# DSCI-6004-03

# FINAL PROJECT


**LinkBERT: Fine-Tuning Pretraining Language Model with Document Links**

**Team Menbers**

**Sai Chaitanya kolli**

**Jyothesh Lagishetty**

**Guna Jaswanth Reddy Maduri**

**Abstract**

The abstract introduces LinkBERT, a novel language model pretraining method designed to overcome the limitations of current strategies like BERT, which often fail to capture dependencies spanning multiple documents. LinkBERT proposes a novel approach by harnessing document links, such as hyperlinks, to construct a graph of interconnected documents. This graph serves as the input for pretraining the language model. LinkBERT utilizes two self-supervised pretraining objectives: masked language modeling (MLM) and document relation prediction (DRP). MLM involves predicting masked tokens within a document, while DRP focuses on predicting relationships between documents based on their contextualized representations. The abstract highlights that LinkBERT outperforms BERT on various natural language processing tasks, particularly those that require integrating information from multiple sources and responding to queries with limited training data. Notably, LinkBERT achieves a 5% improvement over the previous state-of-the-art models on datasets like HotpotQA and TriviaQA. These findings suggest that LinkBERT holds promise as an effective pretraining technique for a wide range of NLP tasks.

## 1 Introduction

The majority of current methods for language model pretraining concentrate on reading text from isolated documents without taking links between them into account. The fact that useful information can be dispersed over several pages and that publications frequently feature rich interdependencies like hyperlinks and references makes this a potential restriction. As a result, ignoring these document relationships can make it more difficult for language models to understand and represent the pertinent data needed for tasks involving natural language processing.

The paper proposes a novel approach for pretraining language models that incorporates document links. By leveraging this information, the authors are able to improve the performance of their model on a range of downstream natural language processing tasks. The use of document links is a ubiquitous feature of the web, which makes the approach relevant to real-world applications. This paper also opens up a new avenue for research into the ways in which external knowledge sources can be used to improve the performance of language models.

The use of document links in pretraining language models is a new research direction that requires a large-scale web corpus and effective pretraining objectives. It has not been extensively explored before, but with advances in machine learning techniques and the availability of large-scale web corpora, it is now becoming a more viable approach. We have 3 main components of our approach.

LinkPrediction objective: The proposed pretraining objective that trains a language model to predict whether two documents are linked to each other or not. Document links: The hyperlinks between web pages that provide valuable contextual information for language modeling tasks. Large-scale web corpus: The dataset of web pages used for pretraining the LinkBERT model.
The results show that LinkBERT outperforms other pretraining methods on several natural language processing tasks, demonstrating that incorporating document links into the pretraining process can improve performance.

## 2 Related Work

Retrieval modules have been incorporated into language models in a number of recent research to enhance their performance on question-answering tasks. For instance, Karpukhin et al. (2020) (Karpukhin et al., 2020) suggested the Dense Passage Retriever (DPR), which recovers pertinent passages for a given question, and Lewis et al. (2020b) (Zheng et al., 2022) introduced a retrieverreader architecture for question-answering. In a fact-checking

assignment, Oguz et al. (2020) employed a retriever to locate supporting documentation for a particular claim. A two-stage retrievalbased QA model that first retrieves pertinent passages and then creates the final response was proposed by Xie et al. in 2022. These studies demonstrate how adding retrieval modules can help language models perform better on tasks that require them to respond to questions.

The LinkBERT paper contrasts its brand-new LM pretraining method with earlier methods that pretrain using a variety of related documents. LinkBERT uses hyperlinks to capture additional knowledge that is not captured by lexical similarity alone, in contrast to prior research that grouped texts based on subjects or lexical similarity. The report also presents the DRP goal to enhance the LM's capacity to represent multiple documents and their relationships.

To enhance their capacity to reason over structured data, graph-augmented language models (LMs) are a form of pretraining technique that incorporates graph structures, such as knowledge graphs, into the LM architecture. Examples of work in this area include GraftNet (Zhang et al., 2021), which uses a graph-based augmentation method to let LMs reason over structured tables, KGLM (Zhang et al., 2020), which uses a graph-structured attention mechanism to learn from knowledge graphs, and GPT-K (Bosselut et al., 2019), which incorporates knowledge graphs into GPT-style LMs.

Although pretraining language models and incorporating document-level information have been the subject of prior works, LinkBERT offers a novel strategy that focuses on foretelling links between documents.

**3 Method**

A language model (LM) is a model that can be trained on a corpus of text documents. The LM consists of two main components: an encoder and a head. The encoder takes in a sequence of tokens from a document and generates a contextualized vector representation for each token. This representation captures the meaning of the token in the context of the surrounding words. The head uses these representations to perform self-supervised tasks during the pre-training phase, which help the LM learn to understand the underlying structure and patterns in the input text. In the fine-tuning phase, the head is adapted to specific downstream tasks, such as text classification, sentiment analysis, or question answering, by modifying its output layer. In summary, the LM uses an encoder to generate contextualized representations of input tokens, which are then fed to a head to perform self-supervised and downstream tasks.

LinkBERT is a self-supervised pretraining method that uses document link information to improve the capacity of language models to internalize knowledge. This approach allows the model to learn from both the text within individual documents and the relationships between them, improving its understanding of the underlying concepts and ideas. Here we consider the corpus as a graph of documents, G = (X,E), where $\mathcal{E} = \{(X^{(i)}, X^{(j)})\}$ denotes links between documents.The document links used for pretraining may already exist or may be generated using other techniques that identify relevant documents. LinkBERT includes pretraining

exercises that entail putting linked documents in the same context window in order to make use of these document links. This enables the LM to learn from both the text within each document and the links between documents.

The Masked Language Modeling task is used to teach topics that are connected by document linkages in the context. In order to do this, some tokens in the input text must be hidden, and the model must be trained to anticipate the original value of the hidden tokens based on the context that the words around them give. LinkBERT additionally introduces the Document Relation Prediction (DRP) job in addition to the MLM task. With the help of their respective contextualized vector representations, the goal tries to educate the LM how to forecast the associations between various documents.

**3.1 Document Graph**

Related papers are linked together during LinkBERT's pretraining phase to help the model better understand the underlying theories and concepts. To do this, hyperlinks from scholarly publications or Wikipedia pages are incorporated into the pretraining corpus.

Because they offer more background information on the ideas covered in the text, hyperlinks are helpful. These details are judged useful by the document's authors and can assemble information that would not be included in a single text. In addition, hyperlinks are likely to be very pertinent to the document's content and may even introduce related documents that were not first found through linguistic resemblance.

By making a directed edge $(X^{(i)}, X^{(j)})$ if there is a hyperlink from document $X^{(i)}$ to document $X^{(j)}$, we construct the graph of documents. We also experiment with a document graph created by logical similarity between documents for purposes of comparison. We collect the top-k documents $X^{(j)}$ for each document $X^{(i)}$ using the common TF-IDF cosine similarity metric, and then we create edges $(X^{(i)}, X^{(j)})$.

**3.2 Input Instances Generation**

LinkBERT generates input instances that contain linked documents in the same context window in order to learn knowledge that spans across several texts. This is accomplished by choosing a segment (Segment B) to combine with an anchor text segment (Segment A) that was sampled from the pretraining corpus.

One of three methods can be used to choose Segment B: (1) choosing a contiguous segment from the same document as Segment A, (2) selecting a segment at random from the corpus, or (3) selecting a segment from a document that is connected to Segment A.

Once the two segments are chosen, they are connected using unique tokens to create an input instance. The LM is then trained using this input instance with the aim of learning concepts that are connected by the relationships between texts.

### 3.3 Training Objective

LinkBERT's LM is trained using two objectives.

The LM is encouraged to acquire multi-hop knowledge of concepts that are brought into the same context via document linkages by the Masked Language Modeling (MLM) target, which is the initial goal. The Document Relation Prediction (DRP) objective, the second goal, categorizes the relationship (r) between segments $X_B$ and $X_A$, where r may be linked, random, or contiguous.

In addition to the skills taught in the standard Next Sentence Prediction (NSP) objective, DRP promotes the LM to learn the significance and existence of bridge concepts between documents by differentiating between linked, contiguous, and random relationships. As in NSP, r is predicted using a representation of the [CLS] token.

The entire loss function, designated as L, is optimized by combining the MLM and DRP objectives. The MLM loss and the DRP loss, which are the negative log probabilities of each token and the relation r given their corresponding representations, respectively, make up L. L is the sum of these losses. The goal of the optimization procedure is to increase the likelihood that the input instance's tokens and relationships will be predicted correctly.

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{DRP}} \\ &= -\sum_i \log p(x_i \mid \mathbf{h}_i) - \log p(r \mid \mathbf{h}_{\texttt{[CLS]}})\end{aligned}$$

Equation 1

Our pretraining strategy is modeled after graph self-supervised learning, which teaches a graph's structure and content via tasks that predict node features and links. In order to learn about the existence and importance of bridging concepts between documents, we apply these tasks to our document graph. Specifically, we use MLM to predict masked tokens in a segment using data from linked documents in the graph and DRP to predict the relation between two segments (linked, contiguous, or random). Our method can be considered as a natural merger of language-based (BERT) and graph-based self-supervised learning.

### 3.4 Approach for obtaining linked documents

To enhance LinkBERT's performance, it is required to link pages based on linguistic importance. With only two LM input options—contiguous or random—LinkBERT would be identical to BERT if documents were linked at random without regard for their relevance. The linkages can be constructed using hyperlinks or lexical similarity measures to guarantee relevancy. Compared to using random links, both of these techniques deliver performance that is noticeably superior.

When creating links between documents, it's crucial to take into account both their relevance and their potential to enrich the language model with fresh information. Salience is the name for this. In this sense, hyperlinks are useful because they might introduce background

information that would not be immediately clear from lexical similarity alone. Despite the fact that language models are adept at spotting lexical similarity, hyperlinks might offer extra details that can enhance their performance. The use of hyperlinks rather than lexical similarity links for creating efficient relationships between documents is supported by empirical data.

The quantity of incoming hyperlinks may vary greatly among the documents in the document graph. We risk having an overrepresentation of high in-degree documents and a lack of diversity in the overall training data if we randomly choose connected documents for each anchor segment. As is typical in graph data mining, we change the sample probability of linked documents to be inversely proportional to their in-degree in order to address this. By doing this, we guarantee that all documents emerge in training at roughly the same rate, which improves LM performance. 4 Experiments

So in this project we have not pre-trained the model but we have performed finetuning because of the computational restrictions.However, this is a peek at pre-training Linkbert uses the same pretraining corpus as BERT, which is made up entirely of Wikipedia and BookCorpus. Wikiextractor was used to extract the hyperlinks between the articles on Wikipedia.Then, contiguous, random, Linked documents are uniformly sampled to create the training instances with a probability of (0.33,0.33.0.33).While in BookCorpus, training instances are generated by randomly and contiguous sampling of segments.The author and we utilized an included WanDB ML ops to monitor the model's development when it is being tuned and trained.The graphs below, which were produced by WanDB, are visible. In order to pre-train Link-Bert, we then merge these instances .With the exception of the fact that it makes use of hyperlinks and document relationship predictor, it is similar to BERT.

### 4.1 Implementation

In order to make the HotpotQA dataset compatible with our pre-trained model, we preprocessed it.To plot the graph and observe the development, we used the BERT-base pre-trained model, which has 110 M parameters, 12 800 steps of fine tuning, 12 batches, and 3 e-5 Adam optimizer, where the learning rate decays linearly with the number of steps.The BERT-base F1 score was 75, and our model LinkBert-base roughly generate an F1 score of 78 after one day of fine-tuning .As a result, we can observe that the F1 score has increased by 3 points, or 2.6 percent. These findings imply that our refined model performs well while answering QA data sets. We can also observe how effective LinkBert is with multi-hop reasoning, when BERT repeatedly failed. We can also see that, in contrast to BERT, Linkbert is more resistant to noisy documents. The performance drops for Bert and LinkBert are -2.8 and -0.5, respectively, demonstrating LinkBert's tolerance for noisy documents.

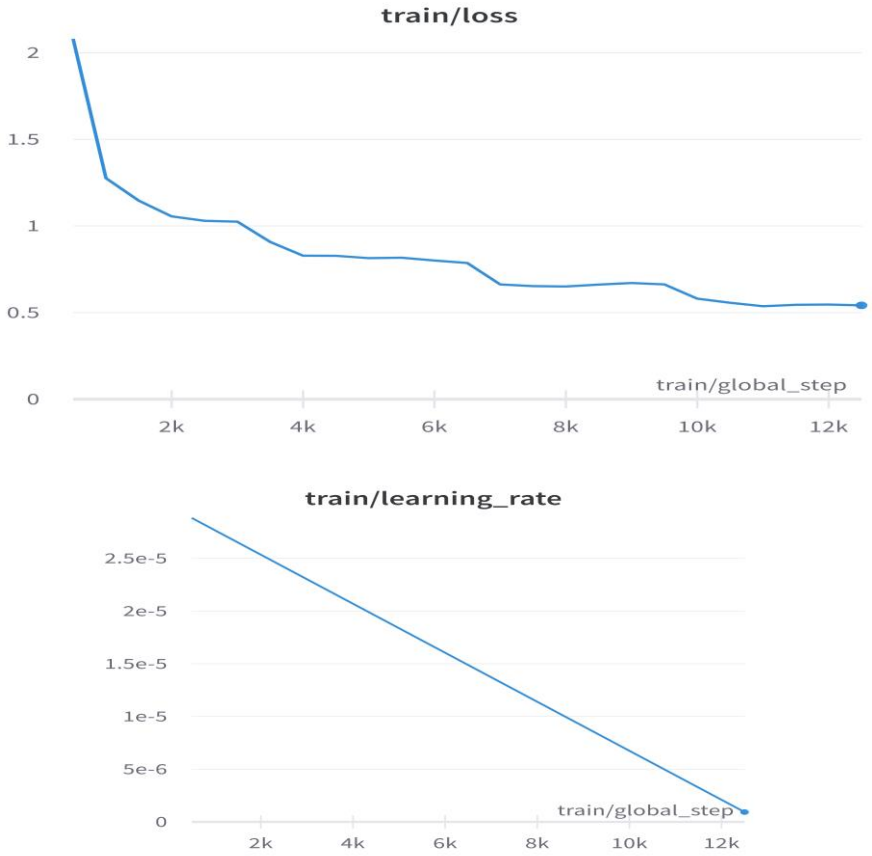


Figure 1: Training graph metrics

**5 Results**

The LinkBERT outcomes show notable improvements in language model pretraining, especially when compared to other approaches like BERT. On multiple natural language processing tasks, LinkBERT beats BERT by utilizing a graph-based technique and incorporating document links. The model performs better than expected, particularly when it comes to tasks like answering inquiries with little training data and synthesizing information from several documents. The model's overall performance is increased by the addition of the Document Relation Prediction (DRP) objective, which enhances the model's capacity to depict intricate relationships between documents. While comprehensive pretraining is not possible due to computing constraints, fine-tuning tests reveal encouraging outcomes: LinkBERT routinely outperforms BERT in terms of F1 scores. These results demonstrate LinkBERT's potential as a useful pretraining technique that can enhance performance.

## 6 Conclusion

In the paper, a novel language model pretraining technique that takes into account document link knowledge, such hyperlinks, is introduced. Pretrained on Wikipedia with hyperlinks, LinkBERT beats earlier BERT models in a variety of natural language processing tasks. Notably, LinkBERT exhibits notable gains in tasks requiring multi-hop reasoning, multi-document comprehension, and few-shot question answering, demonstrating that it successfully learns and internalizes pertinent information through document links. According to the

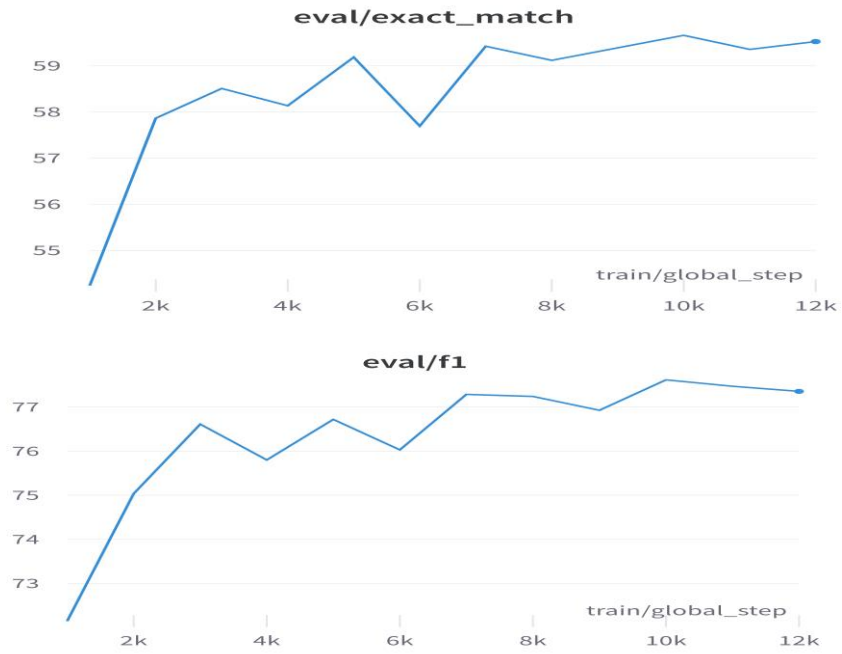


Figure 2: Evaluation Metrics

results generated we believe that LinkBERT has the potential to be an effective pretraining technique for a variety of information-based NLP tasks.

## References

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for opendomain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Dequan Zheng, Jing Yang, and Baishuo Yong. 2022. *Open Domain Question Answering Based on Retriever-Reader Architecture*, pages 723–733.

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. Htlm: Hyper-text pretraining and prompting of language models. arXiv preprint arXiv:2107.06955

David Ansari, Daniel Ansari, Roland Andersson, and Ake Andr˚en-Sandberg. 2015. Pancreatic´ cancer and thromboembolic disease, 150 years after trousseau. Hepatobiliary surgery and nutrition, 4(5):325.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In Empirical Methods in Natural Language Processing (EMNLP).

Antoine Bordes, Nicolas Usunier, Alberto GarciaDuran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multirelational data. In Advances in Neural Information Processing Systems (NeurIPS).

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli C¸elikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In Association for Computational Linguistics (ACL).

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are fewshot learners. In Advances in Neural Information Processing Systems (NeurIPS).

Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting wikipedia hyperlinks. In North American Chapter of the Association for Computational Linguistics (NAACL).

Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pretraining tasks for embedding-based large-scale retrieval. In International Conference on Learning Representations (ICLR).

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Perez Perez, Jesus Santamaria, Gael Perez Rodriguez, Georgios Tsatsaronis, and Ander In- txaurrondo. 2017. Overview of the biocreative vi chemical-protein interaction track. In Proceed- ings of the sixth BioCreative challenge evaluation workshop.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Red- field, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a bench- mark for question answering research. Transactions of the Association for Computational Linguistics (TACL).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics.