A comprehensive description of information extraction tasks on four heterogeneous visually rich document dataset

Ritesh Sarkhel, Arnab Nandi
Department of Computer Science and Engineering
The Ohio State University

The Medical Article Record Groundtruth Dataset

Dataset overview:

No. of document types:	9
No. of documents:	1553
Dataset properties:	Scanned articles of biomedical journals from the NLM
Publicly available?	Yes
Download link:	link

Information extraction task overview:

No. of distinct named entities:	4
List of named entities:	See below

Index	Named entity	Description
1	*"Title"*	Name of the article
2	*"Contribution"*	The main contribution of the scientific article (as claimed by the authors)
3	*"Authors"*	Name of the researchers who authored the article
4	*"Affiliations"*	The institutions each author belonged to during the publication of the article

The Tobacco Litigation Dataset

Dataset overview:

No. of document types:	10
No. of documents:	3482
Dataset properties:	Scanned documents from publicly available records of a lawsuit
Publicly available?	Yes
Download link:	link

Information extraction task overview:

No. of distinct named entities:	34
List of named entities:	See below

Index	Named entity	Description
1	*"Sender"*	The sender of an email/letter/memo
2	*"Receiver"*	The receiver of an email/letter/memo
3	*"Timestamp"*	Date of communication via an email/letter/memo
4	*"Receiver's address"*	The listed address of the receiver of an email/letter
5	*"Sender's affiliation"*	Position and affiliation of the sender of an email/letter
6	*"Subject of communication"*	The subject of an email/letter/memo
7	*"Reported tobacco product"*	Name of the tobacco product mentioned in a news/scientific report
8	*"Affiliated company's name"*	Name of a tobacco product manufacturer mentioned in a news/scientific report
9	*"Action items"*	Todo's as mentioned in an email/letter/memo
10	*"Action deadline"*	Deadlines of the todo list
11	*"News headline"*	Title of a news article/scientific report

Index	Named entity	Description
12	*“Publication date”*	Date of publication of a news article/scientific report
13	*“Reported side effects”*	Harmful effects of tobacco usage as mentioned in a news article/scientific article
14	*“Age group of the affected population”*	The majority age group of the affected population
15	*“Potential lawsuit details”*	Plaintiffs of a potential lawsuit as mentioned in a news article
16	“*Plaintiff’s reported location*”	City/location of the plaintiffs of a potential lawsuit as mentioned in a news article
17	*“Recommended regulatory action”*	Potential prohibitory actions to be introduced as mentioned in a news article
18	*“Author”*	Contributors of a news article/scientific report
19	*“Authors’ affiliations”*	Affiliations of the contributors
20	*“Sources cited”*	References cited by a scientific report
21	*“Key findings”*	Main discoveries of the report
22	*“Date of publication”*	The date when a news article/scientific report was published
23	*“Job applicant”*	Name of the applicant applying for a job
24	*“Age”*	Age of the applicant as mentioned in the resume
25	*“Gender”*	Gender of the job applicant as mentioned in the resume
26	*“Position applied”*	Title of the applied job
27	*“Employment history”*	Past work experience of the applicant
28	*“Highest qualification”*	Highest educational qualification of the job applicant
29	*“Honors”*	Awards received by the applicant
30	*“Membership of societies”*	List of scientific societies the applicant is a member of
31	*“Publications”*	List of publications listed by the applicant in their resume
32	*“Report title”*	Title of a submitted scientific report

Index	Named entity	Description
33	*“Report number”*	Serial no. of the report
34	*“Submission date”*	The date when the report was submitted

The Brains Dataset

Dataset overview:

No. of document types:	36
No. of documents:	~ 1.4M
Dataset properties:	Scanned documents for keeping vital patient information in ER
Publicly available?	Will be made publicly available
Download link:	N/A

Information extraction task overview:

No. of distinct named entities:	10
List of named entities:	See below

Index	Named entity	Description
1	*"Patient's name"*	Name of the patient under medical care
2	*"Age"*	Patient's age when admitted
3	*"Gender"*	Patient's gender
4	*"Code"*	Resuscitation status of the patient
5	*"Admit date"*	The day when the patient was first admitted to the ER
6	*"Room number"*	Room number where the patient is now in the hospital
7	*"Diagnosis"*	Latest diagnosis of the patient made by the medical doctor responsible for the patient
8	*"Medical history"*	Past medical records of the patient
9	*"Dietary restrictions"*	Known food allergies
10	*"Consulting physician"*	Name of the medical doctor responsible for the patient

The NIST Special Dataset

Dataset overview:

No. of document types:	20
No. of documents:	5595
Dataset properties:	Tax documents from the IRS 1040 package of 1988
Publicly available?	Yes
Download link:	link

Information extraction task overview:

No. of distinct named entities:	1379
List of named entities:	See below

Each of the form fields in this dataset corresponds to a named entity in the information extraction task defined on this dataset. A complete list of these named entities can be found at https://s3.amazonaws.com/nist-srd/SD6/SD06_users_guide.pdf