A Report on the First Workshop on Document Intelligence (DI) at NeurIPS 2019

Hamid Motahari, Nigel Duffy
EY AI Lab
Palo Alto, CA
USA
{hamid.motahari,nigel.p.duffy}@ey.com

Paul Bennett
Microsoft Research AI
Redmond, WA
USA
paul.n.bennett@microsoft.com

Tania Bedrax-Weiss
Google Research
Mountain View, CA
USA
tbedrax@google.com

ABSTRACT

The first workshop on Document Intelligence (DI'2019) was held on December 14, 2019 at NeurIPS 2019 conference in Vancouver, Canada. The report summarizes the workshop, with a summary of the talks, papers and posters presented, and discusses common themes, issues and open questions that came up in the workshop.

Keywords

Document reading, document understanding, document analysis, information extraction, question answering, document structure analysis, computer vision, natural text understanding.

1. INTRODUCTION

Business documents are central to the operation of business. Such documents include vendor contracts, sales agreements, loan applications, purchase orders, invoices, financial statements, employment agreements and many more. The information in such business documents is presented in natural language and can be organized in a variety of ways from straight text, multi-column formats, and a wide variety of tables. Understanding these documents is challenging due to inconsistent formats, poor quality scans and OCR (optical character recognition), internal cross references, and complex document structure. Furthermore, these documents often reflect complex legal agreements and reference, explicitly or implicitly, regulations, legislation, case law and standard business practices.

The ability to read, understand and interpret business documents, collectively referred to as "Document Intelligence", is a critical and challenging application of artificial intelligence (AI) in business. While a variety of research has advanced the fundamentals of document understanding [1,2,3,4,5], the majority have focused on document retrieval, image analysis, scientific documents and/or documents on the web which fail to capture the complexity of analysis and types of understanding needed across business documents. Realizing the vision of Document Intelligence remains a research challenge that requires a multi-disciplinary perspective spanning not only natural language processing and understanding, but also computer vision, knowledge representation and reasoning, information retrieval, and more -- all of which have been profoundly impacted and advanced by neural network-based approaches and deep learning in the last few years.

The first Document Intelligence Workshop (DI 2019) was held on December 14, 2019 in Vancouver, Canada, collocated with the Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019). The workshop received 38 high-quality submissions on a wide variety of topics addressing different aspects of Document Intelligence. All submissions were thoroughly reviewed by the DI'2019's Technical Program Committee (TPC) and external reviewers. Based on the reviews and discussions among the TPC and Workshop Chairs, 19 papers were accepted, which results in a 50% acceptance rate. Each paper was presented in the form of a poster, as well as a 5-minute spotlight presentation ahead of the poster session.

The accepted papers are compiled in the form of an online proceedings and are available at https://openreview.net/group?id=NeurIPS.cc/2019/Workshop/Document_Intelligence. The workshop was well-attended with more than 100 people in the audience filling the room to capacity. We selected the best workshop paper based on the reviews, which was "BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding", by Timo I. Denk and Christian Reisswig. The workshop web site is at https://sites.google.com/view/di2019/ where the workshop agenda, program, presentation slides, recording of the talks and the links to the papers are available.

In the rest of this report, we provide an overview of the workshop program and the presentations. Next, we present different dimensions of Document Intelligence, which were discussed in the workshop, namely applications of Document Intelligence, complexity of business documents, horizontal tasks in Document Intelligence, and research and technical challenges. We also touch on issues such as datasets, and privacy implications and considerations in Document Intelligence. We believe Sections 3-7 will be of particular use to the community in understanding motivating applications, gaining an understanding of the complexity of challenges across a range of industries where the type of business documents can vary considerably, identifying horizontal themes that may yield new research agendas, demonstrating the span of research questions in this space, and suggestions of several datasets using which researchers may be able to start investigating some of these issues.

2. Program Overview

After opening remarks by Nigel Duffy, on behalf of workshop organizers, the workshop started with an invited talk by David D. Lewis from Brainspace. He talked about artificial intelligence applications in legal e-discovery, which involves search and retrieval of information from documents sought in the process of a legal investigation. He highlighted the opportunities, successes and research challenges related to Document Intelligence in legal domain including information and knowledge representation of different document types spanning tweets, spreadsheets, emails, manuals, etc., which are subject to e-discovery, the need for language-independent linguistic processing techniques, duplicate

detection, document clustering and visualization, and summarization techniques.

Next, as another invited speaker, Prof. Nakashole from University of California, San Diego, spoke about the problem of models failing to generalize on data drawn from distributions other than that of the training data. She presented a few approaches for improving the generalization of language representations leading to the generalizability of trained AI models. Next on the agenda were the teaser presentations of accepted papers and posters, as follows: Luke de Oliveira et al. discussed decoder-transformer language models for abstractive summarization. Kehinde Aruleba presented an approach for recognition of hand-drawn finite automata from language learners; Wonseok Hwang et al. presented a post-OCR parser to structuralize textual information in images by BIO (Begin, Inside, Outside)-style of tagging text segments extracted from the OCR; Oleg Bakhteev et al. presented a cross-lingual system for plagiarism detection for English-Russian language pair; Timo I. Denk et al. presented BERTGrid language model that takes 2D layout of documents through a grid system; Christian Reisswig et al. presented chargrid-OCR for OCR based on predicting character segmentation mask together with object bounding boxes for characters; Bharat M. et al. presented a spatial recurrent neural network for document deblurring; Haoyu Dong et al. presented a multi-task learning architecture for table structure extraction in spreadsheets; Mehrdad J. Gangeh et al. presented an auto-encoder based learning architecture for denoising images of documents to improve OCR; Seunghyun Park presented a labelled receipt dataset for post-OCR parsing tasks; Kostiantyn Liepieshov et al. presented a dataset for cyrillic handwritten text recognition; Zikri Bayraktar et al. presented GilBERT language model for geological knowledge capturing; Ilias Chalkidis presented experimental results for comparing different neural architectures for contract elements extraction; Song Feng et al. presented a framework for automated dialog generation based on business documents; Petar Stojanov et al. discussed domain transfer issues and presented a transfer gap analysis technique for AI models; Emad Elwany et al. showed fine-tuning BERT (Bidirectional Encoder Representations from Transformers) language model on legal documents provides improvements in analyzing commercial contracts; Kaixuan Zhang et al. presented a Japanese document dataset with complex tabular structure, and methods for information extraction from complex text regions; Yike Qi presented DeepErase method for cleansing ink traces of text in the images of forms, showing improvements in OCR and downstream tasks; Hassan Kané et al. discussed issues with popular BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall Oriented Understudy for Gisting Evaluation) metrics and presented a metric scorecard for measuring semantic similarity.

Next in the program was the invited talk of Rajsekar Krishnamurthy from IBM Research who presented opportunities and challenges of business document understanding in the enterprise including document, task and operational complexities through use cases and examples, a document intelligence stack for tackling some of these challenges and open research problems. Following that Asli Celikyilmaz from Microsoft Research presented the last invited talk of the workshop on the topic of learning structures for text generation. She reviewed state of the art in natural language generation and highlighted the weaknesses of language models and text generators in producing coherent and informative text. She then discussed open questions in long text generation and abstractive text summarization with a glimpse of research on common sense reasoning, and its contributions in these tasks. Finally, there was a discussion session in which all workshop audience participated in outlining key opportunities and challenges of Document Intelligence, some of which is summarized in the rest of the paper.

3. Application Areas

Due to the prevalence of business documents across many business domains, the audience listed a wide range of applications for Document intelligence technologies. These areas include but are not limited to:

- Communication intelligence encompassing conversation and dialog support and other means of communications over emails, messaging, etc.
- Contract/EULA (End User License Agreement) understanding, centered around information extraction, analysis and understanding of documents capturing various business transactions in form of contracts written in natural language;
- E-discovery, as elaborated by the first invited talk;
- Tax/Expense tracking, which is centered around tax forms, attachments and supporting documentations;
- Search over document collections in the enterprise;
- Automated data visualization and analysis over the information in the business documents, and over document collections;
- Automated document-centered customer care, which is also further elaborated in one poster for reading and understanding manuals, FAQs, etc.;
- NUI (Natural User Interface)-driven task completion through automation, ranging from automating simple tasks such as creating and sending a calendar invite based on a user command in natural language, or learning-based task automation, to more complex tasks involving process and workflow automation.

4. Complexity of Business Documents

A set of key issues and challenges in Document Intelligence stems from the variety and complexity of how business documents are structured and represented. In particular, the participants highlighted the following as key set of issues manifesting the complexity of business documents:

- *Heterogeneity and type diversity*: business information, collectively referred as business documents, appear in a wide variety of forms including emails, spreadsheets, contracts, manuals, specifications, regulations, memos, other documents, and cover different formats including native digital, html, printed scanned pdf, handwritten scanned pdf (e.g., invoices), rendered pdf, etc.;
- *Structure and semantics*: The organization of content in business documents is another source of complexity. Documents range from free text, to semi-structured which contains a mix of free form and structured text, and to structured but semantically poor documents that mainly includes tabular content like spreadsheets. The layout of documents may play a role in understanding its content and add another layer of complexity including division into sections, tabs, lists, figures, etc.

Beyond structure, there is an element of domain dependence that may need to be taken into account. While entities and concepts can be identified in a domain-independent manner, their interpretation require understanding the domain and context including domain terminology, concepts and entities such as contractual parties, patent claims, etc.;

- *Relationships and composition in document collections*: business documents may have complex relationships to each other; they may be composed of other documents with sequential and hierarchical relationships, e.g., threads in emails, embedded objects in documents, and multi-part contracts. Another example of relationship is that of multiple versions of a document;

5. Horizontal Tasks

While there are differences and varieties in business documents, participants agreed that there are categories of common horizontal tasks that are applicable for building Document Intelligence solutions across different application areas:

- Key-value extraction from free form text, semi-structured and structured documents, scanned and digital documents of different domains. A prime example is contract element extraction, and information extraction from invoices, both presented in posters;
- Question answering within long documents, e.g., in digital manuals where even current solutions reduce the time people spend answering questions from a five-minute average to nearly automatic and discussed in one of the posters and in Rajsekar's invited talk.
- Summary/report generation from heterogenous collections e.g., collating scientific reports, data, and ongoing conversations to generate geo-energy reports;
- Query-based summary for e-discovery production of materials in the scope of a discovery order, as referred to in David Lewis's talk and text summary evaluation in Asli's talk;
- Document understanding and comparative analysis: A prime example is contract/EULA understanding. Typical examples include analyzing contracts for non-standard terms to answer questions such as shall I sign? Should this be escalated for advice? What is the summary of changes? What language is atypical relative to previous versions or other parties?
- Leveraging linked documents, including label propagation, thread linkage, clustering-assisted analysis, and amendments and change analysis;
- Understanding abbreviations and definitions in a specific domain or in the context of a document collection;
- Analyzing and understanding the language used in communication, including sentiments, tone, emotions to identify abusive language, etc.;
- Another task is explainability and provenance tracking, which is the ability to provide explanations of results produced by AI and provenance tracking of data and reasoning chains, both of which are of interest to many stakeholders including lawyers, judges, etc., as it was referenced in David Lewis's talk. Interestingly, simple explanations such as: "this set contains all documents that mentioned the following keywords" may be preferred by such stakeholders to more complex model-based explanations due to their simplicity to explain and their transparency in evaluation by a 3rd party.

6. Research and Technical Challenges

Document intelligence remains as a research vision and grand challenge. In particular, the participants raised a number of key research challenges remaining to be tackled, from more generic to more specific, domain-dependent ones:

- *Labeling Challenges*: while we may have access to large amounts of documents, in some applications, these are not labelled. This highlights the need for learning from unlabeled data, but also presents challenges related to labelling, which require domain expertise (e.g., in tax, legal, medical, consumer domains). At times data may be labelled by different entities and groups, and there is a need for federated learning over diverse sets of annotated data. Last but not least is the issue of data visibility and representativeness, when labelled by a particular group of labelers impacting generalization, bias and correctness of labels.
- *Universal representation for heterogeneous inputs*: getting a universal and unified representation of data and knowledge on different types of business information (chat, tweet, email, manual, contracts, spreadsheets, PowerPoint, etc.) is a challenge and key to unlocking a lot of downstream tasks in Document Intelligence.
- *Rules, Regulation and knowledge constrained interpretation*: Another key challenge is how to leverage rules, regulations and enterprise and domain knowledge in many horizontal DI tasks. As an example, using rules, regulations and knowledge base to flag issues and violations in contract language, and be able to reason over contract content. A similar challenge is using knowledge bases, domain-specific rules and regulations in training and inference of downstream DI tasks. A key other missing link is the use of common-sense knowledge in understanding and interpreting business documents including entities, and entity relations in the documents.
- *Domain Transfer*: the challenge here lies in AI models being trained on data from domains others than the application domain. The issues include data sparsity which impacts building robust models, and identifying discrepancies between the trained and target domains and distributions and having right metrics for predicting success;
- *Explainability:* this is still an open challenge for the adoption of black-box AI models, in general, and more so in the context of Document Intelligence. There are different threads of work on producing explanations for black-box deep learning models including feature importance, linkage between input and outputs, and surrogate models. Though, the architecture of deep models is complex, and for some that it's challenging to come up with explanations. For example, what are consumable explanations for subword and byte pair encoding (BPE) in deep neural network models?
- *Possible transductive bounds of error*: In some applications such as e-discovery (as referenced in David Lewis's invited talk) it's important to have error bounds. For managing the machine learning process, error bounds are also important.

- There were also discussions on methods and best practices for application of general AI modeling techniques in Document Intelligence space, including: When to do batch-training? Where does applying interactive human-AI practices work, considering privacy, expertise, etc.? When does training image-text model jointly help? When does fine-tuning BERT work? The participants discussed that sharing data, best practices and methods in the community helps in expediting progress through forming communities of interest and organizing such workshops.

7. Datasets

There was a lively discussion in the workshop on the topic of preparation and sharing of document datasets, which could be used for training and benchmarking of AI models in the Document Intelligence space, drawing parallels with the sharing of labelled images that enabled rapid progress in computer vision research, and scientific community [4]. The participants discussed opportunities and challenges, including confidentiality and privacy, for building and sharing datasets, with specific examples including:

- Avocado (see https://catalog.ldc.upenn.edu/LDC2015T03) which includes emails, calendar entries, and attachments to explore more complex interactions of linked data with corporate communications.
- Enron email dataset, available at https://www.cs.cmu.edu/~enron/.
- Parallel scientific articles, used for cross-lingual plagiarism detection, cross-lingual discovery and near duplicate discovery (https://doi.org/10.6084/m9.figshare.5382757.v2).
- FunSD dataset for form understanding (https://guillaumejaume.github.io/FUNSD/).
- Receipt OCR dataset (https://expressexpense.com/blog/free-receipt-images-ocr-machine-learning-dataset/).
- ICDAR Robust Reading Competition (https://rrc.cvc.uab.es).
- PubLayNe which is a dataset of document images, annotated with text bounding box information (https://github.com/ibm-aur-nlp/PubLayNet).
- Tobacco 800 document dataset with images, and the ground truth and handwritten notes (http://tc11.cvc.uab.es/datasets/Tobacco800_1).

In order to facilitate sharing of data from the private sector for the purpose of running academic challenges and reference datasets, the participants raised the need for EULAs that enable sharing the data publicly for creating challenges such as a contract understanding challenge.

8. Privacy

Given that the business documents are created and used in business context, privacy is discussed among the participants, as being an overarching issue impacting access, sharing, labelling, training and inference on business data. In particular, the following were noted by the participants:

- Privacy challenges of unsupervised fine-tuning BERT, e.g., related to information leakage challenges even within an enterprise on business documents from different clients;
- Dataset creation and sharing in scientific community;
- Labeling paradigms e.g., both data and labels may be scarce, and accessible by certain individuals as per contractual requirements with clients;
- Dealing with documents with PII (Personal Identifiable Information) in terms of identifying and learning with documents including PII. A related challenge is dealing with evolving definitions of PII and differences in different jurisdictions;
- Learning schemes that honor GDPR (General Data Protection Regulation), and other national and geographical privacy and data sovereignty requirements.

9. Conclusions

Document Intelligence (DI) Workshop at NeurIPS 2019, the first workshop on Document Intelligence, was extremely successful in raising interest, and bringing together a large audience of researchers and practitioners working on different topics within DI and reaching a number of disciplines and getting high quality papers from people both in academia and industry. All materials from the workshop, including papers, posters, presentations and recording of the talks are available at the DI'19 website at https://sites.google.com/view/di2019/. We hope they will be widely used and extended upon, including future work that builds on and addresses research challenges described in this report.

10. ACKNOWLEDGMENTS

Our special thanks to invited speakers, authors and workshop audience that shared ideas and research challenges in discussions.

11. REFERENCES

[1] Chia-Hui Chang, et al. A Survey of Web Information Extraction Systems, in IEEE Transactions on Knowledge and Data Engineering, 18 (10), pp. 1411-1428, 2006.

[2] ICDAR, International Conference on Document Analysis and Recognition. Link: http://icdar2019.org/.

[3] Emilio Ferrara, et al. Web Data Extraction, Applications and Techniques: A Survey. Knowledge-Based Systems, Vol. 70, Pages 301-323. Nov. 2014.

[4] Zara Nasar, et al. Information extraction from scientific articles: a survey. Scientometrics 117(3): 1931-1990. 2018.

[5] DocEng, ACM Symposium on Document Engineering. Link: https://doceng.org/.

About the authors:

There is a 5th author on the document, which we could not fit in the ACM format, as follows:

Rama Akkiraju from IBM Watson, USA. Contact her at akkiraju@us.ibm.com.