

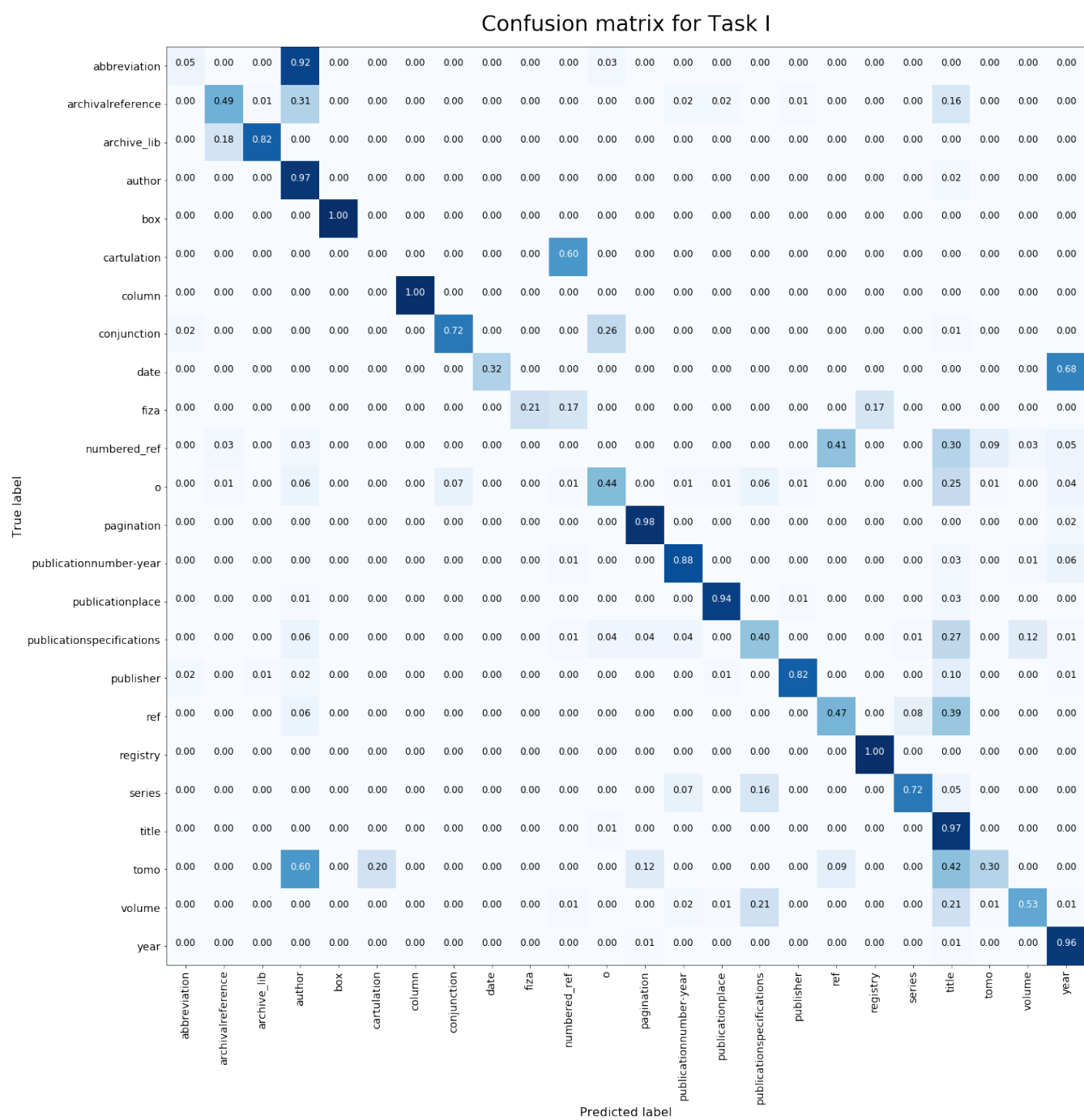


Figure S1. Confusion matrix for Task 1 calculated on the validation dataset. A cell color ranges from white (0%) to dark blue/green (100%).

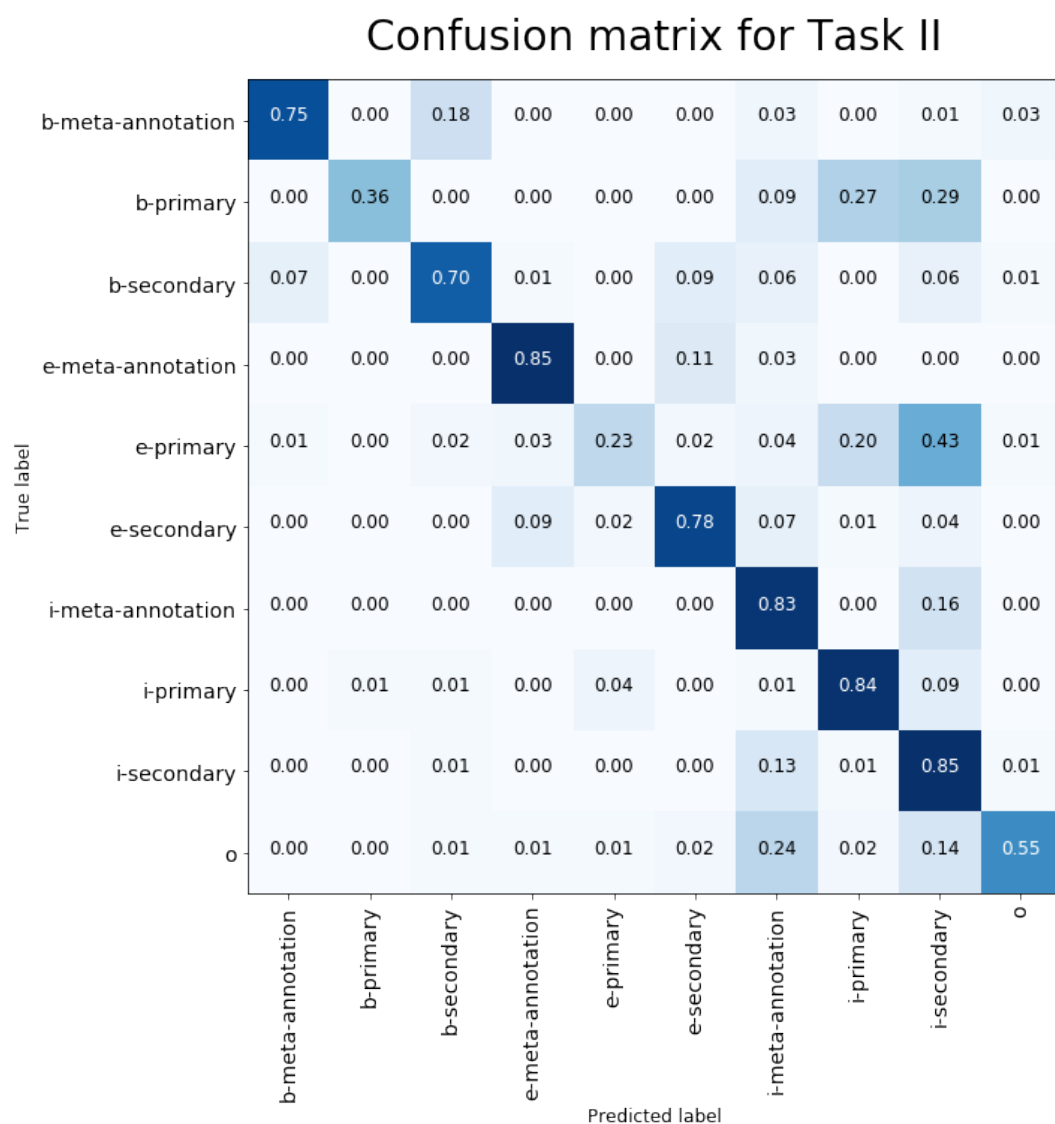


Figure S2. Confusion matrix for Task 2 calculated on the validation dataset. A cell color ranges from white (0%) to dark blue/green (100%).

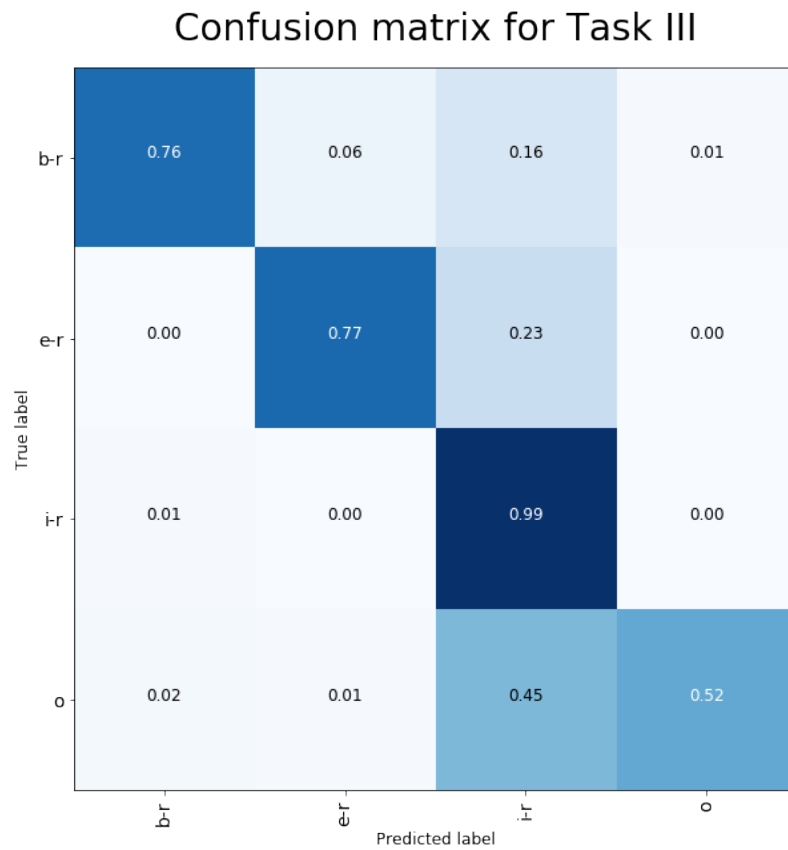


Figure S3. Confusion matrix for Task 3 calculated on the validation dataset. A cell color ranges from white (0%) to dark blue/green (100%).

	NN	CRF	diff	support
archivalreference	0.6107	0.3065	0.3042	328
publisher	0.8552	0.6550	0.2002	937
registry	0.8537	0.7000	0.1537	35
series	0.7561	0.6234	0.1327	43
volume	0.6286	0.5017	0.1269	335
o	0.5732	0.4531	0.1201	1379
publicationspecifications	0.4025	0.3163	0.0862	329
publicationplace	0.9494	0.8692	0.0802	1555
publicationnumber-year	0.8820	0.8106	0.0714	665
abbreviation	0.0684	0.0000	0.0684	87
year	0.9328	0.8677	0.0651	1601
author	0.9317	0.8707	0.0610	4581
title	0.9519	0.8929	0.0590	13744
pagination	0.9650	0.9485	0.0165	1154
attachment	0.0000	0.0000	0.0000	0
box	1.0000	1.0000	0.0000	6
cartulation	0.0000	0.0000	0.0000	10
column	1.0000	1.0000	0.0000	6
numbered_ref	0.0000	0.0000	0.0000	87
filza	0.3409	0.3409	0.0000	70
ref	0.3333	0.3571	-0.0238	36
conjunction	0.5733	0.6036	-0.0303	120
date	0.4286	0.5294	-0.1008	19
tomo	0.3030	0.4746	-0.1716	33
archive_lib	0.4242	0.7179	-0.2937	17
avg / total	0.8966	0.8263	0.0703	27177

Table S1. Differences in the F1 validation scores for Task 1 between the Neural Network and CRF baseline models.

	NN	CRF	diff	support
b-primary	0.4706	0.076923	0.393677	45
b-meta-annotation	0.7487	0.518519	0.230181	280
e-primary	0.3047	0.09375	0.21095	167
o	0.6507	0.47012	0.18058	1035
e-secondary	0.8083	0.667759	0.140541	1583
b-secondary	0.7362	0.614865	0.121335	779
i-meta-annotation	0.7917	0.685087	0.106613	8457
e-meta-annotation	0.8242	0.719266	0.104934	879
i-secondary	0.858	0.769477	0.088523	13682
i-primary	0.6857	0.609982	0.075718	270
avg / total	0.8151	0.710366	0.104734	27177

Table S2. Differences in the F1 validation scores for Task 1 between the Neural Network and CRF baseline models.

	NN	CRF	diff	support
o	0.6463	0.408621	0.237679	1035
b-r	0.8044	0.740454	0.063946	1104
i-r	0.9767	0.959393	0.017307	23893
e-r	0.8286	0.850733	-0.022133	1145
avg / total	0.9509	0.924946	0.025954	27177

Table S3. Differences in the F1 validation scores for Task 1 between the Neural Network and CRF baseline models.