



ADAM MICKIEWICZ
UNIVERSITY
POZNAŃ

Faculty of Mathematics and Computer Science

Department of Artificial Intelligence

**Novel Methods and Datasets
for Intelligent Document Processing**

Dawid Jurkiewicz

PhD thesis in computer and information sciences

written under supervision of

prof. UAM dr hab. Filip Galiński

Poznań 2023



UNIwersytet
IM. ADAMA MICKIEWICZA
W POZNANIU

Wydział Matematyki i Informatyki
Zakład Sztucznej Inteligencji

Nowe metody i zbiory danych
do inteligentnego przetwarzania dokumentów

Dawid Jurkiewicz

Rozprawa doktorska z informatyki
napisana pod kierunkiem
prof. UAM dra hab. Filipa Gralińskiego

Poznań 2023

Abstract

The field of Intelligent Document Processing (IDP) is gaining prominence as organizations struggle to utilize their ever-growing data effectively. This thesis aims to contribute innovative solutions and datasets to the IDP domain. The focus is set on two key areas within IDP: Span Identification (SI) and Document Understanding (DU). Span Identification involves localizing relevant spans of text containing specific information, while Document Understanding encompasses various tasks related to comprehending and extracting meaningful information from visually rich documents.

Significant emphasis is placed on addressing the challenges posed by low-data scenarios, which are prevalent in various business use cases. A few-shot SI dataset and a unique approach for sub-sequence matching with few examples are proposed to address this. Besides the few-shot setting, methods for identifying and classifying propaganda spans are presented.

Furthermore, a multi-modal end-to-end Transformer-based model for Document Understanding is introduced. The model efficiently comprehends layout information, textual semantics, and visual cues present in the document and can answer various document-related questions posed in the natural language. Additionally, the first DU benchmark is proposed, allowing the community to measure the DU field's state accurately. Lastly, a challenging DU competition is showcased. The task features novel question and answer type pairs over multi-domain, multi-industry, and multi-page documents, encouraging the development of solutions with strong generalization capabilities in low-data regimes.

Streszczenie

Dziedzina inteligentnego przetwarzania dokumentów (ang. *Intelligent Document Processing*) zyskuje na znaczeniu, ponieważ organizacje mają trudności z efektywnym wykorzystaniem swoich stale przybywających danych. Niniejsza rozprawa ma na celu wnieść wkład w innowacyjne rozwiązania i zbiory danych dla dziedziny inteligentnego przetwarzania dokumentów. Nacisk kładziony jest na dwa kluczowe obszary w ramach dziedziny inteligentnego przetwarzania dokumentów: identyfikację relewantnych fragmentów tekstu (ang. *Span Identification*) i problematykę rozumienia dokumentów (ang. *Document Understanding*). Identyfikacja relewantnych fragmentów tekstu zajmuje się lokalizacją fragmentów tekstu zawierających określone informacje, podczas gdy problematyka rozumienia dokumentów obejmuje różne zadania związane z pojmowaniem i wydobywaniem istotnych informacji z dokumentów bogatych wizualnie.

Duży nacisk położony jest na zmierzenie się z wyzwaniami związanymi z małą ilością dostępnych danych, które są powszechne w różnych zastosowaniach biznesowych. Aby rozwiązać ten problem, zaproponowano zbiór danych dla identyfikacji relewantnych fragmentów tekstu na podstawie kilku przykładów oraz unikatową metodę do wyszukiwania podsekwencji na podstawie kilku przykładów. Oprócz rozwiązań bazujących na kilku przykładach, przedstawiono metody do identyfikacji i klasyfikacji fragmentów tekstu zawierających propagandę.

Ponadto wprowadzono multimodalny model oparty na architekturze Transformer dla problematyki rozumienia dokumentów. Model rozumie semantykę tekstu, cechy wizualne i strukturę dokumentu oraz potrafi odpowiadać na różne sformułowania w języku naturalnym dotyczące dokumentu. Dodatkowo zaproponowano pierwszy zestaw zbiorów danych pozwalający społeczności na dokładną obserwację postępów w dziedzinie rozumienia dokumentów. Na koniec zaprezentowano wymagający konkurs dla problematyki rozumienia dokumentów. Zadanie zawiera nowatorskie pary typów pytań i odpowiedzi dla wielodomenowych, wielobranżowych i wielostronicowych dokumentów, zachęcając do opracowywania rozwiązań, które mają znaczące możliwości uogólniania w przypadku małej ilości dostępnych danych.

Acknowledgments

First and foremost, I would like to extend my deepest appreciation to my loving wife. Her unwavering support, understanding, and encouragement throughout this journey have been invaluable. This work wouldn't be possible without her patience, time, and love.

I am grateful to prof. UAM dr hab. Filip Graliński, my supervisor, for his support and guidance throughout my academic journey. I would like to express my sincere appreciation for the research opportunities he has provided me, which have been instrumental in shaping my academic development.

Finally, I would like to express my gratitude to the members of my research group from Applica.ai and Snowflake companies who have provided valuable feedback, engaging discussions, and collaborative opportunities. Their contributions have greatly enriched my understanding and perspective in the field.

Contents

1	Introduction	13
1.1	Motivation	13
1.2	Thesis in Brief	14
1.2.1	Span Identification	14
1.2.2	Document Understanding	17
1.3	Objectives	18
	List of publications	20
	References	21
	Span Identification	24
2	Contract Discovery: Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines	25
3	Dynamic Boundary Time Warping for sub-sequence matching with few examples	41
4	ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them	54
	Document Understanding	65
5	Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer	66

6	DUE: End-to-End Document Understanding Benchmark	83
7	ICDAR 2023 Competition on Document UnderstanDing of Every- thing (DUDE)	109
	Appendices	126
A	Competitions and Projects	127
	A.1 Competitions	128
	A.2 Projects	129
B	Declarations of Contribution	131

Chapter 1

Introduction

1.1 Motivation

The abundance of accumulated unstructured data has presented organizations with an increasingly challenging task of effectively utilizing and deriving value from this vast information pool [9]. The process of extracting actionable insights from large, unstructured datasets can be resource-intensive and time-consuming. To address these issues, Intelligent Document Processing (IDP) solutions are employed. IDP utilizes natural language technologies and computer vision to extract data from structured and unstructured documents in order to automate and enhance high-volume, repetitive document processing tasks. The increasing adoption of cloud-based document processing solutions and the shift towards digital transformation are major factors driving the IDP market's growth. This market is expected to experience significant growth in the next few years, with its size projected to increase from around USD 1.1 billion in 2022 to approximately USD 5.2 billion in 2027 [9].

The rapidly growing IDP market with a wide variety of complex client-specific use cases faces many challenging problems [2, 15]. One thing is that documents originate from multiple domains and industries, featuring diverse formats, layouts, and structures. This makes it difficult to develop universal processing techniques that generalize across numerous document types. The problem is particularly pronounced in low-resource scenarios, where effectively generalizing with limited labeled training

data and adapting to new document types becomes exceedingly challenging. This work touches upon several of the mentioned problems, especially those that are underresearched and lack adequate publicly available datasets.

This thesis will be directed toward two focal areas in the field of IDP, namely, Span Identification (SI) and Document Understanding (DU) (Figure 1-1).

1.2 Thesis in Brief

The thesis comprises six papers, the first three address the Span Identification problem, while the remaining three delve into the Document Understanding field.

Chapters 2 and 3 are concerned with few-shot Span Identification. The former proposes a new shared task and baselines, while the latter presents a novel method for sub-sequence matching with few examples. Chapter 4, in contrast to previous chapters, focuses on SI solutions that excel when abundant labeled training data is available. Chapter 5 opens the Document Understanding part and introduces a new end-to-end Transformer-based model for DU. Chapter 6 proposes a Document Understanding Benchmark for end-to-end DU models evaluation, and Chapter 7 describes a novel challenging DU competition.

Sections 1.2.1 and 1.2.2 provide a brief introduction to Span Identification and Document Understanding respectively, along with an overview of the papers included in the thesis.

1.2.1 Span Identification

Span Identification refers to the task of identifying continuous relevant spans of text that contain specific information or entities of interest (see the left side of Figure 1-1). These spans can vary in length, ranging from individual words and phrases to whole sentences or even paragraphs. They could correspond to propaganda spans [1], legal clauses (Chapter 2), stance-taking expressions [3], toxic spans [13] or any other designated text units. Unlike the challenges encountered in Document Understanding (Section 1.2.2) Span Identification is commonly associated with plain text documents.

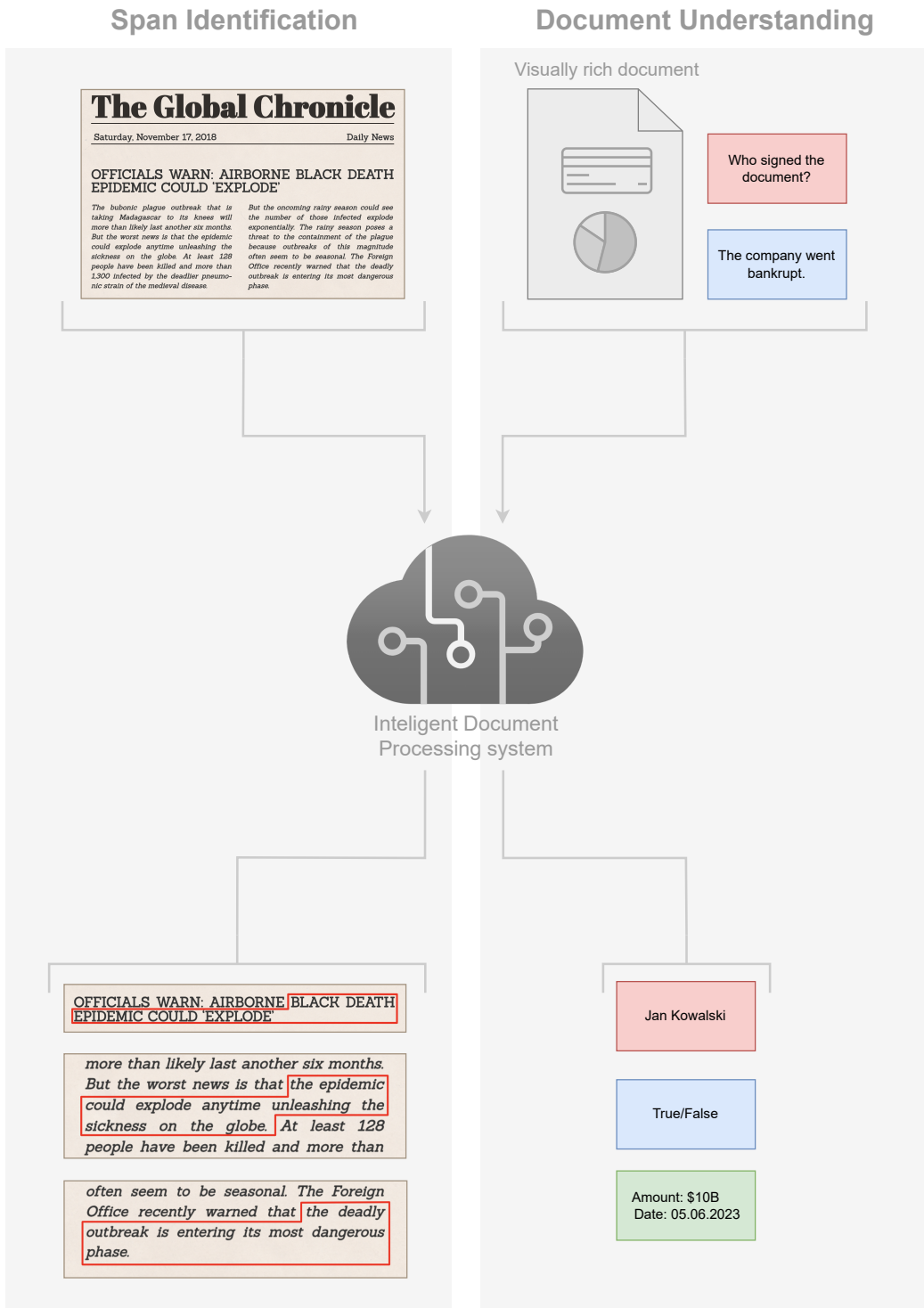


Figure 1-1: Intelligent Document Processing areas considered in this thesis are Span Identification (left) and Document Understanding (right). The IDP system objectives involve ■ localizing relevant text spans, ■ extracting key information, ■ verifying statements and ■ answering open ended questions about the document.

Contract Discovery: Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines (Chapter 2). This work proposes a new shared task of few-shot Span Identification for the legal domain. It was the first paper to introduce such a task. Besides the dataset, a unified framework for the evaluation of textual embedding-based solutions is introduced. Under this framework, baselines are provided for embeddings generated from various methods such as TF-IDF, GloVe, Sentence-BERT, USE, RoBERTa, GPT-1, and GPT-2. The study shows that the GPT-2 model achieves the best scores, while surprisingly state-of-the-art pretrained encoders (Sentence-BERT, USE) are worse than simple TF-IDF solutions. Additionally, the paper studies how the number of available examples impacts the models' accuracy. The dataset, reference results, and language models specialized in the legal domain are opened.

Dynamic Boundary Time Warping for sub-sequence matching with few examples (Chapter 3). This paper is a continuation of work on few-shot Span Identification solutions. A novel algorithm for matching a fragment of a long sequence that is similar to the set of other sequences is presented. The uniqueness of this method lies in not computing an average consensus sequence from the set of example sequences, but instead, using all of them at the same time. On top of that, the method requires much lower computational and memory costs compared to the exact solution. Finally, it is demonstrated that the solution performs very well on two distinct few-shot tasks.

ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them (Chapter 4). Compared with previously mentioned publications, this work deviates from the few-shot setting and introduces solutions that are better suited for situations with a lot of labeled training data. The paper describes the victorious system for the propaganda Technique Classification (TC) task and the second-best system for the propaganda Span Identification (SI) task. The TC task aimed to classify text fragments with propaganda techniques,

while the SI task focused on localizing fragments containing such techniques. The SI problem was addressed as a sequence labeling task, leveraging the RoBERTa-CRF architecture as an approach. For the TC task, an ensemble of RoBERTa-based models was employed. Both systems used self-training, a semi-supervised learning algorithm where a model is at first trained on a labeled dataset and then used to predict labels for unlabeled data. This process helped in expanding the available training data and improved the final model.

1.2.2 Document Understanding

Document Understanding is an overarching concept that gathers various ML tasks dealing with visually rich documents, including Key Information Extraction [5, 6, 7, 12, 16, 17], Classification [4, 18], Document Layout Analysis [8, 14, 19], and Visual Question Answering [10, 11] tasks¹. Document Understanding typically involves comprehending, interpreting, and extracting meaningful information from content, structure, and visual cues present in the document. The right side of the Figure 1-1 illustrates problems that fall within the purview of Document Understanding.

Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer (Chapter 5). This paper introduces a novel DU model that draws from encoder-decoder models, multi-modal transformers, and language models that are able to comprehend spatial connections between words. The model takes advantage of layout information, visual features, and textual semantics modalities that are present in the document. Layout information is initially represented by the positions of the tokens on the page and then converted to the learnable relative positional biases that are utilized in the self-attention mechanism. Regarding visual features, they are obtained from a truncated U-Net network and merged with textual semantics. The generative nature of the model enabled the handling of DU problems in an end-to-end manner, meaning the same architecture and loss function could be used for all tasks. State-of-the-art results were achieved on CORD, SROIE, and

¹The complete landscape of Document Understanding tasks is described in Chapter 6.

DocVQA challenges. Moreover, after the paper’s publication, the model won first place in the Infographics VQA competition (Appendix A.1).

DUE: End-to-End Document Understanding Benchmark (Chapter 6). The work proposes the first Document Understanding benchmark, enabling accurate measuring of the progress in the field. Among the over thirty datasets that were considered, seven datasets were carefully selected based on their adherence to the highest quality, difficulty, and licensing criteria. Some of them were reformulated, corrected, and modified to improve their quality or align them with an end-to-end Document Understanding setup. The benchmark features various multi-domain documents containing lists, tables, charts, and infographics. Apart from the benchmark, competitive baselines were implemented, and both the benchmarks and reference implementations are made publicly available.

ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE) (Chapter 7). This work not only expands upon the DUE benchmark but also raises the bar for DU models. The ICDAR 2023 competition on Document UnderstanDing of Everything is introduced. It consists of 5,000 visually-rich documents with 40,000 questions and covers around 200 diverse document types across 15 different industries, spanning a timeframe from 1900 to 2023. Challenging abstractive and extractive questions are introduced, with some of them requiring comprehension beyond the document content. Additionally, the evidence for the answers can be found on any page within the multi-page documents. The competition was intentionally designed to assess the models’ ability to generalize in low-data scenarios, particularly with unseen questions and domains.

1.3 Objectives

The main objective of the thesis is to propose new language-based methods and datasets for Span Identification and Document Understanding, equipping the Intelligent Document Processing domain with innovative solutions and resources. More

specifically, the primary objective could be divided into the following sub-goals:

1. Dataset and methods for few-shot Span Identification — The few-shot scenario is prevalent in various practical applications, particularly in business use cases, e.g., retrieving relevant legal clauses based only on a few examples (Chapter 2). The scarcity of publicly available datasets and suitable methods for few-shot SI highlights the importance of filling this gap. This need is addressed in Chapters 2 and 3.
2. Propaganda Span Identification and classification system — In the context of Intelligent Document Processing systems, propaganda can introduce biased or false information into documents, which can significantly impact the accuracy and reliability of the extracted data. If propaganda goes undetected, it can influence the decision-making process, leading to incorrect or biased outcomes. Integrating propaganda detection and classification capabilities into IDP systems can enhance the overall quality and trustworthiness of the extracted information. Chapter 4 proposes systems that could help mitigate these issues.
3. Multi-modal generative end-to-end model for Document Understanding — The invention of a model that could comprehend various document modalities and be applied to numerous DU problems without architectural changes while achieving state-of-the-art results. Chapter 5 introduces such a model.
4. Challenging datasets for Document Understanding — Two needs of the Document Understanding community are to be addressed. The first is a widely recognized benchmark for Document Understanding. It is introduced in Chapter 6. The second one is a lack of a real-world scenario shared task requiring strong generalization under a low-resource setting. It should cover diverse domains and industries, featuring complex question-and-answer pairs over multi-page documents. Chapter 7 showcases a DU competition with these characteristics.



Publication	MEiN points	Cited by ¹	Author contribution ²
Ł. Borchmann, D. Wisniewski, A. Gretkowski, I. Kosmala, <u>D. Jurkiewicz</u> , Ł. Szalkiewicz, G. Pałka, K. Kaczmarek, A. Kaliska, and F. Graliński. Contract Discovery: Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines. In <i>Findings of the Association for Computational Linguistics: EMNLP 2020</i> , pages 4254–4268, Online, November 2020. Association for Computational Linguistics	140	9	Implementation of baselines, evaluation of human performance.
Ł. Borchmann*, <u>D. Jurkiewicz</u> *, F. Graliński, and T. Górecki. Dynamic Boundary Time Warping for sub-sequence matching with few examples. <i>Expert Systems with Applications</i> , 169:114344, 2021	140	-	Conceptualization and methodology, improvement of the DBTW prototype, performing experiments, writing the paper, analysis of the results, design and implementation of ACBOW model.
<u>D. Jurkiewicz</u> *, Ł. Borchmann*, I. Kosmala, and F. Graliński. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them. In <i>Proceedings of the Fourteenth Workshop on Semantic Evaluation</i> , pages 1415–1424, COLING: Barcelona (online), December 2020. International Committee for Computational Linguistics 	140	27	Conceptualization and methodology, performing experiments, writing the paper, implementation of model prototypes, analysis of the results, error analysis.
R. Powalski*, Ł. Borchmann*, <u>D. Jurkiewicz</u> †, T. Dwojak†, M. Pietruszka†, and G. Pałka. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. In Josep Lladós, Daniel Lopresti, and Seiichi Uchida, editors, <i>Document Analysis and Recognition – ICDAR 2021</i> , pages 732–747, Cham, 2021. Springer International Publishing	140	75	Running experiments, design and implementation of image token embeddings, review and preparation of datasets, improvements of model prototype, editing the manuscript.
Ł. Borchmann*, M. Pietruszka*, T. Stanisławek*, <u>D. Jurkiewicz</u> , M. Turski, K. Szyndler, and F. Graliński. DUE: End-to-End Document Understanding Benchmark. In J. Vanschoren and S. Yeung, editors, <i>Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks</i> , volume 1. Curran, 2021	200	15	Participated in regular discussions, implementation of baselines, significantly improved results of the baselines (hyperparameter search), performing experiments, preparing code and models for a final release, edition of the paper.
J. Landeghem, R. Tito, Ł. Borchmann, M. Pietruszka, <u>D. Jurkiewicz</u> , R. Powalski, P. Józiak, S. Biswas, M. Coustaty, and T. Stanisławek. ICDAR 2023 Competition on Document Understanding of Everything (DUDE). Accepted for <i>Document Analysis and Recognition – ICDAR 2023</i> .	140	-	Preliminary experiments (T5, T5+2D, TILT models), manual collection of documents (agriculture, manufacturing category), data scraper for manually chosen documents, backtracking archive.org licenses, reviewing and assisting in the paper and proposal.

Table 1.1: List of publications included in the thesis.

 Best Paper Award

*, † denote equal contribution groups.

¹ Citations from Google Scholar, excluding self-citations, accessed on 21-06-2023.

References

- [1] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [2] A.R. Dengel. Making documents work: challenges for document understanding. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 1026–1035, 2003.
- [3] Masaki Eguchi and Kristopher Kyle. Span identification of epistemic stance-taking in academic written english, 2023.
- [4] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995, 2015.
- [5] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. ICDAR2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520, 2019.
- [6] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6, 2019.
- [7] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8580–8594, Online, November 2020. Association for Computational Linguistics.
- [8] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.
- [9] MarketsandMarkets Research Private Ltd. Intelligent document processing market size, share and global market forecast to 2027. <https://web.archive.org/web/20221130001445/https://www.marketsandmarkets.com/Market-Reports/intelligent-document-processing-market-195513136.html>.

- [10] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.
- [11] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
- [12] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. CORD: A consolidated receipt dataset for post-OCR parsing. 2019.
- [13] John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online, August 2021. Association for Computational Linguistics.
- [14] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S Weld, and Doug Downey. VILA: Improving structured content extraction from scientific pdfs using visual layout groups. *Transactions of the Association for Computational Linguistics*, 10:376–392, 2022.
- [15] Matyáš Skalický, Štěpán Šimsa, Michal Uříčář, and Milan Šulc. Business document information extraction: Towards practical benchmarks. In Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 105–117, Cham, 2022. Springer International Publishing.
- [16] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: Key information extraction datasets involving long documents with complex layouts. In Josep Lladós, Daniel Lopresti, and Seiichi Uchida, editors, *Document Analysis and Recognition – ICDAR 2021*, pages 564–579, Cham, 2021. Springer International Publishing.
- [17] Stacey Svetlichnaya. DeepForm: Understand structured documents at scale. https://wandb.ai/stacey/deepform_v1/reports/DeepForm-Understand-Structured-Documents-at-Scale--Vmlldzoy0DQ3Njg, 2020.
- [18] Te-Lin Wu, Shikhar Singh, Sayan Paul, Gully Burns, and Nanyun Peng. MELINDA: A multimodal dataset for biomedical experiment method classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14076–14084, 2021.

- [19] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. PubLayNet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.

Span Identification

Contract Discovery: Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines

Łukasz Borchmann and Dawid Wiśniewski and Andrzej Gretkowski
Izabela Kosmala and Dawid Jurkiewicz and Łukasz Szalkiewicz
Gabriela Pałka and Karol Kaczmarek and Agnieszka Kaliska and Filip Graliński

Applica.ai, Warsaw, Poland
{firstname.surname}@applica.ai

Abstract

We propose a new shared task of semantic retrieval from legal texts, in which a so-called *contract discovery* is to be performed—where legal clauses are extracted from documents, given a few examples of similar clauses from other legal acts. The task differs substantially from conventional NLI and shared tasks on legal information extraction (e.g., one has to identify text span instead of a single document, page, or paragraph). The specification of the proposed task is followed by an evaluation of multiple solutions within the unified framework proposed for this branch of methods. It is shown that state-of-the-art pretrained encoders fail to provide satisfactory results on the task proposed. In contrast, Language Model-based solutions perform better, especially when unsupervised fine-tuning is applied. Besides the ablation studies, we addressed questions regarding detection accuracy for relevant text fragments depending on the number of examples available. In addition to the dataset and reference results, LMs specialized in the legal domain were made publicly available.

1 Introduction

Processing of legal contracts requires significant human resources due to the complexity of documents, the expertise required and the consequences at stake. Therefore, a lot of effort has been made to automate such tasks in order to limit processing costs—notice that law was one of the first areas where electronic information retrieval systems were adopted (Maxwell and Schafer, 2008).

Enterprise solutions referred to as *contract discovery* deal with tasks, such as ensuring the inclusion of relevant clauses or their retrieval for further analysis (e.g., risk assessment). Such processes can consist of a manual definition of a few examples, followed by conventional information

Task	Legal	SI	Few-shot
COLIEE	+	–	–
SNLI	–	–	–
MultiNLI	–	–	–
TREC Legal Track	+	–	–
Propaganda detection	–	+	–
THUMOS (video)	–	+	+
ActivityNet (video)	–	+	+
ALBAYZIN (audio)	–	+	–
Contract Discovery (ours)	+	+	+

Table 1: Comparison of existing shared tasks. Most of the related NLP tasks do not assume Span Identification (SI), even those outside the legal domain (Legal). Moreover, the few-shot setting is not popular within the field of NLP yet.

retrieval. This approach was taken recently by Nagpal et al. (2018) for the extraction of fairness policies spread across agreements and administrative regulations.

2 Review of Existing Datasets

Table 1 summarizes main differences between available challenges. It is shown that most of the related NLP tasks do not assume span identification, even those outside the legal domain. Moreover, the few-shot setting is not popular within the field of NLP yet.

None of existing tasks involving semantic similarity methods, such as SNLI (Bowman et al., 2015) or multi-genre NLI (Bowman et al., 2015), assume span identification. Instead, standalone sentences are provided to determine their entailment. It is also the case of existing shared tasks for legal information extraction, such as COLIEE (Kano et al., 2017), where one has to recognize entailment between articles and queries, as considered in the question answering problem. Obviously, the tasks aimed at retrieving documents consisting of multiple sentences, such as TREC legal track (Baron

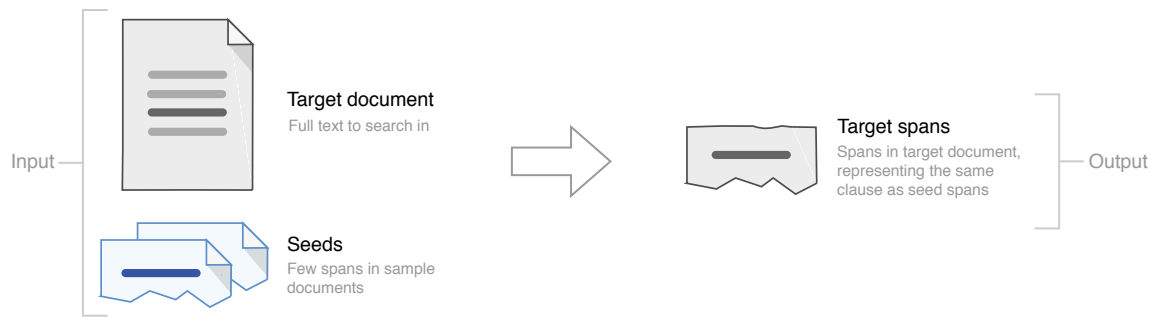


Figure 1: The aim of this task is to identify spans in the requested documents (referred to as *target* documents) representing clauses analogous to the spans selected in other documents (referred to as *seed* documents).

et al., 2006; Oard et al., 2010; Chu, 2011), lack this component.

There are a few NLP tasks where span identification is performed. These include some of plagiarism detection competitions (Potthast et al., 2010) and recently introduced SemEval task of propaganda techniques detection (Da San Martino et al., 2020). When different media are considered, NLP span identification task is equivalent to the action recognition in temporally untrimmed videos where one is expected to provide the start and end times for detected activity. These include THUMOS 14 (Jiang et al., 2014) as well as ActivityNet 1.2 and ActivityNet 1.3 challenges (Fabian Caba Heilbron and Niebles, 2015). Another example is query-by-example spoken term detection, as considered e.g., in ALBAYZIN 2018 challenge (Tejedor et al., 2019).

In a typical business case of *contract discovery* one may expect only a minimal number of examples. The number of available annotations results from the fact that *contract discovery* is performed constantly for different clauses, and it is practically impossible to prepare data in a number required by a conventional classifier every time. When one is interested in the few-shot setting, especially querying by multiple examples, there are no similar shared tasks within the field of NLP. Some authors however experimented recently with few-shot Named Entity Recognition (Fritzler et al., 2019) or few-shot text classification (Bao et al., 2019). The first, however, involves identification of short spans (from one to few words), whereas the second does not assume span identification at all.

What is important, existing tasks aimed at recognizing textual entailment in natural language (Bow-

man et al., 2015), differ in terms of the domain. This also applies to a multi-genre NLI (Williams et al., 2017), since legal texts vary significantly from other genres. As it will be shown later, methods optimal for MultiNLI do not perform well on the proposed task.

3 Contract Discovery: New Dataset and Shared Task

In this section, we introduce a new dataset of *Contract Discovery*, as well as a derived few-shot semantic retrieval shared task.

3.1 Desiderata

We define our desiderata as follows. We wish to construct a dataset for testing the mechanisms that detect various types of regulations in legal documents. Such systems should be able to process unstructured text; that is, no legal documents segmentation into the hierarchy of distinct (sub)sections is to be given in advance. In other words, we want to provide natural language streams lacking formal structure, as in most of the real-word usage scenarios (Vanderbeck et al., 2011). What is more, it is assumed that a searched passage can be any part of the document and not necessarily a complete paragraph, subparagraph, or a clause. Instead, the process should be considered as a span identification task.

We intend to develop a dataset for identifying spans in a query-by-example scenario instead of the setting where articles are being returned as an answer for the question specified in natural language.

We wish to propose using this dataset in a few-shot scenarios, where one queries the system using multiple examples rather than a single one. The

intended form of the challenge following these requirements is presented in Figure 1. Roughly speaking, the task is to identify spans in the requested documents (referred to as *target* documents) representing clauses analogous (i.e. semantically and functionally equivalent) to the examples provided in other documents (referred to as *seed* documents).

3.2 Data Collection and Annotation

Random subsets of bond issue prospectuses and non-disclosure agreement documents from the US EDGAR database¹, as well as annual reports of charitable organizations from the UK Charity Register² were annotated. Note there are no copyright issues and both datasets belong to the public domain.

Annotation was performed in such a way that clauses of the same type were selected (e.g., determining the governing law, merger restrictions, tax changes call, or reserves policy). Clause types depend on the type of a legal act and can consist of a single sentence, multiple sentences or sentence fragments. The exact type of a clause is not important during the evaluation since no full-featured training is allowed and a set of only a few sample clauses can be used during execution.

We restricted ourselves to 21 types as a result of a trade-off between annotation cost and the ability to formulate general remarks. Note that each clause type must be well-understood by the annotator (we described each very carefully in the instructions), and one must have all of the considered clauses in mind when the legal acts are being read during the process. In real-world legal applications, the clauses change in an everyday manner and depend on the problem analyzed by the lawyer at the moment.

Each document was annotated by two experts, and then reviewed (or resolved) by a super-annotator, who also decided the gold standard. An average Soft F_1 score (Section 4.2) of the two primary annotators, when compared to the gold standard (after the super-annotation), was taken to estimate human baseline performance of 0.84.

The inter-annotator agreement was equal to 0.76 in terms of Soft F_1 metric (Section 4.2). It should be treated as an agreement between two randomly

picked annotations since the total number of annotators was 10 (annotators were aligned randomly to a subset of documents in such a way that there would be two annotations and super-annotation per document).

Table 3 presents examples of clauses annotated in the sub-group of Charity Annual Reports documents. The detailed list of clauses and their examples can be found in Appendix C.

The dataset is made publicly available. In addition, we release a large, cleaned, plain-text corpus of legal and financial texts for the purposes of unsupervised model training or fine-tuning. All the available documents of US EDGAR as for November 19, 2018 were crawled. The resulting corpus consists of approx. 1M documents and 2B words in total (1.5G of text after xz compression).

3.3 Core Statistics

More than 2,500 spans were annotated in around 600 documents representing either bond issue prospectuses, non-disclosure agreement documents or annual reports of charitable organizations (the detailed statistics regarding the dataset are presented in Table 2).

Annotated clauses differ substantially from what can be found in existing sentence entailment challenges in terms of sentence length and complexity. SNLI contains less than 1% of sentences longer than 20 words, MultiNLI 5%, whereas in the case of clauses, we expect to return and consider it is 93% (and 77% of all spans in our shared task are longer than 20 words).

3.4 Evaluation Framework

Documents were split into halves to form validation and test sets for the purposes of few-shot semantic retrieval challenge. Evaluation is performed by means of a repeated random sub-sampling validation procedure. Sub-samples (k -combinations for each of 21 clauses, $k \in [2, 6]$) drawn from a particular set of annotations are split into $k - 1$ *seed* documents and 1 *target* document. Thus, clauses similar to the *seed* are expected to be returned from the target. We observed that the choice of input examples have an immense impact on the score. It is thus far more important to evaluate various *seed* configurations that various target documents. On the other hand, we wanted to keep the computational cost of evaluation reasonably small, so either the number of seed configurations had to be

¹<http://www.sec.gov/edgar.shtml>

²<http://www.gov.uk/find-charity-information>

Statistic	
Documents annotated	586
Mean document length (words)	24,284
Clause types	21
Mean clause length (words)	110
Clause instances	2,663

Table 2: Core statistics regarding released dataset.

reduced or the number of target documents for each configuration.

The selected k interval results in 1-shot to 5-shot learning, considered to be few-shot learning (Wang et al., 2019), whereas with the chosen number of sub-samples we expect improvements of 0.01 F_1 to be significant. Note that the 1–5 range denotes the number of annotated documents available, and it is possible that the same clause type appeared twice in one document, resulting in a higher number of clause instances.

Soft F_1 metric on character-level spans is used for the purpose of evaluation, as implemented in *GEval* tool (Graliński et al., 2019). Roughly speaking, this is the conventional F_1 measure, with precision and recall definitions altered to reflect the partial success of returning entities. In the case of the expected clause ranging between [1, 4] characters and the answer with ranges [1, 3], [10, 15] (the system assumes a clause occurs twice within the document), recall equals 0.75 (since this is the part of the relevant item selected) and precision equals ca. 0.33 (since this is the number of selected characters which turned out to be relevant). The Hungarian algorithm (Burkard et al., 2012) is employed to solve the problem of expected and returned range assignments. Soft F_1 has the desired property of being based on the widely utilized F_1 metric while abandoning the binary nature of the match, which is undesirable in the case dealt with in the task described.

4 Competitive Baselines

Solutions based on networks consuming pairs of sequences, such as BERT in sentence pair classification task setting (Devlin et al., 2018a), are considered out of the scope of this paper since they are suboptimal in terms of performance—they require expensive encoding of all combinations from the Cartesian product between seeds and targets, making such solutions unsuitable for semantic similarity search due to the combinatorial explosion (Reimers and Gurevych, 2019). Because of

the aforementioned problem and the fact that conventional classifiers require much more data than available in a few-shot setting, in this section, we describe simple k -NN-based approaches that we propose as baseline solutions to the problem stated.

4.1 Processing Pipeline

Evaluated solutions assume pre-encoding of all candidate segments and can be described within the unified framework consisting of segmenters, vectorizers, projectors, aggregators, scorers, and choosers ordered in a pipeline of transformations.

Segmenter is used to split a text into candidate sub-sequences to be encoded and considered in further steps. All the described solutions rely on a candidate sentence and n-grams of sentences, determined with the *spaCy* CNN model trained on OntoNotes.³ *Vectorizer* produces vector representations of texts on either word, sub-word, or segment (e.g., sentence) level. In our case, vectorization was based on TF-IDF representations, static word embeddings, and neural sentence encoders. *Projector* projects embeddings into a different space (e.g., decomposition methods such as PCA or ICA). *Aggregator* has the capability to use word or sub-word unit embeddings to create a segment embedding (e.g., embedding mean, inverse frequency weighting, autoencoder). *Scorer* compares two or more embeddings and returns computed similarities. Since we often compare multiple seed embeddings with one embedding of a candidate segment, a scorer includes policies to aggregate scores obtained for multiple seeds into the final candidate score (e.g., mean of individual cosine similarities or max-pooling over Word Mover Distances). *Chooser* determines whether to return a candidate segment with a given score (e.g., threshold, one best per document, or a combination thereof). For the sake of simplicity, during the evaluation, we restricted ourselves to the chooser returning only one, the most similar candidate. It is not optimal (because multiple might be expected), but we consider this setting a good reference for further methods.

The proposed taxonomy is consistent with the assumptions made by Gillick et al. (2018). It is presented in order to highlight the similarities and differences between particular solutions when they are introduced and compared within the ablation

³http://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-2.1.0

Clause (Instances)	Example
MAIN OBJECTIVE (195/231) The main objective of a charitable organization.	The aim of the Scout Association is to promote the development of young people in achieving their full physical, intellectual, social and spiritual potentials, as individuals, as responsible citizens and as members of their local, national and international communities. The method of achieving the Aim of the Association is by providing an enjoyable and attractive scheme of progressive training based on the Scout Promise and Law and guided by Adult leadership.
GOVERNING DOCUMENT (160/174) Information about the legal document which represents the rule book for the way in which a charity operates (title, date of creation etc.).	The Open University Students Educational Trust (Ouset) is controlled by its governing document, a deed of trust, dated 22 May 1982 as amended by a scheme dated 9 October 1992 and constitutes an unincorporated charity.
TRUSTEE APPOINTMENT (153/168) Procedure for selecting trustees and the term of office.	As per the governing document, four of the Trustee positions are appointed by virtue of their position within the Open University Students Association (OUSA). One further position is appointed by virtue of their previous position within OUSA. One Trustee is nominated by the Vice Chancellor of the Open University (OU) and there are co-opted positions whereby the Trustees are empowered to approach up to two other persons to act as Trustees. It is envisaged that all Trustees will serve a general term of two years in line with the main election periods within OUSA.
RESERVES POLICY (170/185) What are the current financial reserves of the organization and how much these reserves should be as assumed?	The Trustees regularly reviews the amount of reserves that are required to ensure that they are adequate to fulfill the charities continuing obligations.
INCOME SUMMARY (124/134) General information on income for the last year, sometimes associated with information on expenses.	Excluding the adjustments for FRS17 in respect of Pension Fund the results by way of net incoming resources accumulated £3.85m as against £6.78m in 2014, however last years performance benefited from extraordinary property sales generating a profit of £3.15m.
AUDITOR OPINION (190/192) Summary of the opinion of an independent auditor or inspector, often in the form of a list of points.	In connection with my examination, no matter has come to my attention: 1. which gives me reasonable cause to believe that in any material respect the requirements to keep accounting records in accordance with Section 130 of the Charities Act; and to prepare accounts which accord with the accounting records and comply with the accounting requirements of the Charities Act have not been met; or 2. to which, in my opinion, attention should be drawn in order to enable a proper understanding of the accounts to be reached.

Table 3: Clauses annotated in Charity Annual Reports (one of three groups of documents included in the shared task). The values in parentheses indicate the number of documents with a particular clause and the total number of clause instances, respectively. More examples are available in Appendix C.

studies later in this paper. The next section describes vectorizers, aggregators, and scorers used for evaluation.

4.1.1 Vectorizers

We intend to provide results of TF-IDF representations, as well as two methods that may be considered the state of the art of sentence embedding. The latter include *Universal Sentence Encoder* (USE) and *Sentence-BERT*.

USE is a Transformer-based encoder, where an element-wise sum of word representations is treated as a sentence embedding (Cer et al., 2018), trained with the multi-task objective. *Sentence-BERT* is a modification of the pretrained BERT network, utilizing Siamese and triplet network structures to derive sentence embeddings, trained with

the explicit objective of making them comparable with cosine similarity (Reimers and Gurevych, 2019). In both cases the original models released by the authors were used for the purposes of evaluation.

In addition, multiple contextual embeddings from Transformer-based language models, as well as static (context-less) GloVe word embeddings were tested (Pennington et al., 2014). Many approaches to generating context-dependent vector representations have been proposed in recent years (e.g., Peters et al. (2018); Vaswani et al. (2017)). One important advantage over static embeddings is the fact that every occurrence of the same word is assigned a different embedding vector based on the context in which the word is used. Thus, it is much easier to address issues arising from pre-

trained static embeddings (e.g., taking into consideration polysemy of words). For the purposes of evaluation, we relied on Transformer-based models provided by authors of particular architectures, utilizing the Transformers library (Wolf et al., 2019). These include BERT (Devlin et al., 2018b), GPT-1 (Radford, 2018), GPT-2 (Radford et al., 2018), and RoBERTa (Liu et al., 2019). They differ substantially and introduce many innovations, though they are all based on either the encoder or the decoder from the original model proposed for sequence-to-sequence problems (Vaswani et al., 2017). Selected models were fine-tuned on using the next word prediction task on the Edgar corpus we release and re-evaluated.

4.1.2 Aggregators

In addition to conceptually simple methods such as average or max-polling operations, multiple solutions to utilizing word embeddings for comparing documents can be used. In addition to embeddings mean we evaluated the *Smooth Inverse Frequency* (SIF), *Word Mover’s Distance* (WMD) and *Discrete Cosine Transform* (DCT).

SIF is a method proposed by Arora et al. (2017), where a representation of a document is obtained in two steps. First, each word embedding is weighted by $a/(a + f_r)$, where f_r stands for the underlying word’s relative frequency, and a is the weight parameter. Then, the projections on the first tSVD-calculated principal component are subtracted, providing final representations.

WMD is a method of calculating a similarity between documents. For two documents, embeddings calculated for each word (e.g., with GloVe) are matched between documents, so that semantically similar pairs of words between documents are detected. This matching procedure generally leads to better results than simply averaging over embeddings for documents and calculating similarity between centers of mass of documents as their similarity (Kusner et al., 2015). Recently, Zhao et al. (2019) showed it might be beneficial to use the method with contextual word embeddings.

DCT is a way to generate document-level representations in an order-preserving manner, adapted from image compression to NLP by Almarwani et al. (2019). After mapping an input sequence of real numbers to the coefficients of orthogonal cosine basis functions, low-order coefficients can be used as document embeddings, outperforming vector averaging on most tasks, as shown by the

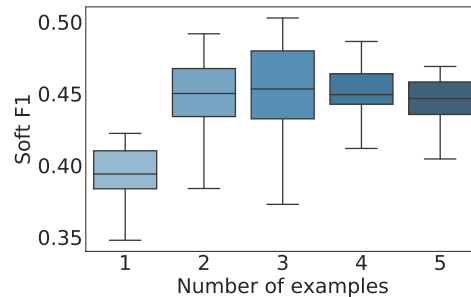


Figure 2: Performance as a function of the number of example documents available (solutions based on LMs). The methods benefit substantially from availability of a second example document and a bigger number leads to a decreased variance.

authors.

4.2 Results

Table 4 recapitulates the most important results of the completed evaluation.

Sentence-BERT and Universal Sentence Encoder could not outperform the simple TF-IDF approach, especially when SVD decomposition was applied (the setting commonly referred to as Latent Semantic Analysis). Static word embeddings with SIF weighting performed similarly to TF-IDF, or better, provided they were trained on a legal text corpus rather than on general English. It could not be clearly confirmed whether the use of WMD or DCT is beneficial. For the latter, the best results were achieved with c^0 , which in the case of the k -NN algorithm leads to the same answers as mean-pooling and thus is not reported in the table. In case of $c^{0:n}$ where $n > 0$ constant decrease of k -NN methods performance was observed (Appendix B).

Interestingly, from all the released USE models, the multilingual ones performed best — for the monolingual *universal-sentence-encoder-large* model, scores were ten percentage points lower. The best Sentence-BERT model performed significantly worse than the best USE—note that the authors of Sentence-BERT compared it to monolingual models released earlier, which they indeed outperform. Moreover, Sentence-BERT does not perform better than BERT trained with whole word masking, although there is no Sentence-BERT equivalent of this model available so far.

⁴TF-IDF with truncated SVD decomposition is commonly referred to as Latent Semantic Analysis (Halko et al., 2011).

⁵SVD in SIF method is used to perform removal of single common component (Arora et al., 2017).

Segmenter	Vectorizer	Projector	Scorer	Aggregator	Soft F_1
sentence	TF-IDF (1–2 grams, binary TF term)	—	mean cosine	—	0.38
		tSVD (500) ⁴	mean cosine	—	0.39
sentence	GloVe (300d, Wikipedia & Gigaword)	—	mean cosine	mean	0.34
		—	mean WMD	—	0.35
		SIF tSVD ⁵	mean cosine	SIF	0.37
sentence	GloVe (300d, EDGAR)	—	mean cosine	mean	0.36
		—	mean WMD	—	0.35
		SIF tSVD	mean cosine	SIF	0.41
sentence	Sentence-BERT (base-nli-stsb-mean *)	—	mean cosine	mean	0.32
sentence	USE (multilingual *)	—	mean cosine	—	0.38
sentence	BERT, last layer (large-uncased-whole...*)	—	mean cosine	mean	0.35
sentence	GPT-1, last layer	—	mean cosine	mean	0.36
sentence	GPT-2, last layer (large *)	—	mean cosine	mean	0.41
sentence	RoBERTa, last layer (large *)	—	mean cosine	mean	0.31
sentence	GPT-1, last layer (fine-tuned)	—	mean cosine	mean	0.43
sentence	GPT-1, last layer (fine-tuned)	fICA (500)	mean cosine	mean	0.44
sentence	GPT-2, last layer (large, fine-tuned)	—	mean cosine	mean	0.44
sentence	GPT-2, last layer (large, fine-tuned)	fICA (400)	mean cosine	mean	0.45
1–3 sen.	GPT-1, last layer (fine-tuned)	—	mean cosine	mean	0.47
1–3 sen.	GPT-1, last layer (fine-tuned)	fICA (500)	mean cosine	mean	0.49
1–3 sen.	GPT-2, last layer (large, fine-tuned)	—	mean cosine	mean	0.46
1–3 sen.	GPT-2, last layer (large, fine-tuned)	fICA (400)	mean cosine	mean	0.51
human					0.84

Table 4: Selected results when returning a single, most similar segment, determined with given segmenters, vectorizers, projectors, scorers and aggregators. The * symbol indicates only the best models from each architecture are presented here (results for the remaining ones are available in Appendix B).

In cases of averaging (sub)word embeddings from the last layer of neural Language Models, the results were either comparable or inferior to TF-IDF. The best-performing language models were GPT-1 and GPT-2. Fine-tuning of these on a sub-sample of a legal text corpus improved the results significantly, by a factor of 3–7 points. LMs seem to benefit neither from SIF nor from the removal of a single common component; their performance can, however, be mildly improved with a conventionally used decomposition, such as ICA (Hyvärinen and Oja, 2000).

Substantial improvement can be achieved by considering segments different from a single sentence, such as n-grams of sentences (meaning that any contiguous sequence of up to n sentences from a given text was scored and could be returned as a result).

Figure 2 presents how the performance of particular methods changes as a function of the number of example documents available within the simple similarity averaging scheme used in all the presented solutions. In general, the methods benefit substantially from the availability of a second exam-

ple. A bigger number leads to a decreased variance but yields no improvement in the median score.

5 Discussion

The brief evaluation presented in the previous section has multiple limitations. First, it assumed retrieval of a single, most similar segment, whereas it appears that multiple clauses might be returned instead. However, we consider this restriction justifiable during a preliminary comparison of applicable methods. Multiple alternative selectors may be proposed in the future.

Secondly, all the evaluated methods assume scoring with the policy of averaging individual similarities. We encourage readers to experiment with different pooling methods or meta-learning strategies. Moreover, even the LM-based methods we had studied the most can be further studied in the proposed shared task. For example, only embeddings from the last layer were evaluated, even though it is possible that the higher layers may capture semantics better.

Finally, it is in principle possible to address the task in entirely different ways, for example, by per-

forming neither segmentation nor aggregation of word embeddings at all, but by matching clauses on the word level instead, which may be an interesting direction for further research. We decided to take the most common and straightforward way, due to fact performed evaluations are to serve as baselines for other methods.

6 Related Work

There is a large and varied body of work related to information retrieval in general; however, following Gillick et al. (2018) we consider the problem stated in an end-to-end manner, where the nearest neighbor search is performed on dense document representations. With this assumption, the main issue is to obtain reliable representations of documents, where by document we mean *any self-contained unit that can be returned to the user as a search result* (Büttcher et al., 2010). We use the term *segment* with the same meaning wherever it aids clarity.

Many approaches considered in the literature rely on word embedding and aggregation strategies. Simple methods proposed include averaging, as in the continuous bag-of-words (CBOW) model (Mikolov et al., 2013) or frequency-weighted averaging with the decomposition method applied (Arora et al., 2017). More sophisticated schemes include utilizing multiple weights, such as a novelty score, a significance score, and a corpus-wise uniqueness (Yang et al., 2018) or computing a vector of locally aggregated descriptors (Ionescu and Butnaru, 2019). Most of the proposed methods are orderless, and their limitations were recently discussed by Mai et al. (2019). However, there are also pooling approaches preserving spatial information, such as a hierarchical pooling operation (Shen et al., 2018). Other methods of obtaining sentence representations from word embeddings include training an autoencoder on a large collection of unlabeled data (Zhang et al., 2018) or utilizing random encoders (Wieting and Kiela, 2019). Despite its shortcomings and the availability of many sophisticated alternatives, the CBOW model is a common choice due to its ability to ensure strong results on many downstream tasks.

Different approaches assume training encoders through document embedding in an unsupervised or supervised manner, without the need for explicit aggregation. The former include Skip-Thought Vectors, trained with the objective of reconstruct-

ing the surrounding sentences of an encoded passage (Kiros et al., 2015). Although this method was outperformed by supervised models trained on a single NLI task (Conneau et al., 2017), paraphrase corpora (Jiao et al., 2018) or multiple tasks (Subramanian et al., 2018), the objective of predicting the next sentence is used as an additional objective in multiple novel models, such as the Universal Sentence Encoder (Cer et al., 2018). Even though many Transformer-based language models implement their own pooling strategy for generating sentence representations (special token pooling), they were shown to yield weak sentence embeddings, as described recently by Reimers and Gurevych (2019). The authors proposed a superior method of fine-tuning a pretrained BERT network with Siamese and triplet network structures to obtain sentence embeddings.

There were attempts to utilize semantic similarity methods explicitly in the legal domain, e.g., for a case law entailment within the COLIEE shared task. In a recent edition, Rabelo et al. (2019) used a BERT model fine-tuned on a provided training set in a supervised manner, and achieved the highest F-score among all teams. However, due to the reasons discussed in Section 4, their approach is not consistent with the nearest neighbor search, which is what we are aiming for.

7 Summary and Conclusions

We have introduced a new shared task of semantic retrieval from legal texts, which differs substantially from conventional NLI. It is heavily inspired by enterprise solutions referred to as *contract discovery*, focused on ensuring the inclusion of relevant clauses or their retrieval for further analysis. The main distinguishing characteristic of Contract Discovery shared task is conceptual, since:

- Candidate sequences are being mined from real texts. It is assumed span identification should be performed (systems should be able to return any document substring without any segmentation given in advance).
- It is suited for few-shot methods, filling the gap between conventional sentence classification and NLI tasks based on sentence pairs.

For the purposes of providing competitive baselines, we considered the problem stated in an end-to-end manner, where the nearest neighbor search is performed on document representations. With

this assumption, the main issue was to obtain representations of text fragments, which we referred to as segments. The description of the task was followed by the evaluation of multiple k -NN-based solutions within the unified framework, which may be used to describe future solutions. Moreover, a practical justification for handling the problem with k -NN was briefly introduced.

It has been shown that in this particular setting, pretrained, *universal* encoders fail to provide satisfactory results. One may suspect that this is a result of the difference between the domain they were trained on and the legal domain. During the evaluation, solutions based on the Language Models performed well, especially when unsupervised fine-tuning was applied. In addition to the aforementioned ability to fine-tune the method on legal texts, the most important indicator of success so far has been the involvement of multiple, sometimes overlapping substrings instead of sentences. Moreover, it has been demonstrated that the methods benefit substantially from the availability of a second example, and the presence of more leads to a decrease in variance, even when a simple similarity averaging scheme is applied.

The discussion regarding the presented methods and their limitations briefly outlined possible measures towards improving the baseline methods. In addition to the dataset and reference results, legal-specialized LMs have been made released to assist the research community in performing further experiments.

The Contract Discovery dataset, Edgar Corpus, we crawled, and all the mentioned models are publicly available on GitHub: <https://github.com/applcaai/contract-discovery>.

Acknowledgements

The Smart Growth Operational Programme supported this research under project no. POIR.01.01.01-00-0605/19 (*Disruptive adoption of Neural Language Modelling for automation of text-intensive work*).

There are no copyright issues regarding the Contract Discovery dataset, as both sources belong to the public domain. Documents were annotated ethically by our co-workers. Moreover, the colleagues who participated in annotation are among the authors of the paper.

References

- Nada Almarwani, Hanan Aldarmaki, and Mona Diab. 2019. [Efficient sentence embedding using discrete cosine transform](#).
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings.
- Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2019. Few-shot text classification with distributional signatures. *arXiv:1908.06039*.
- Jason R. Baron, National Archives, Records Administration, and Office Of General. 2006. Trec-2006 legal track overview. In *In The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Rainer Burkard, Mauro Dell’Amico, and Silvano Martello. 2012. *Assignment Problems. Revised reprint*. SIAM - Society of Industrial and Applied Mathematics. 393 Seiten.
- Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. 2010. *Information Retrieval - Implementing and Evaluating Search Engines*. MIT Press.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. [Universal sentence encoder](#). *CoRR*, abs/1803.11175.
- Heting Chu. 2011. Factors affecting relevance judgment: a report from trec legal track. *Journal of Documentation*, 67:264–278.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. [Supervised learning of universal sentence representations from natural language inference data](#).
- Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation, SemEval 2020, Barcelona, Spain*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.

- Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.
- Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. [Few-shot classification in named entity recognition task](#). In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC '19*, pages 993–1000, New York, NY, USA. ACM.
- Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. [End-to-end retrieval in continuous space](#).
- Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz Górecki. 2019. [GEval: Tool for debugging NLP datasets and models](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy. Association for Computational Linguistics.
- N. Halko, P. G. Martinsson, and J. A. Tropp. 2011. [Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions](#). *SIAM Rev.*, 53(2):217–288.
- Aapo Hyvärinen and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13 4-5:411–30.
- Radu Tudor Ionescu and Andrei M. Butnaru. 2019. [Vector of Locally-Aggregated Word Embeddings \(VLAWE\): A Novel Document-level Representation](#).
- Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. 2014. THUMOS challenge: Action recognition with a large number of classes. <http://crcv.ucf.edu/THUMOS14/>.
- Xiaoqi Jiao, Fang Wang, and Dan Feng. 2018. [Convolutional neural network for universal sentence embeddings](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2470–2481, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yoshinobu Kano, Mi Young Kim, Randy Goebel, and Ken Satoh. 2017. Overview of COLIEE 2017. In *COLIEE@ICAIL*.
- Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. [Skip-thought vectors](#). In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. [From word embeddings to document distances](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#).
- Florian Mai, Lukas Galke, and Ansgar Scherp. 2019. [CBOW is not all you need: Combining CBOW with the compositional matrix space model](#). *CoRR*, abs/1902.06423.
- K. Tamsin Maxwell and Burkhard Schafer. 2008. Concept and context in legal information retrieval. In *JURIX*.
- Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#).
- Rashmi Nagpal, Chetna Wadhwa, Mallika Gupta, Samiulla Shaikh, Sameep Mehta, and Vikram Goyal. 2018. [Extracting fairness policies from legal documents](#). *CoRR*, abs/1809.04262.
- W. Douglas Oard, R. Jason Baron, Bruce Hedin, D. David Lewis, and Stephen Tomlinson. 2010. Evaluation of information retrieval for e-discovery. *Artif. Intell. Law*, pages 347–386.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *In EMNLP*.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.
- Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. [An evaluation framework for plagiarism detection](#). In *Coling 2010: Posters*, pages 997–1005, Beijing, China. Coling 2010 Organizing Committee.
- Juliano Rabelo, Mi-Young Kim, and Randy Goebel. 2019. [Combining similarity and transformer methods for case law entailment](#). In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19*, pages 290–296, New York, NY, USA. ACM.
- Alec Radford. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. [Language models are unsupervised multitask learners](#).

- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. [Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms](#).
- Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J. Pal. 2018. [Learning general purpose distributed sentence representations via large scale multi-task learning](#). *CoRR*, abs/1804.00079.
- Javier Tejedor, Doroteo T. Toledano, Paula Lopez-Otero, Laura Docio-Fernandez, Mikel Peñagarikano, Luis Javier Rodriguez-Fuentes, and Antonio Moreno-Sandoval. 2019. [Search on Speech from Spoken Queries: The Multi-Domain International ALBAYZIN 2018 Query-by-Example Spoken Term Detection Evaluation](#). *EURASIP J. Audio Speech Music Process.*, 2019(1).
- Scott Vanderbeck, Joseph Bockhorst, and Chad Oldfather. 2011. A machine learning approach to identifying sections in legal briefs. In *MAICS*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *CoRR*, abs/1706.03762.
- Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. 2019. Generalizing from a few examples: A survey on few-shot learning.
- John Wieting and Douwe Kiela. 2019. [No training required: Exploring random encoders for sentence classification](#).
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.
- Ziyi Yang, Chenguang Zhu, and Weizhu Chen. 2018. [Zero-training sentence embedding via orthogonal basis](#). *ArXiv*, abs/1810.00438.
- Minghua Zhang, Yunfang Wu, Weikang Li, and Wei Li. 2018. [Learning universal sentence representations with mean-max attention autoencoder](#).
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. [MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

A File Structure

The documents' content can be found in the `reference.tsv` files. The input files `in.tsv` consist of tab-separated fields: Target ID (e.g. 57), Clause considered (e.g. *governing-law*), Example #1 (e.g. 59 15215-15453), ..., Example #N. Each example consists of document ID and characters range. Ranges can be discontinuous. In such a case the sequences are separated with a comma, e.g. 4103-4882,12127-12971. The file with answers (`expected.tsv`) contains one answer per line, consisting of the entity name (to be copied from input) and characters range in the same format as described above. The reference file contains two tab-separated fields: document ID and content.

B Other Evaluation Results

Tables below describe evaluation results which were not included in the paper (or were included without broader context, that is without reference to different results from the same class of solutions).

Table 5 presents results with all the evaluated Sentence-BERT models. Table 6 shows scores achieved by TF-IDF with different settings, including other n-gram ranges. Results of particular Universal Sentence Encoder models are presented in Table 7. Table 8 shows results of Transformer-based Language Models not included in the paper. Finally, Table 9 is devoted to analysis of Discrete Cosine Transform embeddings.

Model	Soft F_1
bert-base-nli-cls-token	0.29
bert-base-nli-max-tokens	0.30
bert-base-nli-mean-tokens	0.31
bert-base-nli-stsb-mean-tokens	0.32
bert-base-wikipedia-sections-mean-tokens	0.25
bert-large-nli-cls-token	0.29
bert-large-nli-max-tokens	0.30
bert-large-nli-mean-tokens	0.30
bert-large-nli-stsb-mean-tokens	0.31
roberta-base-nli-mean-tokens	0.28
roberta-base-nli-stsb-mean-tokens	0.29
roberta-large-nli-mean-tokens	0.31
roberta-large-nli-stsb-mean-tokens	0.31

Table 5: Results of Sentence-BERT models on the *test-A* dataset when returning the most similar sentence. Names as in *sentence-transformers* library: <https://github.com/UKPLab/sentence-transformers>

Range (n-grams)	Binary	Soft F_1
1-1	-	0.32
1-2	-	0.35
1-3	-	0.36
1-1	+	0.36
1-2	+	0.38
1-3	+	0.37

Table 6: Results of TF-IDF on the *test-A* dataset when returning the most similar sentence.

Model	Soft F_1
multilingual/1	0.38
multilingual-large/1	0.33
multilingual-qa/1	0.28
large/3	0.26

Table 7: Results of Universal Sentence Encoder models on the *test-A* dataset when returning the most similar sentence.

Model	Soft F_1
bert-base-cased	0.25
bert-base-multilingual-cased	0.24
bert-base-multilingual-uncased	0.32
bert-base-uncased	0.26
bert-large-cased	0.21
bert-large-cased-whole-word-masking	0.31
bert-large-uncased	0.18
bert-large-uncased-whole-word-masking	0.35
roberta-base	0.25
roberta-large	0.32
openai-gpt	0.36
gpt2	0.16
gpt2-medium	0.11
gpt2-large	0.41

Table 8: Results of particular Transformer-based Language Models (without finetuning) on the *test-A* dataset when returning the most similar sentence. Names as in *transformers* library: <https://github.com/huggingface/transformers>

C	Soft F_1
c^0	0.36
$c^{0:1}$	0.30
$c^{0:2}$	0.25
$c^{0:3}$	0.20
$c^{0:4}$	0.18

Table 9: Results of GloVe embeddings (300d, EDGAR) on the *test-A* dataset when Discrete Cosine Transform sentence embeddings were created. The c^0 is equivalent to embeddings mean when k -NN methods are considered. The similar decrease of performance was observed for other models.

C Rest of the Clauses Considered

Random subsets of bond issue prospectuses and non-disclosure agreement documents from the US EDGAR database⁶, as well as annual reports of charitable organizations from the UK Charity Register⁷ were annotated, in such a way that clauses of the same type were selected (e.g. determining the governing law, merger restrictions, tax changes call or reserves policy). Clause types depend on the type of a legal act and can consist of a single sentence, multiple sentences or sentence fragments. Tables bellow present clause types annotated in each of the document groups.

Clause (Instances)	Example
GOVERNING LAW (152/160) The parties agree on which jurisdiction the contract will be subject to.	This Agreement shall be governed by and construed in accordance with the laws of the State of California without reference to its rules of conflicts of laws.
CONFIDENTIAL PERIOD (108/122) The parties undertake to maintain confidentiality for a certain period of time.	The term of this Agreement during which Confidential Information may be disclosed by one Party to the other Party shall begin on the Effective Date and end five (5) years after the Effective Date, unless extended by mutual agreement.
EFFECTIVE DATE (79/89) Information on the date of entry into force of the contract.	THIS AGREEMENT is entered into as of the 30th of July 2010 and shall be deemed to be effective as of July 23, 2010.
EFFECTIVE DATE REFERENCE (91/111)	This Contract shall become effective (the "Effective Date") upon the date this Contract is signed by both Parties.
NO SOLICITATION (101/117) Prohibition of acquiring employees of the other party (after the contract expires) and maintaining business relations with the customers of the other party.	You agree that for a period of eighteen months (18) from the date hereof you will not directly or indirectly recruit, solicit or hire any regional or district managers, corporate office employee, member of senior management of the Company (including store managers), or other employee of the Company identified to you.
CONFIDENTIAL INFORMATION FORM (152/174) Forms and methods of providing confidential information.	"Confidential Information" means any technical or commercial information or data, trade secrets, know-how, etc., of either Party or their respective Affiliates whether or not marked or stamped as confidential, including without limitation, Technology, Invention(s), Intellectual Property Rights, Independent Technology and any samples of products, materials or formulations including, without limitation, the chemical identity and any properties or specifications related to the foregoing. Any Development Program Technology, MPM Work Product, MSC Work Product, Hybrid Work Product, Prior End-Use Work Product and/or Shared Development Program Technology shall be Confidential Information of the Party that owns the subject matter under the terms set forth in this Agreement.
DISPUTE RESOLUTION (67/68) Arrangements for how to resolve disputes (arbitration, courts).	The Parties will attempt in good faith to resolve any dispute or claim arising out of or in relation to this Agreement through negotiations between a director of each of the Parties with authority to settle the relevant dispute. If the dispute cannot be settled amicably within fourteen (14) days from the date on which either Party has served written notice on the other of the dispute then the remaining provisions of this Clause shall apply.

Table 10: Clauses annotated in Non-disclosure Agreements. The values in parentheses indicate the number of documents with a particular clause and the total number of clause instances, respectively.

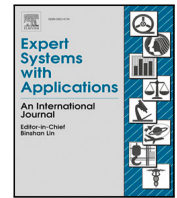
⁶<http://www.www.sec.gov/edgar.shtml>

⁷<http://www.gov.uk/find-charity-information>

Clause (Instances)	Example
CHANGE OF CONTROL COVENANT (88/95) Information about the obligation to redeem bonds for 101% of the price in the event of change of control.	Upon the occurrence of a Change of Control Triggering Event (as defined below with respect to the notes of a series), unless we have exercised our right to redeem the notes of such series as described above under “Optional Redemption,” the indenture provides that each holder of notes of such series will have the right to require us to repurchase all or a portion (equal to \$2,000 or an integral multiple of \$1,000 in excess thereof) of such holder’s notes of such series pursuant to the offer described below (the “Change of Control Offer”), at a purchase price equal to 101% of the principal amount thereof, plus accrued and unpaid interest, if any, to the date of repurchase, subject to the rights of holders of notes of such series on the relevant record date to receive interest due on the relevant interest payment date.
CHANGE OF CONTROL NOTICE (78/79) Information about the obligation to inform bondholders (usually by mail) about the event of change of control. This clause usually follows immediately the above clause.	Within 30 days following any Change of Control, B&G Foods will mail a notice to each holder describing the transaction or transactions that constitute the Change of Control and offering to repurchase notes on the Change of Control Payment Date specified in the notice, which date will be no earlier than 30 days and no later than 60 days from the date such notice is mailed, pursuant to the procedures required by the indenture and described in such notice. Holders electing to have a note purchased pursuant to a Change of Control Offer will be required to surrender the note, with the form entitled “Option of Holder to Elect Purchase” on the reverse of the note completed, to the paying agent at the address specified in the notice of Change of Control Offer prior to the close of business on the third business day prior to the Change of Control Payment Date.
CROSS DEFAULT (96/110) The company does not comply with certain conditions (event of default), so the bonds become due (e.g. when the company does not submit financial statements on time) — our clause was limited to the event of non-repayment, usually the minimum sum is given.	due to our default, we (i) are bound to repay prematurely indebtedness for borrowed moneys with a total outstanding principal amount of \$75,000,000 (or its equivalent in any other currency or currencies) or greater, (ii) have defaulted in the repayment of any such indebtedness at the later of its maturity or the expiration of any applicable grace period or (iii) have failed to pay when properly called on to do so any guarantee of any such indebtedness, and in any such case the acceleration, default or failure to pay is not being contested in good faith and not cured within 15 days of such acceleration, default or failure to pay;
LITIGATION DEFAULT (42/51) Court verdict or administrative decision which charge the company for a significant unpaid amount (another from the series of event of default).	(8) one or more judgments, orders or decrees of any court or regulatory or administrative agency of competent jurisdiction for the payment of money in excess of \$30 million (or its foreign currency equivalent) in each case, either individually or in the aggregate, shall be entered against the Company or any subsidiary of the Company or any of their respective properties and shall not be discharged and there shall have been a period of 60 days after the date on which any period for appeal has expired and during which a stay of enforcement of such judgment, order or decree, shall not be in effect;
MERGER RESTRICTIONS (188/241) A clause preventing the merger or sale of a company, etc., except under certain conditions (generally, the company should not avoid its obligations to its bondholders).	Without the consent of the holders of the outstanding debt securities under the indentures, we may consolidate with or merge into, or convey, transfer or lease our properties and assets to any person and may permit any person to consolidate with or merge into us. However, in such event, any successor person must be a corporation, partnership, or trust organized and validly existing under the laws of any domestic jurisdiction and must assume our obligations on the debt securities and under the applicable indenture. We agree that after giving effect to the transaction, no event of default, and no event which, after notice or lapse of time or both, would become an event of default shall have occurred and be continuing and that certain other conditions are met; provided such provisions will not be applicable to the direct or indirect transfer of the stock, assets or liabilities of our subsidiaries to another of our direct or indirect subsidiaries. (Section 801)

<p>BONDHOLDERS DEFAULT (191/241) A clause on the payment of the principal amount and interest — they become due as a result of an event of default, if such a declaration is made by bondholders.</p>	<p>If an event of default (other than an event of default referred to in clause (5) above with respect to us) occurs and is continuing, the trustee or the holders of at least 25% in aggregate principal amount of the outstanding notes by notice to us and the trustee may, and the trustee at the written request of such holders shall, declare the principal of and accrued and unpaid interest, if any, on all the notes to be due and payable. Upon such a declaration, such principal and accrued and unpaid interest will be due and payable immediately. If an event of default referred to in clause (5) above occurs with respect to us and is continuing, the principal of and accrued and unpaid interest on all the notes will become and be immediately due and payable without any declaration or other act on the part of the trustee or any holders.</p>
<p>TAX CHANGES CALL (48/56) A clause about the possibility of an earlier redemption of the bond by the issuer if the tax law or its interpretation changes.</p>	<p>If, as a result of any change in, or amendment to, the laws (or any regulations or rulings promulgated under the laws) of the Netherlands or the United States or any taxing authority thereof or therein, as applicable, or any change in, or amendments to, an official position regarding the application or interpretation of such laws, regulations or rulings, which change or amendment is announced or becomes effective on or after the date of the issuance of the notes, we become or, based upon a written opinion of independent counsel selected by us, will become obligated to pay additional amounts as described above in “Payment of additional amounts,” then the Issuer may redeem the notes, in whole, but not in part, at 100% of the principal amount thereof together with unpaid interest as described in the accompanying prospectus under the caption “Description of WPC Finance Debt Securities and the Guarantee-Redemption for Tax Reasons.”</p>
<p>FINANCIAL STATEMENTS (201/317) A clause on the obligation to submit (usually to the SEC) annual reports or other reports.</p>	<p>Notwithstanding that the Company may not be subject to the reporting requirements of Section 13 or 15(d) of the Exchange Act, the Company will file with the SEC and provide the Trustee and Holders and prospective Holders (upon request) within 15 days after it files them with the SEC, copies of its annual report and the information, documents and other reports that are specified in Sections 13 and 15(d) of the Exchange Act. In addition, the Company shall furnish to the Trustee and the Holders, promptly upon their becoming available, copies of the annual report to shareholders and any other information provided by the Company to its public shareholders generally. The Company also will comply with the other provisions of Section 314(a) of the TIA.</p>

Table 11: Clauses annotated in Corporate Bonds. The values in parentheses indicate the number of documents with a particular clause and the total number of clause instances, respectively.



Dynamic Boundary Time Warping for sub-sequence matching with few examples

Łukasz Borchmann^{a,*}, Dawid Jurkiewicz^{a,1}, Filip Galiński^a, Tomasz Górecki^b

^a *Applica.ai, Warsaw, Poland*

^b *Adam Mickiewicz University, Poznań, Poland*

ARTICLE INFO

Keywords:

Dynamic Time Warping
Sub-sequence matching
Natural Language Processing
Information Retrieval
Few-shot learning
Semantic retrieval

ABSTRACT

The paper presents a novel method of finding a fragment in a long temporal sequence similar to the set of shorter sequences. We are the first to propose an algorithm for such a search that does not rely on computing the average sequence from query examples. Instead, we use query examples as is, utilizing all of them simultaneously. The introduced method based on the Dynamic Time Warping (DTW) technique is suited explicitly for few-shot query-by-example retrieval tasks. We evaluate it on two different few-shot problems from the field of Natural Language Processing. The results show it either outperforms baselines and previous approaches or achieves comparable results when a low number of examples is available.

1. Introduction

This work bridges Information Retrieval, Natural Language Processing, Dynamic Programming, and Machine Learning, introducing a novel approach to identifying text spans with semantic matching. Although the method can retrieve any sequential information from an untrimmed stream, this paper demonstrates application to diverse problems involving text in natural language.

Let us start by observing that a substantial proportion of retrieval, detection, and sequence labeling tasks can be solved using sub-sequence matching. However, so far, no mainstream methods tackle the problem this way.

Consider the case of Named Entity Recognition (also referred to as entity identification, entity chunking or entity extraction, NER) – a task of locating and classifying spans of text associated with real-world objects, such as person names, organizations, and locations, as well as with abstract temporal and numerical expressions such as dates (Goyal et al., 2018; Li et al., 2018; Yadav & Bethard, 2018).

The problem is commonly solved with trained models for structured prediction (Huang et al., 2015; Lample et al., 2016). In contrast, we propose to solve it in a previously not recognized way: to use word embeddings (see Section 5.2.1) directly, performing semantic sub-sequence matching. In other words, determine a sentence span similar to named entities provided in the train set, with no training required beforehand. In some cases, for instance, when few-shot scenarios are

considered (where only a few examples are available), this approach may be beneficial (problem was investigated in Section 6.2).

Other examples can be found in the field of Information Retrieval (IR). When text documents are considered, the typical IR scenario is a provision of ranked search results for a given text query entered by a user. Search results can be either full documents or spans of texts, and each of the mentioned scenarios poses different challenges (Mitra & Craswell, 2018).

Many modern approaches to Information Retrieval rely on a straightforward comparison of dense embeddings representing query documents and candidate documents, determining optimal results using k -nearest neighbor search (Boytsov et al., 2016; Brokos et al., 2016; Gysel et al., 2018; Kim et al., 2017; Schmidt et al., 2019). When such end-to-end retrieval systems are considered, the main question becomes how to determine reliable representations of documents (Gillick et al., 2018).

To take the approach to Information Retrieval described above, one has to already know the boundaries of units to be returned, e.g., assume sentences or paragraphs should be considered as possible results. A more challenging problem arises when we do not search for a predefined text fragment (e.g., entire document or whole sentence) but are expected to return any possible and adequate sub-sequence in a document (e.g., few sentences, several words, or even one word). This is the case for many real-world scenarios, where documents lack accessible formal structure, and one is expected to determine spans in

* Corresponding author.

E-mail addresses: lukasz.borchmann@applica.ai (Ł. Borchmann), dawid.jurkiewicz@applica.ai (D. Jurkiewicz), filip.galinski@applica.ai (F. Galiński), tomasz.gorecki@amu.edu.pl (T. Górecki).

¹ Equal contribution.

List of Symbols

\mathbb{E}	Set of embeddings, each embedding represent different sequence from set \mathbb{S}
\mathbb{P}	Exponentially explosive set of all possible warping paths through the grid
\mathbb{S}	Set of time-depended sequences $\mathbb{S}; \mathbb{S} := \{\mathcal{X}_1, \dots, \mathcal{X}_h\}$
\mathcal{X}	Time-dependent sequence to align within target sequence $\mathcal{Y}; \mathcal{X} := (x_1, \dots, x_n)$
\mathcal{X}'	Reversed sequence of $\mathcal{X}; \mathcal{X}' := (x_n, \dots, x_1) = (x'_1, \dots, x'_n)$
\mathcal{Y}	Time-dependent target sequence; $\mathcal{Y} := (y_1, \dots, y_m)$
\mathcal{Y}'	Reversed sequence of $\mathcal{Y}; \mathcal{Y}' := (y_m, \dots, y_1) = (y'_1, \dots, y'_m)$
\mathcal{Z}	Consensus sequence at the current iteration; $\mathcal{Z} := (z_1, \dots, z_q)$
\mathcal{Z}^*	Final consensus sequence
a	Hyperparameter of the smooth inverse frequency (SIF) method
b	Number of iterations needed for DTW Barycenter Averaging (DBA) to converge
$c(x_i, y_j)$	Local cost measure for domain-specific objects x_i and y_j e.g., cosine distance between word embeddings
$C_p(\mathcal{X}, \mathcal{Y})$	Cost of the warping path p between \mathcal{X} and $\mathcal{Y}; C_p(\mathcal{X}, \mathcal{Y}) := \sum_{s=1}^k c(x_{i_s}, y_{j_s})$
D	Accumulated cost matrix of size $n \times m$ calculated from \mathcal{X}, \mathcal{Y}
D'	Accumulated cost matrix of size $n \times m$ calculated from $\mathcal{X}', \mathcal{Y}'$
$D'_{i,j}$	Item from i th row and j th column of matrix D calculated from $\mathcal{X}_i, \mathcal{Y}_j$
e	Element of set \mathbb{E}
e_u	Embedding representing sequence u
f_i	Relative frequency of the token t_i
h	Size of set \mathbb{S}
i	Index of i th element of \mathcal{X}
j	Index of j th element of \mathcal{Y}
j_1^*	Index of the beginning of optimal sub-sequence alignment in \mathcal{Y}
j_k^*	Index of the end of optimal sub-sequence alignment in \mathcal{Y}
$j_1^{* \prime}$	Index of the beginning of optimal sub-sequence alignment in \mathcal{Y}' ; $j_1^{* \prime} = m - j_k^* + 1$
$j_k^{* \prime}$	Index of the end of optimal sub-sequence alignment in \mathcal{Y}' ; $j_k^{* \prime} = m - j_1^* + 1$
k	Length of warping path p
l	Index of l th element of set \mathbb{S}
n	Length of sequence \mathcal{X}
n_l	Length of sequence \mathcal{X}_l
m	Length of sequence \mathcal{Y}
p	Warping path; $p := (p_1, \dots, p_s, \dots, p_k)$
p^*	Optimal warping path; $p^* := \operatorname{argmin}_{p \in \mathbb{P}} (C_p(\mathcal{X}, \mathcal{Y}))$
p_1^*	First element of optimal warping path in D ; $p_1^* = (1, j_1^*)$
p_k^*	Last element of optimal warping path in D ; $p_k^* = (n, j_k^*)$
$p_1^{* \prime}$	First element of optimal warping path in D' ; $p_1^{* \prime} = (1, j_1^{* \prime})$
$p_k^{* \prime}$	Last element of optimal warping path in D' ; $p_k^{* \prime} = (n, j_k^{* \prime})$
q	Length of sequence \mathcal{Z}
r	Length of the u sub-sequence
s	Index of s th element of warping path p
t_i	i th token corresponding to i th element of \mathcal{X}

u	Sub-sequence from \mathcal{Y} similar to sequences from set $\mathbb{S}; u := (u_1, \dots, u_r)$
u^*	Sub-sequence from \mathcal{Y} most similar to sequences from set \mathbb{S}
w	Additional weight factor applied to the DTW equation
x_i, y_j	Domain-specific objects e.g., word embeddings

natural language streams (Borchmann et al., 2020; Vanderbeck et al., 2011). Take an example of a lawyer or researcher searching for crucial parts of legal documents to determine whether they contain fairness policies and how these policies look like (Nagpal et al., 2018).

As shown later, it is possible to tackle the problem with a proper sub-sequence matching strategy, which can incorporate all given examples to retrieve suitable text span (Section 6.1).

We solve the problems stated above with unconventionally used Dynamic Programming algorithms and propose their modifications. In particular, the well-known DTW Barycenter Averaging heuristic is evaluated in a new scenario, where word embeddings are used to determine document spans. More importantly, a new sub-sequence matching method is introduced, performing a search by multiple examples simultaneously. This matching method maximizes gain from the availability of a few semantically similar text span examples. Because of the relation of the newly introduced method to the Dynamic Time Warping algorithm, it is referred to as the Dynamic Boundary Time Warping (DBTW).

The rest of this paper is organized as follows. Section 2 summarizes related works in the areas of Information Retrieval, Natural Language Processing, and time-series mining. Section 3 describes the problem we are dealing with. Section 4 introduces the Dynamic Time Warping algorithm and its derivatives. In Section 5, we present our Dynamic Boundary Time Warping algorithm together with complexity study and its adaptation to NLP problems. Section 6 reports evaluation results on two different NLP tasks. Finally, Section 7 concludes the paper and outlines future research directions.

2. Related works

Dynamic Boundary Time Warping with maximum distance limit can be considered a binary non-parametric classifier (Boiman et al., 2008) over all possible document sub-sequences because it determines which of them represents the same class as positive examples. In such a sense, its application to few-shot semantic retrieval is related to the widely studied problem of one- and few-shot learning (e.g., Bart & Ullman, 2005; Fei-Fei et al., 2006; Koch et al., 2015; Snell et al., 2017; Sung et al., 2017). However, these approaches are not directly comparable because, in contrast to DBTW, knowledge obtained during training for previous categories is used.

Many time-series mining problems require subsequence similarity search as a subroutine. While this can be performed with any distance measure, and dozens of distance measures have been proposed in the last years, there is increasing evidence that DTW is the best measure across a wide range of domains (Ding et al., 2008). Subsequence DTW (S-DTW) is a variant of the DTW technique (Müller, 2007), which is designed to find multiple similar subsequences between two templates. One of the most cited methods is SPRING (Sakurai et al., 2007), where a query time series is searched in a larger streaming time series. Examples of subsequence matching applications are sensor network monitoring (Sakurai et al., 2007), spoken keyword spotting (Guo et al., 2012), sensor-based gait analysis (Barth et al., 2015), acoustic (Rosa et al., 2017), motion capture (Chen et al., 2009), or human action recognition in video (Hoai et al., 2011). Additionally, to speed up computations, some hardware implementations of S-DTW-based algorithms were proposed, using GPUs and FPGAs (Huang et al., 2013;

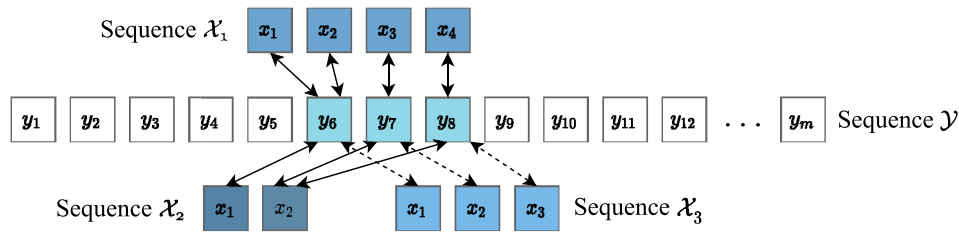


Fig. 1. The problem considered is to align multiple sequences (here $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$) optimally within the target sequence \mathcal{Y} , assuming all have to be matched to the same sub-sequence of \mathcal{Y} . Optimal alignment is one that minimizes the cost over all possible alignments. An example from Natural Language Processing is to locate a named entity within the sentence, given a few examples of other named entities.

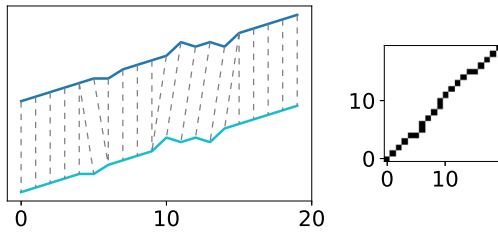


Fig. 2. DTW between two time series and the optimal alignment path. The dashed line connects elements aligned between up and down time series. The plot on the right depicts which time step was aligned to which, with each off-diagonal move indicating warping.

Rakthanmanon et al., 2013; Sart et al., 2010). Further optimizations could be achieved, e.g., by learning a kernel approximating DTW as proposed by Candelieri et al. (2019) or replacing DTW with Pruned-DTW (Silva & Batista, 2016), an exact algorithm for speeding up DTW matrix calculation.

There have been a few attempts to utilize Dynamic Time Warping in Natural Language Processing. Matuschek et al. (2008) explored the earlier idea of Ratanamahatana and Keogh (2004) to treat texts as bit streams for the purposes of measuring text similarity. Liu et al. (2007) utilized DTW with WordNet-based word similarity to decide the semantic similarity of sentences. Zhu et al. (2017) used DTW with word embeddings distances to determine the similarity between paragraphs of text to decide the similarity between whole documents. Although sub-sequence DTW was successfully applied to query-by-example tasks of spoken term detection (e.g., Hazen et al., 2009; Parada et al., 2009), to the best of our knowledge, we are the first to apply it to plain-text query-by-example tasks. Moreover, we are unaware of any existing adaptations of sub-sequence DTW for querying by multiple examples simultaneously.

3. Problem statement

The general problem considered is to align multiple sequences of possibly different lengths from the set \mathbb{S} optimally within some target sequence \mathcal{Y} , assuming all have to be matched to the same sub-sequence of \mathcal{Y} (see Fig. 1).

The total cost of alignment between sequences from \mathbb{S} and sub-sequence of the \mathcal{Y} sequence is the sum of distances between all pairs of matched elements. Distance between two elements is some domain-specific measure, such as the absolute difference between scalars associated with these elements. Optimal alignment is one that finds such sub-sequence of \mathcal{Y} that the cost of aligning all \mathbb{S} within this sub-sequence is minimized over all possible sub-sequences of \mathcal{Y} . Sections 4.1 and 4.2 provide a formal definition of the mentioned objective under additional requirements of monotonicity and continuity.

An example real-word problem from Natural Language Processing is Named Entity Recognition, which may be considered under this paradigm, when one has to locate a named entity within the sentence,

given a few examples of other named entities (Fig. 3). Another case is semantic retrieval of legal clauses from unstructured documents, given examples of clauses covering the same topic of interest from other documents.

Note that the problem mentioned above is a generalization of every problem previously considered as a sub-sequence matching to the cases when multiple examples are available instead of a single one. Problems outside the NLP to be considered under this framework include spoken term detection or temporal activity detection in continuous, untrimmed video streams, which resembles the mentioned approach to semantic retrieval if one realizes it is in principle possible to perform sub-sequence matching on video frames.

4. Dynamic time warping

Let us start with an introduction of a widely used Dynamic Time Warping algorithm since evaluated methods either directly use one of its variants or propose its generalization to multiple alignment scenarios. DTW is a classical and well-established distance measure well suited to the task of comparing time series (Berndt & Clifford, 1994) and was proposed by Vintsyuk (1968).

In general, DTW is based on the calculation of an optimal match between two given sequences, assuming one sequence is a time-warped version of another, that is, the target sequence is either stretched (one-to-many alignment), condensed (many-to-one alignment), or not warped (one-to-one alignment) concerning the source sequence (Fig. 2). The optimal match is the one with the lowest cost computed as the sum of (predominantly Euclidean) distances for each matched pair of points.

4.1. Algorithm

Classic DTW algorithm compares sequences assuming the first elements, and the last elements in both sequences are to be matched. In the case of natural language, this means that given two sentences (or documents), in every case, the first words of these will be linked with each other, as well as the last words. Although this variant is of no use in problems we consider in the present paper (see Section 1), there is a need to introduce it before going further.

The process of determining the optimal match between two time-dependent sequences $\mathcal{X} := (x_1, \dots, x_n)$ and $\mathcal{Y} := (y_1, \dots, y_m)$ (where $x_1, \dots, x_n, y_1, \dots, y_m$ are domain-specific objects, e.g., word embeddings) can be conducted on the $n \times m$ unit grid (Fig. 4). The path through the grid $p = (p_1, \dots, p_s, \dots, p_k)$ where $p_s = (i_s, j_s)$ is referred to as the warping path, whereas the total cost of the warping path p between \mathcal{X} and \mathcal{Y} is given by the sum of the local cost measures for the underlying grid nodes:

$$C_p(\mathcal{X}, \mathcal{Y}) := \sum_{s=1}^k c(x_{i_s}, y_{j_s}).$$

where c is a local cost measure as defined by Müller (2007).²

² In Section 5.2 we propose a local cost measure specifically tailored for problems in the NLP field.

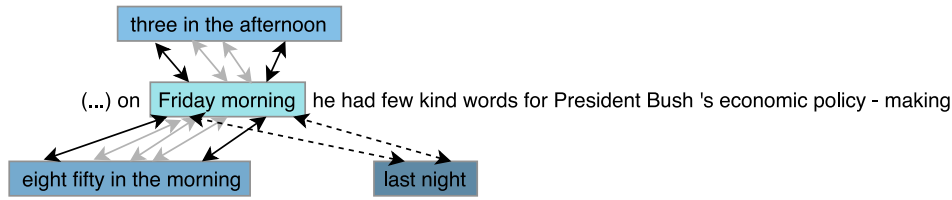


Fig. 3. The DBTW matching using the semantic distance between word embeddings applied to the Named Entity Recognition problem. Here, the three examples of time expressions were matched to the *Friday morning* sub-sequence.

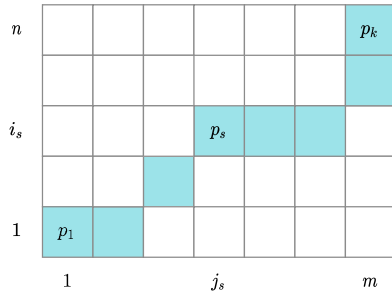


Fig. 4. The problem of determining the optimal match between sequences considered on $n \times m$ unit grid.

It can be further normalized with division by $n + m$, leading to the *time-normalized cost*.

Let \mathbb{P} denote an exponentially explosive set of all possible warping paths through the grid. The Dynamic Time Warping algorithm determines the best alignment path (*optimal warping path*)

$$p^* = \operatorname{argmin}_{p \in \mathbb{P}} (C_p(\mathcal{X}, \mathcal{Y}))$$

in $\mathcal{O}(nm)$ time, assuming:

- the alignment path has to start at the bottom left of the grid ($i_1 = 1$ and $j_1 = 1$), that is the first points in both sequences are matched,
- monotonicity ($i_{s-1} \leq i_s$ and $j_{s-1} \leq j_s$), that is moves to the left (back in time) on the grid are not allowed,
- continuity ($i_s - i_{s-1} \leq 1$ and $j_s - j_{s-1} \leq 1$) that is no node on a path can be skipped,
- the alignment path ends at the top right of the grid ($i_k = n$ and $j_k = m$), that is the last points in both sequences are matched,
- optional conditions regarding the warping window or slope constraint that can be applied in order to improve performance (Sakoe & Chiba, 1990).

Let D denote the $n \times m$ matrix referred to as the *accumulated cost matrix*. The problem stated can be solved with the following initial conditions:

$$\begin{aligned} D_{i,1} &:= \sum_{l=1}^i c(x_l, y_1), \quad \text{for } i \in \{1, \dots, n\}, \\ D_{1,j} &:= \sum_{l=1}^j c(x_1, y_l), \quad \text{for } j \in \{1, \dots, m\}. \end{aligned} \tag{1}$$

and the following dynamic programming equation, calculated recursively in ascending order:

$$D_{i,j} := c(x_i, y_j) + \min \begin{cases} D_{i,j-1}, \\ D_{i-1,j-1}, \\ D_{i-1,j}. \end{cases}$$

The value of $D_{n,m}$ (accumulated cost after reaching the top-right of the grid) is the total cost of the best alignment path:

$$\operatorname{DTW}(\mathcal{X}, \mathcal{Y}) := C_{p^*}(\mathcal{X}, \mathcal{Y}).$$

4.2. Sub-sequence DTW

Mining scenarios considered in the introduction (such as Named Entity Recognition or Information Retrieval from untrimmed text streams) require slightly different behavior, offered by DTW operating on sub-sequences. It was initially introduced for problems such as the detection of spoken terms in audio recording.

In the case of sub-sequence DTW, the constraints on admissible paths are relaxed. Boundary conditions $j_1 = 1$ and $j_k = m$ are withdrawn, so the remaining $i_1 = 1$ and $i_k = n$ guarantee that the shorter sequence \mathcal{X} will be matched entirely within \mathcal{Y} , but not necessarily starting from the beginning of \mathcal{Y} (and not obligatorily ending at the end of it). This behavior is achieved by a modification of the initial conditions described by Eq. (1). Before recursively calculating the remaining values of D the first row and first column, are being set to Müller (2007):

$$\begin{aligned} D_{i,1} &:= \sum_{l=1}^i c(x_l, y_1), \quad \text{for } i \in \{1, \dots, n\}, \\ D_{1,j} &:= c(x_1, y_j), \quad \text{for } j \in \{1, \dots, m\}. \end{aligned} \tag{2}$$

Minimal value from the m th row of D is the total cost of the best alignment path $s\operatorname{DTW}(\mathcal{X}, \mathcal{Y})$, whereas its index points to the i_k .

4.3. Multi-sequence DTW

What if one has to determine a single sub-sequence warping path for a set of short sequences? This is the case we want to consider in the present paper because this applies to few-shot semantic retrieval tasks and Named Entity Recognition. For example, it is expected to align multiple sub-sequences (named entities from train set) optimally within the target sequence (sentence or document to detect new named entities in).

4.3.1. Exact solution

Unfortunately, it is impossible to provide an exact solution due to practical reasons resulting from computational complexity.

As shown by Wang and Jiang (1994), multiple sequence alignment with the *sum of all pairs score*³ is an NP-complete problem. In particular, the problem of aligning h sequences can be solved by applying DTW on the h -dimensional cuboid (see Fig. 5). Assuming sequences are of the lengths n_1, \dots, n_h , the algorithm would take $\Theta(\prod_{l=1}^h n_l)$ operations and would require an exponential space, meaning that calculating it for larger h is not possible in most cases (Petitjean et al., 2011).

4.3.2. Barycenter averaging

A reference heuristic for aligning multiple sub-sequences within the target sequence relies on the construction of an average, consensus sequence, representative for a given set of sentences. The term *consensus sequence* refers to a sequence which represents the most commonly encountered pattern in the set of sequences (Pierce, 2017). To approximate the optimal solution to the problem with multiple sequences, one

³ When SP-score is considered, optimal alignment is one that minimizes the value over all possible alignments (Bonizzoni & Della Vedova, 2001).

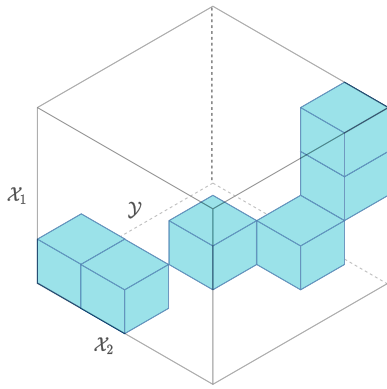


Fig. 5. The problem of determining the optimal match between sequences X_1, X_2, Y considered on the rectangular cuboid. Computing the optimal match would have $\mathcal{O}(n_1 n_2 m)$ time complexity.

can compute sub-sequence DTW between such consensus sequence and target sequence.

Petitjean et al. (2011) proposed the DTW Barycenter Averaging (DBA), the method for constructing consensus sequence inspired by computational biology. According to the authors, it *builds an average sequence around significant states of the data, which is truly representative of the underlying phenomenon.*

The algorithm assumes the iterative computation of an averaged sequence (See lines 2–7 from Algorithm 1). Let $Z = (z_1, \dots, z_q)$ denote the consensus sequence at the current iteration. First, the initial Z is set (e.g., as a randomly selected element of \mathbb{S}). Then, during each iteration:

- for each $X \in \mathbb{S}$, DTW(X, Z) is calculated and underlying associations⁴ resulting from the optimal warping path are stored,
- Z is updated as an average of the associated sequence’s members e.g., word embeddings.

During this process, the initial averaging is being refined since the new Z is closer to the sequences it averages concerning the total cost. The process finishes when a new consensus sequence Z_{new} is almost equal to the previous consensus sequence Z_{old} or when the maximum number of iterations⁵ is reached. For a thorough, detailed description of DBA, please refer to Algorithm 5 from Petitjean et al. (2011).

Strictly speaking, to handle the set of sequences $\mathbb{S} = \{X_1, \dots, X_n\}$ to be aligned within \mathcal{Y} , one can first determine the consensus sequence Z^* from \mathbb{S} using DBA, and then utilize a standard sub-sequence DTW algorithm for two sequences (See Algorithm 1). This approach resembles the nearest centroid classifier (Tibshirani et al., 2002) since one is determining class prototype and rely on distances between it and candidate sequences.

5. Novel solution: Dynamic boundary time warping

Contrary to the DBA, we propose a method that does not average sub-sequences before determining the best match. Simultaneously, there is a low computational cost involved, even though a form of multi-alignment is being performed.

Note that, for Information Retrieval, we are often interested only in approximating the p_1^* and p_k^* (more strictly the j_1^* and j_k^* components),⁶

⁴ We mean DTW associations like in Fig. 1. For example y_6 from Fig. 1 is associated with 4 sequence’s members x_1, x_2 from X_1 , x_1 from X_2 and x_1 from X_3 . Analogously z_1 from Z could also be associated with sequence’s members from each $X \in \mathbb{S}$.

⁵ For simplicity we omitted constraint on a number of maximum iterations criterion in Algorithm 1.

⁶ For instance, when retrieving text spans, we do not care about the alignment with the search query, but only the content (defined by j_1 and j_k).

Algorithm 1 DTW Barycenter Averaging based solution for aligning set of sequences \mathbb{S} within target sequence \mathcal{Y} .

DBA is the Algorithm 5 from Petitjean et al. (2011).

```

1: procedure MATCHUSINGDBA( $\mathbb{S}, \mathcal{Y}$ )
2:    $Z_{new} \leftarrow$  random element from set  $\mathbb{S}$ 
3:   do
4:      $Z_{old} \leftarrow Z_{new}$ 
5:      $Z_{new} \leftarrow$  DBA( $Z_{old}, \mathbb{S}$ )
6:   while  $Z_{old} \not\approx Z_{new}$ 
7:    $Z^* \leftarrow Z_{new}$ 
8:   return sDTW( $Z^*, \mathcal{Y}$ )
9: end procedure

```

that is the beginning and the end of the optimal warping path concerning the set of short sequences \mathbb{S} and long sequence \mathcal{Y} . In other words, we want to find j_1 and j_k that would minimize the sum of warping paths costs between each sequence $X \in \mathbb{S}$ and the long sequence \mathcal{Y} :

$$j_1^*, j_k^* = \underset{j_1, j_k}{\operatorname{argmin}} \left(\sum_{X \in \mathbb{S}} C_p(X, \mathcal{Y}) \right).$$

Note that the final warping paths between considered sequences have the same j_1^*, j_k^* . Calculating such optimal solution is more straightforward than presented in Section 4.3.1, but still too time-consuming for long sequence \mathcal{Y} , because one would have to consider all possible j_1 and j_k pairs (see Section 5.1). The situation changes when we allow either j_1 or j_k to be different among examined warping paths, for instance, as it will be shown later (see Algorithm 2), we can easily find

$$j_k^* = \underset{j_k}{\operatorname{argmin}} \left(\sum_{X \in \mathbb{S}} C_p(X, \mathcal{Y}) \right).$$

Our algorithm exploits this fact, and searches for the j_k first (j_1 being unconstrained), and then for j_1 given previously determined optimal j_k . We will use the name Dynamic Boundary Time Warping to highlight this difference when referring to the proposed solution.

Let us introduce the generalized DTW (or gDTW) first. We will use this term when referring to the DTW that is parameterized by the pre-initialized accumulated cost matrix D . For example, for D initialized from Eq. (1):

$$\text{gDTW}(X, \mathcal{Y}, D_{(1)}) = \text{DTW}(X, \mathcal{Y})$$

and for D initialized from Eq. (2):

$$\text{gDTW}(X, \mathcal{Y}, D_{(2)}) = \text{sDTW}(X, \mathcal{Y}).$$

DBTW degenerates to sDTW in the case of $|\mathbb{S}| = 1$, that is when only one example is available. The complete computation when multiple examples are given is detailed in Algorithm 2 and Algorithm 3. We propose to handle the problem as follows:

- Initialize the accumulated cost matrix D from Eq. (2) for each of the \mathbb{S} elements independently.
- Calculate sDTW for each of the \mathbb{S} elements independently, time-normalize underlying accumulated cost matrices, and sum their m th rows. The result can be used to determine $p_k^* = (i_k^*, j_k^*)$ analogously to the conventional sub-sequence DTW.
- Reverse \mathcal{Y} , as well as all sequences in \mathbb{S} , and initialize D' for each reversed sequence from \mathbb{S} :

$$\begin{aligned}
 D'_{i,1} &:= \sum_{i=1}^n c(x'_i, y'_1) \quad \text{for } i \in \{1, \dots, n\}, \\
 D'_{1,j} &:= \infty \quad \text{for } j \in \{1, \dots, m\} \setminus j_1^{f*}, \\
 D'_{1,j_1^{f*}} &:= c(x_1, y_{j_1^{f*}}),
 \end{aligned} \tag{3}$$

where $j_1^{f*} = m - j_k^* + 1$.

Algorithm 2 Approximation of optimal j_k for the multiple sub-sequences DTW problem.

```

1: procedure MULTIWARPINGEND( $\mathbb{S}, \mathcal{Y}, \text{equation}$ )
2:    $\bar{s}um \leftarrow (0, \dots, 0)$ 
3:   for  $l \leftarrow 1, |\mathbb{S}|$  do
4:      $D^l \leftarrow D^l$  from equation
5:      $gDTW(\mathcal{X}_l, \mathcal{Y}, D^l)$ 
6:      $\bar{s}um \leftarrow \bar{s}um + D^l_{n,*}$ 
7:   end for
8:    $j_k \leftarrow \text{argmin}_i(\bar{s}um_i)$ 
9:   return  $j_k$ 
10: end procedure

```

Algorithm 3 Approximation of optimal j_1 and j_k for the multiple sub-sequences DTW problem.

```

1: procedure REV( $\mathcal{X}$ ) ▷ Sequence  $(x_1, \dots, x_n)$ 
2:   return  $(x_n, x_{n-1}, \dots, x_1)$ 
3: end procedure
4:
5: procedure MATCHUSINGDBTW( $\mathbb{S}, \mathcal{Y}$ )
6:    $j_k \leftarrow \text{MULTIWARPINGEND}(\mathbb{S}, \mathcal{Y}, \text{Eq. (2)})$ 
7:    $\mathcal{Y}' \leftarrow \text{REV}(\mathcal{Y})$ 
8:    $\mathbb{S}' \leftarrow \{\text{REV}(\mathcal{X}) : \mathcal{X} \in \mathbb{S}\}$ 
9:    $j'_k \leftarrow \text{MULTIWARPINGEND}(\mathbb{S}', \mathcal{Y}', \text{Eq. (3)})$ 
10:   $j_1 \leftarrow m - j'_k + 1$ 
11:  return  $j_1, j_k$ 
12: end procedure

```

- Calculate gDTW (using D') on reversed sequences with the constraint that it should start with $p_1^{*'} = (1, m - j_k^* + 1)$, that is p_k^* after reversal. In this way $p_k^{*'}$ is determined, which gives $p_1^{*'} = (1, m - j_k^{*' } + 1)$, that is $p_k^{*'}$ after reversal.

Note that DBTW first finds an optimal, common j_k^* for all sequences in \mathbb{S} (starting indexes could be different). Then, all sequences are reversed, and $j_k^{*'}$ is determined by forcing the algorithm to start from $j_1^{*'}$. This way, such $j_1^{*'}$ and $j_k^{*'}$ are found that approximate an optimal solution.

5.1. Complexity study

Let us assume that the set of short sequences \mathbb{S} consists of h sequences of length n , and long sequence \mathcal{Y} is of length m .

DBA based solution from Algorithm 1 consists of two parts: (1) calculation of consensus sequence using DBA, and (2) calculation of sDTW between consensus sequence and \mathcal{Y} sequence.

As described by Petitjean et al. (2011), the time complexity of Step 1 is equal to $\Theta(bn^2h)$, where b refers to the number of iterations needed for DBA to converge. Since the complexity of Step 2 is $\Theta(nm)$, the complexity of all steps is equal to $\Theta(bn^2h + nm)$.

The most costly operation for DBTW is the MULTIWARPINGEND procedure, which for each sequence in \mathbb{S} computes gDTW with \mathcal{Y} sequence, and it is called twice. Therefore DBTW time complexity is equal to $\Theta(2nmh) = \Theta(nmh)$.

Depending on the problem setup, the time complexity of DBTW can be either smaller or higher than the complexity of the DBA solution.

Note that the optimal solution requires to compute gDTW between \mathcal{Y} and each sequence in \mathbb{S} for every possible j_1 and j_k . Since there are $\frac{m(m+1)}{2}$ such possible unique pairs of j_1 and j_k , the overall complexity is equal to $\Theta(nmh \times \frac{m(m+1)}{2}) = \Theta(nm^3h)$, which is larger than the time complexity of DBTW and in most common cases larger than the DBA solution's complexity.

5.2. Local cost for natural language processing problems

There is a need to propose a suitable local cost function to apply any DTW-based dynamic programming algorithms to problems from the

field of Natural Language Processing. We introduce a novel approach, relying on the distance between contextualized word embeddings.

5.2.1. Contextualized word embeddings

Roughly speaking, the reasoning behind word embeddings is to follow the distributional hypothesis, according to which *difference of meaning correlates with the difference of distribution* (Harris, 1954). This means words sharing context tend to share similar meanings, and one is able to obtain semantic representations of words by optimizing some auxiliary objective in a sizeable unlabeled text corpus.

A famous example is the Continuous Bag of Words (CBOW) model, where an average of vectors representing surrounding words is used as an input to log-linear classifier predicting the target (middle) word (Mikolov et al., 2013). This simple yet effective algorithm and the skip-gram model trained with the opposite objective have taken the world of word embeddings by storm (Young et al., 2018).

Representations provided using CBOW and similar models, however, are static. This means that when the pre-trained word embeddings are used in a downstream task, the representation of a given word is context-invariant: *wound* used as a past tense of *wind* share representation with *wound* denoting *to injure*.

Later approaches of Peters et al. (2018b), and Akbik et al. (2018) assume the use of deep language models' internal states. These, contrary to static word embeddings, are expected to capture context-dependent word semantics. Resulting contextualized word embeddings are a function of the entire input sentence, such as for a sequence of z input tokens, an associated sequence of z vectors is returned.

Early contextualized word embeddings were sourced from language models using Recurrent Neural Networks, and they are currently being replaced by language models based on the architecture of Transformers (Vaswani et al., 2017a) such as BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), or RoBERTa (Liu et al., 2019). In the case of embeddings sourced from Transformer-based language models, the representation is obtained by attending to different tokens of the input sentence (Ethayarajah, 2019).

To the best of our knowledge, only Zhu et al. (2017) used Dynamic Time Warping with word embeddings, and none of the previous attempts were based on contextualized word embeddings.

5.2.2. Distance measure

Many distance measures may be applied as local cost functions. In some domains, simple distance measures such as Euclidean distance are sufficient enough (Shieh & Keogh, 2008), whereas in other, it may be beneficial to use learned distance metric (Gündoğdu & Saraçlar, 2017).

In the case of Natural Language Processing, we propose to rely on the cosine distance between contextualized word embeddings as the local cost, which is defined as:

$$c(\mathbf{x}, \mathbf{y}) = \frac{1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}}{2}$$

where, $\|\mathbf{x}\|$ is ℓ_2 -norm, and $\mathbf{x} \cdot \mathbf{y}$ is the dot product of the two vectors.

It is the most common metric used in NLP tasks when dissimilarity between two word vectors is considered (Faruqui et al., 2016).

5.2.3. Optional weighting

Methods of determining document similarity tend to benefit from the inclusion of frequency or distribution information, such as in Inverse Document Frequency (Metzler, 2008) or Smooth Inverse Frequency (SIF) weighting (Arora et al., 2017). We propose to further extend the algorithm with the additional weight factor w applied to the DTW equation:

$$D_{i,j} := w_i \cdot c(x_i, y_j) + \min \begin{cases} D_{i,j-1}, \\ D_{i-1,j-1}, \\ D_{i-1,j}. \end{cases}$$

The w_i is defined as the SIF of the underlying token t_i :

$$w_i^{SIF} = \frac{a}{a + f_i},$$

where f_i stands for relative frequency of the token t_i and a is the weight parameter, recommended to be between 10^{-3} and 10^{-4} (Arora et al., 2017).

The intuition behind the introduction of such weighting is to capture the importance of the token when calculating an accumulated cost, in such a way that less informative (more probable) words contribute less to the final score.

5.3. Implementation details

The performance of local cost calculations is the primary factor when one is bound by time or resource restrictions in the case of DTW and similar algorithms (Myers et al., 1980). Since a cosine distance between word embeddings is used in our scenario, there is a need to calculate at least $n \times m$ distances (for the one-shot scenario) between vectors of 768 or more components, where n denote the number of words in positive example and m stands for the length of the document.

We were able to compute them efficiently with GPU and CUDA parallel computing platform. In our PyTorch-based implementation (Paszke et al., 2019) for given input matrices representing embeddings of sequences to compare, a matrix of cosine distances is returned. It is further cast to NumPy array (Oliphant, 2006) used in the Dynamic Programming part, which is implemented using Numba (JIT compiler translating Python and NumPy code into fast machine code, see Lam et al. (2015)).

6. Evaluation

The introduced Dynamic Boundary Time Warping algorithm has broad applications in few-shot retrieval tasks from a variety of domains. We restricted ourselves to already established problems within the field of Natural Language Processing. For these, simple albeit specialized proof-of-concept solutions were provided.

In each setting, an addition to DBTW has been proposed to facilitate handling the specific problem and demonstrate the algorithm's extensibility.

6.1. Few-shot semantic retrieval

The recently proposed contract discovery task (Borchmann et al., 2020) aims to provide spans of requested target documents semantically similar to examples of spans from a few other documents. The mentioned dataset is intended to test the mechanisms that detect legal texts' regulations, given a few examples of other clauses regulating the same issue (query-by-multiple-examples scenario). Sample spans often vary in length, and the contained text is written using different vocabulary or syntax. Moreover, the text to search in lacks a formal structure, that is, no segmentation into distinct sections, articles, paragraphs, or points is given in advance.

For example, given two examples of text, where the parties agree on which jurisdiction the contract will be subject to:

This Agreement shall be governed by and construed under the laws of the State of California without reference to its rules of conflicts of laws.

This Agreement is governed by the internal laws of the State of Florida and may be modified or waived only in writing signed by the Party against which such modification or waiver is sought to be enforced.

match the following text span in another document:

Each party hereto consents to exclusive personal jurisdiction in the State of Delaware and voluntarily submits to the jurisdiction of the courts of the State of Delaware in any action or proceeding concerning this Agreement.

Because each word is represented by word embedding that reflects its meaning, and we can compute the distance between any pair of embeddings (Section 5.2), it is in principle possible to state that California is semantically quite similar to Delaware.

As a result, it is possible to attempt matching clauses such as the two shown above into the third one – word by word, embedding by embedding. Due to this fact, the problem of contract discovery is suited for the DBTW algorithm – it can be perceived as an alignment of multiple sequences (examples of desirable text spans from other legal documents) optimally within the target sequence (document in which one wants to determine a text span regulating the same issue).

Contract Discovery is evaluated with Soft F_1 metric calculated on character-level spans, as implemented in *GEval* tool (Graliński et al., 2019). Roughly speaking, this is the conventional F_1 measure, with precision and recall definitions altered to reflect the partial success of returning entities. As a result, identifying half of the correct span does not result in a 0 score.

Experiment. DBA and Adaptive CBOW solutions were evaluated in addition to DBTW. All utilized the same finetuned GPT-1 model, as described by Borchmann et al. (2020). We decided to utilize GPT-1 instead of GPT-2 because the authors achieved comparable results for both of them. At the same time, the latter has more parameters, larger embeddings, and more fine-grained tokenization, while all of these have a significant performance impact.

The GPT-1 Language Model we used was originally introduced by Radford (2018) who proposed to rely on the decoder of multi-layer Transformer (Vaswani et al., 2017b). The authors released a 12-layer model with 768-dimensional states and 12 attention heads. It uses a BPE vocabulary (Sennrich et al., 2016) consisting of 40,000 sub-word units. Borchmann et al. (2020) fine-tuned the model for 40 epochs on a corpus of legal documents, using a standard, next-word prediction objective. The authors used the initial learning rate of $5e-5$, linear learning rate decay, and Adam optimizer with decoupled weight decay (Loshchilov & Hutter, 2019).⁷ We used internal states from the last layer of the model as word embeddings, leading to the dimensionality of 768.

Because of the annotation assumptions made in this shared task, it is often beneficial to return the whole sentence, even though one can find the exact location of the desired clause (within the sentence). Consider an example of the following sentence:

This Agreement shall be governed by and construed and enforced in accordance with the laws of the State of Georgia...

...as to all matters regardless of the laws that might otherwise govern under principles of conflicts of laws applicable thereto.

Here, DBTW selects only the first part, and it would be desirable to highlight it for an end user in the real-world application. Nevertheless, it was preferred to keep the complete sentence as an expected clause during the preparation of Borchmann et al. (2020) dataset. The annotator selected an incomplete sentence only when the remaining, non-important part was of a greater length than the crucial one, which contains the desired information. That is the reason why we were returning results rounded in order to match the entire sentence that “clause core” was found in.

⁷ Both model and the corpus are publicly available at <http://github.com/applicai/contract-discovery>

Algorithm 4 Finding one similar sub-sequence $u = (u_1, \dots, u_r)$ from \mathcal{Y} to \mathbb{S} sequences given starting index j .

```

1: procedure SIM( $\mathbb{S}, u$ )
2:    $\mathbb{E} \leftarrow \{\text{MEAN}(\mathcal{X}) : \mathcal{X} \in \mathbb{S}\}$ 
3:    $e_u \leftarrow \text{MEAN}(u)$ 
4:    $scores \leftarrow \{c(e_u, e) : e \in \mathbb{E}\}$ 
5:   return MEAN(scores)
6: end procedure
7:
8: procedure FINDONE( $\mathbb{S}, \mathcal{Y}, j$ )
9:    $u^* \leftarrow (y_j)$ 
10:   $u \leftarrow ()$ 
11:  while  $j + 1 < m$  and  $u \neq u^*$  do
12:    if  $\text{SIM}(\mathbb{S}, u) < \text{SIM}(\mathbb{S}, (u, y_{j+1}))$  then
13:       $u \leftarrow (u, y_{j+1})$ 
14:       $u^* \leftarrow u$ 
15:    end if
16:     $u' \leftarrow (u_2, \dots, u_r)$ 
17:    if  $\text{SIM}(\mathbb{S}, u) < \text{SIM}(\mathbb{S}, u')$  then
18:       $u \leftarrow u'$ 
19:       $u^* \leftarrow u$ 
20:    end if
21:  end while
22:  return  $u^*, \text{SIM}(\mathbb{S}, u)$ 
23: end procedure

```

Algorithm 5 Finding most similar subsequence $u = (u_1, \dots, u_r)$ from \mathcal{Y} given \mathbb{S} sequences using ACBOW algorithm.

```

1: procedure MATCHUSINGACBOW( $\mathbb{S}, \mathcal{Y}, overlap$ )
2:    $u^*, score^* \leftarrow \text{FINDONE}(\mathbb{S}, \mathcal{Y}, 1)$ 
3:    $u \leftarrow u^*$ 
4:   while  $|u| < m$  do
5:      $j \leftarrow |u| + 1 - |u^*| \times overlap$ 
6:      $u, score \leftarrow \text{FINDONE}(\mathbb{S}, \mathcal{Y}, j)$ 
7:     if  $score > score^*$  then
8:        $u^*, score^* \leftarrow u, score$ 
9:     end if
10:  end while
11:  return  $u^*, score^*$ 
12: end procedure

```

Baseline. In Algorithm 5 we introduce the Adaptive Continuous Bag of Words (ACBOW), a simple and fast algorithm, that represents a straightforward, natural approach to tackling the problem. Roughly speaking, the idea is to move with a constantly changing window over tokens from \mathcal{Y} and determine the best sub-sequence (Algorithm 4). Embeddings for each text fragment are averaged and the resulting vectors compared with cosine similarity. In the case of multiple sequences, an average of individual similarities to the considered window is used (procedure SIM in Algorithm 4).

Note that the ACBOW for which the results were reported in Table 1 differs from the ACBOW Algorithm 5. The former was extended with a possibility to look into the future and check if adding more tokens would improve an overall score, even when some of them temporarily lower the similarity.

Results. Table 1 summarizes the Soft F_1 scores achieved. Contrary to what one might suspect, the Adaptive CBOW baseline was unable to provide satisfactory results. Scores of the sub-sequence DTW with a DBA-determined consensus sequence were substantially higher. The usage of cosine distance instead of Euclidean seems beneficial in the case of DBA used with word embeddings. DBTW performs the best, and its effectiveness can be attributed to both inverse frequency weighting and the proposed way of handling multiple sequences. The new method proposed in this paper slightly outperforms the method presented

Table 1

Results of solutions based on the same finetuned GPT-1 model as described by Borchmann et al. (2020), obtained on test set.

Method	Soft F_1
Borchmann et al. (2020)	
-fICA	.47
+fICA	.49
ACBOW	.35
DBA	
Euclidean	.43
Cosine	.44
DBTW	
-SIF	.47
+SIF ($a = 10^{-3}$)	.50
+SIF + fICA	.51

by Borchmann et al. (2020) even when fICA projection⁸ of embeddings was not applied. It is worth mentioning that SIF weighting does not lead to an improvement in the aforementioned paper. Results were even better when both SIF and fICA projection was used.

There are several distinguishing features the improvement over Borchmann et al. (2020) can be attributed to. First of all, there is a reduction of noise that occurs in DBTW. Recall the example of the governing law clause presented at the beginning of Section. The first part of the sentence contains information required to correctly classify the clause, whereas the rest is a potential noise source. The DBTW considers all the possible sub-sentences and is not restricted to the sentence boundaries, as is the method proposed by Borchmann et al. (2020). Secondly, DBTW is not order-invariant, and thus it can easily capture key phrases and word n-grams. Thirdly, DBTW operates on word-level, whereas other methods rely on averaged representation of multiple, possibly a few hundred words. The latter results in yet additional noise and information loss.

Moreover, note that Borchmann et al. (2020) chose the most similar spans from the sentence n-grams. Although their approach leads to comparable results to those obtained with DBTW, it could be applied to a limited number of problems when the number of considered n-grams is low. In contrast, DBTW is not subject to such constraints and can effectively search for a very long sequence. For example, when word-level (instead of sentence-level) sequences are considered, they often become much longer, and the n-gram based methods would be too expensive computationally.

Most of the mentioned advantages also apply to the DBA. However, one may hypothesize that information loss occurring during the consensus sequence calculation is substantial in long passages from the Contract Discovery dataset. Similarly, ACBOW shares some desired properties of DBTW (e.g., consideration of arbitrary sub-sequence on word-level) but, contrary to the DBTW, is order-invariant and relies on noisy averaged representations of multiple word embeddings.

6.2. Few-shot named entity recognition

Named Entity Recognition is the task of tagging entities in text with their corresponding type. These differ depending on the dataset. In the case of the richly-annotated Ontonotes corpus (Pradhan et al., 2013), tags such as people and organization names, locations, languages, events, monetary values, and more are used.

There were several attempts to the NER problem in a few-shot scenario (Fritzler et al., 2019; Hofer et al., 2018). Since the mentioned setting is in line with our problem statement (Section 3), we

⁸ Borchmann et al. (2020) used decomposition of contextualized word embeddings based on Independent Component Analysis (Hyvärinen & Oja, 2000) and observed it helps to distinguish semantically differing texts. See Table 1 for comparison.

approached it to provide another proof-of-concept from the field of NLP. As outlined in Section 1, we solve the problem of Named Entity Recognition with a new approach of semantic sub-sequence matching.

Named Entity Recognition task differs substantially from Semantic Retrieval discussed in the previous section. To tackle the problem effectively, one has to notice there is a significant variance in lengths of entities to be retrieved—they can range from one word to over a dozen words within the same class. This fact could motivate non-trivial modifications of DBTW such as:

- Normalization of accumulated costs for sequences from \mathbb{S} in order to compensate the impact of longer sequences on the overall score (otherwise the longer individual warping path is, the higher would be its impact when choosing the approximately optimal path for the set of sequences).
- Preference for either contraction or expansion when determining the warping path for a single sequence, e.g., depending on its length in relation to average named entity length.

There are multiple normalization methods to consider in the former, whereas the latter may require the introduction of warping path bands to restrict the upper length of matched sub-sequence. We decided to take a more straightforward, which solves both problems at the same time:

- Given the set of sequences \mathbb{S} , take the length of the longest as a target size.
- Resample shorter sequences to reach the target size using interpolation with the spline of order 1, as implemented in `tslearn TimeSeriesResampler` (Tavenard et al., 2017).

After this step, no further normalization nor weights adjustments may be required to provide satisfactory results.

Because the number of results to be returned for a given sentence varies from zero to few, one cannot simply return the most similar sub-sequence in the case of Named Entity Recognition. We tackle the problem by introducing a threshold and return all non-overlapping paths from the given sentence, with an accumulated cost below the assumed distance level. Given a set of training examples \mathbb{S} , we calculate $\text{DBTW}(\mathbb{S} \setminus \{\mathcal{X}\}, \mathcal{X})$ for each $\mathcal{X} \in \mathbb{S}$. The threshold is calculated as the maximal cost of optimal warping path from such inner-train matches. The threshold for DBA is determined analogously.

Experiment. We roughly followed the procedure for evaluation of a few-shot NER proposed by Fritzler et al. (2019). Authors trained models on subsamples of Ontonotes development set (Pradhan et al., 2013) for each class separately.⁹ For each case, $h = 20$ sentences containing a particular named entity were selected. Besides, sentences without considered entity had all the classes replaced with 0, and part of them were added to the train set, to preserve the original distribution of the currently evaluated class. Note that h is not necessarily equal to the number of annotations available since it is common for one Ontonotes sentence to contain more than one named entity of the same type.

In our case, solutions were evaluated for $h \in [1, 10]$, since we are aiming mainly at good performance for a lower number of examples available. Moreover, ten experiments with different random seeds were conducted for each class, instead of four performed by Fritzler et al. (2019).

⁹ The original train set was used as a source of *out-of-domain* data in part of scenarios, but this does not apply to methods based on DBTW. Similarly, as a baseline, we relied on an approach, which utilizes only *in-domain* training data. See Fritzler et al. (2019) for details regarding this distinction.

Table 2
p-values for permutation t-test comparing DBA and DBTW.

n	1	2	3	4	5	6	7	8	9	10
p-value	0.9339	0.3895	0.8779	0.8803	0.0038	0.4499	0.309	0.2049	0.2161	0.1727

Table 3
p-values for permutation t-test comparing DBTW and LSTM-CRF (+char).

n	1	2	3	4	5	6	7	8	9	10
p-value	0.0001	0.0001	0.1262	0.0025	0.0606	0.0482	0.1114	0.0693	0.0819	0.7024

Baseline. LSTM-CRF used as a reference is a BiLSTM-CRF model trained on ELMo and GloVe embeddings. It follows the specification of Fritzler et al. (2019), but with the difference that trained character embeddings were not used to simplify the comparison with DBTW. Note that otherwise, one had to propose a procedure of training character embeddings compatible with DBTW, which is beyond the scope of this paper. Nevertheless, we report results of LSTM-CRF with trained character-level embeddings for the sake of completeness.

The remaining LSTM-CRF baseline, DBA, and DBTW approaches rely on the same embeddings, resulting from the concatenation of the 1024-dimensional ELMo model released by Peters et al. (2018a) with the original 50-dimensional GloVe embeddings (Pennington et al., 2014). Although Fritzler et al. (2019) trained their baselines for 20 epochs, we found our models undertrained in this setting and decided to enlarge the value to 30 epochs.

Results. Comparison of DBTW, DBA, and LSTM-CRF with the same input embeddings is presented on Fig. 6. *Span F1* score refers to a commonly used $F_{\beta=1}$ variant where exact matches of the corresponding entities are considered (Tjong Kim Sang & De Meulder, 2003).

Both DBA and DBTW outperform the LSTM-CRF baseline in a few-shot setting. Noteworthy, DTW-based methods receive near-identical scores in the experiment. In order to statistically compare methods, we decided to use the permutation t-test. The implemented test corresponds to the proposal of Chung and Romano (2013). While a permutation test requires that we see all possible permutations of the data (which can become quite large), we can easily conduct “approximate permutation tests” by simply conducting a very large number of samples (we used 10,000 permutations instead of 3,628,800 possible permutations). That process should, in expectation, approximate the permutation distribution. Obtained p-values we can find in Tables 2 and 3.

From Table 2 we can see that it is possible to reject ($\alpha = 5\%$) the null hypothesis (about equality of methods DBA and DBTW) only for $n = 5$ (the same we can read from Fig. 6). In such situations, it seems reasonable to assume that methods do not differ significantly.

Comparable results of DBTW and DBA can be potentially attributed to two factors. Firstly, named entities in Ontonotes are usually short: 58% of the test set entities consist of a single word and 21% – of two words. When one-word sub-sequences are to be considered, the methods are roughly equivalent. We expect DBTW to perform better in the case of long sequences because it is where noise related to the calculation of the DBA consensus sequence emerges. Secondly, we found the problem of determining the number of sub-sequences to return, which occurs in both DBA and DBTW, to play an important role. If the sentence contains a named entity of a particular type, the highest-scored sub-sequence can be classified as such with high confidence. E.g., we can maximize recall by withdrawing the threshold and returning the top result. Nevertheless, precision suffers without the threshold, and the simple heuristics we experimented with are unable to provide an optimal cut-off.

LSTM-CRF with character-level embeddings seems to converge faster than the LSTM-CRF baseline. It appears that it achieves scores comparable to DBTW for five and more sentences in the train set (Table 3). However, due to the reasons outlined at the beginning, the methods cannot be directly compared.

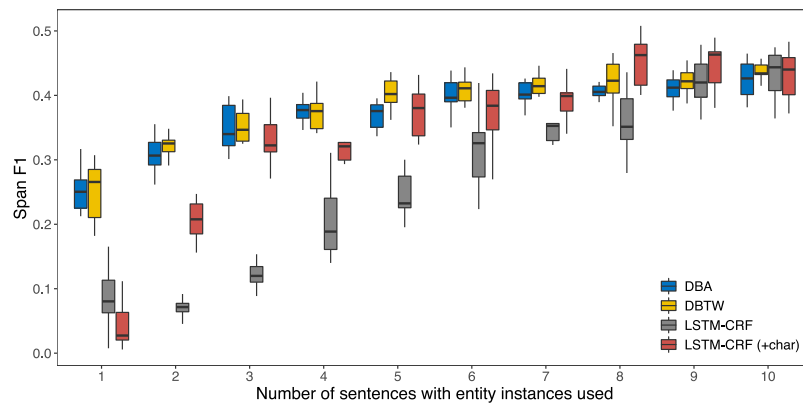


Fig. 6. Performance in Named Entity Recognition as a function of the number of sentences with positive examples available. Note that LSTM-CRF (+ char) model is not directly comparable because, contrary to the LSTM-CRF, DBA, and DBTW, it uses character-level embeddings in addition to ELMo and GloVe.

7. Summary and future work

In this paper, an algorithm inspired by Dynamic Time Warping was proposed, as well as a new application of existing DBA Barycenter Averaging heuristics. It was shown how to adapt it to current problems in the field of Natural Language Processing as a result of cosine distance applied to contextualized word embeddings. Unlike its predecessors, Dynamic Boundary Time Warping can find an approximate solution for the problem of querying by multiple examples. What is crucial, the proposed approach is in some applications substantially better than calculating a consensus sequence and utilizing it to perform sub-sequence DTW search, presumably because there is no unnecessary information loss involved. Due to the inclusion of inverse frequency weighting specific to NLP problems, its effectiveness was further improved. Thus it was able to outperform methods previously proposed for *Few-shot Contract Discovery* with the same Language Model applied.

Applications of the proposed algorithm are not limited to the cases where proof-of-concept solutions were provided, and it can be applied to other few-shot retrieval tasks. Problems outside the NLP to be considered under this framework include temporal activity detection in continuous, untrimmed video streams (Montes et al., 2016; Xu et al., 2019), which resembles mentioned approach to Semantic Retrieval if one realizes it is in principle possible to perform sub-sequence matching on video frame embeddings. Such can be encoded with a pretrained image classification network (i.e., ResNeXt Xie et al., 2016) and processed analogously. Moreover, the DBTW applies to every problem previously considered as a sub-sequence matching when multiple examples are available instead of a single one.

CRedit authorship contribution statement

Łukasz Borchmann: Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Dawid Jurkiewicz:** Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Filip Galiński:** Supervision, Writing - review & editing, Validation. **Tomasz Górecki:** Supervision, Writing - review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The Smart Growth Operational Programme, Poland supported this research under project no. POIR.01.01.01-00-0605/19 (*Disruptive adoption of Neural Language Modelling for automation of text-intensive work*).

References

- Akbik, Alan, Blythe, Duncan, & Vollgraf, Roland (2018). Contextual String Embeddings for Sequence Labeling. In *27th International conference on computational linguistics*. (pp. 1638–1649).
- Arora, Sanjeev, Liang, Yingyu, & Ma, Tengyu (2017). A simple but tough-to-beat baseline for sentence embeddings. In *5th International conference on learning representations*. OpenReview.net.
- Bart, Evgeniy, & Ullman, Shimon (2005). Cross-generalization: learning novel classes from a single example by feature replacement. In *2005 IEEE computer society conference on computer vision and pattern recognition*. (vol. 1). (pp. 672–679).
- Barth, Jens, Oberndorfer, Cécilia, Pasluosta, Cristian Federico, Schüle, Samuel, Gaßner, Heiko, Reinfelder, Samuel, Kugler, Patrick, Schulhaus, Dominik, Winkler, Jürgen, Klucken, Jochen, & Eskofier, Björn (2015). Stride segmentation during free walk movements using multi-dimensional subsequence dynamic time warping on inertial sensor data. *Sensors*, 15, 6419–6440, UnivIS-Import:2015-04-14:Pub.2015.tech.IMMD.IMMD5.stride.
- Berndt, Donald J., & Clifford, James (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining* (pp. 359–370). AAAI Press.
- Boiman, Oren, Shechtman, Eli, & Irani, Michal (2008). In defense of nearest-neighbor based image classification. In *IEEE conference on computer vision and pattern recognition*. (pp. 1–8).
- Bonizzoni, Paola, & Della Vedova, Gianluca (2001). The complexity of multiple sequence alignment with SP-score that is a metric. *Theoretical Computer Science*, 259(1–2), 63–79.
- Borchmann, Łukasz, Wiśniewski, Dawid, Gretkowski, Andrzej, Kosmala, Izabela, Jurkiewicz, Dawid, Szałkiewicz, Łukasz, Pałka, Gabriela, Kaczmarek, Karol, Kaliska, Agnieszka, & Galiński, Filip (2020). Contract discovery: Dataset and a few-shot semantic retrieval challenge with competitive baselines.
- Boytsov, Leonid, Novak, David, Malkov, Yury, & Nyberg, Eric (2016). Off the beaten path: Let's replace term-based retrieval with k-NN search. In *Proceedings of the 25th ACM international conference on information and knowledge management* (pp. 1099–1108). New York, NY, USA: Association for Computing Machinery.
- Brokos, Georgios-Ioannis, Malakasiotis, Prodromos, & Androutsopoulos, Ion (2016). Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In *Proceedings of the 15th workshop on biomedical natural language processing* (pp. 114–118). Berlin, Germany: Association for Computational Linguistics.
- Candelieri, Antonio, Fedorov, Stanislav, & Messina, Vincenzina (2019). Efficient kernel-based subsequence search for enabling health monitoring services in IoT-based home setting. *Sensors*, 19, 5192.
- Chen, Yueguo, Chen, Gang, Chen, Ke, & Ooi, Beng Chin (2009). Efficient processing of warping time series join of motion capture data. In *2009 IEEE 25th international conference on data engineering*. (pp. 1048–1059).
- Chung, EunYi, & Romano, Joseph P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2), 484–507.
- Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.
- Ding, Hui, Trajcevski, Goce, Scheuermann, Peter, Wang, Xiaoyue, & Keogh, Eamonn (2008). Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2), 1542–1552.
- Ethayarajh, Kawin (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv:abs/1909.00512v1.

- Faruqui, Manaal, Tsvetkov, Yulia, Rastogi, Pushpendre, & Dyer, Chris (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP* (pp. 30–35). Berlin, Germany: Association for Computational Linguistics.
- Fei-Fei, Li, Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4), 594–611.
- Fritzler, Alexander, Logacheva, Varvara, & Kretov, Maksim (2019). Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing* (pp. 993–1000). New York, NY, USA: ACM.
- Gillick, Daniel, Presta, Alessandro, & Tomar, Gaurav Singh (2018). End-to-end retrieval in continuous space.
- Goyal, Archana, Gupta, Vishal, & Kumar, Manish (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, 29, 21–43.
- Graliński, Filip, Wróblewska, Anna, Stanistawek, Tomasz, Grabowski, Kamil, & Górecki, Tomasz (2019). GEval: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL workshop blackboxNLP: analyzing and interpreting neural networks for NLP* (pp. 254–262). Florence, Italy: Association for Computational Linguistics.
- Gündođdu, Batuhan, & Saraçlar, Murat (2017). Distance metric learning for posterogram based keyword search. In *2017 IEEE international conference on acoustics, speech and signal processing*. (pp. 5660–5664).
- Guo, Hongyu, Huang, Dongmei, & Zhao, Xiaoqun (2012). An algorithm for spoken keyword spotting via subsequence DTW. In *2012 3rd IEEE international conference on network infrastructure and digital content* (pp. 573–576).
- Gysel, Christophe Van, de Rijke, Maarten, & Kanoulas, Evangelos (2018). Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems*, 36(4).
- Harris, Zellig S. (1954). Distributional structure. *Word*, 10(2–3), 146–162.
- Hazen, Timothy J., Shen, Wade, & White, Christopher M. (2009). Query-by-example spoken term detection using phonetic posteriorgram templates. In *2009 IEEE workshop on automatic speech recognition & understanding*. (pp. 421–426).
- Hoai, Minh, Lan, Zhen-Zhong, & la Torre, Fernando De (2011). Joint segmentation and classification of human actions in video. In *CVPR 2011*. (pp. 3265–3272).
- Hofer, Maximilian, Kormilitzin, Andrey, Goldberg, Paul, & Nevado-Holgado, Alejo (2018). Few-shot learning for named entity recognition in medical text.
- Huang, Sitao, Dai, Guohao, Sun, Yuliang, Wang, Zilong, Wang, Yu, & Yang, Huazhong (2013). DTW-Based Subsequence Similarity Search on AMD Heterogeneous Computing Platform. In *2013 IEEE 10th international conference on high performance computing and communications 2013 IEEE international conference on embedded and ubiquitous computing*. (pp. 1054–1063).
- Huang, Zhiheng, Xu, Wei, & Yu, Kai (2015). Bidirectional LSTM-CRF models for sequence tagging. CoRR, abs/1508.01991.
- Hyvärinen, Aapo, & Oja, Erkki (2000). Independent component analysis: algorithms and applications. *Neural Networks : The Official Journal of the International Neural Network Society*, 13(4-5), 411–430.
- Kim, Sun, Fiorini, Nicolas, Wilbur, W. John, & Lu, Zhiyong (2017). Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. *Journal of Biomedical Informatics*, 75, 122–127.
- Koch, Gregory, Zemel, Richard, & Salakhutdinov, Ruslan (2015). Siamese Neural Networks for One-shot Image Recognition. In *Proceedings of the 32nd international conference on machine learning*. Lille, France.
- Lam, Siu Kwan, Pitrou, Antoine, & Seibert, Stanley (2015). Numba: A LLVM-based python JIT compiler. In *Proceedings of the second workshop on the LLVM compiler infrastructure in HPC*. New York, NY, USA: Association for Computing Machinery.
- Lample, Guillaume, Ballesteros, Miguel, Subramanian, Sandeep, Kawakami, Kazuya, & Dyer, Chris (2016). Neural architectures for named entity recognition. CoRR, abs/1603.01360.
- Li, Jing, Sun, Aixin, Han, Jianglei, & Li, Chenliang (2018). A survey on deep learning for named entity recognition. arXiv:abs/1812.09449.
- Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, & Stoyanov, Veselin (2019). RoBERTa: A robustly optimized BERT pretraining approach.
- Liu, X., Zhou, Y., & Zheng, R. (2007). Sentence Similarity based on Dynamic Time Warping. In *International conference on semantic computing*. (pp. 250–256).
- Loshchilov, Ilya, & Hutter, Frank (2019). Decoupled weight decay regularization.
- Matuschek, Michael, Schlüter, Tim, & Conrad, Stefan (2008). Measuring text similarity with dynamic time warping. In *Proceedings of the 2008 international symposium on database engineering & applications (vol. 299)* (pp. 263–267). ACM.
- Metzler, Donald (2008). Generalized inverse document frequency. In *CIKM*.
- Mikolov, Tomas, Chen, Kai, Corrado, Gregory S., & Dean, Jeffrey (2013). Efficient estimation of word representations in vector space. CoRR, abs/1301.3781.
- Mitra, Bhaskar, & Craswell, Nick (2018). An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, 13, 1–126.
- Montes, Alberto, Salvador, Amaia, Pascual-deLaPuente, Santiago, & i Nieto, Xavier Giró (2016). Temporal activity detection in untrimmed videos with recurrent neural networks. In *1st NIPS workshop on large scale computer vision systems 2016*.
- Müller, Meinard (2007). Dynamic time warping. *Information Retrieval for Music and Motion*, 69–84.
- Myers, Cory, Rabiner, Lawrence, & Rosenberg, Aaron (1980). Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(6), 623–635.
- Nagpal, Rashmi, Wadhwa, Chetna, Gupta, Mallika, Shaikh, Samiulla, Mehta, Sameep, & Goyal, Vikram (2018). Extracting fairness policies from legal documents. CoRR, abs/1809.04262.
- Oliphant, Travis (2006). *A guide to numpy (vol. 1)*. Trelgol Publishing USA.
- Parada, Carolina, Sethy, Abhinav, & Ramabhadran, Bhuvana (2009). Query-by-example spoken term detection for OOV terms. In *2009 IEEE workshop on automatic speech recognition & understanding*. (pp. 404–409).
- Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, Desmaison, Alban, Kopf, Andreas, Yang, Edward, DeVito, Zachary, Raison, Martin, Tejani, Alykhan, Chilamkurthy, Sasank, Steiner, Benoit, Fang, Lu, Chintala, Soumith (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Álché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems (vol. 32)* (pp. 8024–8035). Curran Associates, Inc.
- Pennington, Jeffrey, Socher, Richard, & Manning, Christopher D. (2014). Glove: Global vectors for word representation. In *EMNLP*.
- Peters, Matthew E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, & Zettlemoyer, Luke (2018). Deep contextualized word representations. In *Proc. of NAACL*.
- Peters, Matthew E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, & Zettlemoyer, Luke (2018). Deep contextualized word representations. CoRR, abs/1802.05365.
- Petitjean, François, Ketterlin, Alain, & Gançarski, Pierre (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44, 678–693.
- Pierce, Benjamin (2017). *Genetics. A conceptual approach*. W. H. Freeman.
- Pradhan, Sameer, Moschitti, Alessandro, Xue, Nianwen, Ng, Hwee Tou, Björkelund, Anders, Uryupina, Olga, Zhang, Yuchen, & Zhong, Zhi (2013). Towards robust linguistic analysis using ontotones. In *Proceedings of the seventeenth conference on computational natural language learning* (pp. 143–152). Sofia, Bulgaria: Association for Computational Linguistics.
- Radford, A. (2018). Improving language understanding by generative pre-training.
- Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, & Sutskever, Ilya (2019). *Language models are unsupervised multitask learners: Technical report*, OpenAI.
- Rakthanmanon, Thanawin, Campana, Bilson, Mueen, Abdullah, Batista, Gustavo, Westover, Brandon, Zhu, Qiang, Zakaria, Jesin, & Keogh, Eamonn (2013). Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Transactions on Knowledge Discovery from Data*, 7(3).
- Ratanamahatana, Chotirat Ann, & Keogh, Eamonn (2004). Everything you know about dynamic time warping is wrong. In *Third workshop on mining temporal and sequential data*.
- Rosa, Marcelo, Fugmann, Elmar, Pinto, Gisele, & Nunes, Maria (2017). An anchored dynamic time-warping for alignment and comparison of swallowing acoustic signals. In *2017 39th annual international conference of the IEEE engineering in medicine and biology society*. (pp. 2749–2752).
- Sakoe, Hiroaki, & Chiba, Seibi (1990). Dynamic programming algorithm optimization for spoken word recognition. In Alex Waibel, & Kai-Fu Lee (Eds.), *Readings in speech recognition* (pp. 159–165). San Francisco: Morgan Kaufmann.
- Sakurai, Yasushi, Faloutsos, Christos, & Yamamuro, Masashi (2007). Stream monitoring under the time warping distance. In *2007 IEEE 23rd international conference on data engineering*. (pp. 1046–1055).
- Sart, Doruk, Mueen, Abdullah, Najjar, Walid, Keogh, Eamonn, & Niennattrakul, Vit (2010). Accelerating dynamic time warping subsequence search with GPUs and FPGAs. In *2010 IEEE international conference on data mining*. (pp. 1001–1006).
- Schmidt, Fabian David, Dietsche, Markus, Ponzetto, Simone Paolo, & Glavaš, Goran (2019). SEAGLE: A platform for comparative evaluation of semantic encoders for information retrieval. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing : System demonstrations* (pp. 199–204). Hong Kong, China: Association for Computational Linguistics.
- Sennrich, Rico, Haddow, Barry, & Birch, Alexandra (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (vol. 1: long papers)* (pp. 1715–1725). Berlin, Germany: Association for Computational Linguistics.
- Shieh, Jin, & Keogh, Eamonn (2008). Isax: Indexing and mining terabyte sized time series. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 623–631). New York, NY, USA: Association for Computing Machinery.
- Silva, Diego, & Batista, Gustavo (2016). Speeding up all-pairwise dynamic time warping matrix calculation. In *Proceedings of the 2016 SIAM international conference on data mining*. (pp. 837–845).
- Snell, Jake, Swersky, Kevin, & Zemel, Richard S. (2017). Prototypical networks for few-shot learning. In *NIPS*.
- Sung, Flood, Yang, Yongxin, Zhang, Li, Xiang, Tao, Torr, Philip H. S., & Hospedales, Timothy M. (2017). Learning to compare: Relation network for few-shot learning. In *2017 IEEE/CVF conference on computer vision and pattern recognition*. (pp. 1199–1208).

- Tavenard, Romain, Faouzi, Johann, & Vandewiele, Gilles (2017). Tsllearn: A machine learning toolkit dedicated to time-series data. <https://github.com/rtavenar/tslearn>.
- Tibshirani, Robert, Hastie, Trevor, Narasimhan, Balasubramanian, & Chu, Gilbert (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), 6567–6572.
- Tjong Kim Sang, Erik F., & De Meulder, Fien (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on natural language learning at HLT (vol. 4)* (pp. 142–147). USA: Association for Computational Linguistics.
- Vanderbeck, Scott, Bockhorst, Joseph, & Oldfather, Chad (2011). A machine learning approach to identifying sections in legal briefs. In *MAICS*.
- Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, & Polosukhin, Illia (2017). Attention is all you need. CoRR, abs/1706.03762.
- Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, & Polosukhin, Illia (2017). Attention is all you need.
- Vintsyuk, Taras K. (1968). Speech discrimination by dynamic programming. *Kibernetika*, 4(1), 81–88.
- Wang, Lusheng, & Jiang, Tao (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4), 337–348.
- Xie, Saining, Girshick, Ross B., Dollár, Piotr, Tu, Zhuowen, & He, Kaiming (2016). Aggregated residual transformations for deep neural networks. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 5987–5995).
- Xu, Huijuan, Das, Abir, & Saenko, Kate (2019). Two-stream region convolutional 3D network for temporal activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 2319–2332.
- Yadav, Vikas, & Bethard, Steven (2018). A survey on recent advances in named entity recognition from deep learning models. In *COLING*.
- Young, Tom, Hazarika, Devamanyu, Poria, Soujanya, & Cambria, Erik (2018). Recent trends in deep learning based natural language processing [Review Article]. *IEEE Computational Intelligence Magazine*, 13, 55–75.
- Zhu, Xiaofeng, Klabjan, Diego, & Bless, Patrick (2017). Semantic document distance measures and unsupervised document revision detection. In *Proceedings of the eighth international joint conference on natural language processing (vol. 1: long papers)* (pp. 947–956). Taipei, Taiwan: Asian Federation of Natural Language Processing.

ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them

Dawid Jurkiewicz* Łukasz Borchmann* Izabela Kosmala Filip Graliński

Applica.ai Zajęcza 15, 00-351 Warsaw, Poland
firstname.lastname@applica.ai

Abstract

This paper presents the winning system for the propaganda Technique Classification (TC) task and the second-placed system for the propaganda Span Identification (SI) task. The purpose of the TC task was to identify an applied propaganda technique given propaganda text fragment. The goal of SI task was to find specific text fragments which contain at least one propaganda technique. Both of the developed solutions used semi-supervised learning technique of self-training. Interestingly, although CRF is barely used with transformer-based language models, the SI task was approached with RoBERTa-CRF architecture. An ensemble of RoBERTa-based models was proposed for the TC task, with one of them making use of Span CLS layers we introduce in the present paper. In addition to describing the submitted systems, an impact of architectural decisions and training schemes is investigated along with remarks regarding training models of the same or better quality with lower computational budget. Finally, the results of error analysis are presented.

1 Introduction

The idea of fine-grained propaganda detection was introduced by Da San Martino et al. (2019), whose intention was to facilitate research on this topic by publishing a corpus with detailed annotations of high reliability. There was a chance to propose NLP systems solving this task automatically as a part of this year’s SemEval series. It was expected to detect all fragments of news articles that contain propaganda techniques, and to identify the exact type of used technique (Da San Martino et al., 2020).

The authors decided to evaluate Technique Classification (TC) and Span Identification (SI) tasks separately. The purpose of the TC task was to identify an applied propaganda technique given the propaganda text fragment. In contrast, the goal of the SI task was to find specific text fragments that contain at least one propaganda technique. This paper presents the winning system for the propaganda Technique Classification task and the second-placed system for the propaganda Span Identification task.

2 Systems Description

Systems proposed for both SI and TC tasks were based on RoBERTa model (Liu et al., 2019) with task-specific modifications and training schemes applied.

The central motif behind our submissions is a commonly used semi-supervised learning technique of self-training (Yarowsky, 1995; Liao and Veeramachaneni, 2009; Liu et al., 2011; Wang et al., 2020), sometimes referred to as incremental semi-supervised training (Rosenberg et al., 2005) or self-learning (Lin et al., 2010). In general, these terms stand for a process of training an initial model on a manually annotated dataset first and using it to further extend the train set by automatically annotating other dataset. Usually, only a selected subset of auto-annotated data is used, however neither selection of high-confidence examples nor loss correction for noisy annotations is performed in our case. This is why it can be considered a simplification of mainstream approaches—the *naïve* self-training.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

* Equal contribution. Author order determined by a coin flip.

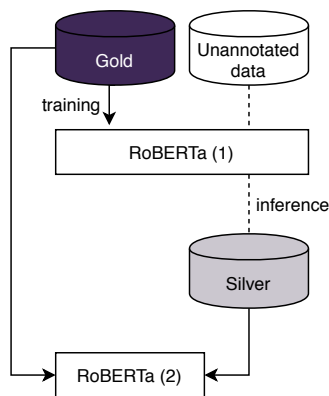


Figure 1: Self-training stands for a process of training an initial model on manually annotated dataset first and using it to further extend train by means of annotating other dataset automatically.

Hparam	SI	TC
Dropout	.1	
Attention dropout	.1	
Max sequence length	256	256
Batch size	8	16
Learning rate	5e-4	2e-5
Number of steps	60k	20k
Learning rate decay	-	-
Weight decay	-	.01
Momentum	.9	-
Optimizer	SGD	AdamW
Loss	Viterbi	BCE

Table 1: Optimizers and hyperparameters used for both fine-tuning RoBERTa and training additional parameters.

Hparam	SI	TC
Dropout	.0	
Attention dropout	.0	
Batch size	16	16

Table 2: Hyperparameter overwrites for self-training.

2.1 Span Identification

The problem of span identification was treated as a sequence labeling task, which in the case of Transformer-based language models is often solved by means of classifying selected sub-tokens (e.g., first BPE of each word considered) with or without applying LSTM before the classification layer (Devlin et al., 2019).

Although pre-Transformer sequence labeling solutions exploited CRF layer in the output (Huang et al., 2015; Lample et al., 2016), this practice was abandoned by the authors of BERT (Devlin et al., 2019) and subsequent researchers developing the idea of bidirectional Transformers, with rare exceptions, such as Souza et al. (2019) who used BERT-CRF for Portuguese NER. Contrary to the above, we approached Span Identification task with RoBERTa-CRF architecture.

The impact of this decision will be discussed in Section 3 along with remarks regarding training models of the same or better quality with a lower computational budget in an orderly fashion. In contrast, the following narrative aims at a faithful reflection of the actual way the model which we used was trained.

Recipe Take one pretrained RoBERTa_{LARGE} model, add CRF layer and train on original (gold) dataset until progress is no longer achieved with Viterbi loss, SGD optimizer, and hyperparameters defined in Table 1. Use the best-performing model to annotate random 500k OpenWebText¹ sentences automatically. Train the second model on both original (gold) dataset and autotagged (silver) one with hyperparameters defined in Table 1. Repeat the procedure two more times with the best model from the previous step, hyperparameters from Table 2, and other OpenWebText sentences.

Note that hyperparameters were indeed not overwritten during the first self-training iteration. Scores achieved by the best-performing models were respectively 50.91 (without self-training) and 50.98, 51.45, 52.24 in consecutive self-training iterations.

Many questions may arise regarding this procedure and the role of purely random factors. It is not a problem when rather the best score than its explanation is desired. In a leaderboard-driven exploration, one can simply conduct a broad set of experiments and choose the best-performing model without reflection, whether it is a byproduct of training instability. What actually happened here was investigated afterward and will be discussed in Section 3.

¹See: <https://github.com/jcpeterson/openwebtext> OpenWebText is a project aimed at the reconstruction of OpenAI’s unreleased WebText dataset.

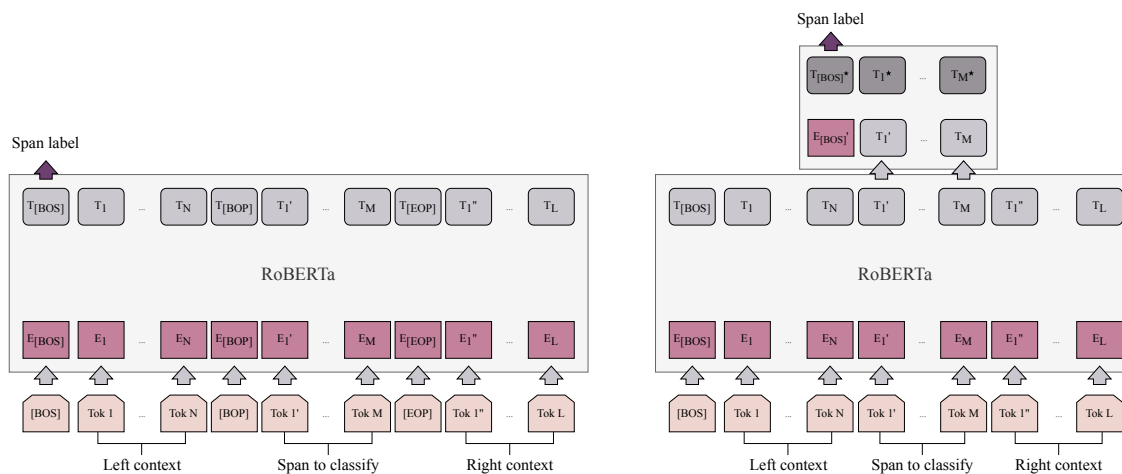


Figure 2: Comparison of span classification by means of special tokens (left) and in Span CLS approach (right). On the left, special [BOP] and [EOP] tokens are introduced, and the span is further classified as in the usual Transformer-based sentence classification task. On the right, an additional, small Transformer is stacked only over the selected tokens. It has no own embeddings apart from one for the [BOS] token, but uses representations provided by the host model instead.

2.2 Technique Classification

Transformer-based language models used in the sentence classification setting assume that representations of special tokens (such as [CLS] or [BOS]) are passed to the classification layer. Since TC task is aimed at the classification of spans, it might be beneficial to introduce information about the text fragment to be classified. We experimented with two approaches addressing this requirement.

The first assumes an injection of special tokens indicating the beginning and the end of the text marked as propaganda, such as a sample sentence before BPE applied appears as:

[BOS] Democrats acted like [BOP] babies [EOP] at the SOTU [EOS]

In this approach we continue with representation of [BOS], as in the usual sentence classification task. The second approach is to stack a small Transformer only on the selected tokens.² This one has no own embeddings apart from the ones for [BOS] but uses the host model’s representations instead. This technique is roughly equivalent to adding consecutive layers and masking attention outside the selected span and will be referred to as Span CLS. Figure 2 summarizes differences between Span CLS and classification using special [BOP] and [EOP] tokens.

The initial experiments have shown that underrepresented classes achieve lower scores. To overcome this problem, we experimented with class-dependent rescaling applied to binary cross-entropy. In this setting (further referred to as *re-weighting*) factor for each class was determined as its inverse frequency multiplied by the frequency of the most popular class. The modified loss is equal to:

$$\ell(\mathbf{x}, \mathbf{y}) = -\frac{1}{Nd} \sum_{n=1}^N \sum_{k=1}^d [p^k y_n^k \log x_n^k + (1 - y_n^k) \log(1 - x_n^k)]$$

$$p^k = \frac{1}{f^k} \max(\mathbf{f})$$

where N is the batch size, n index denotes n th batch element, d is the number of classes, \mathbf{f} stands for a vector of class absolute frequencies calculated on the train set, \mathbf{x} is the output vector from the last sigmoid layer and \mathbf{y} is a vector of multi-hot encoded ground truth labels. Note that the only difference from the original binary cross entropy for multi-label classification is the addition of the p^k class weights.

In addition to the above, a part of the tested models took the use of the self-training approach. In the case of TC task one had to identify spans first and then predict their classes to generate silver train

²The Transformer we used in our experiment had 3 hidden layers, 4 attention heads and an intermediate layer of size 512. Note that hidden size depends on host model, since we are using external embeddings.

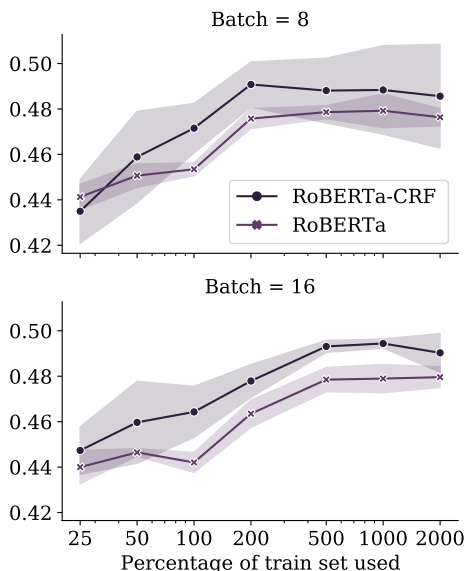


Figure 3: Performance of RoBERTa with and without CRF as a function of percentage of train set available. Values above 100% indicate self-training was performed. Mean FLC-F1 and standard deviation across 5 runs for each percentage.

CRF	Self-train	FLC-F1 (std, max)	
-	-	45.2 ± 0.3	45.6
+	-	47.4 ± 0.8	48.2
-	+	48.9 ± 0.5	50.2
+	+	49.1 ± 3.0	51.7
+	+(2)	49.7 ± 2.0	51.6
+	+(3)	50.0 ± 1.8	51.8

Table 3: Best scores on the dev set achieved with RoBERTa large model on SI task. Mean, standard deviation and maximum across 10 runs with different random seeds. Numbers in brackets indicate how many self-training iterations were used.

Batch	Dropouts	Self-train	CRF	Δ FLC-F1
16 → 8	.0 → .1	-	-	-1.1
		+	+	-1.6
	.0	-	-	-0.4
8 → 16		-	+	-3.9
	.1 → .0	-	+	-7.0
		-	-	-0.7
	.1	-	+	-1.3

Table 4: Impact of hypothetical lowering batch size during self training or enlarging batch size during initial training, as well as of enabling or disabling both hidden and attention dropouts. Change between means across 10 runs with different random seeds.

set (Figure 1). We reused our best-performing model from SI task to identify spans, and the TC model trained on ground truth to automatically annotate these spans.

Regardless of the approach taken, context as broad as possible within the 256 subword units limit was provided on both sides of the span to be classified. Note that it was a maximum equal extension of the span text in both directions, and we did not limit the extension to the sentence boundaries.

The winning TC model (described in the recipe below) was an ensemble of three models. Each of them used a different mix of previously described approaches with hyperparameters defined in Table 1 for first and second model, and those from Table 2 in case of the third model.

Recipe Add classification layer (described in Figure 2 on the left) to the pretrained RoBERTa_{LARGE} model in order to obtain the first model and train until no score gain is observed on development set. Train the second model in the same manner, but this time using the *re-weighting*. Combine *re-weighting*, Span CLS and self-training approaches to get the third model, and again train until no score improvement on development set is observed. Finally, ensemble all three models by averaging class probabilities from their final layers.

As shown later, the approach we took and reported above turned out to be sub-optimal. An in-depth analysis of this system and a better one is proposed in Section 3.2.

3 Ablation Studies

Since different random initialization or data order can result in considerably higher scores,³ models with different random seeds were trained for the purposes of ablation studies. In the case of the SI task, results were evaluated on the original development set. In contrast, in the case of TC, where fewer data points are available, we decided to use cross-validation instead.

#	Re-weight	Span CLS	Self-train	Micro-F1 (std)
(1)	−	−	−	71.9 ± 1.5
(2)	−	−	+	71.4 ± 1.4
(3)	−	+	−	72.2 ± 1.3
(4)	−	+	+	71.8 ± 1.7
(5)	+	−	−	71.8 ± 1.6
(6)	+	−	+	70.9 ± 1.7
(7)	+	+	−	72.4 ± 1.5
(8)	+	+	+	71.3 ± 1.5

Table 5: Average of 6-fold cross-validation score on TC task with micro-averaged F1 metric.

Ensemble	Micro-F1 (std)
(1) (6)	72.3 ± 1.7
(1) (2)	72.9 ± 1.8
(3) (5)	73.6 ± 1.5
(1) (5) (8)	74.1 ± 1.7
(2) (4) (7)	74.4 ± 1.5
(1) (4) (7)	74.6 ± 1.4
(1) (4) (7) (8)	74.9 ± 1.2
(1) (2) (4) (5) (7)	75.1 ± 1.5

Table 6: Average scores achieved with ensembles of individual models described in Table 5. Micro-averaged F1 metric.

3.1 Span Identification

Models with different random seeds were trained for 60K steps with an evaluation performed every 2K steps. This is equivalent to approximately 30 epochs, and per-epoch validation in a scenario without data generated during the self-training procedure. Table 3 summarizes the best scores achieved across 10 runs for each configuration.

CRF has a noticeable positive impact on FLC-F1 (Da San Martino et al., 2020) scores achieved without self-training in the setting we consider. The presence of the CRF layer is correlated positively with the score ($\rho = 0.27$, $p < 0.001$). The difference is significant ($p < 0.001$), according to the Kruskal–Wallis test (Kruskal and Wallis, 1952). Unless said otherwise, all further statistical statements within this section were confirmed with statistically significant positive Spearman rank correlation and Kruskal–Wallis test results. Differences in variance were confirmed using Bartlett’s test (Snedecor and Cochran, 1989). The 0.05 significance level was assumed.

The statistically significant influence of CRF disappears when the self-training is investigated. In the case of first self-training, regardless of whether or not CRF was used, a considerable increase in median score can be observed. Self-trained models with and without the CRF layer, however, are indistinguishable.

Improvement offered by further self-training iterations is not so evident but is statistically significant. In particular, they slightly improve mean scores and decrease variance (see Table 3). As it comes to the latter, CRF-extended models generally have higher variance and scores achieved across the runs.

Table 4 analyzes the importance of using different hyperparameters. Whereas use of a smaller batch size and dropout is beneficial for the initial training without noisy data, it negatively impacts the self-training phase. The most substantial negative impact is observed when dropout is disabled during training on the small amount of manually annotated data.

Figure 3 illustrates scores achieved by models trained for the same number of steps on subsets or supersets of manually annotated data. CRF layer has a positive impact regardless of the percentage of train set available. Once again, a large variance in scores of CRF-equipped models can be observed, however, it is substantially reduced with the increase of a batch size. Interestingly, figures suggest the proportion of automatically annotated data we used might be suboptimal since it was an equivalent of around 3000% in line with the chart’s convention. One may hypothesize better scores would be achieved by models trained with 1 : 4 gold to silver proportion.

3.2 Technique Classification

6-fold cross-validation was conducted. The results are presented in Table 5. Folds were created by mixing training and development datasets, then shuffling them and splitting into even folds. Parameters were set according to Table 1 and Table 2, whereas experiments were carried out as follows. Each approach from Table 5 was separately evaluated on each fold using the micro-averaged F1 metric. Then, for each approach, the average score and the standard deviation were obtained using six scores from every fold.

³See e.g., Junczys-Dowmunt et al. (2018) or recent analysis of Dodge et al. (2020).

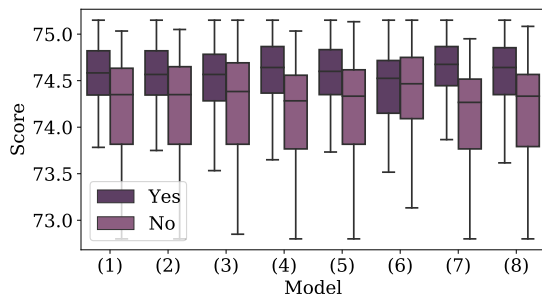


Figure 4: Impact of adding a particular model to the ensemble has on mean scores from different folds. Comparison of results with and without it present in tested combination.

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ρ	.28	.30	.20	.41	.32	.05*	.50	.36

Table 7: Spearman’s ρ between presence of ensemble component (models from Table 5) and score achieved by ensemble. * indicate results were not significant, assuming 0.05 significance level.

Moreover, all the 247 possible ensembles⁴ were evaluated in the same fashion as in experiments from Table 5. Table 6 shows the performance achieved by selected combinations when simple averaging of the probabilities returned by individual models was used as the final prediction.

Due to a large number of available results, it is beneficial to conduct a statistical analysis to formulate remarks regarding the general observed trends. Each component model of the ensemble was treated as a categorical variable with respect to the ensemble score. Spearman rank correlation between the presence of an ensemble component (approaches from Table 5) and achieved scores shows that adding model to the ensemble correlates with a significant increase in score, except for (6) model (see Table 7). Boxplots from Figure 4 lead to the same conclusions.⁵

Re-weighting seems to be beneficial only when ensembled with other models. An interesting finding is that Span CLS offers a small but consistent increase of performance both in models from Table 5 and when used in ensembles. Bear in mind, we outperformed the second-placed team by ε , so an improvement of a point or half is not negligible.

What is most conspicuous, however, is that self-training based solutions from Table 5 seem to be detrimental in the case of TC task. This damaging effect can be potentially attributed to the fact that automatically generated data accumulate errors from both Span Identification and Classification. Another possible explanation is that much fewer data points are available for span classification task than for span identification attempted as a sequence labeling task. The latter would be somehow consistent with what was found in the field of Neural Machine Translation, where the use of the back-translation technique in low-resource setting was determined to be harmful (Edunov et al., 2018).

On the other hand, self-training has a positive, statistically significant impact on the score when used in ensembles (see Figure 4 and Table 7). It is not surprising as the beneficial impact of combining individual estimates was observed in many disciplines and is known since the times of Laplace (Clemen, 1989).

4 Error analysis

In addition to providing an overview of problematic classes, the question of which shallow features influence score and worsen the results was addressed. This problem was analyzed in a *no-box* manner, as proposed by Graliński et al. (2019). The main idea is to create two dataset subsets for each feature considered (one for data points with the feature present and one for data points without the feature), rank subsets by per-item scores, and use Mann-Whitney rank U (Mann and Whitney, 1947) to determine whether there is a non-accidental difference between subsets. A low p-value indicates that feature reduces the evaluation score of the model.

⁴It is the number of all subsets with cardinality greater than one, drawn from an 8-element set.

⁵Kruskal-Wallis test and Boruta algorithm (Kursa et al., 2010) were used in addition to support these findings too.

	Authority	Fear	Bandwagon	B&W	Simplification	Doubt	Minimization	Flag-Waving	Loaded	Labeling	Repetition	Slogans	Clichés	Strawman	Overall	
Identified subsequence	57	56	20	36	50	42	48	40	44	45	26	62	41	41	43	
Fully identified	%	7	18	0	18	5	6	11	50	25	21	33	7	23	10	23
Not identified		35	25	80	45	44	51	39	9	29	33	40	30	35	48	33
Number of instances		14	44	5	22	18	66	68	87	325	183	145	40	17	29	1063

Table 8: Proportion of partially and fully identified spans (SI task) depending on the propaganda technique used. All the experiments conducted on the original development set.

4.1 Span Identification

Since the FLC-F1 metric used in the SI task gives non-zero scores for partial matches; it is interesting to analyze what was the proportion of entirely missed (partially identified) spans. Table 8 investigates this question broken down by the propaganda technique used.

Our system was unable to identify one-third of expected spans, whereas a majority of those correctly identified were the partial matches. The spans the easiest to identify in the text represented *Flag-Waving*, *Appeal to fear/prejudice*, and *Slogans* techniques. In contrast, *Bandwagon*, *Doubt*, and the group of *{Whataboutism, Strawman, Red Herring}* turned out to be the hardest. The highest proportion of fully identified spans was achieved for *Flag-Waving*, *Repetition*, and *Loaded Language*. Unfortunately, it is not possible to investigate precision in this manner, without training separate models for each label or estimating one-to-one alignments between output and expected spans.

Further investigation of problematic cases in a paradigm of no-box debugging with the GEval tool (Graliński et al., 2019) revealed the most worsening features, that are features whose presence impacts span identification evaluation metrics negatively (Table 9). It seems that our system tends to return ranges without adjacent punctuation. This is the case of sentences such as *The new CIA Director Haspel, who ‘tortured some folks,’ probably can’t travel to the EU*, where only the quoted text was returned, whereas annotation assumes it should be returned with apostrophes and commas. This remark can be used to improve overall results with simple post-processing slightly. Returned *and* conjunction refers to the cases where it connects two propaganda spans. The system frequently returns them as a single span, contrary to what is expected in the gold standard.

4.2 Technique Classification

Figure 5 presents the normalized confusion matrix of the submitted system predictions. Interestingly, there are a few commonly confused pairs. *Loaded Language* and *Black-and-white Fallacy* were frequently misclassified as *Appeal to fear/prejudice*. Similarly, *Causal Oversimplification* was often predicted as *Doubt* and *Clichés* as *Loaded Language*.

The most worsening features are presented in Table 10. One of the frequent predictors of low accuracy is a comma character present within the span to be classified. It can probably be attributed to the fact that its presence is a good indicator of span linguistic complexity. Another determinant of inefficiency turned out to be a negation—around half of the sentences containing word *not* were misclassified by the system. Suggested features of a quotation mark before the span and the digram *according to* after the span are related to reported or indirect speech. The explanation of the worsening effect of other features is not as evident as in the case mentioned above. Moreover, it seems there is no obvious way of improving the final results with our findings, and a more detailed analysis might be required.

5 Discussion and Summary

The winning system for the propaganda Technique Classification (TC) task and the second-placed system for the propaganda Span Identification (SI) task has been described. Both of the developed solutions used a semi-supervised learning technique of self-training. Although CRF is barely used with Transformer-based language models, the SI task was approached with RoBERTa-CRF architecture. An ensemble of

Authority	.43	.07				.14	.07	.07	.07	.07			.07	
Fear	.02	.52				.02	.07	.02	.23	.07	.02			
Bandwagon			.8			.2								
B&W	.05	.32		.14		.05	.18	.05	.14	.09				
Simplification	.06	.06			.44	.22	.06			.06	.11			
Doubt	.02	.08			.03	.62	.08		.08	.03	.05	.02	.02	
Minimisation	.06	.04				.01	.66		.1	.06	.03	.01	.01	
Flag-Waving		.02			.01	.06		.79	.02	.02	.01	.06		
Loaded		.03				.01	.04		.81	.03	.04	.02		
Labeling					.01		.02	.01	.15	.74	.05		.02	
Repetition		.01						.02	.13	.14	.66			
Slogans		.03						.12	.03	.05	.12	.62	.03	
Clichés				.06		.06	.12	.12	.24			.06	.29	.06
Strawman		.03		.03	.07	.17	.07	.03	.07	.1	.07			.34
	Authority	Fear	Bandwagon	B&W	Simplification	Doubt	Minimisation	Flag-Waving	Loaded	Labeling	Repetition	Slogans	Clichés	Strawman

Figure 5: Confusion matrix of the submitted system predictions normalized over the number of correct labels. Rows represent the correct labels and columns – the predicted ones (TC).

Feature	Count	P-value
<i>question</i>	expected	21 0.036
<i>dot</i>		36 0.037
<i>quotation</i>		58 0.050
<i>exclamation</i>		15 0.064
and	output	14 0.070

Table 9: Selected shallow features one may hypothesize impact evaluation scores negatively (SI).

Feature	Count	P-value
<i>comma</i>	inside	119 < 0.001
we		15 0.002
this		28 0.007
will		40 0.008
not		62 0.013
<i>exclamation</i>		16 0.014
CIA	before	25 < 0.001
according to	after	8 < 0.001
<i>quotation</i>	before	65 0.004

Table 10: Selected shallow features one may hypothesize impact evaluation scores negatively (TC).

RoBERTa-based models has been proposed for the TC task, with one of them making use of Span CLS layers we introduce in the present paper.

Analysis conducted afterward can be applied in a rather straightforward manner to further improve the scores for both SI and TC tasks. It is because some of the decisions we have made given lack of or uncertain information, during the post-hoc inquiry turned out to be sub-optimal. These include the proportion of data from self-training in the SI task, and the possibility of providing a better ensemble in the case of TC.

The ablation studies conducted, however, have some limitations. The same subset of OpenWebText was used in experiments conducted within one self-training iteration. This means a random seed did not impact which sentences were used during the first, second, and third self-training phase, and in each, we were manipulating only the data order. Moreover, an analysis we reported was limited to few hyperparameter combinations and no extensive hyperparameter space search was performed. Finally, only one and a rather simple method of cost-sensitive re-weighting was tested, and there is a great chance it was sub-optimal. It would be interesting to investigate other schemes, such as the one proposed by Cui et al. (2019).

The error analysis revealed propaganda techniques commonly confused in TC task, and the techniques we were unable to detect effectively within the SI input articles. In addition to providing an overview of problematic classes, the question of which shallow features influence score and worsen the results was addressed. A few of these were identified and our remarks can be used to slightly improve results on SI task with simple post-processing. This is not the case for TC task, where one is unable to propose how to improve the final results with our findings.

An interesting future research direction seems to be the application of the CRF layer and Span CLS to Transformer-based language models when dealing with other tasks outside the propaganda detection problem. These may include Named Entity Recognition in the case of RoBERTa-CRF, and an aspect-based sentiment analysis that can be viewed through the lens of span classification with Span CLS we proposed.

6 Outro

Developed systems were used to identify and classify spans in the present paper to detect fragments one may suspect to represent one or more propaganda techniques. Unfortunately for the entertaining value of this work, none of such were identified by our SI model.

References

- Robert T. Clemen. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559 – 583.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-balanced loss based on effective number of samples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269.
- Giovanni Da San Martino, Seunghak Yu, Alberto Barr’o-n-Cede no, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, EMNLP-IJCNLP 2019, Hong Kong, China, November*.
- Giovanni Da San Martino, Alberto Barr’o-n-Cede no, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation, SemEval 2020, Barcelona, Spain, September*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.
- Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz Górecki. 2019. GEval: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy, August. Association for Computational Linguistics.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana, June. Association for Computational Linguistics.
- William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.
- Miron B Kursa, Aleksander Jankowski, and Witold R Rudnicki. 2010. Boruta—a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.
- Wenhui Liao and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*, pages 58–65, Boulder, Colorado, June. Association for Computational Linguistics.
- Yao Lin, Chengjie Sun, Wang Xiaolong, and Wang Xuan. 2010. Combining self learning and active learning for chinese named entity recognition. *Journal of Software*, 5, 05.
- Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, Portland, Oregon, USA, June. Association for Computational Linguistics.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.
- Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models. In *WACV/MOTION*, pages 29–36.
- George W. Snedecor and William G. Cochran. 1989. *Statistical Methods*. Iowa State University Press, eighth edition.
- Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *CoRR*, abs/1909.10649.
- Lei Wang, Qing Qian, Qiang Zhang, Jishuai Wang, Wenbo Cheng, and Wei Yan. 2020. Classification Model on Big Data in Medical Diagnosis Based on Semi-Supervised Learning. *The Computer Journal*, 03. bxaa006.
- David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA, June. Association for Computational Linguistics.

Document Understanding



Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer

Rafał Powalski¹, Łukasz Borchmann^{1,2(✉)}, Dawid Jurkiewicz^{1,3},
Tomasz Dwojak^{1,3}, Michał Pietruszka^{1,4}, and Gabriela Pałka^{1,3}

¹ Applica.ai, Warsaw, Poland

² Poznan University of Technology, Poznań, Poland

{rafal.powalski, lukasz.borchmann, dawid.jurkiewicz, tomasz.dwojak,
michal.pietruszka, gabriela.palka}@applica.ai

³ Adam Mickiewicz University, Poznań, Poland

⁴ Jagiellonian University, Cracow, Poland

Abstract. We address the challenging problem of Natural Language Comprehension beyond plain-text documents by introducing the TILT neural network architecture which simultaneously learns layout information, visual features, and textual semantics. Contrary to previous approaches, we rely on a decoder capable of unifying a variety of problems involving natural language. The layout is represented as an attention bias and complemented with contextualized visual information, while the core of our model is a pretrained encoder-decoder Transformer. Our novel approach achieves state-of-the-art results in extracting information from documents and answering questions which demand layout understanding (DocVQA, CORD, SROIE). At the same time, we simplify the process by employing an end-to-end model.

Keywords: Natural Language Processing · Transfer learning · Document understanding · Layout analysis · Deep learning · Transformer

1 Introduction

Most tasks in Natural Language Processing (NLP) can be unified under one framework by casting them as triplets of the question, context, and answer [26, 29, 39]. We consider such unification of Document Classification, Key Information Extraction, and Question Answering in a demanding scenario where context extends beyond the text layer. This challenge is prevalent in business cases since contracts, forms, applications, and invoices cover a wide selection of document types and complex spatial layouts.

R. Powalski, L. Borchmann, D. Jurkiewicz, T. Dwojak and M. Pietruszk—Contributed equally.

Importance of Spatio-visual Relations. The most remarkable successes achieved in NLP involved models that map raw textual input into raw textual output, which usually were provided in a digital form. An important aspect of real-world oriented problems is the presence of scanned paper records and other analog materials that became digital.

Consequently, there is no easily accessible information regarding the document layout or reading order, and these are to be determined as part of the process. Furthermore, interpretation of shapes and charts beyond the layout may help answer the stated questions. A system cannot rely solely on text but requires incorporating information from the structure and image.

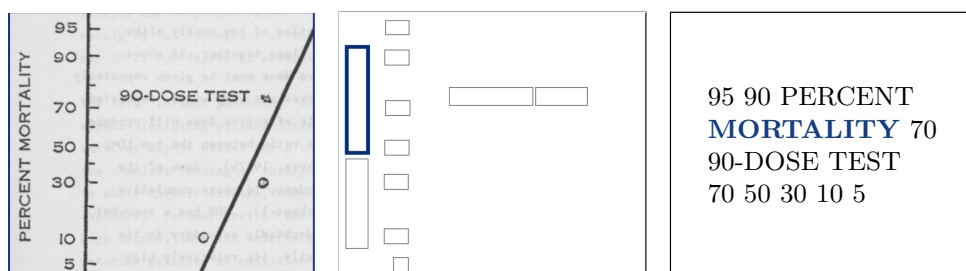


Fig. 1. The same document perceived differently depending on modalities. Respectively: its visual aspect, spatial relationships between the bounding boxes of detected words, and unstructured text returned by OCR under the detected reading order.

Thus, it takes three to solve this fundamental challenge—the extraction of key information from richly formatted documents lies precisely at the intersection of NLP, Computer Vision, and Layout Analysis (Fig. 1). These challenges impose extra conditions beyond NLP that we sidestep by formulating layout-aware models within an encoder-decoder framework.

Limitations of Sequence Labeling. Sequence labeling models can be trained in all cases where the token-level annotation is available or can be easily obtained. Limitations of this approach are strikingly visible on tasks framed in either key information extraction or property extraction paradigms [9, 19]. Here, no annotated spans are available, and only property-value pairs are assigned to the document. Occasionally, it is expected from the model to mark some particular subsequence of the document. However, problems where the expected value is not a substring of the considered text are unsolvable assuming sequence labeling methods.¹ As a result, authors applying state-of-the-art entity recognition models were forced to rely on human-made heuristics and time-consuming rule engineering.

Take, for example, the total amount assigned to a receipt in the SROIE dataset [19]. Suppose there is no exact match for the expected value in the

¹ Expected values have always an exact match in CoNLL, but not elsewhere, e.g., it is the case for 20% WikiReading, 27% Kleister, and 93% of SROIE values.

document, e.g., due to an OCR error, incorrect reading order or the use of a different decimal separator. Unfortunately, a sequence labeling model cannot be applied off-the-shelf. Authors dealing with property extraction rely on either manual annotation or the heuristic-based tagging procedure that impacts the overall end-to-end results [11, 18, 36, 52, 55, 56]. Moreover, when receipts with one item listed are considered, the total amount is equal to a single item price, which is the source of yet another problem. Precisely, if there are multiple matches for the value in the document, it is ambiguous whether to tag all of them, part or none.

Another problem one has to solve is which and how many of the detected entities to return, and whether to normalize the output somehow. Consequently, the authors of Kleister proposed a set of handcrafted rules for the final selection of the entity values [52]. These and similar rules are either labor-intensive or prone to errors [40].

Finally, the property extraction paradigm does not assume the requested value appeared in the article in any form since it is sufficient for it to be inferable from the content, as in document classification or non-extractive question answering [9].

Resorting to Encoder-Decoder Models. Since sequence labeling-based extraction is disconnected from the final purpose the detected information is used for, a typical real-world scenario demands the setting of Key Information Extraction.

To address this issue, we focus on the applicability of the encoder-decoder architecture since it can generate values not included in the input text explicitly [16] and performs reasonably well on all text-based problems involving natural language [44]. Additionally, it eliminates the limitation prevalent in sequence labeling, where the model output is restricted by the detected word order, previously addressed by complex architectural changes (Sect. 2).

Furthermore, this approach potentially solves all identified problems of sequence labeling architectures and ties various tasks, such as Question Answering or Text Classification, into the same framework. For example, the model may deduce to answer *yes* or *no* depending on the question form only. Its end-to-end elegance and ease of use allows one to not rely on human-made heuristics and to get rid of time-consuming rule engineering required in the sequence labeling paradigm.

Obviously, employing a decoder instead of a classification head comes with some known drawbacks related to the autoregressive nature of answer generation. This is currently investigated, e.g., in the Neural Machine Translation context, and can be alleviated by methods such as lowering the depth of the decoder [24, 47]. However, the datasets we consider have target sequences of low length; thus, the mentioned decoding overhead is mitigated.

The specific contribution of this work can be better understood in the context of related works (Fig. 2).

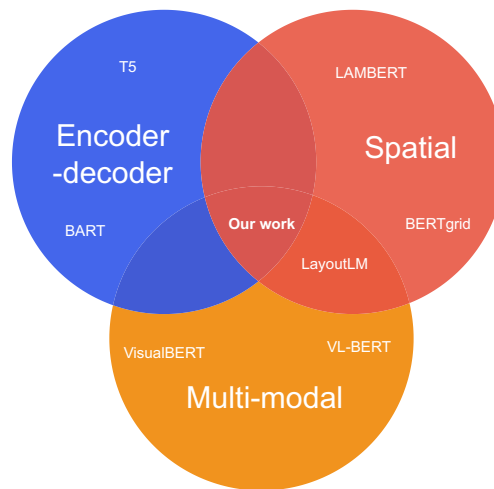


Fig. 2. Our work in relation to encoder-decoder models, multi-modal transformers, and models for text that are able to comprehend spatial relationships between words.

2 Related Works

We aim to bridge several fields, with each of them having long-lasting research programs; thus, there is a large and varied body of related works. We restrict ourselves to approaches rooted in the architecture of Transformer [54] and focus on the inclusion of spatial information or different modalities in text-processing systems, as well as on the applicability of encoder-decoder models to Information Extraction and Question Answering.

Spatial-Aware Transformers. Several authors have shown that, when tasks involving 2D documents are considered, sequential models can be outperformed by considering layout information either directly as positional embeddings [11, 17, 56] or indirectly by allowing them to be contextualized on their spatial neighborhood [6, 15, 57]. Further improvements focused on the training and inference aspects by the inclusion of the area masking loss function or achieving independence from sequential order in decoding respectively [18, 20]. In contrast to the mentioned methods, we rely on a bias added to self-attention instead of positional embeddings and propose its generalization to distances on the 2D plane. Additionally, we introduce a novel word-centric masking method concerning both images and text. Moreover, by resorting to an encoder-decoder, the independence from sequential order in decoding is granted without dedicated architectural changes.

Encoder-Decoder for IE and QA. Most NLP tasks can be unified under one framework by casting them as Language Modeling, Sequence Labeling or Question Answering [25, 43]. The QA program of unifying NLP frames all the problems as triplets of question, context and answer [26, 29, 39] or item, property name and answer [16]. Although this does not necessarily lead to the use of

encoder-decoder models, several successful solutions relied on variants of Transformer architecture [9, 34, 44, 54]. The T5 is a prominent example of large-scale Transformers achieving state-of-the-art results on varied NLP benchmarks [44]. We extend this approach beyond the text-to-text scenario by making it possible to consume a multimodal input.

Multimodal Transformers. The relationships between text and other media have been previously studied in Visual Commonsense Reasoning, Video-Grounded Dialogue, Speech, and Visual Question Answering [3, 13, 32]. In the context of images, this niche was previously approached with an image-to-text cross-attention mechanism, alternatively, by adding visual features to word embeddings or concatenating them [33, 35, 37, 53, 56]. We differ from the mentioned approaches, as in our model, visual features added to word embeddings are already contextualized on an image’s multiple resolution levels (see Sect. 3).

3 Model Architecture

Our starting point is the architecture of the Transformer, initially proposed for Neural Machine Translation, which has proven to be a solid baseline for all generative tasks involving natural language [54].

Let us begin from the general view on attention in the first layer of the Transformer. If n denotes the number of input tokens, resulting in a matrix of embeddings X , then self-attention can be seen as:

$$\text{softmax} \left(\frac{Q_X K_X^\top}{\sqrt{n}} + B \right) V_X \quad (1)$$

where Q_X , K_X and V_X are projections of X onto query, keys, and value spaces, whereas B stands for an optional attention bias. There is no B term in the original Transformer, and information about the order of tokens is provided explicitly to the model, that is:

$$X = S + P \quad B = 0_{n \times d}$$

where S and P are respectively the semantic embeddings of tokens and positional embedding resulting from their positions [54]. $0_{n \times d}$ denote a zero matrix.

In contrast to the original formulation, we rely on relative attention biases instead of positional embeddings. These are further extended to take into account spatial relationships between tokens (Fig. 3).

Spatial Bias. Authors of the T5 architecture disregarded positional embeddings [44], by setting $X = S$. They used relative bias by extending self-attention’s equation with the sequential bias term $B = B^{1D}$, a simplified form of positional signal inclusion. Here, each logit used for computing the attention head weights has some learned scalar added, resulting from corresponding token-to-token offsets.

We extended this approach to spatial dimensions. In our approach, biases for relative horizontal and vertical distances between each pair of tokens are calculated and added to the original sequential bias, i.e.:

$$B = B^{1D} + B^H + B^V$$

Such bias falls into one of 32 buckets, which group similarly-distanced token-pairs. The size of the buckets grows logarithmically so that greater token pair distances are grouped into larger buckets (Fig. 4).

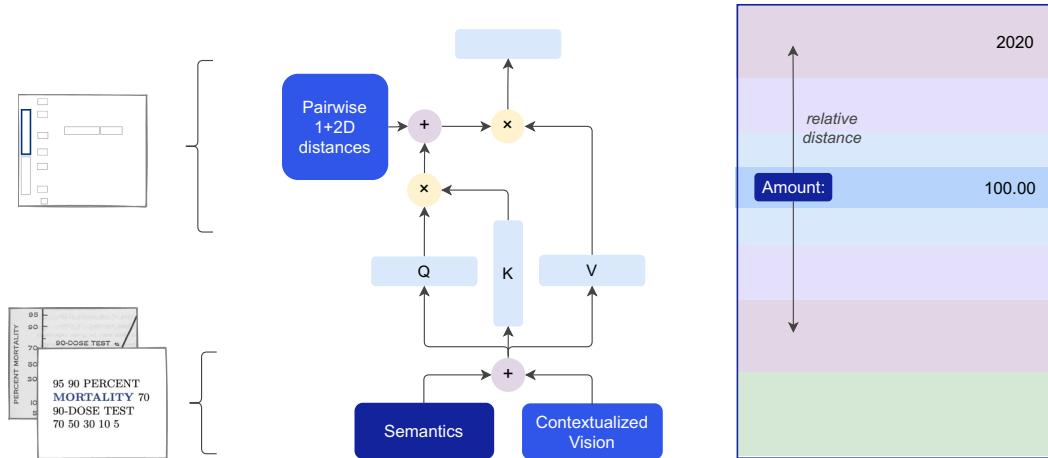


Fig. 3. T5 introduces sequential bias, separating semantics from sequential distances. We maintain this clear distinction, extending biases with spatial relationships and providing additional *image semantics* at the input.

Fig. 4. Document excerpt with distinguished vertical buckets for the *Amount* token.

Contextualized Image Embeddings. Contextualized *Word* Embeddings are expected to capture context-dependent semantics and return a sequence of vectors associated with an entire input sequence [10]. We designed Contextualized *Image* Embeddings with the same objective, i.e., they cover the image region semantics in the context of its entire visual neighborhood.

To produce image embeddings, we use a convolutional network that consumes the whole page image of size 512×384 and produces a feature map of $64 \times 48 \times 128$. We rely on U-Net as a backbone visual encoder network [48] since this architecture provides access to not only the information in the near neighborhood of the token, such as font and style but also to more distant regions of the page, which is useful in cases where the text is related to other structures, i.e., is the description of a picture. This multi-scale property emerges from the skip connections within chosen architecture (Fig. 5). Then, each token’s bounding box is used to extract features from U-Net’s feature map with ROI pooling [5]. The obtained vector is then fed into a linear layer which projects it to the model embedding dimension.

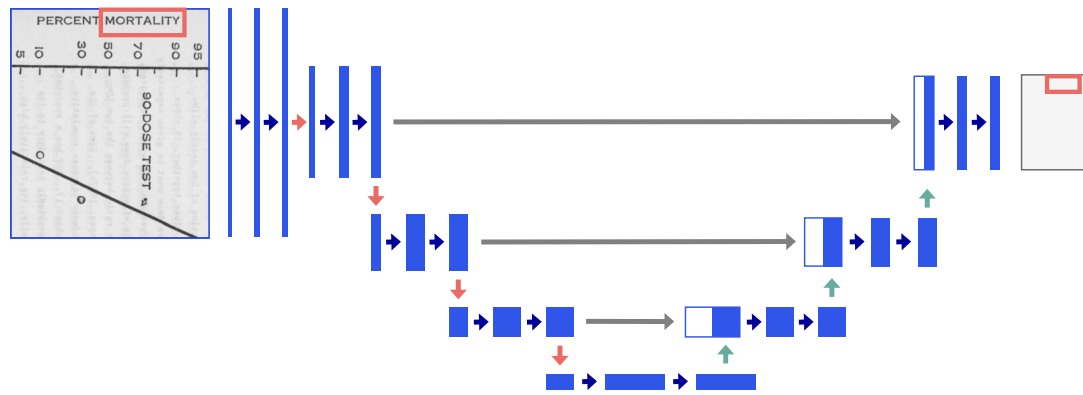


Fig. 5. Truncated U-Net network. ■ conv ■ max-pool ■ up-conv ■ residual (Color figure online)

In order to inject visual information to the Transformer, a matrix of contextualized image-region embeddings U is added to semantic embeddings, i.e. we define

$$X = S + U$$

in line with the convention from Sect. 3 (see Fig. 3).

4 Regularization Techniques

In the sequence labeling scenario, each document leads to multiple training instances (token classification), whereas in Transformer sequence-to-sequence models, the same document results in one training instance with feature space of higher dimension (decoding from multiple tokens).

Since most of the tokens are irrelevant in the case of Key Information Extraction and contextualized word embeddings are correlated by design, one can suspect our approach to overfit easier than its sequence labeling counterparts. To improve the model’s robustness, we introduced a regularization technique for each modality.

Case Augmentation. Subword tokenization [28, 49] was proposed to solve the word sparsity problem and keep the vocabulary at a reasonable size. Although the algorithm proved its efficiency in many NLP fields, the recent work showed that it performs poorly in the case of an unusual casing of text [42], for instance, when all words are uppercased. The problem occurs more frequently in formatted documents (FUNSD, CORD, DocVQA), where the casing is an important visual aspect. We overcome both problems with a straightforward regularization strategy, i.e., produce augmented copies of data instances by lower-casing or upper-casing both the document and target text simultaneously.

Spatial Bias Augmentation. Analogously to Computer Vision practices of randomly transforming training images, we augment spatial biases by multiplying the horizontal and vertical distances between tokens by a random factor.

Such transformation resembles stretching or squeezing document pages in horizontal and vertical dimensions. Factors used for scaling each dimension were sampled uniformly from range $[0.8, 1.25]$.

Affine Vision Augmentation. To account for visual deformations of real-world documents, we augment images with affine transformation, preserving parallel lines within an image but modifying its position, angle, size, and shear. When we perform such modification to the image, the bounding box of every token is updated accordingly. The exact hyperparameters were subject to an optimization. We use 0.9 probability of augmenting and report the following boundaries for uniform sampling work best: $[-5, 5]$ degrees for rotation angle, $[-5\%, 5\%]$ for translation amplitude, $[0.9, 1.1]$ for scaling multiplier, $[-5, 5]$ degrees for the shearing angle.

5 Experiments

Our model was validated on series of experiments involving Key Information Extraction, Visual Question Answering, classification of rich documents, and Question Answering from layout-rich texts. The following datasets represented the broad spectrum of tasks and were selected for the evaluation process (see Table 1 for additional statistics).

The CORD dataset [41] includes images of Indonesian receipts collected from shops and restaurants. The dataset is prepared for the information extraction task and consists of four categories, which fall into thirty subclasses. The main goal of the SROIE dataset [19] is to extract values for four categories (company, date, address, total) from scanned receipts. The DocVQA dataset [38] is focused on the visual question answering task. The RVL-CDIP dataset [14] contains gray-scale images and assumes classification into 16 categories such as letter, form, invoice, news article, and scientific publication. For DocVQA, we relied on Amazon Textract OCR; for RVL-CDIP, we used Microsoft Azure OCR, for SROIE and CORD, we depended on the original OCR.

5.1 Training Procedure

The training procedure consists of three steps. First, the model is initialized with vanilla T5 model weights and is pretrained on numerous documents in an unsupervised manner. It is followed by training on a set of selected supervised tasks. Finally, the model is finetuned solely on the dataset of interest. We trained two size variants of TILT models, starting from T5-Base and T5-Large models. Our models grew to 230M and 780M parameters due to the addition of Visual Encoder weights.

Unsupervised Pretraining. We constructed a corpus of documents with rich structure, based on RVL-CDIP (275k docs), UCSF Industry Documents Library

Table 1. Comparison of datasets considered for supervised pretraining and evaluation process. Statistics given in thousands of documents or questions.

Dataset	Data type	Image	Docs (k)	Questions (k)
CORD [41]	Receipts	+	1.0	—
SROIE [19]	Receipts	+	0.9	—
DocVQA [38]	Industry documents	+	12.7	50.0
RVL-CDIP [14]	Industry documents	+	400.0	—
DROP [8]	} Wikipedia pages	—	6.7	96.5
QuAC [2]		—	13.6	98.4
SQuAD 1.1 [45]		—	23.2	107.8
TyDi QA [4]		—	204.3	204.3
Natural Questions [30]		—	91.2	111.2
WikiOps [1]	Wikipedia tables	—	24.2	80.7
CoQA [46]	Various sources	—	8.4	127.0
RACE [31]	English exams	—	27.9	97.7
QASC [27]	School-level science	—	—	10.0
FUNSD [21]	RVL-CDIP forms	+	0.1	—
Infographics VQA	infographics	+	4.4	23.9
TextCaps [50]	Open Images	+	28.4	—
DVQA [22]	Synthetic bar charts	+	300.0	3487.2
FigureQA [23]	Synthetic, scientific	+	140.0	1800.0
TextVQA [51]	Open Images	+	28.4	45.3

(480k),² and PDF files from Common Crawl (350k). The latter were filtered according to the score obtained from a simple SVM business document classifier.

Then, a T5-like masked language model pretraining objective is used, but in a salient span masking scheme, i.e., named entities are preferred rather than random tokens [12, 44]. Additionally, regions in the image corresponding to the randomly selected text tokens are masked with the probability of 80%. Models are trained for 100,000 steps with batch size of 64, AdamW optimizer and linear scheduler with an initial learning rate of $2e-4$.

Supervised Training. To obtain a general-purpose model which can reason about documents with rich layout features, we constructed a dataset relying on a large group of tasks, representing diverse types of information conveyed by a document (see Table 1 for datasets comparison). Datasets, which initially had been plain-text, had their layout produced, assuming some arbitrary font size and document dimensions. Some datasets, such as *WikiTable Questions*, come

² <http://www.industrydocuments.ucsf.edu/>.

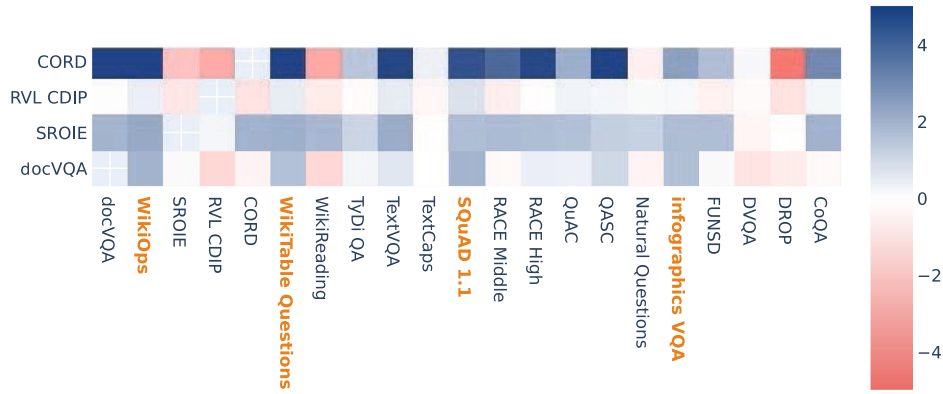


Fig. 6. Scores on CORD, DocVQA, SROIE and RVL-CDIP compared to the baseline without supervised pretraining. The numbers represent the differences in the metrics, orange text denote datasets chosen for the final supervised pretraining run.

with original HTML code – for the others, we render text alike. Finally, an image and computed bounding boxes of all words are used.

At this stage, the model is trained on each dataset for 10,000 steps or 5 epochs, depending on the dataset size: the goal of the latter condition was to avoid a quick overfitting.

We estimated each dataset’s value concerning a downstream task, assuming a fixed number of pretraining steps followed by finetuning. The results of this investigation are demonstrated in Fig. 6, where the group of WikiTable, WikiOps, SQuAD, and infographicsVQA performed robustly, convincing us to rely on them as a solid foundation for further experiments.

Model pretrained in unsupervised, and then supervised manner, is at the end finetuned for two epochs on a downstream task with AdamW optimizer and hyperparameters presented in Table 2.

Table 2. Parameters used during the finetuning on a downstream task. Batch size, learning rate and scheduler were subject of hyperparameter search with considered values of respectively $\{8, 16, \dots, 2048\}$, $\{5e-5, 2e-5, 1e-5, 5e-4, \dots, 1e-3\}$, $\{\text{constant}, \text{linear}\}$. We have noticed that the classification task of RVL-CDIP requires a significantly larger bath size. The model with the highest validation score within the specified steps number limit was used.

Dataset	Batch size	Steps	Learning rate	Scheduler
SROIE	8	6,200	$1e-4$	Constant
DocVQA	64	100,000	$2e-4$	Linear
CORD	8	36,000	$2e-4$	Linear
RVL-CDIP	1,024	12,000	$1e-3$	Linear

Table 3. Results of selected methods in relation to our base and large models. Bold indicates the best score in each category. All results on the test set, using the metrics proposed by dataset’s authors. The number of parameters given for completeness though encoder-decoder and LMs cannot be directly compared under this criterion.

Model	CORD F1	SROIE F1	DocVQA ANLS	RVL-CDIP Accuracy	Size variant (Parameters)
LayoutLM [56]	94.72	94.38	69.79	94.42	Base (113–160M)
	94.93	95.24	72.59	94.43	Large (343M)
LayoutLMv2 [55]	94.95	96.25	78.08	95.25	Base (200M)
	96.01	97.81	86.72	95.64	Large (426M)
LAMBERT [11]	96.06	98.17	—	—	Base (125M)
TILT (our)	95.11	97.65	83.92	95.25	Base (230M)
	96.33	98.10	87.05	95.52	Large (780M)

5.2 Results

The TILT model achieved state-of-the-art results on three out of four considered tasks (Table 3). We have confirmed that unsupervised layout- and vision-aware pretraining leads to good performance on downstream tasks that require comprehension of tables and other structures within the documents. Additionally, we successfully leveraged supervised training from both plain-text datasets and these involving layout information.

DocVQA. We improved SOTA results on this dataset by 0.33 points. Moreover, detailed results show that model gained the most in table-like categories, i.e., forms (89.5 \rightarrow 94.6) and tables (87.7 \rightarrow 89.8), which proved its ability to understand the spatial structure of the document. Besides, we see a vast improvement in the yes/no category (55.2 \rightarrow 69.0).³ In such a case, our architecture generates simply *yes* or *no* answer, while sequence labeling based models require additional components such as an extra classification head. We noticed that model achieved lower results in the image/photo category, which can be explained by the low presence of image-rich documents in our datasets.

RVL-CDIP. Part of the documents to classify does not contain any readable text. Because of this shortcoming, we decided to guarantee there are at least 16 image tokens that would carry general image information. Precisely, we act as there were tokens with bounding boxes covering 16 adjacent parts of the document. These have representations from U-Net, exactly as they were regular text tokens. Our model places second, 0.12 below the best model, achieving the similar accuracy of 95.52.

³ Per-category test set scores are available after submission on the competition web page: <https://rrc.cvc.uab.es/?ch=17&com=evaluation&task=1>.

Table 4. Results of ablation study. The minus sign indicates removal of the mentioned part from the base model.

Model	Score	Relative change
TILT-Base	82.9 ± 0.3	—
– Spatial Bias	81.1 ± 0.2	–1.8
– Visual Embeddings	81.2 ± 0.3	–1.7
– Case Augmentation	82.2 ± 0.3	–0.7
– Spatial Augmentation	82.6 ± 0.4	–0.3
– Vision Augmentation	82.8 ± 0.2	–0.1
– Supervised Pretraining	81.2 ± 0.1	–1.7

CORD. Since the complete inventory of entities is not present in all examples, we force the model to generate a *None* output for missing entities. Our model achieved SOTA results on this challenge and improved the previous best score by 0.3 points. Moreover, after the manual review of the model errors, we noticed that model’s score could be higher since the model output and the reference differ insignificantly e.g. “2.00 ITEMS” and “2.00”.

SROIE. We excluded OCR mismatches and fixed total entity annotations discrepancies following the same evaluation procedure as Garncarek et al. [11].⁴ We achieved results indistinguishable from the SOTA (98.10 vs. 98.17). Significantly better results are impossible due to OCR mismatches in the test-set.

Though we report the number of parameters near the name of the model size variant, note it is impossible to compare the TILT encoder-decoder model to language models such as LayoutLMs and LAMBERT under this criterion. In particular, it does not reflect computational cost, which may be similar for encoder-decoders twice as big as some language model [44, Section 3.2.2]. Nevertheless, it is worth noting that our Base model outperformed models with comparable parameter count.

6 Ablation Study

In the following section, we analyze the design choices in our architecture, considering the base model pretrained in an unsupervised manner and the same hyperparameters for each run. The DocVQA was used as the most representative and challenging for Document Intelligence since its leaderboard reveals a large gap to human performance. We report average results over two runs of each model varying only in the initial random seed to account for the impact of different initialization and data order [7].

⁴ Corrections can be obtained by comparing their two public submissions.

Significance of Modalities. We start with the removal of the 2D layout positional bias. Table 4 demonstrates that information that allows models to recognize spatial relations between tokens is a crucial part of our architecture. It is consistent with the previous works on layout understanding [11, 55]. Removal of the UNet-based convolutional feature extractor results in a less significant ANLS decrease than the 2D bias. This permits the conclusion that contextualized image embeddings are beneficial to the encoder-decoder.

Justifying Regularization. Aside from removing modalities from the network, we can also exclude regularization techniques. To our surprise, the results suggest that the removal of case augmentation decreases performance most severely. Our baseline is almost one point better than the equivalent non-augmented model. Simultaneously, model performance tends to be reasonably insensitive to the bounding boxes’ and image alterations. It was confirmed that other modalities are essential for the model’s success on real-world data, whereas regularization techniques we propose slightly improve the results, as they prevent overfitting.

Impact of Pretraining. As we exploited supervised pretraining similarly to previous authors, it is worth considering its impact on the overall score. In our ablation study, the model pretreated in an unsupervised manner achieved significantly lower scores. The impact of this change is comparable to the removal of spatial bias or visual embeddings. Since authors of the T5 argued that pretraining on a mixture of unsupervised and supervised tasks perform equally good with higher parameter count, this gap may vanish with larger variants of TILT we did not consider in the present paper [44].

7 Summary

In the present paper, we introduced a novel encoder-decoder framework for layout-aware models. Compared to the sequence labeling approach, the proposed method achieves better results while operating in an end-to-end manner. It can handle various tasks such as Key Information Extraction, Question Answering or Document Classification, while the need for complicated preprocessing and postprocessing steps is eliminated.

Although encoder-decoder models are commonly applied to generative tasks, both DocVQA, SROIE, and CORD we considered are extractive. We argue that better results were achieved partially due to the independence from the detected word order and resistance to OCR errors that the proposed architecture possesses. Consequently, we were able to achieve state-of-the-art results on two datasets (DocVQA, CORD) and performed on par with the previous best scores on SROIE and RVL-CDIP, albeit having a much simpler workflow.

Spatial and image enrichment of the Transformer model allowed the TILT to combine information from text, layout, and image modalities. We showed that the proposed regularization methods significantly improve the results.

Acknowledgments. The authors would like to thank Filip Graliński, Tomasz Stanisławek, and Łukasz Garncarek for fruitful discussions regarding the paper and our managing directors at Applica.ai. Moreover, Dawid Jurkiewicz pays due thanks to his son for minding the deadline and generously coming into the world a day after.

The Smart Growth Operational Programme supported this research under project no. POIR.01.01.01-00-0877/19-00 (*A universal platform for robotic automation of processes requiring text comprehension, with a unique level of implementation and service automation*).

References

1. Cho, M., Amplayo, R., Hwang, S.W., Park, J.: Adversarial TableQA: attention supervision for question answering on tables. In: PMLR (2018)
2. Choi, E., et al.: QuAC: question answering in context. In: EMNLP (2018)
3. Chuang, Y., Liu, C., Lee, H., Lee, L.: SpeechBERT: an audio-and-text jointly learned language model for end-to-end spoken question answering. In: ISCA (2020)
4. Clark, J.H., et al.: TyDi QA: a benchmark for information-seeking question answering in typologically diverse languages. *TACL* **8**, 454–470 (2020)
5. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: NeurIPS (2016)
6. Denk, T.I., Reisswig, C.: BERTgrid: contextualized embedding for 2d document representation and understanding. arXiv preprint (2019)
7. Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., Smith, N.A.: Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. arXiv preprint (2020)
8. Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., Gardner, M.: DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs. In: NAACL-HLT (2019)
9. Dwojak, T., Pietruszka, M., Borchmann, L., Chłedowski, J., Graliński, F.: From dataset recycling to multi-property extraction and beyond. In: CoNLL (2020)
10. Ethayarajh, K.: How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: EMNLP-IJCNLP (2019)
11. Garncarek, Ł., et al.: LAMBERT: layout-aware (language) modeling using BERT for information extraction. In: Llad, J. et al. (eds.) ICDAR 2021. LNCS, vol. 12822, pp. 532–547 (2021). Accepted to ICDAR 2021
12. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: ICML (2020)
13. Han, K., et al.: A survey on visual transformer. arXiv preprint (2021)
14. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: ICDAR (2015)
15. Herzig, J., Nowak, P.K., Müller, T., Piccinno, F., Eisenschlos, J.: TaPas: weakly supervised table parsing via pre-training. In: ACL (2020)
16. Hewlett, D., et al.: WikiReading: a novel large-scale language understanding task over Wikipedia. In: ACL (2016)
17. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multi-dimensional transformers. arXiv preprint (2019)
18. Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., Park, S.: BROS: a pre-trained language model for understanding texts in document. openreview.net preprint (2021)

19. Huang, Z., et al.: ICDAR2019 competition on scanned receipt OCR and information extraction. In: ICDAR (2019)
20. Hwang, W., Yim, J., Park, S., Yang, S., Seo, M.: Spatial dependency parsing for semi-structured document information extraction. arXiv preprint (2020)
21. Jaume, G., Ekenel, H.K., Thiran, J.P.: FUNSD: a dataset for form understanding in noisy scanned documents. In: ICDAR-OST (2019)
22. Kafle, K., Price, B.L., Cohen, S., Kanan, C.: DVQA: understanding data visualizations via question answering. In: CVPR (2018)
23. Kahou, S.E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., Bengio, Y.: FigureQA: an annotated figure dataset for visual reasoning. In: ICLR (2018)
24. Kasai, J., Pappas, N., Peng, H., Cross, J., Smith, N.A.: Deep encoder, shallow decoder: reevaluating the speed-quality tradeoff in machine translation. arXiv preprint (2020)
25. Keskar, N., McCann, B., Xiong, C., Socher, R.: Unifying question answering and text classification via span extraction. arXiv preprint (2019)
26. Khashabi, D., et al.: UnifiedQA: crossing format boundaries with a single QA system. In: EMNLP-Findings (2020)
27. Khot, T., Clark, P., Guerquin, M., Jansen, P., Sabharwal, A.: QASC: a dataset for question answering via sentence composition. In: AAAI (2020)
28. Kudo, T.: Subword regularization: improving neural network translation models with multiple subword candidates. In: ACL (2018)
29. Kumar, A., et al.: Ask me anything: dynamic memory networks for natural language processing. In: ICML (2016)
30. Kwiatkowski, T., et al.: Natural questions: a benchmark for question answering research. *TACL* **7**, 453–466 (2019)
31. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: large-scale ReAiding comprehension dataset from examinations. In: EMNLP (2017)
32. Le, H., Sahoo, D., Chen, N., Hoi, S.: Multimodal transformer networks for end-to-end video-grounded dialogue systems. In: ACL (2019)
33. Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 212–228. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_13
34. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL (2020)
35. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: a simple and performant baseline for vision and language. arXiv preprint (2019)
36. Liu, X., Gao, F., Zhang, Q., Zhao, H.: Graph convolution for multimodal information extraction from visually rich documents. In: NAACL-HLT (2019)
37. Ma, J., Qin, S., Su, L., Li, X., Xiao, L.: Fusion of image-text attention for transformer-based multimodal machine translation. In: IALP (2019)
38. Mathew, M., Karatzas, D., Jawahar, C.: DocVQA: a dataset for VQA on document images. In: WACV, pp. 2200–2209 (2021)
39. McCann, B., Keskar, N.S., Xiong, C., Socher, R.: The natural language decathlon: multitask learning as question answering. arXiv preprint (2018)
40. Palm, R.B., Winther, O., Laws, F.: CloudScan - a configuration-free invoice analysis system using recurrent neural networks. In: ICDAR (2017)
41. Park, S., et al.: CORD: a consolidated receipt dataset for post-OCR parsing. In: Document Intelligence Workshop at NeurIPS (2019)
42. Powalski, R., Stanislawek, T.: UniCase - rethinking casing in language models. arXiv preprint (2020)

43. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. Technical report (2019)
44. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *JMRL* (2020)
45. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: *EMNLP* (2016)
46. Reddy, S., Chen, D., Manning, C.D.: CoQA: a conversational question answering challenge. *TACL* **7**, 249–266 (2019)
47. Ren, Y., Liu, J., Tan, X., Zhao, Z., Zhao, S., Liu, T.Y.: A study of non-autoregressive model for sequence generation. In: *ACL* (2020)
48. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
49. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: *ACL* (2016)
50. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: TextCaps: a dataset for image captioning with reading comprehension. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12347, pp. 742–758. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_44
51. Singh, A., et al.: Towards VQA models that can read. In: *CVPR* (2019)
52. Stanisławek, T., et al.: Kleister: key information extraction datasets involving long documents with complex layouts. In: Llads, J. et al. (eds.) *ICDAR 2021*. LNCS, vol. 12822, pp. 564–579 (2021). Accepted to *ICDAR 2021*
53. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: pre-training of generic visual-linguistic representations. In: *ICLR* (2020)
54. Vaswani, A., et al.: Attention is all you need. In: *NeurIPS* (2017)
55. Xu, Y., et al.: LayoutLMv2: multi-modal pre-training for visually-rich document understanding. *arXiv preprint* (2020)
56. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: pre-training of text and layout for document image understanding. In: *KDD* (2020)
57. Yin, P., Neubig, G., Yih, W.t., Riedel, S.: TaBERT: pretraining for joint understanding of textual and tabular data. In: *ACL* (2020)

DUE: End-to-End Document Understanding Benchmark

Łukasz Borchmann*[†]

Michał Pietruszka*[‡]

Tomasz Stanisławek*[§]

Dawid Jurkiewicz[¶]

Michał Turski[¶]

Karolina Szyndler[¶]

Filip Graliński[¶]

[¶]Applica.ai

firstname.lastname@applica.ai

[†]Poznan University of Technology

[‡]Jagiellonian University, Krakow

[§]Warsaw University of Technology

[¶]Adam Mickiewicz University, Poznan

Abstract

Understanding documents with rich layouts plays a vital role in digitization and hyper-automation but remains a challenging topic in the NLP research community. Additionally, the lack of a commonly accepted benchmark made it difficult to quantify progress in the domain. To empower research in this field, we introduce the Document Understanding Evaluation (DUE) benchmark consisting of both available and reformulated datasets to measure the end-to-end capabilities of systems in real-world scenarios. The benchmark includes Visual Question Answering, Key Information Extraction, and Machine Reading Comprehension tasks over various document domains and layouts featuring tables, graphs, lists, and infographics. In addition, the current study reports systematic baselines and analyzes challenges in currently available datasets using recent advances in layout-aware language modeling. We open both the benchmarks and reference implementations and make them available at <https://duebenchmark.com> and <https://github.com/due-benchmark>.

1 Introduction

While mainstream Natural Language Processing focuses on plain text documents, the content one encounters when reading, e.g., scientific articles, company announcements, or even personal notes, is seldom plain and purely sequential. In particular, the document’s visual and layout aspects that guide our reading process and carry non-textual information appear to be an essential aspect that requires comprehension. These layout aspects, as we understand them, are prevalent in tasks that can be much better solved when given not only text sequence on the input but pieces of multimodal information covering aspects such as text-positioning (i.e. location of words on the 2D plane), text-formatting (e.g., different font sizes, colors), and graphical elements (e.g., lines, bars, presence of figures) among others. Over the decades, systems dealing with document understanding developed an inherent aspect of multi-modality that nowadays revolves around the problems of integrating visual information with spatial relationships and text [36, 2, 50, 13].

*Equal contribution

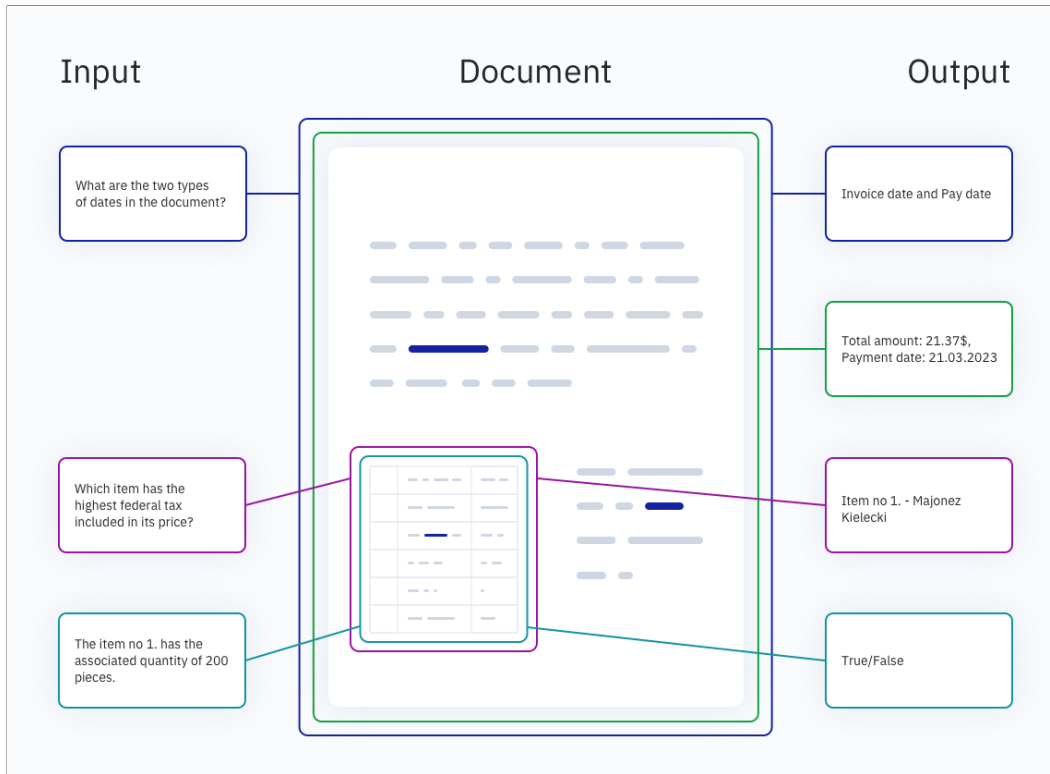


Figure 1: Document Understanding covers problems ranging from the extraction of key information, through verification statements related to rich content, to answering open questions regarding an entire file. It may involve the comprehension of multi-modal information conveyed by a document.

In general, when document processing systems are considered, the term *understanding* is thought of specifically as the capacity to convert a document into meaningful information [10, 57, 16]. This fits into the rapidly growing market of hyperautomation-enabling technologies, estimated to reach nearly \$600 billion in 2022, up 24% from 2020 [42]. Considering that unstructured data is orders of magnitude more abundant than structured data, the lack of tools necessary to analyze unstructured data and extract structured information can limit the performance of these intelligent services. The process of structuring data and content must be robust to various document domains and tasks.

Despite its importance for digital transformation, the problem of measuring how well available models obtain information from a wide range of tasks and document types and how suitable they are for freeing workers from paperwork through process automation is not yet addressed. Meanwhile, in other research communities, there are well-established progress measuring methods, like the most recognizable NLP benchmarks of GLUE and SuperGLUE covering a wide range of problems related to plain-text language understanding [53, 52] or VTAB and ImageNet in the computer vision domain [59, 11]. We intend to bridge this major gap by introducing the first Document Understanding benchmark (available at <https://duebenchmark.com>).

It includes tasks that either originally had a vital layout understanding component or were reformulated in such a way that after our modification, they require layout understanding. In particular, there is no structured representation of the underlying text, such as a database-like table given in advance, and it has to be determined from the input file as a part of the end-to-end process. Every time, there is only a PDF file provided as an input. Additionally, for the convenience of other researchers, we provide information about textual tokens and their locations (bounding boxes) which are coming from the OCR system or directly from the born-digital PDF file (see Section 4).

Contribution. The idea of the paper is to gather, reformulate and unify a set of intuitively dissimilar tasks that we found to share the same underlying requirement of understanding layout concepts. In order to organize them in a useful benchmark, we contributed by performing the following steps:

1. We reviewed and selected the available datasets. Additionally, we reformulated three tasks to a document understanding setting and obtained original documents for all of them (PWC, WTQ, TabFact).
2. We performed data cleaning, including the improvements of data splits (DeepForm, WTQ), data deduplication, manual annotation (PWC, DeepForm), and converted data to a unified format (all datasets).
3. We implemented competitive baselines and measured human performance where it was required (PWC, DeepForm, WTQ).
4. We identified challenges related to the current progress in the DU domain’s tasks and provided manually annotated diagnostic sets (all datasets).

These contributions are organized and described in Table 2. Additionally, a wider review of available tasks is described in Appendix A.

2 The state of Document Understanding

We treat Document Understanding as an umbrella term covering problems of Key Information Extraction, Classification, Document Layout Analysis, Question Answering, and Machine Reading Comprehension whenever they involve rich documents in contrast to plain texts or image-text pairs (Figure 1).

In addition to the problems strictly classified as Document Understanding, several related tasks can be reformulated as such. These provide either text-figure pairs instead of real-world documents or parsed tables given in their structured form. Since both can be rendered as synthetic documents with some loss of information involved, they are worth considering bearing in mind the low availability of proper Document Understanding tasks.

2.1 Landscape of Document Understanding tasks

KIE. Key Information Extraction, also referred to as Property Extraction, is a task where tuple values of the form (property, document) are to be provided. Contrary to QA problems, there is no question in natural language but rather a phrase or keyword, such as *total amount*, or *place of birth*. Public datasets in the field include extraction performed on receipts [20, 38], invoices, reports [45], and forms [24]. Documents within each of the mentioned tasks are homogeneous, whereas the set of properties to extract is limited and known in advance – in particular, the same type-specific property names appear in both test and train sets. In contrast to Name Entity Recognition, KIE typically does not assume that token-level annotations are available, and may require normalization of values found within the document.

Classification. Classification in our context involves rich content, where comprehension of both visual and textual aspects is required since unimodal models underperform. Though document image classification was initially approached using solely the methods of Computer Vision, it has recently become evident that multi-modal models can achieve significantly higher accuracy [55, 56, 40]. Similar conclusions were recently reached in other tasks, e.g., assigning labels to excerpts from biomedical papers [54].

DLA. Document Layout Analysis, performed to determine a document’s components, was initially motivated by the need to optimize storage and the transmission of large information volumes [36]. Even though its motivation has changed over the years, it is rarely an end in itself but rather a means to achieve a different goal, such as improving OCR systems. A typical dataset in the field assumes detection and classification of page regions or tokens [61, 30].

QA and MRC. At first glance, Question Answering and Machine Reading Comprehension over Documents is simply the KIE scenario where a question in natural language replaced a property name. More differences become evident when one notices that QA and MRC involve an open set of questions and various document types. Consequently, there is pressure to interpret the question and

to possess better generalization abilities. Furthermore, a specific content to analyze demands a much stronger comprehension of visual aspects, as the questions commonly relate to figures and graphics accompanying the formatted text [33, 32, 49].

QA over figures. Question Answering over Figures is, to some extent, comparable with QA and MRC over documents described above. The difference is that a ‘document’ here consists of a single born-digital plot, reflecting information from chosen, desirably real-world data. Since questions in this category are typically templated and figures are synthetically generated by authors of the task, datasets in this category contain as many as millions of examples [34, 4].

QA and NLI over tables. Question Answering and Natural Language Inference over Tables are similar, though in the case of NLI, there is a statement to verify instead of a question to answer. There is never a need to analyze the actual layout, as both assume comprehension of a provided data structure in a way that is equivalent to a database table. Consequently, the methods proposed here are distinct from those used in Document Understanding [39, 7].

2.2 Gaps and mistakes in Document Understanding evaluation

Currently available datasets and previous work in the field cannot on their own provide enough information that would allow researchers to generalize results to other tasks within the Document Understanding paradigm. It is crucial to validate models on many tasks with a variety of characteristics a Document Understanding system may encounter in real-world applications. Notably, the scope of the challenges in a single dataset is limited to a specific task (e.g., Key Information Extraction, Question Answering) or to a particular (sub)problem (e.g., processing long documents in Kleister [45], layout understanding in DocBank [30]).

Simultaneously, a common practice in the community is to evaluate models on private data [27, 12, 37, 31] or task-specific datasets selected by authors independently [55, 56, 63, 40, 1, 19], making fair comparison difficult. Many publicly available datasets are too small to enable reliable comparison (FUNSD [24], Kleister NDA [45]) or are almost solved, i.e., there is no room for improvement due to annotation errors and near-perfect scores achieved by models nowadays (SROIE [21], CORD [38], RVL-CDIP [17]).

In light of the above circumstances, the review and selection of representative and reliable tasks is of great importance.

3 End-to-end Document Understanding benchmark

The primary motivation for proposing this benchmark was to select datasets covering the broad range of tasks and DU-related problems satisfying the highest quality, difficulty, and licensing criteria.

Importantly, we opt for an end-to-end nature of tasks as opposed to, e.g., problems assuming some prior information on document layout. In particular, there is no structured representation of the underlying text, such as a database-like table given in advance, and it has to be determined from the raw input file as part of the end-to-end process.

We consider the aforementioned principle of end-to-end nature crucial because it ensures measurement to which degree manual workers can be supported in their repetitive tasks, i.e., how the ultimate goal of document understanding systems is supported in real-world applications. The said *alignment with real applications* is a vital characteristic of a good benchmark [29, 43].

3.1 Selected datasets

Extensive documentation of the selection process, including the datasheet, is available in Appendices A-H and in the supplementary materials. Table 1 summarizes the selected tasks described in detail below, whereas Appendix A covers the complete list of considered datasets and reasons we omitted them.

Lack of the classification, layout analysis and figure QA tasks in this selection results from the fact that none of the available sets fulfills the assumed selection criteria.

Table 1: Comparison of selected datasets with their base characteristics, including information regarding whether an input is an entire document (Doc.) or document excerpt (Exc.)

Task	Size (k documents)			Mean samples per document	Type	Metric	Features		Domain
	Train	Dev	Test				Input	Scanned	
DocVQA	10.2	1.3	1.3	3.9	Visual QA	ANLS	} Doc.	+	Business
InfographicsVQA	4.4	0.5	0.6	5.5	Visual QA	ANLS		-	Open
Kleister Charity	1.7	0.4	0.6	7.8	KIE	F1		+/-	Legal
PWC	0.2	0.06	0.12	25.5	KIE	ANLS ²		-	Scientific
DeepForm [★]	0.7	0.1	0.3	4.8	KIE	F1	} Exc.	+/-	Finances
WikiTableQuestions [★]	1.4	0.3	0.4	11.3	Table QA	Acc.		-	Open
TabFact [★]	13.2	1.7	1.7	7.1	Table NLI	Acc.		-	Open

The [★] symbol denotes that the dataset was reformulated or modified to improve its quality or align with the Document Understanding paradigm (see Table 2 and Appendix C). This symbol is not used to distinguish minor changes, such as data deduplication introduced in multiple datasets (Appendix B).

DocVQA. Dataset for Question Answering over single-page excerpts from various real-world industry documents. Typical questions present here might require comprehension of images, free text, tables, lists, forms, or their combination [33]. The best-performing solutions so far make use of layout-aware multi-modal models employing either encoder-decoder or sequence labeling architectures [40, 56].

InfographicsVQA. The task of answering questions about visualized data from a diverse collection of infographics, where the information needed to answer a question may be conveyed by text, plots, graphical or layout elements. Currently, the best result is obtained by an encoder-decoder model [32, 40].

Kleister Charity. A task for extracting information about charity organizations from their published reports is considered, as it is characterized by careful manual annotation by linguists and a significant gap to human performance [45]. It addresses important areas, namely high layout variability (lack of templates), need for performing an OCR, the appearance of long documents, and multiple spatial features (e.g., tables, lists, and titles).

PWC[★]. Papers with Code Leaderboards dataset was designed to extract result tuples from machine learning papers, including information on task, dataset, metric name, score. The best performing approach involves a multi-step pipeline, with modules trained separately on identified subproblems [26]. In contrast to the original formulation, we provide a complete paper as input instead of the table. This approach allows us to treat the problem as an end-to-end Key Information Extraction task with grouped variables (Appendix C).

DeepForm[★]. KIE dataset consisting of socially important documents related to election spending. The task is to extract contract number, advertiser name, amount paid, and air dates from advertising disclosure forms submitted to the Federal Communications Commission [47]. We use a subset of distributed datasets and improve annotations errors and make the annotations between subsets for different years consistent (Appendix C).

WikiTableQuestions (WTQ)[★]. Dataset for QA over semi-structured HTML tables sourced from Wikipedia. The authors intended to provide complex questions, demanding multi-step reasoning on a series of entries in the given table, including comparison and arithmetic operations [39]. The problem is commonly approached assuming a semantic parsing paradigm, with an intermediate state of formal meaning representation, e.g., inferred query or predicted operand to apply on selected cells [58, 18]. We reformulate the task as document QA by rendering the original HTML and restrict available information to layout given by visible lines and token positions (Appendix C).

TabFact[★]. To study fact verification with semi-structured evidence over relatively clean and simple tables collected from Wikipedia, entailed and refuted statements corresponding to a single row or cell were prepared by the authors of TabFact [7]. Without being affected by the simplicity of binary classification, this task poses challenges due to the complex linguistic and symbolic reasoning

²The ANLS metric used in PWC, representing KIE with property groups, differs from one used in VQA. Since it is not known how many groups are to be returned, the basis of the metric is the F1 score (in contrast to accuracy). Moreover, we require exact math for numerical variables. See implementation in the repository.

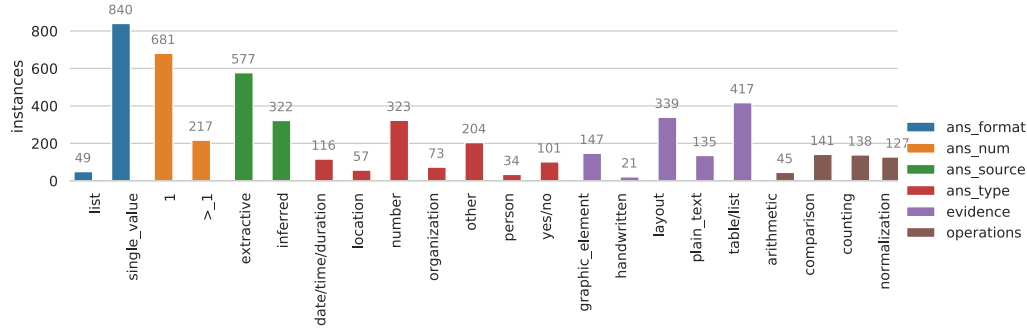


Figure 2: Number of annotated instances in each diagnostic subset category. All datasets in total.

required to perform with high accuracy. Analogously to WTQ, we render tables and reformulate the task as document NLI (Appendix C).

3.2 Diagnostic subsets

As pointed out by Ruder, *to better understand the strengths and weaknesses of our models, we furthermore require more fine-grained evaluation* [43]. We propose several auxiliary validation subsets, spanning across all the tasks, to improve result analysis and aid the community in identifying where to focus its efforts. A detailed description of these categories and related annotation procedures is provided in Appendix F.

Answer characteristic. We consider four features regarding the shallow characteristic of the answer. First, we indicate whether the answer is provided in the text explicitly in exact form (*extractive* data point) or has to be inferred from the document content (*abstractive* one). The second category includes, e.g., all the cases where value requires normalization before being returned (e.g., changing the date format). Next, we distinguish expected answers depending on whether they contain a *single value* or *list* of values. Finally, we decided to recognize several popular data types depending on shapes or class of expected named entity, i.e., to distinguish *date*, *number*, *yes/no*, *organization*, *location*, and *person* classes.

Evidence form. As we intend to analyze systems dealing with rich data, it is natural to study the performance w.r.t. the form that evidence is presented within the analyzed document. We distinguished *table/list*, *plain text*, *graphic element*, *layout*, and *handwritten* categories.

Required operation. Finally, we distinguish whether i.e., *arithmetic operation*, *counting*, *normalization* or some form of *comparison* has to be performed to answer correctly.

Table 2: Brief characteristics of our contribution, major fixes and modifications introduced to particular datasets. The enhancements of "Reformulation as DU" or "Improving data splits" are marked with \star and are sufficient to consider the dataset unique; hence, achieved results are not comparable to the previously reported. See Appendix C for a full description of tasks processing.

Dataset	Diagnostic sets	Unified format	Human performance	Manual annotation	Reformulation as DU	Improved split
DocVQA	+	+	-	-	-	-
InfographicsVQA	+	+	-	-	-	-
Kleister Charity	+	+	-	-	-	-
PWC \star	+	+	+	+	+	+
DeepForm \star	+	+	+	+	-	+
WikiTableQuestions \star	+	+	+	-	+	+
TabFact \star	+	+	-	-	+	-

Datasets included in the benchmark differ in task type, origin, and answer form. As their random samples were annotated, diagnostic categories are not distributed uniformly and reflect the character of the problems encountered in a particular task (see Figures 10–11 in the Appendix). For example, the requirement of answer normalization is prevalent in KIE tasks of DeepForm, PWC, and Kleister Charity but not elsewhere. Consequently, the general framework of diagnostic subsets we designed can be used not only to analyze model performance but also to characterize the datasets themselves.

3.3 Intended use

Data. We propose a unified data format for storing information in the Document Understanding domain and deliver converted datasets as part of the released benchmark (all selected datasets are hosted on the <https://duebenchmark.com/data> and can be downloaded from there). It assumes three interconnected dataset, document annotation and document content levels. The dataset level is intended for storing the general metadata, e.g., name, version, license, and source. The documents annotation level is intended to store annotations available for individual documents within datasets and related metadata (e.g., external identifiers). The content level store information about output and metadata from a particular OCR engine that was used to process documents (Appendix G).

Evaluation protocol. To evaluate a system on the DUE benchmark, one must create a JSON file with the results (in the data format mentioned above) based on the provided test data for each dataset and then upload all of the data to the website. Moreover, we establish a set of rules (Appendix H) which guarantees that all the benchmark submissions will be fair to compare, reproducible, and transparent (e.g., training performed on a development set is not allowed).

Leaderboard. We provide an online platform for the evaluation of Document Understanding models. To keep an objective means of comparison with the previously published results, we decided to retain the initially formulated metrics. To calculate the global score we resort to an arithmetic mean of different metrics due to its simplicity and straightforward calculation.³ In our platform we focus on customization, e.g., multiple leaderboards are available, and it is up to the participant to decide whether to evaluate the model on an entire benchmark or particular category. Moreover, we pay attention to the explanation by providing means to analyze the performance concerning document or problem types (e.g., using the diagnostic sets we provide).⁴

4 Experiments

Following the evaluation protocol, the training is run three times for each configuration of model size, architecture, and OCR engine. We performed OCR pre-processing stage for DocVQA, InfographicsVQA, Kleister Charity, and DeepForm datasets since they have PDF (mix of scans and born-digital documents) or image files as an input. PWC, WikiTableQuestions and TabFact datasets contain all born-digital documents so the ground true data are available and there is no need to run OCR engine (see Appendix C). In both cases, the pre-processing stage as an output return textual tokens and their locations (bounding boxes and page number) as a list (as a result the reading order is also provided).

4.1 Baselines

The focus of the experiments was to calculate baseline performance using a simple and popular model capable of solving all tasks without introducing any task-specific alterations. Employed methods were based on the previously released T5 model with a generic layout-modeling modification and pretraining.

T5. Text-to-text Transformer is particularly useful in studying performance on a variety of sequential tasks. We decided to rely on its extended version to identify the current level of performance on the chosen tasks and to facilitate future research by providing extendable architecture with a straightforward training procedure that can be applied to all of the proposed tasks in an end-to-end manner [41].

³Scores on the DocVQA and InfographicsVQA test sets are calculated using the official website.

⁴We intend to gather datasets not included in the present version of the benchmark to facilitate evaluations in an entire field of DU, regardless of if they are included in the current version of the leaderboard.

Table 3: Best results of particular model configuration in relation to human performance and external best. The external bests marked with — were omitted due to the significant changes in the data sets. *U* stands for unsupervised pretraining.

Dataset / Task type	Score (task-specific metric)					Human
	T5	T5+2D	T5+U	T5+2D+U	External best	
DocVQA	70.4 \pm 2.1	69.8 \pm 0.7	76.3 \pm 0.3	81.0 \pm 0.2	87.1 [40]	98.1
InfographicsVQA	36.7 \pm 0.6	39.2 \pm 1.0	37.1 \pm 0.2	46.1 \pm 0.1	61.2 [40]	98.0
Kleister Charity	74.3 \pm 0.3	72.6 \pm 1.1	76.0 \pm 0.1	75.9 \pm 0.7	83.6 [63]	97.5
PWC \star	25.3 \pm 3.3	25.7 \pm 1.0	27.6 \pm 0.6	26.8 \pm 1.8	—	69.3
DeepForm \star	74.4 \pm 0.6	74.0 \pm 0.7	82.9 \pm 0.9	83.3 \pm 0.3	—	98.5
WikiTableQuestions \star	33.3 \pm 0.7	30.8 \pm 1.9	38.1 \pm 0.1	43.3 \pm 0.4	—	76.7
TabFact \star	58.9 \pm 0.5	58.0 \pm 0.3	76.0 \pm 0.1	78.6 \pm 0.1	—	92.1
Visual QA	53.6	54.5	56.7	63.5	n/a	98.1
KIE	69.1	67.7	74.8	76.4	n/a	88.4
Table QA/NLI	29.4	29.0	38.0	39.3	n/a	84.4
Overall	50.7	50.4	56.5	59.8	n/a	90.3

T5+2D. Extension of the model we propose assumes the introduction of 2D positional bias that has been shown to perform well on tasks that demand layout understanding [56, 40, 63]. We rely on 2D bias in a form introduced in TILT model [40] and provide its first open-source implementation (available in supplementary materials). We expect that comprehension of spatial relationships achieved in this way will be sufficient to demonstrate that methods from the plain-text NLP can be easily outperformed in the DUE benchmark.

Unsupervised pretraining. We constructed a corpus of documents with a visually rich structure, based on 480k PDF files from the UCSF Industry Documents Library. It is used with a T5-like masked language model pretraining objective but in a salient span masking scheme where named entities are preferred over random tokens [41, 15]. An expected gain from its use is to tune 2D biases and become more robust to OCR errors and incorrect reading order.⁵

Human performance. We relied on the original estimation for DocVQA, InfographicsVQA, Charity, and TabFact datasets. For the PWC, WTQ and DeepForm estimation of human performance, we used the help of professional in-house annotators who are full-time employees of our company (see Appendix E). Each dataset was handled by two annotators; the average of their scores, when validated against the gold standard, is treated as the human performance (see Table 3). Interestingly, human scores on PWC are relatively low in terms of ANLS value – we explained this and justified keeping the task in Appendix C.

4.2 Results

Comparison of the best-performing baselines to human performance and top results reported in the literature is presented in Table 3. In several cases, there is a small difference between the performance of our baselines and the external best. It can be attributed to several factors. First, the best results previously obtained on the tasks were task-specific, i.e., were explicitly designed for a particular task and did not support processing other datasets within the benchmark. Secondly, there are differences between the evaluation protocol that we assume and what the previous authors assumed (e.g., we do not allow training models on the development sets, we require reporting an average of multiple runs, we disallow pretraining on datasets that might lead to information leak). Thirdly, our baseline could not address examples demanding vision comprehension as it does not process image inputs. Finally, there is the case of Kleister Charity. An encoder-decoder model we relied on as a one-to-fit-all baseline cannot process an entire document due to memory limitations. As a result, the score was lower as we consumed only a part of the document. Note that external bests for reformulated tasks are no longer applicable to the benchmark in its present, more demanding form.

⁵Details of the training procedure, such as used hyperparameters and source code, are available in the repository accompanying the paper.

Irrespective of the task and whether our competitive baselines or external results are considered, there is still a large gap to humans, which is desired for novel baselines. Moreover, one can notice that the addition of 2D positional bias to the T5 architecture leads to better scores assuming the prior pretraining step, which is yet another result we anticipated as it suggests that considered tasks have an essential component of layout comprehension.

Interestingly, the performance of the model can be significantly enhanced (up to 20.6 points difference for TabFact dataset and T5+2D+U model) by providing additional data for the said unsupervised pretraining. Thus, the results not only support the premise that understanding 2D features demand more unlabeled data than the chosen datasets can offer but also lay a common ground between them, as the same layout-specific pretraining improved performance on all of them independently. This observation confirms that the notion of layout is a vital part of the chosen datasets.

4.3 Challenges of the Document Understanding domain

Owing to its end-to-end nature and heterogeneity, Document Understanding is the touchstone of Machine Learning. However, the challenges begin to pile up due to the mere form a document is available in, as there is a widespread presence of analog materials such as scanned paper records. In the analysis below, we aim to explore the field of DU from the perspective of the model’s development and point out the most critical limiting factors for achieving satisfying results.

Impact of OCR quality. We present detailed results for Azure CV and Tesseract OCR engine in Table 5. The differences in scores are huge for most of the datasets (up to 18.4% in DocVQA) with a clean advantage for Azure CV. Consequently, we see that architectures evaluated with different OCR engines are incomparable, e.g., the choice of an OCR engine may impact results more than the choice of model architecture. Moreover, with the usage of our diagnostic datasets we can observe that Tesseract struggle the most with *Handwritten* and *Table/list* categories in comparison to *Plain text* category. It is worth noting that we see a bigger difference in the results between Azure CV and Tesseract for *Extractive* category, which suggest that we should use better OCR engines especially for that kind of problems.

Requirement of multi-modal comprehension. In addition to layout and textual semantics, part of the covered problems demand a Computer Vision component, e.g., to detect a logo, analyze a figure, recognize text style, determine whether the document was signed or the checkbox nearby was selected. This has been confirmed by ablation studies performed by Powalski et al. [40] for the DocVQA and by the fact that models with vision component achieve better performance on leaderboards for datasets such as DocVQA and the InfographicsVQA datasets [40, 56, 23, 22]. Thus, Document Understanding naturally incorporates challenges of both multi-modality and each modality individually (but not for all tasks equally, see Figures 10–11 in the Appendix). Since none of our baselines contain a vision component, we underperform on the category of problems requiring multi-modality, as is visible on the diagnostic dataset we proposed. Nevertheless, better performance of the T5+2D model suggests that part of the problems considered as *visual*, can be to some extent approximated by solely using the words’ spatial relationships (e.g., text curved around a circle, located in the top-left corner of the page presumably has the logo inside).

Single architecture for all datasets. It is common that token-level annotation is not available, and one receives merely key-value or question-answer pairs assigned to the document. Even in problems of extractive nature, token spans cannot be easily obtained, and consequently, the application of state-of-the-art architectures from other tasks is not straightforward. In particular, authors attempting Document Understanding problems in sequence labeling paradigms were forced to rely on faulty handcrafted heuristics [40]. In the case of our baseline models, this problem is addressed straightforwardly by assuming a sequence-to-sequence paradigm that does not make use of token-level annotation. This solution, however, comes with a trade-off of low performance on datasets requiring comprehension of long documents, such as Kleister Charity (see Table 4).

Table 4: F1 score on the Kleister Charity challenge with various maximum input sequence lengths.

Dataset	Maximum input sequence length			
	1024	2048	4096	6144 (max)
Kleister Charity	56.6	66	73.2	75.9

Table 5: Scores for different OCR engines and datasets with T5+2D model performing on 1024 tokens.

OCR	DocVQA	IVQA	Charity	DeepForm	Average	Average scores for different diagnostic categories				
						Extractive	Inferred	Handwritten	Table/list	Plain text
Azure CV (v3.2)	71.8	40.0	57.7	74.8	61.1	51.3	33.0	31.3	46.0	65.3
Tesseract (v4.0)	55.7	28.3	55.7	66.8	51.6	43.1	29.5	12.5	27.2	61.1

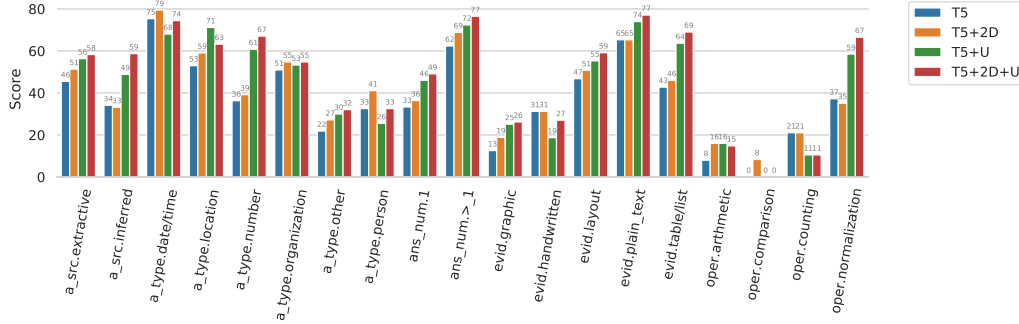


Figure 3: Results for diagnostic subsets. See Appendix F for detailed description of these categories.

Diagnostic dataset. Our diagnostic datasets are an important part of the analysis of different challenges in general (e.g., OCR quality or multi-modal comprehension, as we mentioned above) and for debugging different types of architectural decisions (see Figure 3). For example, we can observe a big advantage of unsupervised pretraining in the *inferred*, *number*, *table/list* categories, which shows the importance of a good dataset for specific problems (dataset used for pretraining the original T5 model has a small number of documents containing tables). The most problematic categories for all models were those related to complex logic operations: *arithmetic*, *counting*, *comparison*.

5 Conclusions

To efficiently pass information to the reader, writers often assume that structured forms such as tables, graphs, or infographics are more accessible than sequential text due to human visual perception and our ability to understand a text’s spatial surroundings. We investigate the problem of correctly measuring the progress of models able to comprehend such complex documents and propose a benchmark – a suite of tasks that balance factors such as quality of a document, importance of layout information, type and source of documents, task goal, and the potential usability in modern applications.

We aim to track the future progress on them with the website prepared for transparent verification and analysis of the results. The former is facilitated by the diagnostics subsets we derived to measure vital features of the Document Understanding systems. Finally, we provide a set of solid baselines, datasets in the unified format, and released source code to bootstrap the research on the topic.

Acknowledgments and disclosure of funding

The authors would like to thank Samuel Bowman, Łukasz Garncarek, Dimosthenis Karatzas, Minesh Mathew, Zofia Prochoroff, and Rubèn Pérez Tito for the helpful discussions on the draft of the paper. Moreover, we thank the reviewers of both rounds of the NeurIPS 2021 Datasets and Benchmarks Track for their comments and suggestions that helped improve the paper.

The Smart Growth Operational Programme supported this research under project no. POIR.01.01.01-00-0877/19-00 (*A universal platform for robotic automation of processes requiring text comprehension, with a unique level of implementation and service automation*).

References

- [1] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha. DocFormer: End-to-end transformer for document understanding, 2021.
- [2] T. Bayer, J. Franke, U. Kressel, E. Mandler, M. Oberländer, and J. Schürmann. Towards the understanding of printed documents. In H. S. Baird, H. Bunke, and K. Yamamoto, editors, *Structured Document Image Analysis*, pages 3–35. Springer Berlin Heidelberg, Berlin, Heidelberg, 1992.
- [3] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy, July 2019. Association for Computational Linguistics.
- [4] R. Chaudhry, S. Shekhar, U. Gupta, P. Maneriker, P. Bansal, and A. Joshi. Leaf-qa: Locate, encode attend for figure question answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3501–3510, 2020.
- [5] L. Chen, X. Chen, Z. Zhao, D. Zhang, J. Ji, A. Luo, Y. Xiong, and K. Yu. WebSRC: A dataset for web-based structural reading comprehension, 2021.
- [6] W. Chen, M. Chang, E. Schlinger, W. Wang, and W. Cohen. Open question answering over tables and text. *Proceedings of ICLR 2021*, 2021.
- [7] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang. TabFact : A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, April 2020.
- [8] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data, 2021.
- [9] M. Cho, R. K. Amplayo, S. won Hwang, and J. Park. Adversarial TableQA: Attention supervision for question answering on tables, 2018.
- [10] M. Dehghani. Toward document understanding for information retrieval. *SIGIR Forum*, 51(3):27–31, Feb. 2018.
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [12] T. I. Denk and C. Reisswig. BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.
- [13] F. Esposito, D. Malerba, G. Semeraro, and S. Ferilli. Knowledge revision for document understanding. In *ISMIS*, 1997.
- [14] V. Gupta, M. Mehta, P. Nokhiz, and V. Srikumar. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online, July 2020. Association for Computational Linguistics.
- [15] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. Retrieval augmented language model pre-training. In *ICML*, 2020.
- [16] R. M. Haralick. Document image understanding: Geometric and logical layout. In *CVPR*, volume 94, pages 385–390, 1994.
- [17] A. W. Harley, A. Ufkes, and K. G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

- [18] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, and J. Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online, July 2020. Association for Computational Linguistics.
- [19] T. Hong, D. Kim, M. Ji, W. Hwang, D. Nam, and S. Park. BROS: A layout-aware pre-trained language model for understanding documents, 2021.
- [20] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. Jawahar. ICDAR2019 competition on scanned receipt OCR and information extraction. In *ICDAR*, 2019.
- [21] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520, 2019.
- [22] ICDAR. Leaderboard of the Document Visual Question Answering - Infographics VQA. <https://rrc.cvc.uab.es/?ch=17&com=evaluation&task=3> (accessed September 30, 2021), 2021.
- [23] ICDAR. Leaderboard of the Document Visual Question Answering - Single Document VQA. <https://rrc.cvc.uab.es/?ch=17&com=evaluation&task=1> (accessed September 30, 2021), 2021.
- [24] G. Jaume, H. K. Ekenel, and J.-P. Thiran. FUNSD: A dataset for form understanding in noisy scanned documents, 2019.
- [25] K. V. Jobin, A. Mondal, and C. V. Jawahar. DocFigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79, 2019.
- [26] M. Kardas, P. Czaplá, P. Stenetorp, S. Ruder, S. Riedel, R. Taylor, and R. Stojnic. AxCell: Automatic extraction of results from machine learning papers, 2020.
- [27] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. B. Faddoul. Chargrid: Towards Understanding 2D Documents. *ArXiv*, abs/1809.08799, 2018.
- [28] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi. Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017.
- [29] S. Kounev, K. Lange, and J. von Kistowski. *Systems Benchmarking: For Scientists and Engineers*. Springer International Publishing, 2020.
- [30] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou. DocBank: A benchmark dataset for document layout analysis, 2020.
- [31] B. P. Majumder, N. Potti, S. Tata, J. B. Wendt, Q. Zhao, and M. Najork. Representation learning for information extraction from form-like documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6504, Online, July 2020. Association for Computational Linguistics.
- [32] M. Mathew, V. Bagal, R. P. Tito, D. Karatzas, E. Valveny, and C. V. Jawahar. InfographicVQA, 2021.
- [33] M. Mathew, D. Karatzas, and C. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209, January 2021.
- [34] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. PlotQA: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [35] L. Nan, C. Hsieh, Z. Mao, X. V. Lin, N. Verma, R. Zhang, W. Kryściński, N. Schoelkopf, R. Kong, X. Tang, M. Mutuma, B. Rosand, I. Trindade, R. Bandaru, J. Cunningham, C. Xiong, and D. Radev. FeTaQA: Free-form table question answering, 2021.

- [36] D. Niyogi and S. N. Srihari. A rule-based system for document understanding. In *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence*, pages 789–793, 1986.
- [37] R. B. Palm, F. Laws, and O. Winther. Attend, copy, parse end-to-end information extraction from documents. *International Conference on Document Analysis and Recognition (ICDAR)*, 2019.
- [38] S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, and H. Lee. CORD: A consolidated receipt dataset for post-ocr parsing. In *Document Intelligence Workshop at NeurIPS*, 2019.
- [39] P. Pasupat and P. Liang. Compositional semantic parsing on semi-structured tables. *CoRR*, abs/1508.00305, 2015.
- [40] R. Powalski, Ł. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer. In J. Lladós, D. Lopresti, and S. Uchida, editors, *Document Analysis and Recognition – ICDAR 2021*, pages 732–747, Cham, 2021. Springer International Publishing.
- [41] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [42] M. Rimol. Gartner Forecasts Worldwide Hyperautomation-Enabling Software Market to Reach Nearly \$600 Billion by 2022. <https://www.gartner.com/en/newsroom/press-releases/2021-04-28-gartner-forecasts-worldwide-hyperautomation-enabling-software-market-to-reach-nearly-600-billion-by-2022>, 2021.
- [43] S. Ruder. Challenges and Opportunities in NLP Benchmarking. <http://ruder.io/nlp-benchmarking>, 2021.
- [44] Z. Shen, K. Lo, L. L. Wang, B. Kuehl, D. S. Weld, and D. Downey. Incorporating visual layout structures for scientific text classification, 2021.
- [45] T. Stanisławek, F. Graliński, A. Wróblewska, D. Lipiński, A. Kaliska, P. Rosalska, B. Topolski, and P. Biecek. Kleister: Key information extraction datasets involving long documents with complex layouts, 2021.
- [46] H. Sun, Z. Kuang, X. Yue, C. Lin, and W. Zhang. Spatial dual-modality graph reasoning for key information extraction, 2021.
- [47] S. Svetlichnaya. DeepForm: Understand structured documents at scale. https://wandb.ai/stacey/deepform_v1/reports/DeepForm-Understand-Structured-Documents-at-Scale--VmlldzoyODQ3Njg, 2020.
- [48] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi, and J. Berant. Multimodalqa: Complex question answering over text, tables and images. *CoRR*, abs/2104.06039, 2021.
- [49] R. Tanaka, K. Nishida, and S. Yoshida. VisualMRC: Machine reading comprehension on document images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13878–13888, May 2021.
- [50] S. L. Taylor, D. Dahl, M. Lipshutz, C. Weir, L. M. Norton, R. Nilson, and M. Linebarger. Integrated text and image understanding for document understanding. In *HLT*, 1994.
- [51] H. M. Vu and D. T. Nguyen. Revising FUNSD dataset for key-value detection in document images. *CoRR*, abs/2010.05322, 2020.
- [52] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

- [53] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.
- [54] T.-L. Wu, S. Singh, S. Paul, G. Burns, and N. Peng. MELINDA: A multimodal dataset for biomedical experiment method classification. *ArXiv*, abs/2012.09216, 2020.
- [55] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou. LayoutLM: Pre-training of text and layout for document image understanding, 2019.
- [56] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding, 2020.
- [57] S. Yacoub. Automated quality assurance for document understanding systems. *IEEE Software*, 20(3):76–82, 2003.
- [58] P. Yin, G. Neubig, W.-t. Yih, and S. Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online, July 2020. Association for Computational Linguistics.
- [59] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruysen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, and N. Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2020.
- [60] X. Zheng, D. Burdick, L. Popa, and N. X. R. Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 697–706, 2021.
- [61] X. Zhong, J. Tang, and A. J. Yepes. PubLayNet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, Sep. 2019.
- [62] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online, Aug. 2021. Association for Computational Linguistics.
- [63] Łukasz Garncarek, R. Powalski, T. Stanisławek, B. Topolski, P. Halama, and F. Graliński. LAMBERT: Layout-aware (language) modeling using bert for information extraction, 2020.

A Considered datasets

A.1 Desired characteristics

End-to-end nature. As the value and importance of Document Understanding result from its application to process automation, a good benchmark should measure to which degree workers can be supported in their tasks. Though Layout Analysis is oldest of the Document Understanding problems, its output is often not an end in itself but rather a half-measure disconnected from the final information the system is used for. We also remove all tasks which as an input takes collection of documents.

Quality. Availability of high-quality annotation was a condition *sine qua non* for a task to qualify. To ensure the highest annotation quality, we excluded resources prepared using a distant annotation procedure, e.g., classification tasks where entire sources were labeled instead of individual instances, or templated question-answer pairs.

Difficulty. As it makes no sense to measure progress on solved problems, only tasks with a substantial gap between human performance and state-of-the-art models were considered. In the case of promising tasks lacking a human baseline, we provided our estimation. Moreover, we remove all tasks where free text was dominated in documents (we don't need to use layout or visual features).

Licensing. In publishing our benchmark, we are making efforts to ensure the highest standards for the future of the machine learning community. Only tasks with a permissive license to use annotations and data for further research can be considered.

At the same time, we recognized it is essential to approach the benchmark construction holistically, i.e., to carefully select tasks from diverse domains and types in the rare cases where datasets are abundant.

A.2 Datasets selection process

The review protocol consisted of a manual search in specific databases, repositories and distribution services. The scientific resources included in the search were:

- <https://paperswithcode.com/datasets/>
- <https://datasetsearch.research.google.com/>
- <https://data.mendeley.com/>
- <https://arxiv.org/search/>
- <https://github.com/>
- <https://allenai.org/data/>
- <https://www.semanticscholar.org/>
- <https://scholar.google.com/>
- <https://academic.microsoft.com/home>

Results were reviewed by one of authors of the present paper and the resources related to classification, KIE, QA, MRC, and NLI over complex documents, figures, and tables were identified as potentially relevant (in accordance with inclusion criteria described in Section A.1).

The initial search assumed use of the following keywords: *Question Answering, Visual Question Answering, Document Question Answering, Document Classification, Document Dataset, Information Extraction*. Additionally, we used *Machine Reading Comprehension, Question Answering, VQA* in combination with *Document*, and *Visual, Document, Table, Figure, Plot, Chart, Hybrid* in combination with *Question Answering* or *Information Extraction*.

Table 6 presents list of relevant datasets and results of their assessment according to the criteria of end-to-end nature, quality, difficulty, and licensing. Candidate tasks resulted from an extensive review of both literature and data science challenges without accompanying publication and their basic characteristics.

Table 6: Comparison of selected and considered datasets with their base characteristic, including information regarding whether an input is a collection of documents (Col.), entire document (Doc.) or document excerpt (Exc.).

Dataset	Type	Size (thousands)			Selection criteria			Input	Domain	Comment	
		Train	Dev	Test	End-to-end	Quality	Difficulty				Licensing
Kleister Charity [45]	KIE	1.73	.44	.61	+	+	+	+	Doc.	Finances	
PWC [26]	KIE	.2	.06	.12	+	+	+	+	Doc.	Scientific	
DeepForm [47]	KIE	.7	.1	.3	+	+	+	+	Doc.	Finances	
DocVQA [33]	Visual QA	10.2	1.3	1.3	+	+	+	+	Doc.	Business	
InfographicsVQA [32]	Visual QA	4.4	.5	.6	+	+	+	+	Doc.	Open	
TabFact [7]	Table NLI	13.2	1.7	1.7	+	+	+	+	Exc.	Open	
WTQ [39]	Table QA	1.4	.3	.4	+	+	+	+	Exc.	Open	
Kleister NDA [45]	KIE	.25	.08	.2	+	+	-	+	Doc.	Legal	Dominated by extraction from free text
SROIE [20]	KIE	.63	-	.35	+	+	-	+	Doc.	Finances	No room for improvement
CORD [38]	KIE	.8	.1	.1	+	+	-	+	Doc.	Finances	No room for improvement
Wildreceipt [46]	KIE	1.27	-	.47	+	+	-	+	Doc.	Finances	No room for improvement
WebSRC [5]	KIE	4.55	.9	1.0	+	-	+	+	Doc.	Open	Templated input data
FUNSD [24]	KIE	.15	-	.05	+	-	+	+	Doc.	Finances	Known disadvantages [51]
DocVQA [32]	Visual QA	4.4	.5	.6	-	+	+	+	Col.	Open	Document Collection Question Answering
TextbookQA [28]	Visual QA	.67	.2	.21	+	-	+	+	Doc.	Educational	Source files are not available
MultiModalQA [48]	Visual QA	23.82	2.44	3.66	+	-	+	+	Doc.	Open	Automatically generated questions
VisualMRC [49]	Visual MRC	7	1	2	+	+	-	+	Doc.	Open	Human performance reached
RVL-CDIP [17]	Classification	320	40	40	+	+	-	+	Doc.	Finances	No room for improvement
DocFigure [25]	Classification	19.8	-	13.1	+	+	-	+	Doc.	Scientific	No room for improvement
EURLEX57K [3]	Classification	45	6	6	+	+	-	+	Doc.	Legal	Dominated by extraction from free text
MELINDA [54]	Classification	4.34	.45	.58	+	-	+	+	Doc.	Scientific	Semi-supervised annotation
S2-VL [44]	DLA	1.3	-	-	-	+	+	+	Doc.	Scientific	Cross-validation for training and testing
DocBank [30]	DLA	398	50	50	-	-	+	+	Doc.	Scientific	Automatic annotation
Publaynet [61]	DLA	340.4	11.9	12	-	-	+	+	Doc.	Scientific	Automatic annotation
FinTabNet [60]	DLA	61.8	7.19	7.01	-	+	+	+	Doc.	Finances	Different styles in comparison to sci./gov. docs
PlotQA [34]	Figure QA	157	33.7	33.7	+	-	+	+	Exc.	Open	Synthetic
Leaf-QA [4]	Figure QA	200	40	8.15	+	-	+	+	Exc.	Open	Templated questions
TAT-QA [62]	Table QA	2.2	.28	.28	+	-	+	+	Exc.	Finances	Source files are not available
WikiOPS [9]	Table QA	17.28	2.47	4.67	+	+	-	+	Exc.	Open	No room for improvement
FeTaQA [35]	Table QA	7.33	1.0	2.0	+	-	+	+	Exc.	Open	Answers as a free-form text
HybridQA [8]	Table QA	62.68	3.47	3.46	-	+	+	+	Col.	Open	Multihop Question Answering
OTT-QA [6]	Table QA	41.46	2.24	2.16	-	+	+	+	Col.	Open	Multihop Question Answering
INFOTABS [14]	Table NLI	1.74	.2	.6	+	+	+	+	Col.	Open	TabFact is very similar

B Minor dataset modifications

Deduplication. Through the systematic analysis and validation of the chosen datasets, we noticed one of the commonly appearing defects is the presence of duplicated annotations. We decided to remove these duplicates from InfographicsVQA (14 annotations from train, two from the dev set), DocVQA (four from train and test sets each), TabFact (309 from train, 53 from dev, and 52 the test set), and WikiTableQuestions (one annotation from each train and test sets).

C Tasks processing and reformulation

Since part of the datasets were reformulated or modified to improve the benchmark quality or align the task with the Document Understanding paradigm, we describe the introduced changes in detail below.

WikiTableQuestions*. We prepare input documents by rendering table-related HTML distributed by authors in *wkhtmltopdf* and crop the resulting files with *pdfcrop*. As these code excerpts do not contain *head* tag with JavaScript and stylesheet references, we use the header from the present version of the Wikipedia website.

Approximately 10% of tables contained at least one *img* tag with a source that is no longer reachable. It results in a question mark icon displayed instead of the image and does not impact the evaluation procedure since the questions here do not require image comprehension.

The original WTQ dataset consists of *training*, *pristine-seen-tables*, and *pristine-unseen-tables* subsets. We treat *pristine-unseen-tables* as a test set and create new training and development sets by rearranging data from *training* and *pristine-seen-tables*. The latter operation is dictated by the leakage of documents in the original formulation, i.e., we consider it undesirable for a document to appear in different splits, even if the question differs. The resulting dataset consists of approximately

Year	Venue	Winners	Runner-up	3rd place
2005	Pardubice	Poland (41 pts)	Sweden (35 pts)	Denmark (24 pts)
2006	Rybnik	Poland (41 pts)	Sweden (27 pts)	Denmark (26 pts)
2007	Abensberg	Poland (40 pts)	Great Britain (36 pts)	Czech Republic (30 pts)
2008	Holsted	Poland (40 pts)	Denmark (39 pts)	Sweden (38 pts)
2009	Gorzów Wlkp.	Poland (57 pts)	Denmark (45 pts)	Sweden (32 pts)
2010	Rye House	Denmark (51 pts)	Sweden (37 pts)	Poland (35 pts)
2011	Balakovo	Russia (61 pts)	Denmark (31 pts)	Ukraine (29+3 pts)
2012	Gniezno	Poland (61 pts)	Australia (44 pts)	Sweden (26 pts)
Year	Venue	Winners	Runner-up	3rd place

Figure 4: Document in WikiTableQuestions reformulated as Document Understanding.

(Question) After their first place win in 2009, how did Poland place the next year at the speedway junior world championship? (Answer) 3rd place

2100 documents divided in the proportion of 65%, 15%, 20% into training, development, and test sets.

We rely on the original WTQ metric which is a form of Accuracy with normalization (see Pasupat et al. [39] and accompanying implementation).

TabFact★. As the authors of TabFact distribute only CSV files, we resorted to HTML from the WikiTables dump their CSV were presumably generated from.⁶ As Chen et al. [7] dropped some of the columns present in used WikiTable tables, we remove them too, to ensure compatibility with the original TabFact. Rendered files are used analogously to the case of WTQ.

	Nation	v t e Games				Points			Table points
		Played	Won	Drawn	Lost	For	Against	Difference	
1	VVA-Podmoskovye Monino	10	9	0	1	374	119	+255	37
2	Krasny Yar Krasnoyarsk	10	6	0	4	198	255	-57	28
3	Slava Moscow	10	5	1	4	211	226	-15	26
4	Yenisey-STM Krasnoyarsk	10	5	0	5	257	158	+99	25
5	RC Novokuznetsk	10	4	1	5	168	194	-26	23
6	Imperia-Dynamo Penza	10	0	0	10	138	395	-257	10

Figure 5: Document in TabFact reformulated as Document Understanding.

(Claim) To calculate table point, a win be worth 3, a tie be worth 1 and a loss be worth 0

Results differ from TabFact in several aspects, i.e., text in our variant is not normalized, it includes the original formatting, and the tables are more complex due to restoring the original cell merges. All mentioned differences are desired, as we intended to consider raw, unprocessed files without any heuristics or normalization applied.

Another difference we noticed is that tables in the original TabFact are sometimes one row shorter, i.e., they do not contain the last row present in the WikiTable dump. As it should not impact expected answers, we decided to maintain the fidelity to Wikipedia and use the complete table.

We use the original splits into training, development, and test sets and the original Accuracy metric.

DeepForm★. The original DeepForm dataset consists of 2012, 2014, and 2020 subsets differing in terms of annotation quality and documents' diversity. We decided to use only the 2020 subset as for 2014, and 2020 annotations were prepared either automatically or by volunteers, leading to questionable quality. The selected subset was randomly divided into training, development and test set.

We noticed several inconsistencies during the initial analysis that lead us to the manual correction of autodetected: (1) invalid date format; (2) flight start dates earlier than flight end; (3) documents lacking one or more data points.

In addition to the improved 2020 subset, we manually annotated one hundred 2012 documents, as they can pose different challenges (contain different document templates, handwriting, have lower

⁶<http://websail-fe.cs.northwestern.edu/TabEL/tables.json.gz>

COXREPS		REP BUYLINES													
Rep: TELEREP, INC. Order: 0000000000 Contract: 0000000000 Agency: TELEREP, INC. Advertiser: TELEREP, INC. Product: TELEREP, INC.		Order: 0000000000 Contract: 0000000000 Agency: TELEREP, INC. Advertiser: TELEREP, INC. Product: TELEREP, INC.													
Mod Code	Buy Line	Day/Time	Length	Starting Rate	Starting Date	Ending Date	# of Wks	Spt/Week	Total Spots	Total Dollars	Program Name	Rating RAAS-	Imprsn ASB-	Rep: Last RAAS+ Activity	Last Mod/Rev
	1	Tue 5-6A	:30S	\$10	May12/20	May12/20	1	1	1	\$10	NEWS10 GOOD MORN -5A	0.9	2.1	0.9 May04/20	Rev #0: A
	Contract Comment: NEWS10 GOOD MORN -5A														
	2	Wed 5-6A	:30S	\$10	May13/20	May13/20	1	1	1	\$10	NEWS10 GOOD MORN -5A	0.9	2.1	0.9 May04/20	Rev #0: A
	Contract Comment: NEWS10 GOOD MORN -5A														
	3	Thu 5-6A	:30S	\$10	May14/20	May14/20	1	1	1	\$10	NEWS10 GOOD MORN -5A	0.9	2.1	0.9 May04/20	Rev #0: A
	Contract Comment: NEWS10 GOOD MORN -5A														
	4	Mon 5-6A	:30S	\$10	May18/20	May18/20	1	1	1	\$10	NEWS10 GOOD MORN -5A	0.9	2.1	0.9 May04/20	Rev #0: A
	Contract Comment: NEWS10 GOOD MORN -5A														
	5	Wed 6-7A	:30S	\$15	May13/20	May13/20	1	1	1	\$15	NEWS10 GOOD MORN -6A	2.2	5.3	2.2 May04/20	Rev #0: A
	Contract Comment: NEWS10 GOOD MORN -6A														
	6	Thu 6-7A	:30S	\$15	May14/20	May14/20	1	1	1	\$15	NEWS10 GOOD MORN -6A	2.2	5.3	2.2 May04/20	Rev #0: A
	Contract Comment: NEWS10 GOOD MORN -6A														
	7	Fri 6-7A	:30S	\$15	May15/20	May15/20	1	1	1	\$15	NEWS10 GOOD MORN -6A	2.2	5.3	2.2 May04/20	Rev #0: A
	Contract Comment: NEWS10 GOOD MORN -6A														
	8	Mon 6-7A	:30S	\$15	May18/20	May18/20	1	1	1	\$15	NEWS10 GOOD MORN -6A	2.2	5.3	2.2 May04/20	Rev #0: A
	Contract Comment: NEWS10 GOOD MORN -6A														
	9	Tue 7-9A	:30S	\$20	May12/20	May12/20	1	1	1	\$20	CBS THIS MORNING	3.0	7.3	3.0 May04/20	Rev #0: A
	Contract Comment: CBS THIS MORNING														
	10	Thu 7-9A	:30S	\$20	May14/20	May14/20	1	1	1	\$20	CBS THIS MORNING	3.0	7.3	3.0 May04/20	Rev #0: A
	Contract Comment: CBS THIS MORNING														
	11	Mon 7-9A	:30S	\$20	May18/20	May18/20	1	1	1	\$20	CBS THIS MORNING	3.0	7.3	3.0 May04/20	Rev #0: A
	Contract Comment: CBS THIS MORNING														
	12	Tue 9-10A	:30S	\$10	May12/20	May12/20	1	1	1	\$10	FAMILY FEUD/ AMERICA SAYS	2.0	4.8	2.0 May04/20	Rev #2: NZ
	Contract Comment: FAMILY FEUD/ AMERICA SAYS														
	13	Thu 9-10A	:30S	\$10	May14/20	May14/20	1	1	1	\$10	FAMILY FEUD/ AMERICA SAYS	2.0	4.8	2.0 May04/20	Rev #2: NZ
	Contract Comment: FAMILY FEUD/ AMERICA SAYS														
	14	Fri 9-10A	:30S	\$10	May15/20	May15/20	1	1	1	\$10	FAMILY FEUD/ AMERICA SAYS	2.0	4.8	2.0 May04/20	Rev #2: NZ
	Contract Comment: FAMILY FEUD/ AMERICA SAYS														

Figure 6: Single page from document in DeepForm.

image quality). They were used to extend development and test set. The final dataset consists of 700 training, 100 development, and 300 test set documents. We rely on the standard F1 score for the purposes of DeepForm evaluation.

PWC*. The authors of AxCell relied on PWC Leaderboards and LinkedResults datasets [26]. The original formulation assumes extraction of *(task, dataset, metric, model, score)* tuples from a provided table. In contrast, we reformulate the task as Document Understanding and provide a complete paper as input instead. These are obtained using arXiv identifiers available in the PWC metadata. Consequently, the resulting task is an end-to-end Key Information Extraction from real-world scientific documents.

Whereas LinkedResults was annotated consistently, the PWC is of questionable quality as it was obtained from leaderboards filled by Papers with Code visitors without a clear guideline or annotation rules. The difference between the two is substantial, i.e., the agreement in terms of F1 score between publications present in both PWC and LinkedResults is lower than 0.35. We attribute this mainly to flaws in the PWC dataset, such as missing records, inconsistent normalization and the difficulty of the task itself.

Consequently, we decided to perform its manual re-annotation assuming that: (1) The best result for a proposed model variant on the single dataset has to be annotated, e.g., if two models with different parameter sizes were present in the table, we report only the best one. (2) Single number is preferred (we take the average over multiple split or parts of the dataset if possible). (3) When results from the test set are available, we prefer them and don't report results from the validation set. (4) We add multiple value variants when possible. (5) We include information on used validation/dev/test split in the dataset description wherever applicable. (6) We don't report results on the train set. (7) We don't annotate results not appearing in the table. (8) We filter out publications that are hard to annotate even for a human.

Interestingly, human scores on PWC are relatively low in terms of ANLS value. This can be attributed to unrestricted nature of particular properties, e.g., *accuracy* and *average accuracy* are equally valid metric values. Similarly, *Action Recognition*, *Action Classification*, and *Action Recognition* are equally valid task names. We mitigated this problem by using ANLS-like comparison used in the F1 metric and providing multiple acceptable value variants, i.e., it is enough to provide half of the string representing one of the valid answers.⁷

Nevertheless, it is impossible to provide all answer variants during the preparation of the gold standard. We decided to keep the dataset in the benchmark as it is extremely demanding, and there is still a large gap between humans' and models' performance (See Table 3).

⁷Please refer to the metric implementation in the Github repository for a detailed description.

As the expected answer in PWC consists of a list of groups (property tuples that represent a complete record of the method, dataset, and results), the F1 metric here has to take into account the miss-placement of properties in another group. We assume the value is incorrect if placed in the wrong group (see reference implementation in supplementary materials).

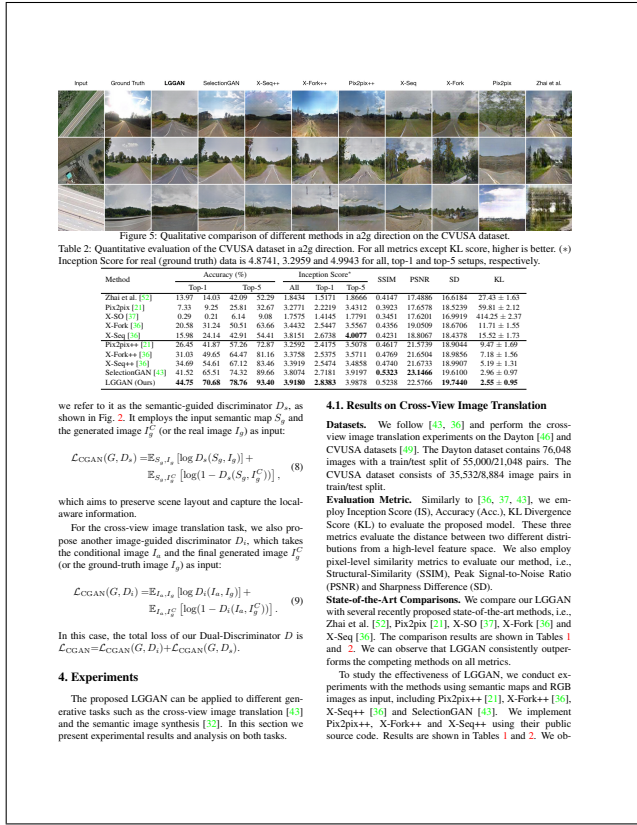


Figure 7: Single page from document in PWC.

D Dataset statistics

Chosen datasets represent the plethora of domains, lengths, and document types. This appendix covers the critical aspects of particular tasks at the population level.

Though part of the datasets is limited to one-pagers, the remaining documents range from a few to few hundred pages (Figure 8). At the same time, there is a great variety in how much text is present on a single page – we have both densely packed scientific documents and concise document excerpts or infographics. This diversity allows us to measure the ability to comprehend documents depending on their length.

E Details of human performance estimation

Estimation of human performance for PWC, WikiTableQuestions, DeepForm was performed in-house by professional annotators who are full-time employees of Applica.ai. Before approaching the process, each of them has to participate in the task-specific training described below.

Number of annotated samples depended on task difficulty and the variance of the resulting scores. We relied on 50 fully annotated papers for the PWC dataset (approx. 150 tuples with five values each), 109 DeepForm documents (532 values), and 300 questions asked to different WikiTableQuestion tables.

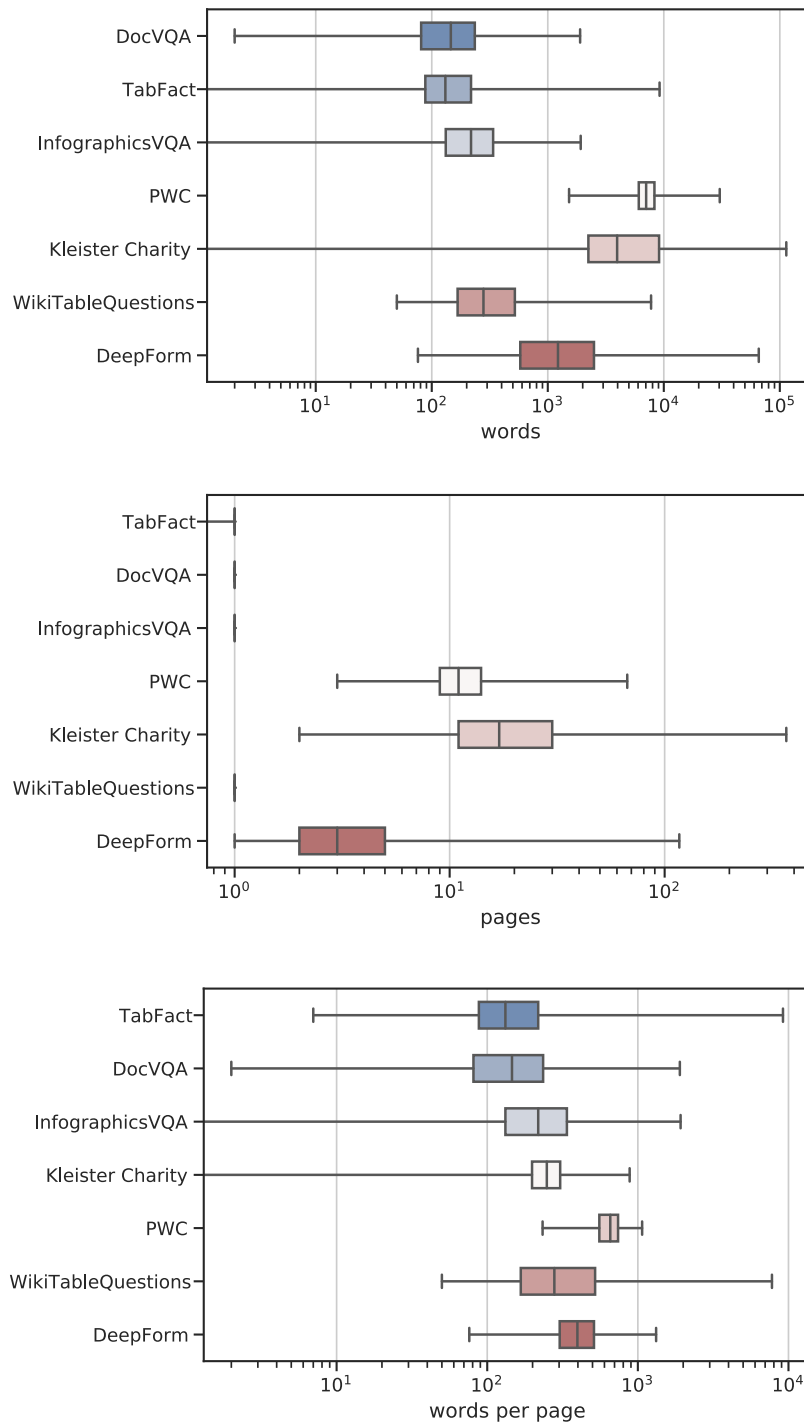


Figure 8: Number of words, pages, and words per page in particular datasets (log scale). Part of the datasets consist only of one-pagers.

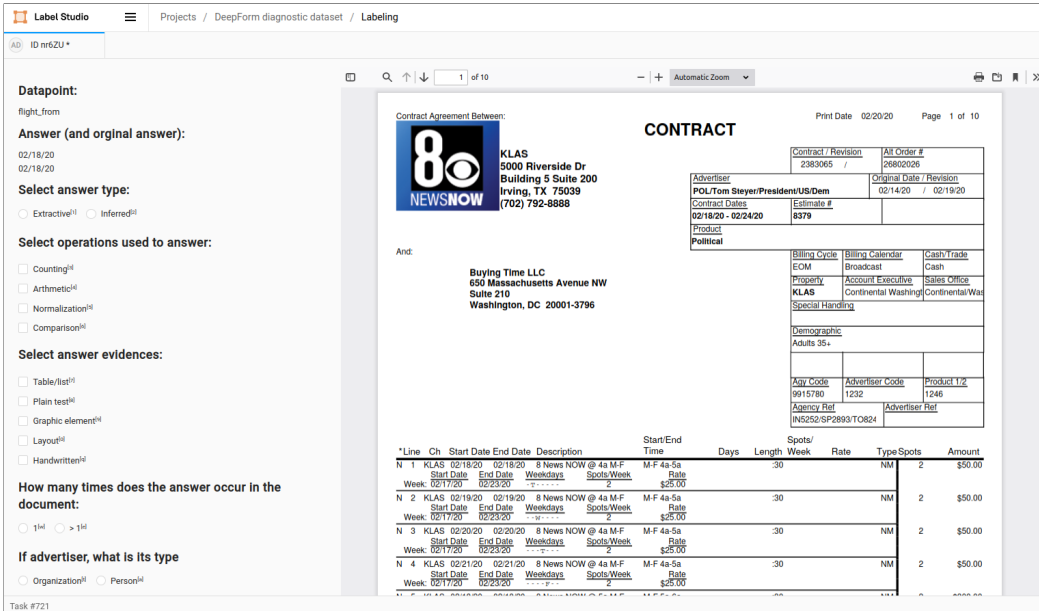


Figure 9: An example of an interface for annotating diagnostic subsets based on document from DeepForm dataset.

Each dataset was approached with two annotators in the LabelStudio tool. Human performance is the average of their scores when validated against the gold standard.

Training. Each person participating in the annotation process completed the training consisting of four stages: (1) Annotation of five random documents from the task-specific development set. (2) Comparative analysis of differences between their annotations and the gold standard. (3) Annotation of ten random documents from the task-specific development set and subsequent comparative analysis. (4) Discussion between annotators aimed at agreeing on the shared, coherent annotation rules.

F Annotation of diagnostic subsets

In order to analyze the prepared benchmark and the results of individual models, diagnostic sets were prepared. These diagnostic sets are subsets of examples selected from the testset for all datasets.

When building a taxonomy for diagnostic sets, we adopted two basic assumptions: (1) It must be consistent across all selected tasks so that at least two tasks can be noted with a given category (2) It should include as many aspects as possible that are relevant from the perspective of document understanding problem.

Initially, we adopted the taxonomies proposed in DocVQA, Infographics, and TabFact as potential categories [33, 32, 7]. In the next step, we adjusted our taxonomy to all datasets following the previously adopted assumptions, distinguishing seven main categories with 25 subcategories (for a more detailed description of the category (see the section F.1). Then, for each dataset, we prepared an annotation task in the LabelStudio tool ⁸ (see example 9) along with an annotation instruction. Finally, to determine Human performance, the annotation was carried out by a team of specialists from Applica.ai, where the selected example was noted only by one person.

F.1 Taxonomy description

The taxonomy is based on multiple aspects of documents, inputs, and answers and was designed to be sufficiently generic for future adaptation to other tasks. Here, in each category, we describe the predicates that annotators followed when classified an example into specific subcategories.

⁸<https://labelstud.io/>

Answer source. This category is based on the relation between answer and text in the document.

- Extractive – after lowercasing and white-characters removing, the answer can be exact-matched in the document.
- Inferred – other non-extractive cases.

Output format This category is based on the shape of an output.

- Single value – the answer consists of only one item.
- List – multiple outputs are to be provided.

Output type. This category is based on the semantic of an output.

- Organization – the answer is a name of an organization or institution.
- Location – the answer is a geographic location globally (e.g., a country, continent, city) or locally (building or street, among others).
- Person – the answer is a personal identifier (name, surname, pseudonym) or its composition. It can have a title prefix or suffix (e.g., Mrs., Mr., Ph.D.) or have a shortened or informal version.
- Number – numerical values given with the unit or percent. Values written in the free text do not comply with this class's definition.
- Date/Time/Duration – the answer represents the date, time, or the difference between two dates or times.
- Yes/No – the answer is a textual output of binary classification, such as Yes/No pairs, and Positive/Negative, 0/1 among others.

Evidence. This category is based on the source of information that allows the correct answer to be generated. When there are multiple justifications based on different pieces of evidence (for example, the address is in a table and block text), it is required to select all the pieces of evidence.

- Table or List – a *table* is a fragment of the document organized into columns and rows. The distinguishing feature of the table is consistency within rows and columns (usually the same data type). Moreover, it may have a header. In that sense, the form is not a table (or at least it does not have to be). A *list* is a table degenerated into one column or row containing a header.
- Plain text – the answer is based on plain text if there is an immediate need to understand a longer fragment of the text while answering.
- Graphic element – the answer is based on graphic evidence when understanding graphically rich, non-text fragments of documents (e.g., graphics, photos, logos (non-text)) are necessary for generating a correct answer.
- Layout – it is evidence when comprehending the placement of text on the page (e.g., titles, headers, footers, forms) is needed to generate the correct answer. This type does not include tables.
- Handwritten – when the text written by hand is crucial for an answer.

Operation. This category is based on the type of operations that are to be performed on the document before reaching to the correct answer.

- Counting – when there is a need to count the occurrences or determine the position on the list.
- Arithmetic – when there is an arithmetic operation applied before answering, or a sequence of arithmetic operations (e.g., averaging).
- Comparison – a comparison in the sense of lesser/greater. Other procedures that a comparison operation can express (e.g., approximation) may be chosen. Here, the operation "is equal" is not a comparison since it is sufficient to match sequences without a semantic understanding.
- Normalization – when we are to return something in the document but in a different form. It may only apply to the output; we do not acknowledge this operation when it is required to normalize a question fragment to match it in the document.

Answer number. This category is based on the number of occurrences of an answer in the document.

- 1 – when there is one path of logical reasoning to find the correct answer in the document. We treat it as one justification for two different reasoning paths based on the same data from the document.
- > 1 – the other cases.

G Unified format

We propose a unified format for storing information in the Document Understanding domain and deliver converted datasets as part of the released benchmark. It assumes three interconnected levels: dataset, document-annotation and document-content. Please refer to the repository for examples and formal specifications of the schemes.

Dataset. The dataset level is intended for storing the general metadata, e.g., name, version, license, and source. Here, the JSON-LD format based on the well-known schema.org web standard is used.⁹

Document. The documents annotation level is intended to store annotations available for individual documents within datasets and related metadata (e.g., external identifiers). Our format, valid for all the Document Understanding tasks, is specified using the JSON-Schema standard. This ensures that every record is well-documented and makes automatic validation possible. Additionally, to make the processing of large datasets efficient, we provide JSON Lines file for each split, thus it is possible to read one record at a time.

Content. As part of the original annotation or additional data we provide is related to document content (e.g., the output of a particular OCR engine), we introduce the document’s content level. Similarly to the document level, we propose an adequate JSON Schema and provide the JSON Lines files in addition. PDF files with the source document accompany dataset -, document-, and content-level annotations. If the source PDF was not available, a lossless conversion was performed.

H Evaluation protocol

Evaluation protocol. All the benchmark submissions are expected to conform to the following rules to guarantee fair comparison, reproducibility, and transparency:

- All results should be automatically obtainable starting from either raw PDF documents or the JSON files we provide. In particular, it is not permitted to rely on the potentially available source file that our PDFs were generated from or in-house manual annotation.
- Despite the fact that we provide an output of various OCR mechanisms wherever applicable, it is allowed to use software from outside the list. In such cases, participants are highly encouraged to donate OCR results to the community, and we declare to host them along with other variants. It is expected to provide detailed information on used software and its version.
- Any dataset can be used for unsupervised pretraining. The use of supervised pretraining is limited to datasets where there is no risk of information leakage, e.g., one cannot train models on datasets constructed from Wikipedia tables unless it is guaranteed that the same data does not appear in WikiTableQuestions and TabFact.
- It is encouraged to use datasets already publicly available or to release data used for pretraining.
- Training performed on a development set is not allowed. We assume participants select the model to submit using training loss or validation score. We do not release test sets and keep them secret by introducing a daily limit of evaluations performed on the benchmark’s website.
- Although we allow submissions limited to one category, e.g., QA or KIE, complete evaluations of models that are able to comprehend all the tasks with one architecture are highly encouraged.
- Since different random initialization or data order can result in considerably higher scores, we require the bulk submission of at least three results with different random seeds.

⁹See <https://json-ld.org/> for information on the JSON-LD standard, and <https://developers.google.com/search/docs/data-types/dataset> for the description of adapted schema.

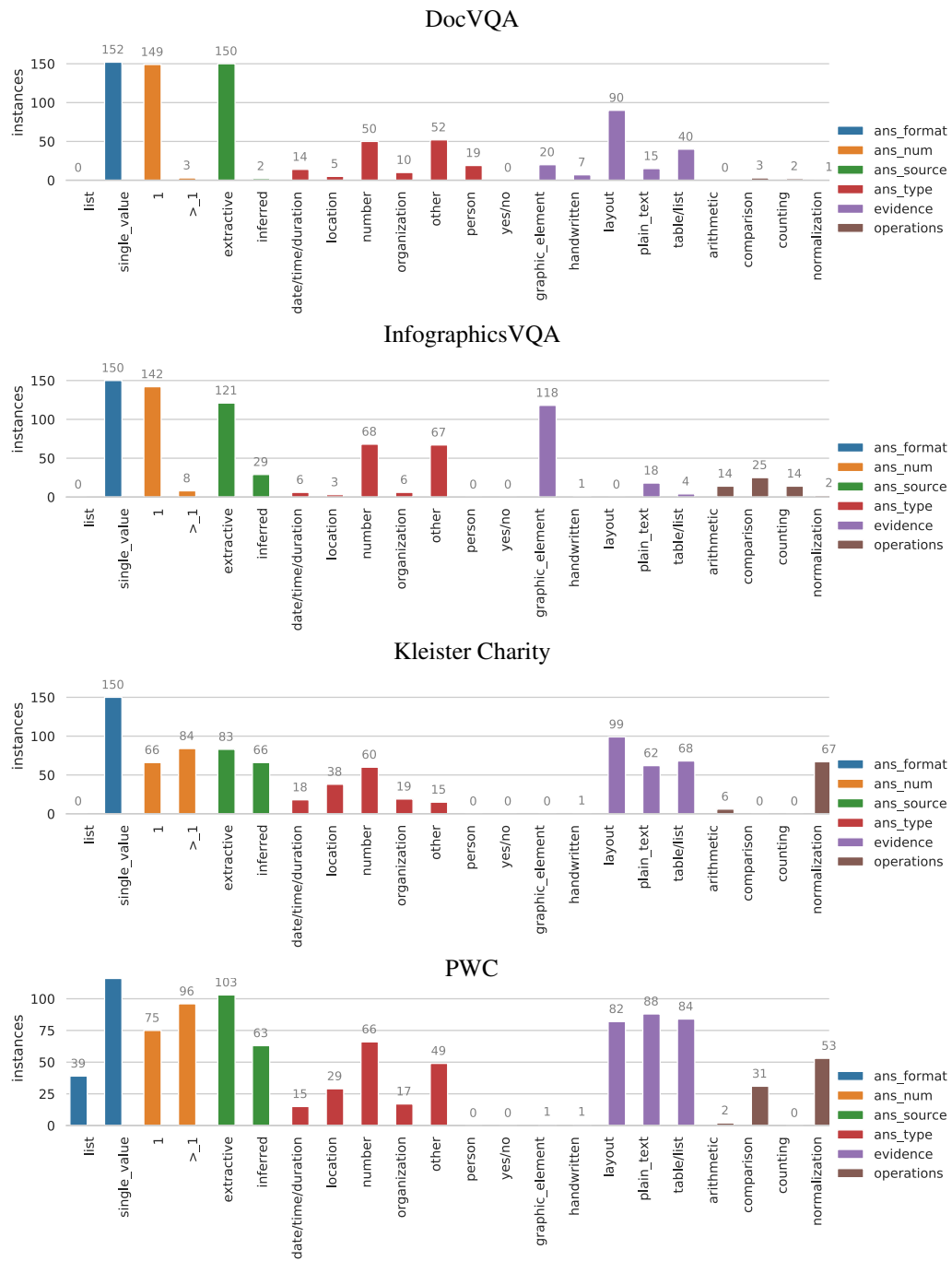


Figure 10: Number of annotated instances in each diagnostic subset category. DocVQA, InfographicsVQA, Kleister Charity, and PWC considered separately.

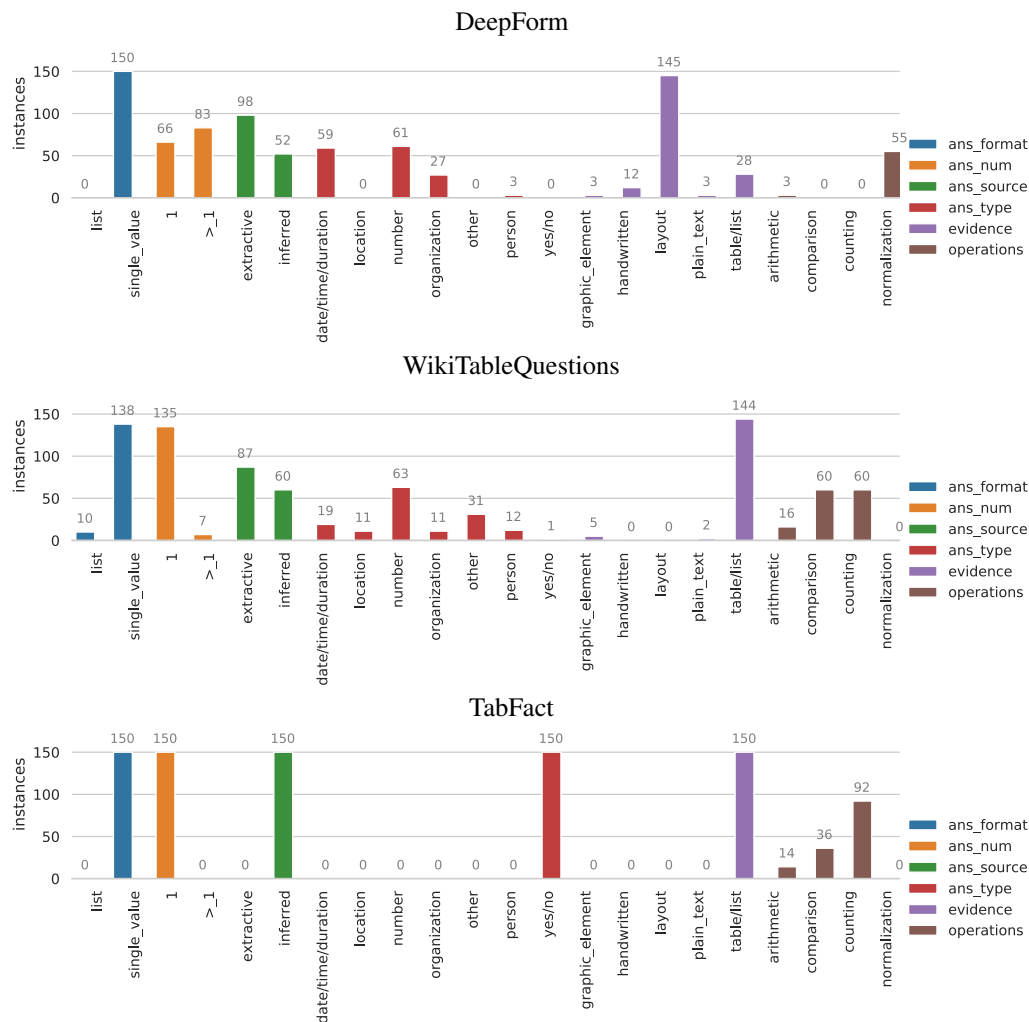


Figure 11: Number of annotated instances in each diagnostic subset category. DeepForm, WikiTableQuestions, and TabFact considered separately.

- Every submission is required to have an accompanying description. It is recommended to include the link to the source code.

I Experiments — training details

The experiments were carried out in an environment with NVIDIA A100-40G cards, PyTorch version 1.8.1, and the *transformers* library in version 4.2.2.

The parameters were selected through empirical experiments with T5-Base model on DocVQA and InfographicsVQA collections. The T5-Large model was used as the basis for finetuning.

The training lasted up to 30 epochs at batch 64 in training, the default optimizer AdamW (lr = $2e-4$), and warmup set to 100 updates. Validation was performed five times per epoch, and when no improvement was seen for 20 validation steps (4 epochs), the training was stopped. The length of the input documents has been truncated to 6144 tokens for all datasets (only Kleister Charity and PWC benefited from that change, for the rest of them 1024 tokens is sufficient)¹⁰ and the responses to 256 tokens. Dropout was set to 0.15, gradient clipping to 1.0, and weight decay to $1e-5$.

¹⁰The hard limit of 6k tokens results from the memory limitation of the used GPU.

ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE)

Jordy Van Landeghem^{1,2}, Rubèn Tito⁵, Łukasz Borchmann³, Michał Pietruszka^{3,6},
Dawid Jurkiewicz^{3,7}, Rafał Powalski⁸, Paweł Józiać^{3,4}, Sanket Biswas⁵, Mickaël
Coustaty⁹, Tomasz Stanisławek^{3,4}

¹ KU Leuven

² Contract.fit jordy@contract.fit

³ Snowflake tomasz.stanislawek@snowflake.com

⁴ Warsaw University of Technology

⁵ Computer Vision Center, Universitat Autònoma de Barcelona

⁶ Jagiellonian University

⁷ Adam Mickiewicz University

⁸ Instabase

⁹ University of La Rochelle

Abstract. This paper presents the results of the ICDAR 2023 competition on Document UnderstanDing of Everything. DUDE introduces a new dataset comprising 5K visually-rich documents (VRDs) with 40K questions with novelties related to types of questions, answers, and document layouts based on **multi-industry**, **multi-domain**, and **multi-page** VRDs of various origins and dates. The competition was structured as a single task with a multi-phased evaluation protocol that assesses the few-shot capabilities of models by testing generalization to previously unseen questions and domains, a condition essential to business use cases prevailing in the field. A new and independent diagnostic test set is additionally constructed for fine-grained performance analysis. A thorough analysis of results from different participant methods is presented. Under the newly studied settings, current state-of-the-art models show a significant performance gap, even when improving visual evidence and handling multi-page documents. We conclude that the DUDE dataset proposed in this competition will be an essential, long-standing benchmark to further explore for achieving improved generalization and adaptation under low-resource fine-tuning, as desired in the real world.

1 Introduction

Document UnderstanDing of Everything (DUDE) is a concept rooted in both machine learning and philosophy, seeking to *expand* the boundaries of document AI systems by creating highly challenging datasets that encompass a diverse range of topics, disciplines, and complexities. Inspired by the philosophical ‘Theory of Everything’, which aims to provide a comprehensive explanation of the nature of reality, DUDE endeavors to stimulate the development of AI models that can effectively comprehend, analyze, and respond to *any* question on *any* complex document.

Incorporating philosophical perspectives into DUDE enriches the approach by engaging with fundamental questions about knowledge, understanding, and the nature of

documents. By addressing these dimensions, researchers can develop AI systems that not only exhibit advanced problem-solving skills but also demonstrate a deeper understanding of the context, nuances, and implications of the information they process.

Over the past few years, the field of Document Analysis and Recognition (DAR) has embraced multi-modality with contributions from both Natural Language Processing (NLP) and Computer Vision (CV). This has given rise to Document Understanding (DU) as the all-encompassing solution [1,24,10] for handling Visually Rich Documents (VRDs), where layout and visual information is decisive in understanding a document.

This umbrella term subsumes multiple subtasks ranging from key-value information extraction (KIE) [12,29], document layout analysis (DLA) [36], visual question answering (VQA) [33,21], table recognition [13,25], and so on. For each of these subtasks, influential challenges have been proposed, e.g., the ICDAR 2019 Scene Text VQA [2,3] and ICDAR 2021 Document VQA (DocVQA) [22,33] challenges, which in turn have generated novel ideas that have impacted the new wave of architectures that are currently transforming the DAR field.

Nevertheless, we argue that the DAR community must encompass the future challenges (multi-domain, multi-task, multi-page, low-resource settings) that naturally juxtapose the previous competitions with pragmatic feedback attained via its business-driven applications.

Challenge objectives. We aim to support the emergence of models with strong multi-domain layout reasoning abilities by adopting a diversified setting where multiple document types with different properties are present (Figure 2). Moreover, a low-resource setting (number of samples) is assumed for every domain provided, which formulated as a DocVQA competition allows us to measure progress with regard to the desired generalization (Section 2). Additionally, we strive for the development of confidence estimation methods that can not only improve predictive performance but also adjust the calibration of model outputs, leading to more practical and reliable DU solutions. We believe that DUDE’s emphasis on task adaptation and the capability of handling a wide range of document types, layouts, and complexities will encourage researchers to push the boundaries of current DU techniques, fostering innovation in areas such as multi-modal learning, transfer learning, and zero-shot generalization.

Challenge contributions. DUDE answers the call for measuring improvements closer to the real-world applicability of DU models. By design of the dataset and competition, participants were forced to make novel contributions in order to make a significant impact on the DU task. Competitors showcased intriguing model extensions, such as combining models that learn strong document representations with the strengths of recent large language or vision-language models (ChatGPT [4] and BLIP2 [17,18]) to better understand questions and extract information from a document context more effectively. HiVT5 + modules extended Hi-VT5 [32] with token/object embeddings for various DU subtasks, while MMT5 employed a two-stage pre-training process and multiple objectives to enhance performance. These innovative extensions highlight the ingenuity in addressing the complex challenges of document understanding.

2 Motivation and Scope

We posit that progress in DU is determined not only by the improvements in each of its related predecessor fields (CV, NLP) but even more by the factors connecting to document intelligence, as explicitly understood in business settings. To improve the real-world applicability of DU models, one must consider (i) the availability and variety of types of documents in a dataset, as well as (ii) the problem-framing methods.

Currently, publicly available datasets avoid **multi-page** documents, are not concerned with **multi-task** settings, nor provide **multi-domain** documents of sufficiently different types. These limitations hinder real-world DU systems, given the ever-increasing number of document types occurring in various business scenarios. This problem is often bypassed by building systems based on private datasets, which leads to a situation where datasets cannot be shared, documents of interest are not covered in benchmarks, and published methods cannot be compared objectively. DUDE counters these limitations by explicitly incorporating a large variety of multi-page documents and document types (see e.g., [Figure 2](#)). Furthermore, the adaptability of DU to the real world is slowed down by a low-resource setting, since only a limited number of training examples can be provided, involving unpleasant manual labor, and subsequently costly model development. Anytime a new dataset is produced in the scientific or commercial context, a new model must be specifically designed and trained on it to achieve satisfactory performance. At the same time, transfer learning is the most promising solution for rapid model improvements, while zero- and few-shot performance still needs to be addressed in evaluation benchmarks.

Bearing in mind the characteristics outlined above, we formulated the DUDE dataset as an instance of *DocVQA* to evaluate how well current solutions can simultaneously handle the complexity and variety of real-world documents and all subtasks that can be expected. Optimally, a DU model should understand layout in a way that allows for zero-shot performance through attaining "desired generalization", i.e., generalization to *any documents* (e.g., drawn from previously unseen distributions of layouts, domains, and types) and *any questions* (e.g., regarding document elements, their properties, and compositions). Therefore, we incorporated these criteria while designing our dataset, which may stand as a common starting point and a cooperative path toward progress in this emerging area.

Desired Generalization. The challenge presented by DUDE is an instance of a Multi-Domain Long-Tailed Recognition (*MDLT*) problem [34].

Definition 1 (Multi-Domain Long-Tailed Recognition). *MDLT focuses on learning from multi-domain imbalanced data whilst addressing label imbalance, divergent label distributions across domains, and potential train-test domain shift. This framework naturally motivates targeting estimators that generalize to all domain-label pairs.*

A domain $D = \{(x_i, y_i)\}_{i=1}^N$ is composed of data sampled from a distribution P_{XY} , where \mathcal{X} denotes an input space (documents) and \mathcal{Y} the output space (QA pairs). Each $x \in \mathcal{X}$ represents a document, forming a tuple of (v, l, t) , expressing a complex composition of visual, layout and textual elements. For simplicity, consider that each

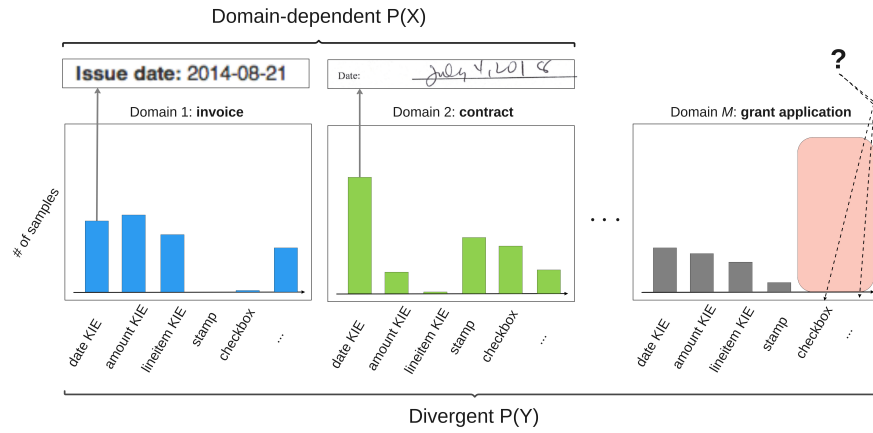


Fig. 1: Illustration of MDLT as applicable to the DUDE problem setting. The y-axis aggregates skills related to specific KIE or reasoning tasks over document elements (checkbox, signature, logo, footnote, ...). The x-axis denotes the obtained samples (QA pairs) per task. Each domain has a different label distribution $P(Y)$, typically relating to within-domain document properties $P(X)$. This training data exhibits label distribution shifts across domains, often requiring zero-shot generalization (marked red).

‘label’ $y \in \mathcal{Y}$ represents a question-answer pair, relating to implicit tasks to be completed (such as date KIE in *What is the document date?*). Due to the potentially compositional nature of QA, the label distribution is evidently *long-tailed*. During training, we are given MM domains (*document types*) on which we expect a solution to generalize (Figure 1), both within (different number of samples for each unique task) and across domains (even without examples of a task in a given domain).

What sets apart domains is any difference in their joint distributions $P_{XY}^j \neq P_{XY}^k$. For example, an invoice is less similar (in terms of language use, visual appearance, and layout) to a contract than to a receipt or credit note. Yet, a credit note naturally contains a stamp stating information such as “invoice paid”, whereas receipts rarely contain stamps. This might require a system to transfer ‘stamp detection’ learned within another domain, say on notary deeds.

Notably, it will be ‘organic’ to obtain more examples of certain questions (*tasks*) in a given domain. This should also encourage models to learn a certain skill in the domains where they have more training examples. Put plainly, it is better to learn checkbox detection on contracts than on invoices, which rarely contain any. This MDLT framework allows us to create a lasting, challenging benchmark that can be easily extended in the future with more tasks (formulated as QA pairs) and domains (relating to document types). In the first iteration of the DUDE competition, we have targeted specific skills by guiding annotators with focused instructions, which we share for future extensions.

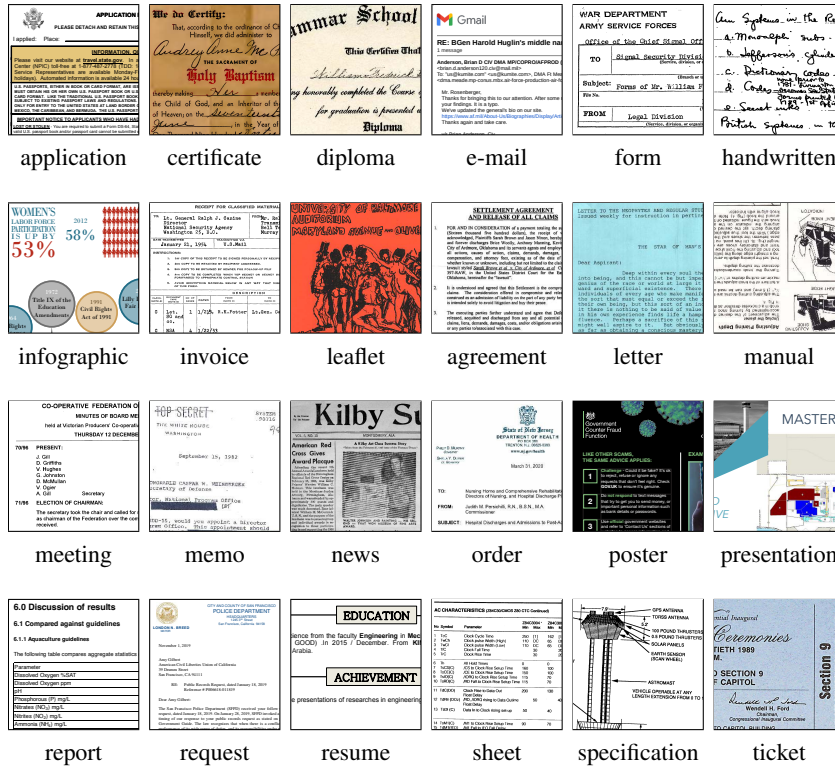


Fig. 2: Excerpts from DUDE documents (one visualized per type). Note that it is not an exhaustive list of document types collected.

3 DUDE Dataset

As part of the ICDAR 2023 DUDE competition, the authors constructed a novel dataset from scratch. A separate publication [16] describes the dataset in more detail, together with how it is different from related VQA datasets with an analysis of baseline methods. As part of the report, we will summarize the most important statistics and provide more insight into how the dataset and diagnostic subset were annotated and controlled for data quality.

The DUDE dataset is diverse, covering a wide range of document types (± 200), sources, dates (1900-2023), and industries (± 15). It contains documents with varying layouts and font styles, targeting diverse questions that require comprehension beyond document content. It includes abstractive and extractive questions, covering various answer types like textual, numerical, dates, yes/no, lists, or 'no answer'.

Annotation Process To create the dataset, diverse documents were manually collected from websites such as [Archive](#), [Wikimedia Commons](#), and [DocumentCloud](#). The se-

lection ensured that the documents were visually distinct and free from controversial content, privacy, or legal concerns. A total of 5,000 multi-page English documents were gathered.

The annotation process involved in-house annotators and Amazon Mechanical Turk freelancers. The process consisted of four stages: generating candidate QA pairs, verifying QA pairs, selecting the best answers, and an optional review by Qualified Linguists for test set annotations. The total cost of annotation was estimated at \$20,000.

Our multi-stage annotation process started with freelancers and in-house annotators proposing QA pairs, which were semi-automatically filtered for length, non-typical character combinations, and type-specific criteria. This was followed by the stage in which freelancers answered the accepted questions. Cases with an inter-answer agreement (ANLS) above 0.8 were added to the final dataset; otherwise, they were directed to further investigation. This stage employed freelancers with the highest historic quality score, who evaluated document, question, and answer variants, making corrections when necessary. Outliers were assessed by Qualified Linguists and corrected if needed (see Van Landeghem et al. [16] for the detailed description).

Future Extensions. To extend the dataset, one could follow the document collection and annotation process outlined in the original description [16]. This involves manually gathering diverse documents from various sources, ensuring they meet the dataset’s criteria, and then following the multi-phase annotation process to generate and verify new QA pairs.

4 DUDE Competition Protocol

The ICDAR 2023 competition on Document UnderstanDing of Everything took place from February to May of 2023. A *training-validation* set with 30k QA annotations on 3.7k documents was given to participants at the beginning of February. The 11.4k questions on 12.1k documents for the *test set* were only made accessible for a window between March and May. Participants were asked to submit results obtained on the public, blind test set documents rather than deliver model executables, although they were encouraged to open-source their implementations. We relied on the scientific integrity of the participants to adhere to the competition’s guidelines specified on The Robust Reading Competition (RRC) portal¹⁰.

Task Formulation. Given an input consisting of a PDF with multiple pages and a natural language question, the objective is to provide a natural language answer together with an assessment of the answer confidence (a float value scaled between 0 and 1). Each unique document is annotated with multiple questions of different types, including extractive, abstractive, list, and non-answerable. Annotated QA pairs are not restricted to the answer being explicitly present in the document. Instead, any question on aspect, form, or visual/layout appearance relative to the document under review is allowed.

¹⁰ <https://rrc.cvc.uab.es/?ch=23>

Additionally, competitors were allowed to submit results for only a specific answer type (provided in annotations) such that, for example for extractive questions, encoder-only architectures could compete in DUDE. Another important subtask is to obtain a *calibrated* and *selective* DocVQA system, which lowers answer confidence when unsure about its answers and does not hallucinate in case of non-answerable questions. Regardless of the number of answers (zero in the case of non-answerable or multiple in list-questions), we expect a single confidence estimate for the whole answer to guarantee consistency in calibration evaluation. To promote fair competition, we provided for each document three OCR versions obtained from one open-source (Tesseract) and two commercial engines (Azure, AWS).

Evaluation Protocol. The first evaluation phase assumes only independently and identically distributed (iid) data containing a similar mixture of document and question-answer types for the train-validation-test splits. To support scoring all possible answer types, the evaluation metric is the Average Normalized Levenshtein Similarity (ANLS) metric, modified for non-answerable questions (0/1 loss) and made invariant to the order of provided answers for list answers (ANLSL [31]). To assess the calibration and ranking of answer confidence, we applied two metrics, Expected Calibration Error (ECE) [23,8] (ℓ_2 norm, equal-mass binning with 100 bins) and Area-Under-Risk-Coverage-Curve (AURC) [7,15,11], respectively.

The (implicit) second evaluation phase created a mixture of seen and unseen domain test data. This was launched jointly with the first evaluation phase, as otherwise, one would be able to already detect the novel unseen domain test samples. To score how gracefully a system deals with unseen domain data, the evaluation metric is AUROC [19], which roughly corresponds to the probability that a positive example (in-domain) is assigned a higher detection score than a negative example (out-of-domain). A system is expected to either lower its confidence or abstain from giving an answer.

There is a strict difference between a non-answerable question and an unseen domain question. For the former, the document is from a domain that was included during training, yet the question cannot be solved with the document content, e.g., asking about who signed the document without any signatures present. For the latter, the question is apt for the document content, yet the document is from a domain that was not included during training and validation, which we would expect the system to pick up on.

For an in-depth explanation of these metrics and design choices, we refer the reader to [16, Appendix B.4.]. All metric implementations and evaluation scripts are made available as a standalone repository to allow participants to evaluate close to official blind test evaluations¹¹.

All submitted predictions are automatically evaluated, and the competition site provides ranking tables and visualization tools newly adapted to PDF inputs to examine the results. After the formal competition period, it will serve as an open archive of results. The main competition winner will be decided based on the aggregate high scores for ANLS, AURC, and AUROC.

To ensure proper validation and interpretability of competitor method results, we have created a diagnostic hold-out test set, where each instance is expert-annotated with

¹¹ <https://github.com/Jordy-VL/DUDEeval>

specific metadata (QA type, document category, expected answer form and type, visual evidence) or operations (counting, normalization, arithmetic) required to answer). Furthermore, we sourced an independent human expert baseline on this diagnostic subset (see [16, Section 3.4]) to further perform a ceiling analysis on the submitted methods.

5 Results and Analysis

Together with the creation of the DUDE dataset, we did a preliminary study with some reference baseline methods [16]. These will not be covered in the competition report, unless relevant for comparison or analysis.

Submitted Methods. Overall, 6 methods from 3 different participants were submitted for the proposed tasks in the DUDE competition. To avoid cherry-picking from considering all submissions of individual participants, we consider only the last submission (accentuated) for the final ranking. All the methods followed an encoder-decoder architecture, which is a standard choice for VQA when abstractive questions are involved. Specifically, the submitted methods are mostly based on T5-base [26] as the decoder. For this reason, we include the *T5-base* baseline to compare how the participant methods improved on it. A short description of each method can be found in Table 1.

Two very recent state-of-the-art architectures, UDOP and HiVT5, have been extensively leveraged by participants. The former is geared toward improved document page representations, while the latter targets multi-page document representations. In their method reports, the UDOP-based models by LENOVO RESEARCH mention calculating confidence by multiplying the maximum softmax score of decoded output tokens with two additional post-processing rules: a) predicted not-answerable questions confidence is set to 1, b) when abstaining, confidence is set to 0.

Performance Analysis. Table 2 reports the competition results ranking comparing the submitted methods’ performance on the test set. Higher ANLS and AUROC values indicate better performance, while lower ECE and AURC values signify improved calibration and confidence ranking. According to the findings, the UDOP+BLIP2+GPT approach attains the highest ANLS score (50.02), achieving the best calibration and OOD (out-of-distribution) detection performance. In a direct comparison of the MMT5 and HiVT5+modules methods, the former shows a higher ANLS score, yet did not provide any confidence estimates.

Thus, the overall winner is UDOP+BLIP2+GPT by LENOVO RESEARCH. Their submitted methods (ranked by highest ANLS) also differentiate themselves by their additional attention to confidence estimation. Based on the numbers in the table, several interesting observations can be made to support the suggested future directions and propose additional experiments:

- **ANLS.** The integration of UDOP, BLIP2, and ChatGPT contributes to the method’s superior overall performance in answering different question types.

Method	Description
<i>T5-base</i> (ours)	T5-base [26] fine-tuned on DUDE (AWS OCR), with a delimiter combining list answers into a single string, and replacing not-answerable questions with 'none'.
LENOVO RESEARCH	
UDOP(M)	Ensemble (M=10) of UDOP [30] (794M each) models without self-supervised pre-training, only fine-tuned in two stages: 1) SP-DocVQA [33] and MP-DocVQA [32], and 2) DUDE (switching between Azure and AWS OCR).
UDOP +BLIP2	UDOP(M=1) with integrated BLIP2 [17] predictions to optimize the image encoder and additional page number features.
UDOP +BLIP2+GPT	UDOP(M=1) and BLIP2 visual encoder with ChatGPT to generate Python-like modular programs to decompose questions for improved predictions [9,6].
UPSTAGE AI	
MMT5	Multimodal T5 pre-trained in two stages: single-page (ScienceQA [28], VQAonBD2023 [27], HotpotQA [35], SP-DocVQA) with objectives (masked language modeling (MLM) and next sentence prediction (NSP)), multi-page (MP-DocVQA and DUDE) with three objectives (MLM, NSP, page order matching). Fine-tuning on DUDE with answers per page combined for final output.
INFRD.AI	
HiVT5	Hi-VT5 [32] with 20 <PAGE> tokens pre-trained with private document collection (<i>no information provided</i>) using span masking objective [14]. Fine-tuned with MP-DocVQA and DUDE.
HiVT5 +mod-ules	Hi-VT5 extended with token/object embeddings for a variety of modular document understanding subtasks (detection: table structure, signatures, logo, stamp, checkbox; KIE: generic named entities; classification: font style).

Table 1: Short descriptions of the methods participating to the DUDE competition, in order of submission. The last submitted method is considered for the final ranking.

- **ECE, AURC.** Integrating UDOP, BLIP2 visual encoder, and ChatGPT for question decomposition contributes to the method’s performance in handling uncertainty across various question types.
- **Abstractive.** The top performance of UDOP+BLIP2+GPT in abstractive questions reveals the potential of combining the UDOP ensemble, BLIP2 visual encoder, and ChatGPT to enable abstract reasoning and synthesis of information beyond simple extraction.
- **List.** The performance of UDOP+BLIP2+GPT in list-based questions suggests that incorporating page number features can enhance the model’s capability to process and generate list information, which might be spread across pages.

Figure 3 visualizes an overview of the performance of each submitted method respective to diagnostic subset samples matching a certain diagnostic category. The models generally struggle with operations involving *counting*, *arithmetic*, *normalization*, and *comparisons*. As expected, models have higher performance when dealing with

<i>Method</i>	Answer		Calibration		OOD Detection		ANLS / answer type			
	ANLS \uparrow	ECE \downarrow	AURC \downarrow	AUROC \uparrow	<i>Ex</i>	<i>Abs</i>	<i>Li</i>	<i>NA</i>		
UDOP+BLIP+GPT	50.02	22.40	42.10	87.44	51.86	48.32	28.22	62.04		
MMT5	37.90	59.31	59.31	50.00	41.55	40.24	20.21	34.67		
HiVT5+modules	35.59	28.03	46.03	51.24	30.95	35.15	11.76	52.50		

Table 2: Summary of Method performance on the DUDE test set. Average ANLS results per question/answer type are abbreviated as (Abs)tractive, (Ex)tractive, (N)ot-(A)nswerable, (Li)st. (*) All scalars are scaled between 0 and 100 for readability.

simpler questions (*complexity simple*) compared to more complex questions (*complexity multi-hop*, *complexity other hard*, and *complexity meta*). Models tend to perform better when handling evidence in the form of plain text (*evidence plain*) compared to other forms of evidence, such as visual charts, maps, or signatures. Performance across models is notably lower for tasks involving lists compared to other question types. Models show varying performance when dealing with different types of forms (e.g., *date*, *numeric*, *other*, *proper*).

Figure 5 studies the ability of the competitors’ methods to answer questions respective to increasingly longer documents. We observe a significant drop in ANLS when aggregating scores over gradually longer documents. This is expected as the longer the document is, the more probable that the answer will either be located on a later page or rely on a long-range dependency between the tokens (e.g., a multi-hop question). Strikingly, all methods’ scores, except Hi-VT5+modules, drop significantly for questions on 2-page documents. This is likely to have the root cause in the standard input size of T5-based methods equal to 512 tokens, covering roughly 1 page.

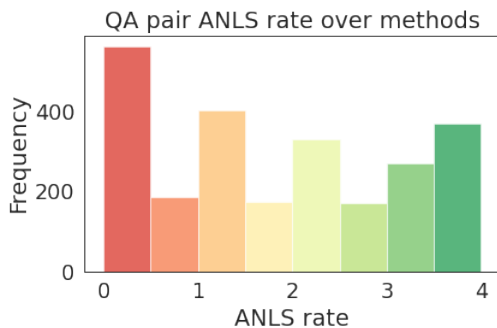


Fig. 4: A histogram (bins=8, matching ANLS-threshold of 0.5) of the average ANLS rate per QA pair when summing ANLS scores over competitor methods.

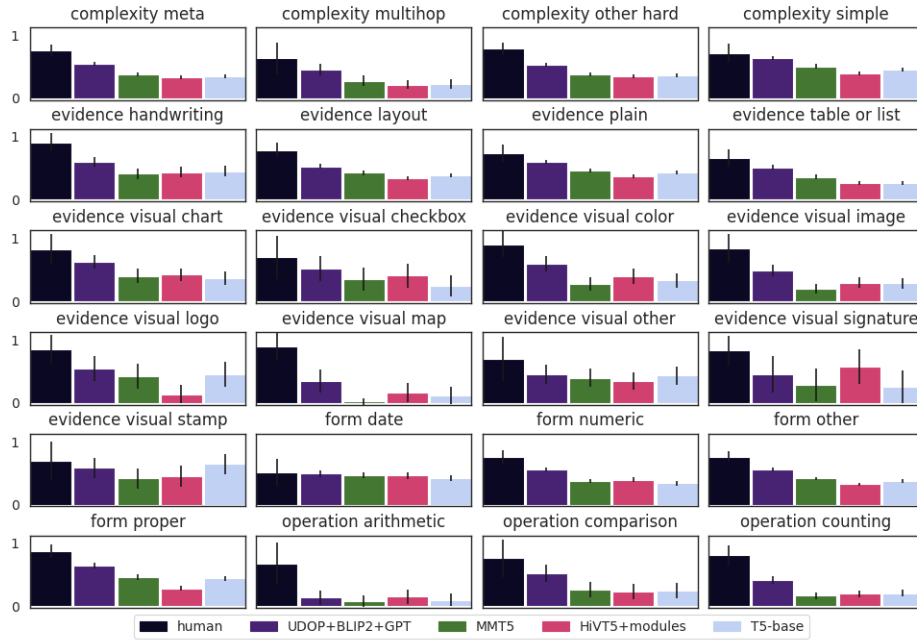


Fig. 3: We report the average ANLS per diagnostic category for each of the submitted methods vs. **human** and a baseline method **T5-base**. Since the diagnostic dataset contains a different number of samples per diagnostic category, we added error bars representing 95% confidence intervals. This helps visually determine statistically significant differences.

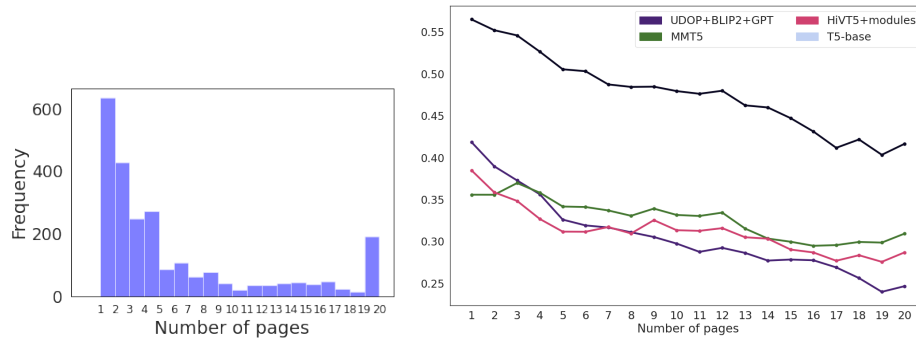


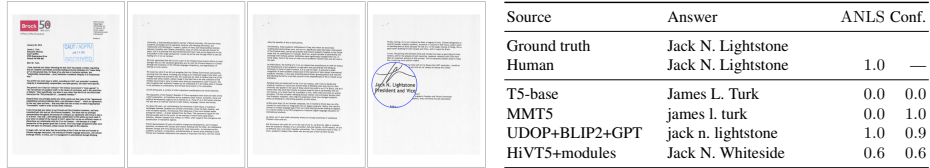
Fig. 5: Left: A histogram over the number of questions relative to the number of pages in the document (limited to 20 pages). Right: A line plot of the average ANLS score per QA pair – documents of length *at least* (x-axis) pages.

Figure 4 analyzes the correlation of errors over competitor methods. A large portion of QA pairs is predicted completely wrong (ANLS-rate = 0) by all competitor methods. This can have many plausible causes: a) by all sharing a similar decoder (T5), methods suffer from similar deficiencies, b) some QA pairs are too complex for current state-of-the-art competitor methods, particularly questions requiring more complex reasoning or unique document-specific layout processing. To further analyze this phenomenon, we will sample qualitative examples with different ANLS rates.

5.1 Qualitative Examples

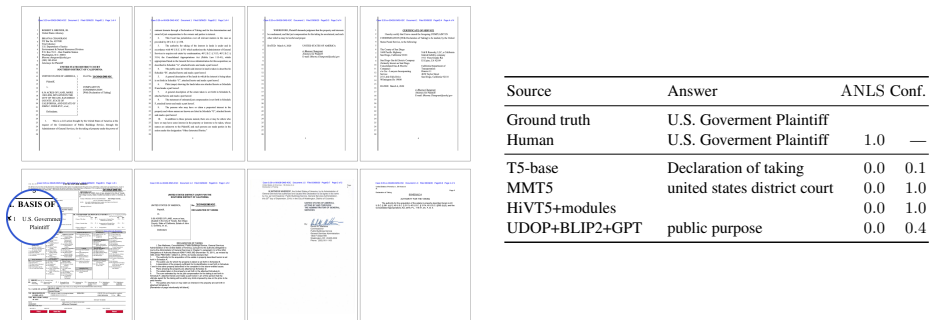
We provide some interesting, hand-picked test set examples with predictions from the submitted competition methods.

Low complexity. *Who is the president and vice-chancellor?* Despite the question’s relatively straightforward nature, some systems struggle with providing the appropriate answer. One can hypothesize it is the result of limited context (the answer is located at the end of the document), i.e., models either hallucinate a value or provide a name found earlier within the document.



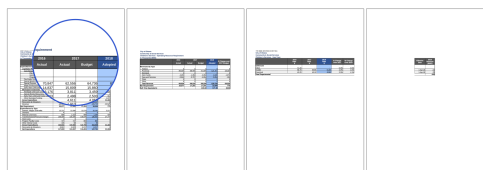
Source	Answer	ANLS Conf.	
Ground truth	Jack N. Lightstone		
Human	Jack N. Lightstone	1.0	—
T5-base	James L. Turk	0.0	0.0
MMT5	james l. turk	0.0	1.0
UDOP+BLIP2+GPT	jack n. lightstone	1.0	0.9
HiVT5+modules	Jack N. Whiteside	0.6	0.6

Requires graphical comprehension. *Which is the basis for jurisdiction?* To provide a valid answer, the model needs to comprehend the meaning of the form field and recognize the selected checkbox. None of the participating systems was able to spot the answer correctly.



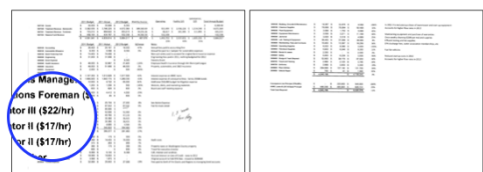
Source	Answer	ANLS Conf.	
Ground truth	U.S. Government Plaintiff		
Human	U.S. Government Plaintiff	1.0	—
T5-base	Declaration of taking	0.0	0.1
MMT5	united states district court	0.0	1.0
HiVT5+modules		0.0	1.0
UDOP+BLIP2+GPT	public purpose	0.0	0.4

Requires comparison. *In which year does the Net Requirement exceed 25,000?* The question requires comprehending a multi-page table and spotting if any values fulfill the posed condition. Some of the models resort to plausible answers (one of the three dates that the document covers), whereas others correctly decide there is no value exceeding the provided amount.




Source	Answer	ANLS Conf.	
Ground truth	[Unanswerable]		
Human	[Unanswerable]	1.0	—
T5-base	[Unanswerable]	1.0	0.2
MMT5	2018	0.0	1.0
UDOP+BLIP2+GPT	[Unanswerable]	1.0	1.0
HiVT5+modules	2017	0.0	0.8

Requires arithmetic. *What is the difference between how much Operator II and Operator III make per hour?* The question requires table comprehension, determining relevant values, and dividing extracted integers. None of the participating models was able to fulfill this requirement.



Source	Answer	ANLS Conf.	
Ground truth	\$5		
Human	\$5	1.0	—
T5-base	\$0.00	0.0	0.0
MMT5	65%	0.0	1.0
UDOP+BLIP2+GPT	-1.5 mile	0.0	0.0
HiVT5+modules	\$5,700.00	0.0	0.4

Requires counting and list output. *What are the first two behavioral and intellectual disabilities of people with FASDs?* It seems most of the models correctly recognized that this type of question requires a list answer but either failed to comprehend the question or provided a list with incorrect length (incomplete or with too many values).



Source	Answer	ANLS Conf.	
Ground truth	Learning disabilities Hyperactivity		
Human	learning disabilities	0.5	—
T5-base	Early embryo brain development External Genitals	0.0	0.0
MMT5	heart beats difficulty with attention lung function hyperactivity problem with judgment speech and language delays	0.2	1.0
UDOP+BLIP2+GPT	hyperactivity speech and language delays	0.5	0.2
HiVT5+modules	HIV/AIDS	0.0	0.6

6 Conclusion and Future Work

As a core contribution of DUDE, we wanted to emphasize the importance of evaluation beyond mere predictive performance. DUDE offers an interesting and varied test bed for the evaluation of novel calibration and selective QA approaches (e.g., [5,20]). While this was not explicitly attempted in this iteration of the competition, we hope that future work will consider testing their methods against DUDE.

Future of the Shared Task. As the competition evolves, we hope that DUDE will serve as an essential platform for pushing the frontiers of research and driving innovation in the DU field. Currently, our competition focuses on English language documents, which means we miss out on the potential of incorporating *multilingual* data. An ideal extension for future iterations of the shared task would be to introduce multilingualism, which our framework can accommodate, provided that source documents are readily available. However, this would also require specifying language qualifications for annotation experts. Moreover, one could automate part of the data collection process and annotation process by allowing the best-performing competition system to validate the aptitude and complexity of human-proposed QA pairs.

Acknowledgements

Jordy Van Landeghem acknowledges the financial support of VLAIO (Flemish Innovation & Entrepreneurship) through the Baekeland Ph.D. mandate (HBC.2019.2604). The Smart Growth Operational Programme partially supported this research under projects no. POIR.01.01.01-00-1624/20 (*Hiper-OCR - an innovative solution for information extraction from scanned documents*) and POIR.01.01.01-00-0605/19 (*Disruptive adoption of Neural Language Modelling for automation of text-intensive work*).

References

1. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 993–1003 (2021) [2](#)
2. Biten, A.F., Tito, R., Mafra, A., Gomez, L., Rusinol, M., Mathew, M., Jawahar, C., Valveny, E., Karatzas, D.: ICDAR 2019 competition on scene text visual question answering. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1563–1570. IEEE (2019) [2](#)
3. Biten, A.F., Tito, R., Mafra, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision (2019) [2](#)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020) [2](#)
5. Dhuliawala, S., Adolphs, L., Das, R., Sachan, M.: Calibration of machine reading systems at scale. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1682–1693. Association for Computational Linguistics, Dublin, Ireland (May 2022). <https://doi.org/10.18653/v1/2022.findings-acl.133>, <https://aclanthology.org/2022.findings-acl.133> [14](#)

6. D'Ádac, S., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. arXiv preprint arXiv:2303.08128 (2023) [9](#)
7. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. *Advances in neural information processing systems* **30** (2017) [7](#)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. pp. 1321–1330. ICML'17 (2017) [7](#)
9. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. arXiv preprint arXiv:2211.11559 (2022) [9](#)
10. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking. p. 4083–4091. *MM '22*, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3503161.3548112>, <https://doi.org/10.1145/3503161.3548112> [2](#)
11. Jaeger, P.F., Lüth, C.T., Klein, L., Bungert, T.J.: A call to reflect on evaluation practices for failure detection in image classification. In: *International Conference on Learning Representations (2023)*, <https://openreview.net/forum?id=YnkGMih0gvX> [7](#)
12. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. vol. 2, pp. 1–6. IEEE (2019) [2](#)
13. Jimeno Yepes, A., Zhong, P., Burdick, D.: ICDAR 2021 competition on scientific literature parsing. In: *Document Analysis and Recognition—ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV* 16. pp. 605–617. Springer (2021) [2](#)
14. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* **8**, 64–77 (2020) [9](#)
15. Kamath, A., Jia, R., Liang, P.: Selective question answering under domain shift. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 5684–5696 (2020) [7](#)
16. Landeghem, J., Tito, R., Borchmann, L., Pietruszka, M., Józiak, P., Powalski, R., Jurkiewicz, D., Coustaty, M., Ackaert, B., Valveny, E., Blaschko, M., Moens, S., Stanisławek, T.: Document Understanding Dataset and Evaluation (DUDE). arXiv (2023), <https://arxiv.org/abs/2305.08455> [5](#), [6](#), [7](#), [8](#)
17. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) [2](#), [9](#)
18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International Conference on Machine Learning*. pp. 12888–12900. PMLR (2022) [2](#)
19. Liang, S., Li, Y., Srikant, R.: Enhancing the Reliability of Out-of-distribution Image Detection in Neural Networks. In: *International Conference on Learning Representations (2018)*, <https://openreview.net/forum?id=H1VGkIxRZ> [7](#)
20. Lin, S., Hilton, J., Evans, O.: Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research* (2022), <https://openreview.net/forum?id=8s8K2UZGTZ> [14](#)
21. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: InfographicVQA. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 1697–1706 (2022) [2](#)
22. Mathew, M., Tito, R., Karatzas, D., Manmatha, R., Jawahar, C.: Document visual question answering challenge 2020. arXiv preprint arXiv:2008.08899 (2020) [2](#)

23. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015) [7](#)
24. Powalski, R., Łukasz Borchmann, Jurkiewicz, D., Dwojak, T., Pietruszka, M., Pałka, G.: Going full-tilt boogie on document understanding with text-image-layout transformer. In: ICDAR (2021) [2](#)
25. Qiao, L., Li, Z., Cheng, Z., Zhang, P., Pu, S., Niu, Y., Ren, W., Tan, W., Wu, F.: LGPMA: Complicated table structure recognition with local and global pyramid mask alignment. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I. pp. 99–114. Springer (2021) [2](#)
26. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**(140), 1–67 (2020) [8](#), [9](#)
27. Raja, S., Mondal, A., Jawahar, C.: ICDAR 2023 competition on visual question answering on business document images (02 2023) [9](#)
28. Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., Bhattacharyya, P.: ScienceQA: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries* **23**(3), 289–301 (2022) [9](#)
29. Stanislawek, T., Gralinski, F., Wróblewska, A., Lipinski, D., Kaliska, A., Rosalska, P., Topolski, B., Biecek, P.: Kleister: Key information extraction datasets involving long documents with complex layouts. In: ICDAR. Lecture Notes in Computer Science, vol. 12821, pp. 564–579. Springer (2021). https://doi.org/10.1007/978-3-030-86549-8_36 [2](#)
30. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., Bansal, M.: Unifying vision, text, and layout for universal document processing. arXiv preprint arXiv:2212.02623 (2022) [9](#)
31. Tito, R., Karatzas, D., Valveny, E.: Document collection visual question answering. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16. pp. 778–792. Springer (2021) [7](#)
32. Tito, R., Karatzas, D., Valveny, E.: Hierarchical multimodal transformers for multi-page DocVQA. arXiv preprint arXiv:2212.05935 (2022) [2](#), [9](#)
33. Tito, R., Mathew, M., Jawahar, C., Valveny, E., Karatzas, D.: ICDAR 2021 competition on document visual question answering. In: International Conference on Document Analysis and Recognition. pp. 635–649. Springer (2021) [2](#), [9](#)
34. Yang, Y., Wang, H., Katabi, D.: On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX. p. 57–75. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20044-1_4, https://doi.org/10.1007/978-3-031-20044-1_4 [3](#)
35. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380. Association for Computational Linguistics, Brussels, Belgium (Oct–Nov 2018). <https://doi.org/10.18653/v1/D18-1259>, <https://aclanthology.org/D18-1259> [9](#)
36. Zhong, X., Tang, J., Yepes, A.J.: PubLayNet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1015–1022. IEEE (2019) [2](#)

Appendices

Appendix A

Competitions and Projects

A.1 Competitions

SemEval 2020 Task 11

Place (Span Identification task): Second

Place (Technique Classification task): First

Team name: ApplicaAI

Team: Dawid Jurkiewicz*, Łukasz Borchmann*, Izabela Kosmala and Filip Graliński

Leaderboard: <https://propaganda.qcri.org/semEval2020-task11/leaderboard.php>

Presentation

Type: Oral presentation

Date: 12.12.2020

Presenters: Dawid Jurkiewicz, Łukasz Borchmann

Venue: International Workshop on Semantic Evaluation (SemEval) collocated with International Conference on Computational Linguistics (COLING)

ICDAR 2021 Infographics VQA

Place: First

Team name: Applica.ai TILT

Team: Dawid Jurkiewicz, Rafał Powalski, Gabriela Pałka, Łukasz Borchmann, Tomasz Dwojak and Michał Pietruszka

Leaderboard: <https://rrc.cvc.uab.es/?ch=17&com=evaluation&task=3>

Presentation

Type: Oral presentation

Date: 06.09.2021

Presenters: Dawid Jurkiewicz

Venue: ICDAR 2021 Workshop on Document Visual Question Answering (DocVQA)

2021, 1st edition) collocated with International Conference on Document Analysis and Recognition (ICDAR)



Figure A-1: Certificate of winning the competition.

A.2 Projects

I took part in the following research projects:

1. Badania w zakresie przetwarzania języka naturalnego (Samsung Electronics and Adam Mickiewicz University project)
2. Robotyzacja procesów biznesowych opartych o tekst z wykorzystaniem metod sztucznej inteligencji i głębokich sieci neuronowych (POIR.01.01.01-00-0144/17)
3. Przełomowe wykorzystanie Neuronowego Modelowania Języka w celu automatyzacji

cji pracochłonnych zadań wymagających przetwarzania danych tekstowych (POIR.01.01.01-00-0605/19-00)

4. Uniwersalna platforma robotyzacji procesów wymagających rozumienia tekstu o unikalnym poziomie automatyzacji wdrożenia i obsługi (POIR.01.01.01-00-0877/19)
5. Hiper-OCR (POIR.01.01.01-00-1624/20)

Appendix B

Declarations of Contribution

Poznań, June 25, 2021

Declaration

I hereby declare that the contribution to the following manuscript:

Łukasz Borchmann, Dawid Wiśniewski, Andrzej Gretkowski, Izabela Kosmala, Dawid Jurkiewicz, Łukasz Szałkiewicz, Gabriela Pałka, Karol Kaczmarek, Agnieszka Kaliska, and Filip Graliński, *Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines*, Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.

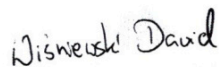
is correctly characterized in the table below.

Contributor	Description of main tasks
Łukasz Borchmann	Conceptualization and methodology, leading and running the experiments, writing the paper, implementation and evaluation of baselines, results' analysis.
Dawid Wiśniewski	Implementation of baselines, writing the paper.
Andrzej Gretkowski	Implementation of baselines, edition of the manuscript and improvements of its initial version.
Izabela Kosmala	Implementation of baselines, running the experiments.
Dawid Jurkiewicz	Implementation of baselines, evaluation of human performance.
Łukasz Szałkiewicz	Curation of human-annotation process, annotation of datasets.
Gabriela Pałka	Implementation of baselines.
Karol Kaczmarek	Implementation of baselines.
Agnieszka Kaliska	Annotation of datasets, writing, and edition of the manuscript.
Filip Graliński	Supervision external to the core team, evaluation methodology.

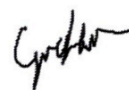
Łukasz Borchmann



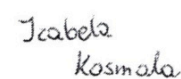
Dawid Wiśniewski



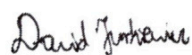
Andrzej Gretkowski



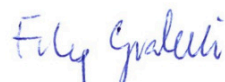
Izabela Kosmala



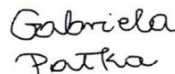
Dawid Jurkiewicz



Filip Graliński



Gabriela Pałka



Karol Kaczmarek



Declaration

I hereby declare that the contribution to the following manuscript:

Łukasz Borchmann, Dawid Jurkiewicz, Filip Graliński, and Tomasz Górecki. *Dynamic Boundary Time Warping for Sub-Sequence Matching with Few Examples*, Expert Systems with Applications 169, 2021.

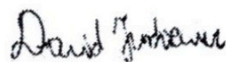
is correctly characterized in the table below (* denotes equal contribution).

Contributor	Description of main tasks
Łukasz Borchmann*	Conceptualization and methodology, design and implementation of the DBTW prototype and DBA baseline, implementation of LSTM-CRF baseline, writing the paper, performing experiments, analysis of the results.
Dawid Jurkiewicz*	Conceptualization and methodology, improvement of the DBTW prototype, performing experiments, writing the paper, analysis of the results, design and implementation of ACBOW model.
Filip Graliński	Supervision, review of the initial manuscript, evaluation methodology.
Tomasz Górecki	Supervision, review of the initial manuscript, description of related works, statistical analysis.

Łukasz Borchmann



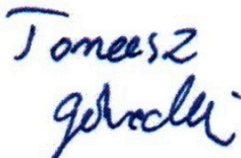
Dawid Jurkiewicz



Filip Graliński



Tomasz Górecki



Poznań, June 25, 2021

Declaration

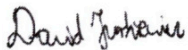
I hereby declare that the contribution to the following manuscript:

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński, *ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them*, Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020.

is correctly characterized in the table below (* denotes equal contribution).

Contributor	Description of main tasks
Dawid Jurkiewicz*	Conceptualization and methodology, performing experiments, writing the paper, implementation of model prototypes, analysis of the results, error analysis.
Łukasz Borchmann*	Conceptualization and methodology, writing the paper, implementation of RoBERTa-CRF model, performing experiments, design and implementation of Span CLS architecture, analysis of the results, error analysis.
Izabela Kosmala	Statistical analysis, implementation of baselines.
Filip Graliński	Supervision external to the core team, metric implementation.

Dawid Jurkiewicz



Łukasz Borchmann



Izabela Kosmala



Filip Graliński



Warsaw, June 25, 2021

Declaration

I hereby declare that the contribution to the following manuscript:

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka, *Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer*, Proceedings of the International Conference on Document Analysis and Recognition ICDAR 2021, 2021.

is correctly characterized in the table below (* and ^ denote groups of equal contribution).

Contributor	Description of main tasks
Rafał Powalski*	Conceptualization and methodology, design and implementation of model prototype, running experiments, writing the paper, design and implementation of case and spatial augmentation.
Łukasz Borchmann*	Conceptualization and methodology, implementation and experiments with model prototypes, running experiments with the final model, writing the paper, review and preparation of the datasets.
Dawid Jurkiewicz^	Running experiments, design and implementation of image token embeddings, review and preparation of the datasets, improvements of model prototype, editing the manuscript.
Tomasz Dwojak^	Running experiments, ablation of the pretraining strategies, editing the manuscript, hyperparameter search, review and preparation of the datasets.
Michał Pietruszka^	Writing the manuscript, running experiments, design and implementation of vision augmentation methods, review and preparation of the multimodal QA datasets.
Gabriela Pałka	Review and preparation of the datasets, running experiments, editing the manuscript.

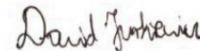
Rafał Powalski

(podpis)

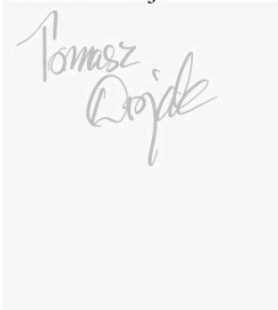

Łukasz Borchmann



Dawid Jurkiewicz



Tomasz Dwojak



Michał Pietruszka



Gabriela Pałka



Declaration

I hereby declare that the contribution to the following paper:
 Lukasz Borchmann, Michał Pietruszka, Tomasz Stanisławek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Galiński. “DUE: End-to-End Document Understanding Benchmark”. In: *Under review in NeurIPS 2021*. 2021. URL: <https://openreview.net/forum?id=rNs2FvJGDK>
 is correctly characterized in the table below (* denotes equal contributions).

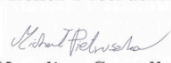
Contributor	Description of main tasks
Lukasz Borchmann*	<ul style="list-style-type: none"> - conceptualization and methodology (participated in regular discussions) - methodology of the considered datasets for DUE benchmark - implementation of baselines - create DUE benchmark webpage - create scripts for evaluation - convert documents from TabFact and WTQ datasets into pdf files - result analysis - writing the paper - organizing and controlling the process of human annotation
Michał Pietruszka*	<ul style="list-style-type: none"> - conceptualization and methodology (participated in regular discussions) - methodology and preparation list of the considered datasets for DUE benchmark - implementation of baselines - preparation of datasets (DocVQA, InfographicsVQA, WikiTableQuestions, PWC) - preparing code, models and datasets for final release - result analysis - writing the paper - organizing and controlling the process of human annotation
Tomasz Stanisławek*	<ul style="list-style-type: none"> - conceptualization and methodology (participated in regular discussions) - methodology and preparation list of the considered datasets for DUE benchmark - prepare schema for storing benchmark datasets in unified data format - preparation of datasets (Kleister Charity, DeepForm, TabFact) - curation of PWC and DeepForm datasets - methodology for creation of the diagnostic subsets - result analysis - improved the first version of the paper / edition of the manuscript - organizing and controlling the process of human annotation
Dawid Jurkiewicz	<ul style="list-style-type: none"> - participated in regular discussions - implementation of baselines - significantly improved results of the baselines (hyper-param search for it) - performing experiments - preparing code and models for final release - edition of the paper
Michał Turski	<ul style="list-style-type: none"> - methodology and preparation of the diagnostic subsets - organizing and controlling the process of human annotation - controlling the process of measuring human performance where it was required (PWC, DeepForm, WTQ) - edition of the paper
Karolina Szyndler	<ul style="list-style-type: none"> - improving schema for storing benchmark datasets in unified data format - code for reading datasets by the baselines
Filip Galiński	<ul style="list-style-type: none"> - participated in regular discussions - edition of the paper

Lukasz Borchmann



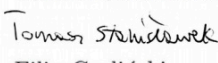
Michał Turski

Michał Pietruszka




Karolina Szyndler

Tomasz Stanisławek



Filip Galiński

Dawid Jurkiewicz



Declaration

I hereby declare that the contribution to the following paper:
 Lukasz Borchmann, Michał Pietruszka, Tomasz Stanisławek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. "DUE: End-to-End Document Understanding Benchmark". In: *Under review in NeurIPS 2021*. 2021. URL: <https://openreview.net/forum?id=rHs2FvJGDK>
 is correctly characterized in the table below (* denotes equal contributions).

Contributor	Description of main tasks
Lukasz Borchmann*	<ul style="list-style-type: none"> - conceptualization and methodology (participated in regular discussions) - methodology of the considered datasets for DUE benchmark - implementation of baselines - create DUE benchmark webpage - create scripts for evaluation - convert documents from TabFact and WTQ datasets into pdf files - result analysis - writing the paper - organizing and controlling the process of human annotation
Michał Pietruszka*	<ul style="list-style-type: none"> - conceptualization and methodology (participated in regular discussions) - methodology and preparation list of the considered datasets for DUE benchmark - implementation of baselines - preparation of datasets (DocVQA, InfographicsVQA, WikiTableQuestions, PWC) - preparing code, models and datasets for final release - result analysis - writing the paper - organizing and controlling the process of human annotation
Tomasz Stanisławek*	<ul style="list-style-type: none"> - conceptualization and methodology (participated in regular discussions) - methodology and preparation list of the considered datasets for DUE benchmark - prepare schema for storing benchmark datasets in unified data format - preparation of datasets (Kleister Charity, DeepForm, TabFact) - curation of PWC and DeepForm datasets - methodology for creation of the diagnostic subsets - result analysis - improved the first version of the paper / edition of the manuscript - organizing and controlling the process of human annotation
Dawid Jurkiewicz	<ul style="list-style-type: none"> - participated in regular discussions - implementation of baselines - significantly improved results of the baselines (hyper-param search for it) - performing experiments - preparing code and models for final release - edition of the paper
Michał Turski	<ul style="list-style-type: none"> - methodology and preparation of the diagnostic subsets - organizing and controlling the process of human annotation - controlling the process of measuring human performance where it was required (PWC, DeepForm, WTQ) - edition of the paper
Karolina Szyndler	<ul style="list-style-type: none"> - improving schema for storing benchmark datasets in unified data format - code for reading datasets by the baselines
Filip Graliński	<ul style="list-style-type: none"> - participated in regular discussions - edition of the paper

Lukasz Borchmann

Michał Pietruszka

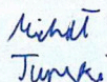
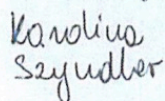
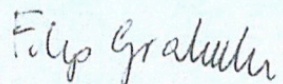
Tomasz Stanisławek

Dawid Jurkiewicz

Michał Turski

Karolina Szyndler

Filip Graliński

Declaration






I hereby declare that the contribution to the following manuscript:

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józia, Sanket Biswas, Mickaël Coustaty, Tomasz Stanisławek, *ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE)*, Proceedings of the International Conference on Document Analysis and Recognition ICDAR 2023, 2023.

is correctly characterized in the table below.

Contributor	Description of main tasks
Jordy Van Landeghem	<p>Conceptualization and methodology Preliminary experiments (Donut, LayoutLM) Manual collection of documents (insurance, legal) * Responsible for finding and preparing DocumentCloud and RVL-CDIP-N/O as potential sources of data; backtracing PDF versions * Designed and implemented equal-spaced grid representation for multi-page PDF annotations with bounding box annotations * Designed annotation schemes for different QA types and annotation phases * Managed 3 AWS accounts with Mechanical Turk annotation processes, including uploading and reviewing batches of HITs. (together with Tomasz) * Master dataset quality control and publication (Zenodo and HuggingFace Hub)</p> <p>Writing the competition proposal Writing the competition report Provided first draft for all Sections followed by complete write-out Revising the competition report (full rework of Table 1 and Section 5 Results)</p> <p>Conceptualization of MDLT framework Public data loading implementation on HuggingFace * Set-up of competition website and evaluation framework, including calibration and confidence ranking metric implementations (ECE, AURC) Communications with competition participants (Q&A) * Evaluation, analysis and compilation of competition submitted results Financial administration Project management as first author</p>
Rubèn Tito	<p>Conceptualization and methodology</p> <p>Manual collection of documents (health) Implemented first version to create equal-spaced grid representation for multi-page PDF annotations with bounding box annotations. Implemented first version of AMT annotation framework.</p> <p>Implemented base framework for baselines. - Baseline experiments (LayoutLMX, T5-based methods) - ANLS and accuracy metrics implementation</p> <p>Manager and setup of the the DUDE Challenge in the RRC portal.</p>

	Writing the competition ICDAR Proposal and Report.
Lukasz Borchmann	Conceptualization and methodology Review and minor role in writing the competition report, figures preparation Writing the competition proposal Qualitative analysis of considered models and human references Proposition and overview of the diagnostic annotation process Analysis of documents similarity Manual collection of documents (transportation, education)
Michał Pietruszka	Conceptualization and methodology Manual collection of documents (personal, real estates) Improving implementation (T5, T5 + 2D), experimenting with improving the baselines and achieving state-of-the-art results. Participated in writing the proposal
Dawid Jurkiewicz	Preliminary experiments (T5, T5 + 2D, TILT) Manual collection of documents (agriculture, manufacturing category) Data scraper for manually chosen documents Backtracking archive.org licenses Reviewing and assisting in the paper and proposal
Rafał Powalski	Experiments related to generative models testing. Testing few-shot prompt variation in order to optimize model performance. Manual collection of documents. Gather datasets statistics
Paweł Józiać	Prepared annotation normalization tool. Participated in manual collection of the documents. Prepared dataset statistics. Participated in analysis of candidate solutions performance. Contributed to document rewrite.
Sanket Biswas	Manual collection of documents, Setting up dataloader, Setting up official github competition repo, Public visibility of competition in social media, Reviewing and assisting in the report and proposal
Mickaël Coustaty	Reviewing of the paper
Tomasz Stanisławek	Conceptualization and methodology Writing the competition proposal Reviewing of the competition report Manual collection of documents (finance, business) Responsible for finding and preparing wikimedia as potential sources of data; Designed annotation schemes for different QA types and annotation phases Managed 3 AWS accounts with Mechanical Turk annotation processes, including uploading and reviewing batches of HITs. (together with Jordy)

Jordy Van Landeghem	Rubèn Tito	Lukasz Borchmann	Michał Pietruszka	Dawid Jurkiewicz
				
Rafał Powalski	Paweł Józiać	Sanket Biswas	Mickaël Coustaty	Tomasz Stanisławek
