DLUE: Benchmarking Document Language Understanding

Ruoxi Xu[1,2], **Hongyu Lin**[2]*, **Xinyan Guan**[1,2], **Yingfei Sun**[1]*, **Le Sun**[2,3]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Chinese Information Processing Laboratory [3]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
{ruoxi2021,hongyu,guanxinyan2022,sunle}@iscas.ac.cn
yfsun@ucas.ac.cn

Abstract

Understanding documents is central to many real-world tasks but remains a challenging topic. Unfortunately, there is no well-established consensus on how to comprehensively evaluate document understanding abilities, which significantly hinders the fair comparison and measuring the progress of the field. To benchmark document understanding researches, this paper summarizes four representative abilities, i.e., document classification, document structural analysis, document information extraction, and document transcription. Under the new evaluation framework, we propose **Document Language Understanding Evaluation – DLUE**, a new task suite which covers a wide-range of tasks in various forms, domains and document genres. We also systematically evaluate six well-established transformer models and representative LLMs on DLUE, and find that due to the lengthy content, complicated underlying structure and dispersed knowledge, document understanding is still far from being solved in complex real-world scenarios.

1 Introduction

Documents are basic units of the organization of natural language (Buckland, 1997). Understanding the structures and the semantics of documents is the foundation for understanding news articles (Kiesel et al., 2019), scientific papers (Dasigi et al., 2021), government reports (Huang et al., 2021b) , stories (Kočiskỳ et al., 2018), etc. Evaluating how a machine intelligence system can read, analyze and generate documents is an important part of evaluating its natural language abilities, which has long been a critical direction in NLP field.

While standard benchmarks like GLUE (Shanahan et al., 2016) and SuperGLUE (Wang et al., 2019) have become a critical part of NLP community, they primarily focused on short utterances like sentences or paragraphs. However, documents are much more than bag-of-sentences, which usually focus on the central theme (Benamara et al., 2017), with underlying structure bound by complex linguistic elements (Parsing, 2009) and dependent knowledge dispersed across the whole text (Huang et al., 2021a). Therefore, these benchmarks cannot be used to evaluate document understanding due to its unique challenges: *First, documents usually have lengthy content*, i.e., they usually are much longer than sentences/paragraphs thus it is difficult to process them due to the computational memory/runtime limitation of current NN models. *Second, documents have underlying structures*, which play a critical role during understanding. For example, mining arguments from a document needs to model both the local coherence between sentences and the global interactions between claims and arguments, which cannot be accomplished by only exploiting sentence- or paragraph-level information. *Third, the knowledge in a document is usually dispersed beyond sentences/paragraphs*, which makes it necessary to model and explore document-level context. For example, long-distance within-document coreference resolution needs to integrate global information to find all expressions that refer to the same entity distributed in full text.

Recently, an increasing number of researches have paid attention to the evaluation of document understanding. Yi Tay et al. (2021) proposes the Long Range Arena (LRA), which contains two syn-

*Corresponding Authors

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1257-1269, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

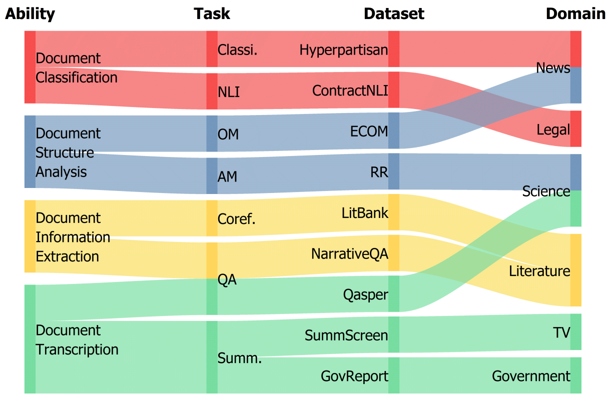


Figure 1: Overview of DLUE, which covers a wide-range of **tasks** and **datasets** in diverse **domains** to evaluate four representative document understanding **abilities**.

thetic text-related tasks and evaluates model quality of understanding multi-modal long contexts. Hudson and Moubayed (2022) proposes MuLD, which focuses on merged sequences over 10,000 tokens. SCROLLS (Shaham et al., 2022) is a benchmark that contains NLI, QA and summarization tasks and focuses on long language sequences, with ZeroSCROLLS (Shaham et al., 2023) as its extension for large language models. However, these benchmarks mainly focus on the lengthy content challenge, while ignoring other important challenges. As a result, almost all tasks in these benchmarks can be resolved via an retrieval-answering paradigm, i.e., retrieving a very limited number of sentences that contains critical information and then resolving the task. Furthermore, these benchmarks only cover limited tasks, which makes them unable to thoroughly evaluate the document understanding abilities of models.

To systematically evaluate document language understanding abilities, this paper proposes **Document Language Understanding Evaluation – DLUE**, a new task suite which covers a wide-range of tasks in various forms, different domains and document genres. Figure 1 shows the overview of DLUE. Specifically, we summarize 4 kinds of document understanding abilities, including 1) Document Classification, which evaluates whether a model can understand the overall semantics of a document, e.g., its topic (Zhang et al., 2015) and standpoint (Kiesel et al., 2019); 2) Document Structural Analysis, which evaluates whether a model can analyze and leverage the underlying structure of a document, e.g., its discourse structure (Zeldes, 2017) and argument structure (Cheng et al., 2020); 3) Document Information Extraction, which evaluates whether a model can recognize and aggregate associated information spanning cross whole document, e.g., long-distance within-document coreference (Bamman et al., 2020); 4) Document Transcription, which evaluates whether a model can capture and transcript important information of a document, e.g., summarization (Huang et al., 2021b; Chen et al., 2022) and abstractive QA (Dasigi et al., 2021). Then we collect 9 datasets and align them with the above abilities. Datasets of the same group are converted to a unified format. In this way, DLUE provides a comprehensive benchmark for document language understanding, which enables the research community to fairly compare and measure the progress of this field.

To better understand the challenges of document language understanding and analyze the performance of current approaches, we experiment on DLUE using several advanced document understanding models, including 1) Memory-based approaches, which includes XLNet (Yang et al., 2019); 2) Pattern-based approaches, which includes Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), Sparse transformer (Child et al., 2019); 3) Low-rank/Kernel-based approaches, which includes Linformer (Wang et al., 2020) and Performer (Choromanski et al., 2020); 4) Large language models, which includes GPT-4, GPT-3.5-turbo[0], Vicuna-7B (Chiang et al., 2023), LLaMA-2-7B-chat (Touvron et al., 2023), Mistral-7B-instruct (Jiang et al., 2023) and ChatGLM3 (Du et al., 2022). Results show that document understanding is still far from being solved due to lengthy content, complex structure and dispersed knowledge, and currently there is no neural architecture that dominates all tasks, raising needs for a universal document

[0] https://openai.com/blog/openai-api

understanding architecture.

Generally, the contributions[1] of this paper are:

- We summarize 4 representative abilities for lengthy, structural and global document understanding, including document classification, document structural analysis, document information extraction and document transcription.
- We propose a comprehensive benchmark for document language understanding – DLUE, which is built on established annotated datasets and selected to cover a diverse range of text genres, dataset sizes, and degrees of difficulty.
- We evaluate current state-of-the-art document understanding models on DLUE, which provides a novel reference to assess current models' abilities and properties when they handle different kinds of document understanding tasks.

2 Background

NLP Benchmarks The development of natural language understanding (NLU) evaluation benchmarks has helped drive the progress of pretraining and transfer learning in NLP. Benchmarks proposed in the early stage mostly aim at general tasks, such as SentEval (Conneau and Kiela, 2018) for universal sentence representations, DecaNLP (McCann et al., 2018) for ten diversified NLP tasks cast as a general question-answering format, GLUE (Wang et al., 2018) for NLU in English and SuperGLUE (Wang et al., 2019) as a harder counterpart of GLUE. Besides, benchmarks for more specific tasks have also been proposed, such as DialoGLUE (Mehri et al., 2020) for task-oriented dialogue, DiscoEval (Chen et al., 2019) for discourse-aware sentence representations, GLUECoS (Khanuja et al., 2020) for code-switched NLP, KILT (Petroni et al., 2020) for knowledge-intensive language tasks, and etc.

Above benchmarks mostly focus on sentences or short texts. However, documents are also very common for complex tasks or real-world textual data. Single-task benchmarks mostly use summarization (Cohan et al., 2018a) or QA tasks (Dasigi et al., 2021), but struggle to fully assess document modeling due to the single nature of the task and data distribution. There are also some multi-task benchmarks for document understanding, such as the Long Range Arena (LRA) (Tay et al., 2020a), SCROLLS (Shaham et al., 2022), ZeroSCROLLS (Shaham et al., 2023), MuLD (Hudson and Moubayed, 2022) and LOT (Guan et al., 2022). Most long inputs of LRA and MuLD are either automatically generated or artificially lengthened. Tasks in SCROLLS also tend to focus on a few sentences or paragraphs, which can be solved by retrieval-based or chunk-based approaches. LOT only focuses on Chinese text. In this paper, distinct from existing benchmarks focusing on long sequences instead of documents, we focus on challenges posed by document understanding, including lengthy content, complicated underlying structure and dispersed knowledge.

Document Understanding Models There have been numerous attempts to improve both memory footprint and computational cost of transformers, thus allowing use of longer inputs. A natural way is to connect blocks via recurrence, such as XLNet (Yang et al., 2019). Another way of tackling the high complexity of full attention is to sparsify attention matrix. Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020) and Sparse transformer (Child et al., 2019) simply sparsify the attention matrix by limiting the field of view to fixed, predefined patterns such as local windows and block patterns of fixed strides. Reformer (Kitaev et al., 2020) uses learnable ones, an extension to fixed, pre-determined pattern. Besides, low-rank approximations or kernelization of the self-attention matrix can be used as well to reduce the complexity. Linformer (Wang et al., 2020) and Performer (Choromanski et al., 2020) are representative low-rank/kernel-based transformers. In this paper, we conduct experiments on the above three document understanding architectures to explore challenges posed by document understanding.

3 DLUE: Document Language Understanding Evaluation Benchmark

The section describes the DLUE benchmark, used to evaluate 4 representative document understanding abilities. Specifically, DLUE is centered on 9 English document understanding datasets, which cover

[1] https://github.com/RossiXu/DLUEvaluation.git

Corpus	Task	Domain	Metric	Avg #Words		#Examples
				Input	Output	
Classification						
Hyperpartisan	Classifi.	News	acc.	588	1	1273
ContractNLI	NLI	Legal	acc.	1708	1	10319
Structure Analysis						
ECOM	OP	News	F_1	488	20	2000
RR	AM	Science	F_1	793	47	4764
Extraction						
LitBank	Coref.	Literature	F_1	2115	7.4	7214
NarrativeQA	QA	Literature	F_1	51790	4.6	71187
Transcription						
GovReport	Summ.	Government	ROUGE	7897	492.7	19402
SummScreen	Summ.	TV	ROUGE	5639	100.0	4348
Qasper	QA	Science	F_1	3671	11.5	5692

Table 1: Task descriptions and statistics of DLUE.

a wide-range of tasks in various forms, different domains and document genres. Table 1 provides an overview of the datasets included in the benchmark. In the following, we describe the details of DLUE.

3.1 Overview

As described above, a document understanding system should resolve the lengthy content, complicated underlying structure, and dispersed knowledge challenges. To effectively evaluate the above abilities and challenges, DLUE selects datasets based on these criteria: First, documents should have lengthy content. Second, tasks require leveraging dispersed document knowledge; hence, document-level datasets aren't preferred if most of them can be tackled through chunk- or retrieval-based methods. Third, documents should be natural, like literature, scientific articles, or government reports. Synthesized documents lack structure information and relation links among different sections. Fourth, tasks should exceed current system capabilities but remain solvable by most educated English speakers. Based on the above desiderata and with the permission of licenses, we collect as diverse datasets as possible to increase the coverage on capabilities. The overview of DLUE is shown in Figure 1, and their statistics are shown in Table 1. In the following, we describe all datasets according to the their target ability.

3.2 Document Classification

A document usually narrow focus on a single central theme (Benamara et al., 2017). We aim to evaluate document classification ability, specifically the ability to understand the overall semantics of documents in this section. To do this, we select two datasets that rely on full-text to make judgements and reformulate every dataset as document classification tasks. Specifically, given single sequence s or sequence pairs (s_1, s_2), the goal is to classify the input into a single label l.

Hyperpartisan (Kiesel et al., 2019) is a document classification dataset for classifying news based on left-wing or right-wing standpoints. A few words or sentences aren't sufficient to gauge the political leaning, which are toned by the full text. This task provides two datasets, one labeled manually and the other labeled semi-automatically via distant supervision at the publisher level. We use the first one to pursue higher evaluation accuracy and keep the same train/test split as the original work.

ContractNLI (Koreeda and Manning, 2021) is a natural language inference dataset in the legal domain, with non-disclosure agreements (NDAs) as premises and legal statements as hypothesizes. NDAs are collected from Internet search engines and Electronic Data Gathering, Analysis, and Retrieval system. To correctly predict whether the hypothesis is entailed, neutral, or contradictory from the contract, we need to refer to non-continuous sentences across the contract with hundreds of tokens. The dataset contains 607 contracts and 17 hypothesizes, which we combine to produce 10319 instances.

3.3 Document Structure Analysis

A document is composed of structured groups of sentences, paragraphs and sections. Analyzing document structure can be very useful in indexing and organizing the information contained in the document.

Tasks in this section aim to evaluate document structure analysis ability, specifically the ability to capture and leverage structure information. We select three datasets as follows and reformulate every dataset as sentence-level sequence labeling tasks. Specifically, given a document $d = \{s_1, s_2, ..., s_n\}$, the goal is to output a tag sequence $t = \{t_1, t_2, ..., t_n\}$ for sentences.

ECOM (Xu et al., 2022) is an event-centric opinion mining corpus in which a system takes in an event descriptor and related news articles to extract event-related opinion segments from articles. An opinion segment is composed of continuous sentences targeting at the same argument. We select the dataset to evaluate the ability to utilize local structure information, which is important for identifying opinion boundaries unambiguously.

RR (Cheng et al., 2020) is an argument mining corpus for extracting arguments from reviews and rebuttals, sourced from ICLR spanning 2013 to 2020 (except 2015). Peer reviews and rebuttals on scientific works are a data source of rich structures and long passages. It's considered suitable because the original paper's experiments highlight the significance of internal structure information for this task.

3.4 Document Information Extraction

Dependent knowledge in a document is usually dispersed across the full text, which plays an important role in the transmission of the intended meaning. Tasks in this section aim to evaluate document information extraction ability, specifically the ability to identify long-distance related mentions and relations. We select two datasets as follows and reformulate every dataset as multi-answer question answering tasks. Specifically, given a document d and a question q, the goal is to extract correct answer spans $a = \{a_1, a_2, ..., a_n\}$ from d for q.

LitBank (Bamman et al., 2020) is a coreference resolution dataset on English literature. Its documents are several times longer than those in other benchmarks (e.g. 463.7 tokens for OntoNotes), making them rich in long-distance within-document coreference. For each coreference link, we transform the sentence of one mention into a question, take all mentions as answers, resulting in 7214 question-answer pairs.

NarrativeQA (Kočiský et al., 2018) is a reading comprehension dataset on books and movie scripts. Its questions are written based on summaries, requiring readers to grasp long-distance associated information across several parts or a larger span of the context document for comprehension and answering.

3.5 Document Transcription

Tasks in this section aim to evaluate document transcription ability, specifically the ability to capture and transcript key information of documents. We select three datasets that need to contextualize across different sections and reformulate every dataset as sequence-to-sequence tasks. Specifically, given a sequence s, the goal is to output a concise and fluent new sequence s_N.

GOVREPORT (Huang et al., 2021b) is a summarization dataset of long reports on national policy issues and paired expert-written summaries. Its documents and summaries are significantly longer than prior datasets, such as 1.5 times longer than Arxiv (Cohan et al., 2018b). Moreover, new salient bigrams are steadily added as more content is consumed, suggesting information is spread throughout documents in the dataset.

SummScreen (Chen et al., 2022) is a summarization dataset of TV series transcripts paired with human written recaps. Unlike official documentation like GOVREPORT (Huang et al., 2021b), its language expression is informal and structure is more unclear. Whole documents need to be combined to understand plots often expressed indirectly in character dialogues and scattered throughout the transcript.

Qasper (Dasigi et al., 2021) is a QA dataset in the research domain focusing on entire papers, in which both questions and answers are handed-written by NLP practitioners. Over half of questions require multiple paragraphs as evidence to answer. We prepend the query to the document, using two newlines as a natural separator to construct the input sequence.

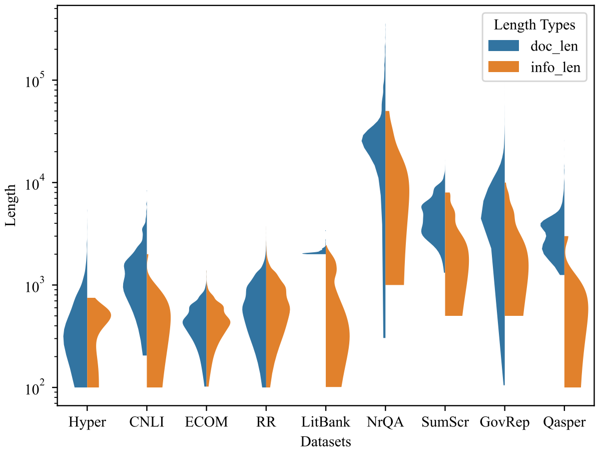


Figure 2: The distribution of document lengths and information spread in DLUE datasets. The spread of information measured by the standard deviation of the position of each bigram's first occurrence in the input document like (Shaham et al., 2022).

3.6 Data Analysis

Datasets in DLUE are analyzed, including documents' length, and distribution of information. In Figure 2, we can observe that the benchmark consist of documents with varying lengths and essential information required for document understanding is distributed throughout the entire text.

4 Experiments and Analysis

4.1 Benchmarking Architectures

This section describes models and architectures we evaluate on DLUE. Specifically, we compare three efficient transformer architectures following the general taxonomy (Tay et al., 2020b) including memory-based models (Yang et al., 2019), pattern-based models (Zaheer et al., 2020; Beltagy et al., 2020; Child et al., 2019), low-rank/Kernel-based models (Wang et al., 2020; Wang et al., 2020) and also large language models.

For all kinds of task formulations described in Section 3, we implement unified model architectures. For document classification tasks, we use the special classification token ([CLS]) for prediction. The document structure analysis tasks are reformulated into sentence-level sequence labeling tasks. We use the classical Transformer-CRF architecture as in named entity recognition (Devlin et al., 2019). The document information extraction tasks are reformulated into multi-span question answering tasks. Following Hu et al. (2019), we expand traditional MRC architecture by adding a span number prediction and search component. For transcription tasks, we use the basic encoder-decoder architecture (Vaswani et al., 2017).

4.2 Implementations

Our models are implemented by PyTorch framework[2][3]. For transformers with public pretrained models, we use the base version, including XLNet-base, Longformer-base, BigBird-base. The learning rate is 1e-5 for pretrained models and 1e-3 for classifier heads. For other models, we follow the same setup as Long range arena (Tay et al., 2020a), a widely recognized benchmark for efficient transformers to minimize the influence of hyper-parameter settings. These transformer models are parameterized by the same number of layers, heads and hidden dimensions, namely 6 layers, 8 heads, 512 hidden dimensions and d = 2048 for positional FFN layers. We use Adam with warmup. All models are trained for 10 epochs. Across datasets and models, we run three repetitions with different random seeds and report averaged scores.

[2] https://pytorch.org/

[3] In practice, these models often use a combination of the proposed approximate global attention and simple local attention.

Model	#param	pretrain	Classification		Structure Analysis		Extraction		Transcription			Avg	Inference Speed
			Hyper	CNLI	ECOM	RR	LitBank	NrQA	SummScr	GovRep	Qasper		(steps per sec)
			Vanilla Transformer										
BERT	110M	yes	80.1	72.3	37.3	57.3	-	14.5	-	-	-	-	-
			Memory-based										
XLNet	110M	yes	81.4	80.2	**39.1**	**74.0**	78.1	15.2	18.9	22.7	24.2	41.1	0.95
			Pattern-based										
Longformer	148M	yes	83.8	71.6	37.9	72.9	**79.1**	18.3	**20.9**	25.7	26.4	42.1	1.7
BigBird	127M	yes	**85.9**	82.8	37.0	71.1	77.8	18.2	20.6	27.3	26.2	**42.7**	1.6
Sparse Trans.	46M	no	64.6	67.7	21.9	45.6	56.7	11.1	21.4	17.6	17.6	30.5	3.8
			Low-rank/Kernel-based										
Linformer	45M	no	67.1	65.5	22.6	44.3	63.8	12.4	18.9	25.8	17.5	30.5	6.4
Performer	44M	no	67.9	69.5	18.6	48.6	51.6	10.1	20.1	15.6	21.5	30.9	6.7
			Large Language Models										
GPT-4	-	yes	79.4	**88.0**	6.9	7.5	74.3	5.1	18.0	**31.1**	**53.5**	40.4	-
GPT-3.5-turbo	-	yes	66.7	74.0	9.9	4.4	60.6	**21.8**	22.8	24.0	40.0	36.0	-
Vicuna-7b	7B	yes	46.0	41.5	10.0	0.5	28.2	1.7	11.4	21.3	11.2	19.1	-
Llama-2-7b-chat	7B	yes	40.9	41.5	7.9	1.6	24.9	2.3	11.9	17.3	6.7	17.2	-
Mistral-7b-instruct	7B	yes	41.7	42.9	10.8	6.5	46.1	2.6	8.5	25.2	8.0	21.4	-
ChatGLM3-6b	6B	yes	57.4	58.1	13.2	6.1	22.2	3.0	11.1	28.4	10.6	23.3	-

Table 2: Overall experimental results on DLUE. Best model is in boldface. ”-” denotes the model can’t handle this task. LLMs are tested in in a zero-shot setting to ensure a fair and unbiased comparison.

4.3 Overall Results

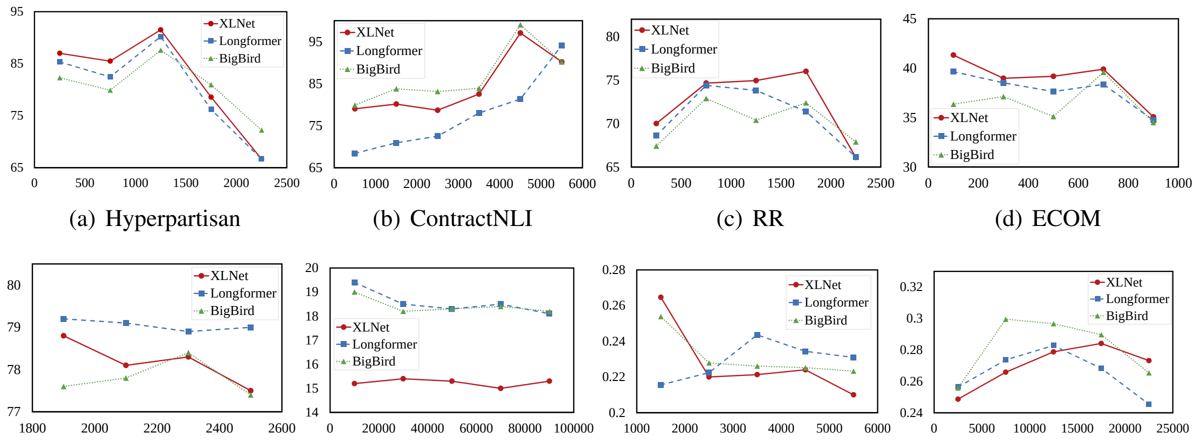
Table 2 shows the overall results on DLUE. From this table, we can see that:

1) Document understanding is far from being solved. From Table 2, we can see that the best benchmark system can only achieve 42.7 average score. While it’s difficult to establish an accurate human performance ceiling in DLUE, we can take some indicators to prove that the performance gap between human and models are huge. For example, human agreement on ECOM was measured at 80.8% F1 (Xu et al., 2022), much higher than our best baseline of 39.1% F1. Likewise, Dasigi et al. (2021) study a subset of Qasper that has multiple annotated answers, and find their overlap to be 60.9% F1, more than double our best baseline. This indicates that contemporary off-the-shelf models struggle with documents, challenging future work to make progress on DLUE.

2) Different tasks have different advantageous architectures, raising a need for an universal document understanding architecture which can dominate all tasks in one architecture. From Table 2, we can see that different model architectures seem to be good at processing different tasks. Specifically, the performance of XLNet ranks first on the structure analysis tasks, while Longformer and BigBird perform better on the other tasks. Linformer and Performer do well on document classification tasks. This shows that recurrence-based models may have advantages over hierarchically structured data and pattern-based models may be more effective on flat data. Contrary to the other tasks, fast low-rank/kernel-based models do better on document classification tasks. No architecture dominates all tasks, which indicates that more universal models are needed.

3) Lengthy content is the critical, but not the only, challenge for document understanding. From Table 2 and Table 1, we can see that models perform poorly with too long inputs, such as the 18.5 best F_1 score in NarrativeQA dataset with 51790 average input length. However, even for those structure analysis and extraction tasks where documents can be taken in completely by long-range transformer models, the model performances still fail to meet expectations. Obviously, there exist other challenges for document understanding apart from lengthy input, such as complex structures and dispersed knowledge.

4) It is critical to take global context into consideration. From Table 2, we can see that long-range transformers that can take in more contexts achieve a higher score than vanilla transformer in most datasets. This demonstrates longer contexts are necessary to understand documents. Document-level tasks can’t be solved in the same way as short-text tasks.



(a) Hyperpartisan (b) ContractNLI (c) RR (d) ECOM

(e) LitBank (f) NarrativeQA (g) SummScreen (h) GovReport

Figure 3: Performance (y axis) on DLUE datasets with different document lengths (x axis).

4.4 Computational Efficiency

The last column of Table 2 shows inference speeds of models. For a fair comparison, we use the standard test datasets of DLUE as testbed. Based on our implementation, the low-rank/kernel-based models are the fastest. The results are consistent with model complexity, which has a significant impact on inference speed. The low-rank/ kernel-based models decompose the $N \times N$ self-attention matrix to a lower-dimensional representation and thus usually have a $O(N)$ time complexity. Pattern-based models sparsify the attention matrix according to predefined or learnable patterns and the time complexity is usually between $O(N)$ and $O(N^2)$. Recurrence-based models connect multiple segments and blocks via recurrence and the representative XLNet has a $O(N^2)$ time complexity.

4.5 Effect Of Document Length

To investigate how the document length will impact the performance, we cluster the documents into buckets for each task according to their document lengths and run the evaluation on each bucket. The breakdown analysis is shown in Figure 3. We find that: 1) On the whole, understanding longer documents faces more challenges. We notice that the performances on most datasets decrease when document lengths increase, with ContractNLI dataset as an exception. This maybe because there exists label bias related to document lengths in ContractNLI datasets. We find that a longer contract tends to entail a hypothesis, with 34% probability for documents shorter than 1000 words and 76% probability for documents longer than 5000 words. 2) The performance of pattern-based models seems to be more stable when the document lengths increase. We can see that Longformer and BigBird obtain a greater advantage when documents get longer. We think there are two reasons. First, the global token mechanism in Longformer and BigBird could help models focus on important information and be less distracted by noise in long contexts. Second, the maximum input length of XLNet is smaller due to the segment-level recurrence mechanism. 3) The performance is relatively stable on datasets where document lengths far exceed input limits. Figure 3(f) shows performance on NarrativeQA dataset. When the document length exceeds 20,000 tokens, the result remains around 18 F_1 for Longformer, BigBird and 15 F_1 for XLNet. This indicates the ability of efficient transformers to understand long documents are limited.

4.6 Effect of dispersed knowledge Exploition

In this section, we aim to validate the importance of recognizing and aggregating dispersed knowledge for document understanding, with much room for current models to improve. We analyze from two perspectives: 1) effect of mention distance, which can be viewed as the measure of dispersion; 2) performance comparison between long-range transformers and short-text models without global information.

Effect of Mention Distance To quantify the impact of dispersed knowledge to document understanding, we analyze the performance of coreference resolution with different mention distances on LitBank.

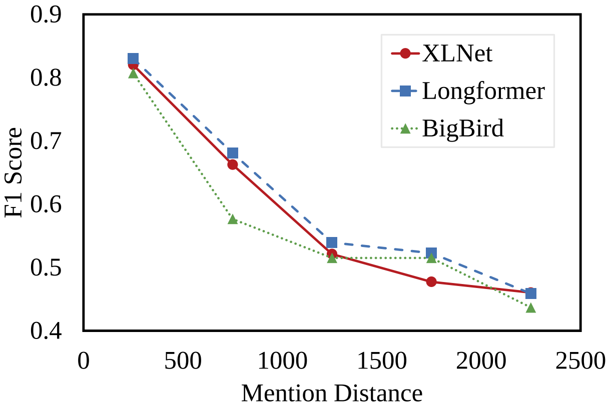


Figure 4: Performance on LitBank dataset with different mention distances. Mention distances can reflect the degree of knowledge dispersion in a document.

Model	Hyperpartisan	Qasper
XLNet	81.4	24.2
Longformer	83.8	**26.4**
BigBird	**85.9**	26.2
CogLTX	82.9	18.9
ToBERT	78.4	16.6

Table 3: Performance comparison between long-range transformers and short-text models.

In Figure 4, we can see performance of all models drops sharply with greater mention distances, indicating long-distance coreference is more challenging than within-sentence coreference. It's easy to understand because it puts forward higher requirements for ability to capture and aggregate dispersed information. We can also notice huge performance gap between short and long mention distances, implying there is still much room for further improvements in models' global information integration ability.

Comparison with Short-text Models To verify the importance of global information to document understanding, we compare the performance of long-range transformers with two existing short-text models, including CogLTX (Ding et al., 2020) and ToBERT (Pappagari et al., 2019). CogLTX jointly trains two BERT models to select key sentences from documents. ToBERT divides documents into smaller chunks and uses a transformer layer over BERT-based chunk representations. We select Hyperpartisan and Qasper datasets, whose tasks can be solved by CogLTX and ToBERT, and in which documents can be completely taken in by long-range transformers to eliminate interference caused by more contexts.

From Table 3, we can see that long-range transformers do have advantages over IR-based and chunking-based methods. Intuitively, the reason behind is that the performance of long-range transformers benefits from the contextual representation with a broader view of the document. These findings emphasize the need for future studies in document understanding to integrate global information. The results also indicate that DLUE effectively covers the assessment of the ability to recognize and aggregate dispersed knowledge across the whole text.

5 Conclusions

We propose a new benchmark DLUE that places the spot on documents and their lengthy content, complex underlying structure and dispersed knowledge challenges. DLUE covers diverse document-level tasks to evaluate four basic abilities required by document understanding, including document classification, document structure analysis, document information extraction and document transcription. Based on DLUE, we conduct an extensive side-by-side comparison of three document understanding architectures. Experiments demonstrate document understanding is far from being solved, and there exists a need for a universal architecture that can dominate all tasks.

Limitations

DLUE now focuses on plain text documents, while the documents one encounter, e.g., scientific articles, company announcements, or even personal notes, may also contain multi-modal information and with non-sequential structure. In future work, we intend to integrate these multi-modal, complex structure information into our document understanding benchmark.

Besides, due to the huge cost of computing resources, we didn't pretrain models specialized for document understanding, but directly use the public pretrained versions or train from scratch. We believe an unified pretraining incorporating document-related tasks will enhance understanding performance.

Ethics Statement

In consideration of ethical concerns, we provide the following detailed description:

1. We believe that this work is beneficial to develop universal document understanding architectures, which can help people quickly get information from business documents, legal statements and so on, saving time and money.

2. We standardize and put together ten datasets, which are all already publicly available under CC-BY-(NC-)SA-4.0 licenses[4]. For all the datasets, we have referenced the original work and encouraged DLUE users to do so.

3. All DLUE benchmark datasets have low ethical risks and do not expose any sensitive or personally identifiable information.

Acknowledgements

We thank all reviewers for their valuable comments. Moreover, this work was supported by the National Natural Science Foundation of China under Grants no. 62122077, 62106251 and 62306303.

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May. European Language Resources Association.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Farah Benamara, Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.

Michael K Buckland. 1997. What is a "document"? *Journal of the American society for information science*, 48(9):804–809.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. *arXiv preprint arXiv:1909.00142*.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland, May. Association for Computational Linguistics.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Ape: argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011.

[4] https://creativecommons.org/licenses/by/4.0/

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018a. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018b. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, June. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online, June. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. Lot: A story-centric benchmark for evaluating chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics*, 10:434–451.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. *arXiv preprint arXiv:1908.05514*.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021a. Document-level entity-based extraction as template generation. *arXiv preprint arXiv:2109.04901*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021b. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online, June. Association for Computational Linguistics.

G Thomas Hudson and Noura Al Moubayed. 2022. Muld: The multitask long document benchmark. *arXiv preprint arXiv:2202.07362*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Constituency Parsing. 2009. Speech and language processing.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. 2022. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989.

Timothy Shanahan, Douglas Fisher, and Nancy Frey. 2016. The challenge of challenging text. *On developing readers: Readings from educational leadership (EL Essentials)*, 100.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020a. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020b. Efficient transformers: A survey. *ACM Computing Surveys (CSUR)*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Ruoxi Xu, Hongyu Lin, Meng Liao, Xianpei Han, Jin Xu, Wei Tan, Yingfei Sun, and Le Sun. 2022. Eco v1: Towards event-centric opinion mining. *arXiv preprint arXiv:2203.12264*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Lang. Resour. Eval.*, 51(3):581–612, sep.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.