Workshop Proposal: Document Intelligence 2022

Workshop Title

Document Intelligence 2022

Topic

The operation of organizations revolves around documents: project reports, planning documents, technical specifications, financial statements, meeting minutes, legal agreements, contracts, resumes, purchase orders, invoices, and many more. Cultural heritage from recent and far away past is also locked in document images. The ability to automatically read, understand and interpret these documents, referred to here as *Document Intelligence* (DI), is challenging due to not only many domains of knowledge involved, but also their complex formats and structures, internal and external cross references deployed, and even less-than-ideal quality of scans and OCR oftentimes performed on them. This workshop aims to explore and advance the current state of research and practice in answering these challenges.

Audience

Researchers, practitioners, and students of any relevant discipline needed for document understanding, including but not limited to data mining, knowledge discovery, machine learning, information retrieval, natural language processing, layout recognition, computer vision, and other technologies are welcome.

Relevance to KDD

While many research areas relevant to DI are separately covered in other conferences such as ICML, NeurIPS, SIGIR, ACL, ICDAR, CVPR and AAAI, the interdisciplinary nature of the pursuit calls for a venue where all interested parties can freely exchange novel ideas, discuss open problems, and compare competing approaches. We believe KDD is the perfect venue for this discussion, and the fruit of the workshop can also benefit the core missions of the conference – to facilitate more effective data mining and knowledge discovery.

Why Now in KDD 2022?

Many key aspects relevant to document understanding have seen rapid progress in the recent years, such as the large deep learning models that are proven successful on computer vision and natural language processing tasks, the effective transfer learning that enables rapid knowledge transfer in multimodal and multilingual settings, and service maturity in democratizing these technologies for end users. Many document-related tasks of long-standing interest are now re-examined with the new computational capabilities, including table/chart understanding and generation, image captioning, visual question answering. This is the perfect time to reexamine our past roadblocks, confront new challenges and opportunities, and outline multidisciplinary solutions in a highly relevant venue such as KDD 2022. The location in Washington, DC is especially appealing because we will be able to invite funders like DARPA to talk about document intelligence-related needs.

Organizing Committee

Douglas Burdick is a Research Staff Member at **IBM Research** - Almaden currently working on the application of AI and machine learning to document understanding, which includes table extraction and understanding in addition to inferring document structure. His document understanding work is incorporated into the IBM Watson Compare & Comply and IBM Watson Discovery products. His other research focuses on the creation of financial knowledge graphs from unstructured data sources such as regulatory filings and analyst reports, which includes interpretation of tabular data from these documents. He has contributed to Apache SystemML and OpenII data integration toolkit, and co-organizes the DSMM workshop series (co-located with SIGMOD). He received his PhD in Computer Science from the University of Wisconsin - Madison.

Benjamin Han is the Principal Science Manager leading the research and development of the natural language services on **Microsoft Azure Cognitive Services**. His current focus is to democratize the state-of-the-art NLP research to serve customers at scale. His research interests include language detection, key phrase extraction, sentiment analysis, named entity recognition, entity linking, coreference resolution, relation extraction, knowledge base construction, summarization, and question answering. During his time at Microsoft he has been a Principal Scientist in Satori (knowledge graph) and Bot Framework (conversational AI). Before that he was a Research Staff Member in the Multilingual NLP Technologies group at IBM TJ Watson Research Center for over a decade, working on all stages of information extraction technologies that power products such as IBM Watson Knowledge Studio and Watson NLU. He had participated in many government organized projects/competitions such as TREC, RADAR, ACE, GALE and TACKBP, published in conferences such as ICME, ICoS, NAACL, IJCAI, AAAI and SIGIR, and organized the Knowledge Graph tutorial in KDD 2018.

Dave Lewis is an Executive Vice President for AI Research, Development, and Ethics at **Reveal**-**Brainspace.** Prior to joining Brainspace, he was variously a freelance consultant, corporate researcher (Bell Labs, AT&T Labs), research professor, and software company co-founder. Dave has published more than 40 peer-reviewed scientific publications and 9 patents. He was elected a Fellow of the American Association for Advancement of Science in 2006 for foundational work in text categorization, and won a Test of Time Award from ACM SIGIR in 2017 for his paper w/ Gale introducing uncertainty sampling.

Sandeep Tata is a Software Engineer at **Google Research** and leads a research group on information extraction. Sandeep has published dozens of peer-reviewed research articles across a variety of disciplines including data management, data mining, natural language processing, and information extraction. Sandeep's research work has impacted billions of people through research-focused enhancements to products like Google Drive, Gmail, and Google Assistant. He has served on the program committees for VLDB, ICDE, CIKM, and as a senior program committee member for KDD. He served on the organizing committee for WSDM 2016. Prior to Google Research, Sandeep was a Research Staff Member at IBM's Almaden Research Center. He has a PhD from the University of Michigan.

Program Committee Chair/Main Contact

Ani Nenkova is a Principal Scientist at **Adobe Research**, leading the Adobe-Maryland part of the Document Intelligence Lab. Ani's work is broadly in the area of language technology, including text quality prediction, summarization and named entity recognition. She has co-organized several workshops on summarization and improving text readability. Ani was program co-chair of NAACL 2016 and currently serves as editor-in-chief for TACL.

Invited Speakers

William Wang, Duncan and Suzanne Mellichamp Chair in Artificial Intelligence and Designs
Representation, reasoning and question answering from tables
https://sites.cs.ucsb.edu/~william/index.html

Shruthi Rijhwani, CMU, Forbes 30 under 30 in Science awardee
OCR, entity extraction and linking for low resource scenarios
https://shrutirij.github.io

Panelists

We will organize two panels. In addition to the researchers below, we plan to invite two program mangers from funding agencies.

(1) Representation learning for document intelligence
  a. Edouard Grave (Facebook)
  b. Vlad Morariu (Adobe Research)
  c. Furu Wei (Microsoft Research Asia)
  d. Dongwei Zhang (Microsoft Research Asia)

(2) Document intelligence for cultural preservation
  a. David Smith (Northeastern University)
  b. Antonis Anastasopoulos (George Mason University)
  c. Taylor Berg-Kirkpatrick (University of California at San Diego)
  d. Ryan Cordell (University of Illinois at Urbana-Champaign)

Program Committee/Reviewers (confirmed)

Freddy Chua (Ernst & Young)
Daniel Campos (UIUC)
Jonathan Degange (Ernst & Young)
Yasuhisa Fujii (Google)
Sean Goldberg (Microsoft)
Jiuxiang Gu (Adobe Research)
Beliz Gunel (Stanford University)
Rajiv Jain (Adobe Research)
Amanda Jones (H5)
Chen-Yu Lee (Google)
James Mayfield (Johns Hopkins University)
Graham McDonald (University of Glasgow)
Lesly Miculicich (Microsoft)
Feifei Pan (Rensselaer Polytechnic Institute)
Brian Price (Adobe Research)
Herbert Roitblat (Mimecast)
Baoguang Shi (Microsoft)
Jyothi Vinjumur (Walmart)
Guoxin Wang (Microsoft)
Li Yang (Google Research)
Qi Zheng (UIUC)
Chris Tensmeyer (Adobe Research)

Length of Workshop

Full day.

Program

We propose a full-day workshop containing the following elements:

1. Invited talks presenting research results, practices, and challenges on DI.
2. Paper sessions of contributed and reviewed work on the topics of DI.
3. Panel discussion reflecting on the challenges and directions on DI.

We will hold a morning and an afternoon session, each consisting of an invited talk, workshop lightning talk presentations and posters, and a panel discussion.

Past Workshops

Document Intelligence Workshop 2019 (https://sites.google.com/view/di2019) was organized along NeurIPS 2019 in Vancouver, BC, Canada, and it was a huge success. The workshop was well-attended, filling the room to the 150 capacity with people standing. The workshop received close to 50 initial abstract submissions and finally 38 paper submissions. Based on the reviews and discussions, 19 papers were accepted resulting in a 50% acceptance rate. The accepted papers are available at https://openreview.net/group?id=NeurIPS.cc/2019/Workshop/Document_Intelligence. There was a Best Paper selected, and the discussion on the challenges and opportunities in the space has resulted in a paper published in KDD Explorations (PDF).

Document Intelligence Workshop 2021, virtual, Singapore (https://document-intelligence.github.io/DI-2021/) was collocated with KDD 2021. There were 10 accepted papers, 5 posters and 6 invited talks at the workshop.

Estimated Participation for DI 2022

We estimate DI 2022 will receive the following participation:

- Number of attendees: 50-70.
- Number of submissions: 40 papers, 20 accepted.

Possible Venues for Publicity

- Mailing lists: SIG-IR list, Google Groups
- Social media: LinkedIn, Twitter, Facebook, etc.
- Company websites: Microsoft, Google, IBM

Past CFP

See below

https://sites.google.com/view/di2019/call-for-papers-cfp

https://document-intelligence.github.io/DI-2021/