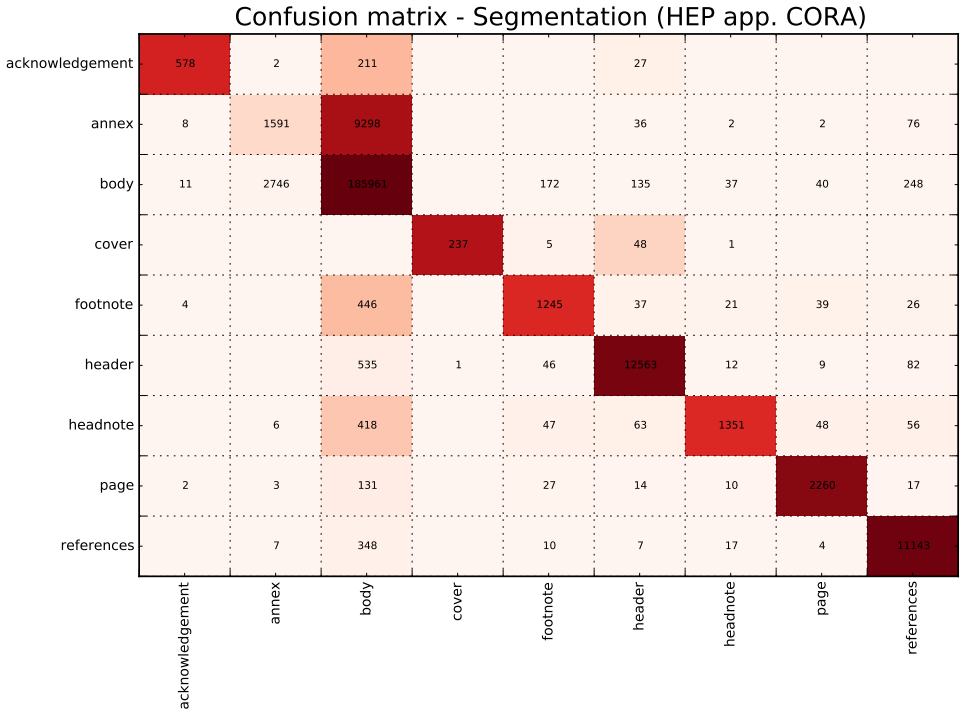
(1.90)



(1.91)


Confusion matrix - Segmentation (HEP app. CORA)
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
acknowledgement
annex
body
cover
footnote
header
headnote
page
references

(1.92)


Confusion matrix - Segmentation (HEP app. CORA)
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
acknowledgement
annex
body
cover
footnote
header
headnote
page
references

(1.93)