VoBERTal: Variant Objective BERT Pretraining Approaches with HyperLinks

Stanford CS224N Custom Project

Junyi(Joey) Ji
Department of Computer Science
Stanford University
junyiji3@stanford.edu

Yangyi Shen
Department of Mathematics
Stanford University
pyyshen@stanford.edu

Abstract

Existing pretrained Language Models (LM) only learn knowledge via single-edge relationships between documents in corpus to help downstream tasks. While BERT do not capture dependencies or knowledge that span across documents, LinkBERT identify single-edge linked dependency. In this research, we propose four types of upgraded LinkBERT: DeepLinkBERT, BiLinkBERT, BroadLinkBERT, and MultiLinkBERT, which captures deeper, bidirectional, broader, and comprehensive linked document dependencies. We aim to utilize multi-edge document relationships extracted via hyperlinks in Wikipedia Corpus. Given a text corpus, we view it as a graph of documents and create LM inputs by placing various types of linked documents in the same context. We then pretrain the LM with two joint self-supervised objectives: masked language modeling and our new proposal, document relation prediction. We find under similar convergence level LinkBERT is outperformance by our DeepLinkBERT in GLUE (+0.8% absolute improvement), WikiText (+0.5%); by BiLinkBERT in WikiQA (+11%), TweetQA (+2.8%), and ComQA (+0.6%); by BroadLinkBERT in average QA task (+2.6%); by MultiLinkBERT in GLUE (+0.9%) and WikiText (+0.5%).

1 Key Information to include

- Mentor: Ansh Khurana

2 Introduction

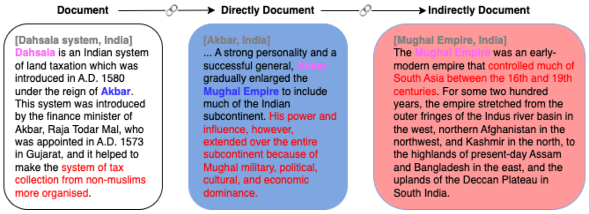
The ability of pretrained language models to effectively incorporate knowledge from human civilization remains a question of interest. Incorporating structured forms of knowledge, such as knowledge graphs, into these models may enable easier processing and manipulation by natural language processing (NLP) algorithms. In the case of the Wikipedia corpus, nodes represent topics of each web page, and relationships between these nodes are left to be defined.

This project focuses on incorporating rich dependencies between documents into BERT (Devlin et al., 2019). Current language model pretraining methods typically only focus on text from single documents, such as the Next Sentence Prediction (NSP) task in BERT. NSP learns to predict whether segments are contiguous within the same document or randomly from two different documents. However, we believe that language models can understand more nuanced relationships between documents and the incorporation of document dependencies can symbolize more general links between human knowledge. Human beings also learn from connecting and disconnecting knowledge from different fields. By training BERT to learn different relationships between human knowledge, we believe BERT can attain a more human-like method of understanding texts by linking the knowledge within the texts with other related texts and thus perform better at downstream tasks.

Existing work, such as LinkBERT (Yasunaga et al., 2021), has modified the NSP task into a three-class prediction (i.e. predicting whether the second segment is "contiguous", "random", or "linked" to the first anchored segment). LinkBERT is capable of learning document dependencies and thus has a better understanding of how knowledge is related in the world. However, we aim to build a more precise representation of the relevance between documents or knowledge, rather than simply predicting "linked" or "not linked."

To accomplish this goal, we challenge ourselves to find new ways of representing the relationship spectrum between documents and create a suitable dataset for this information retrieval task. Our approach includes retrieving deeper-linked documents through different times of hopping via hyperlinks; retrieving bidirectional relationships between documents; increasing the broadness of linked segments via exhaustive sampling in each pair of the linked documents. We modify the NSP training objective by adding more classes to allow BERT to predict a more precise document relationship between two segments.

In this paper, we present our approach and results for incorporating document dependencies into BERT, and how our modified dataset and training objective can improve the language model's ability to understand relationships between documents and knowledge.



[Dahsala system, India]
Dahsala is an Indian system of land taxation which was introduced in A.D. 1580 under the reign of Akbar. This system was introduced by the finance minister of Akbar, Raja Todar Mal, who was appointed in A.D. 1573 in Gujarat, and it helped to make the system of tax collection from non-muslims more organised.

[Akbar, India]
... A strong personality and a successful general, Akbar gradually enlarged the Mughal Empire to include much of the Indian subcontinent. His power and influence, however, extended over the entire subcontinent because of Mughal military, political, cultural, and economic dominance.

[Mughal Empire, India]
The Mughal Empire was an early-modern empire that controlled much of South Asia between the 16th and 19th centuries. For some two hundred years, the empire stretched from the outer fringes of the Indus river basin in the west, northern Afghanistan in the northwest, and Kashmir in the north, to the highlands of present-day Assam and Bangladesh in the east, and the uplands of the Deccan Plateau in South India.

Figure 1: Example of directly linked and indirectly linked documents: Document links (e.g. hyperlinks) can provide salient multi-hop knowledge. For instance, the Wikipedia article “Dahsala system” (left) describes “the system of tax collection from non-muslims more organised". The directly hyperlinked article (middle) reveals that the implementation is created because the emperor's "power and influence extended over the entire subcontinent...". Meanwhile, the indirectly linked document further suggests that the empire "controlled much of South Asia between the 16th and 19th centuries". Taken together, the link suggests new knowledge not available in a single document (e.g. “tax was imposed because of long-term control in surrounding regions”), which can be useful for various applications, including answering the question “During which period of time did Akbar impose Dahasala system?”. We aim to leverage document links to incorporate more knowledge into language model pretraining.

3 Related Work

As mentioned above, LinkBERT was also proposed because current LM pretraining methods do not model links between documents and therefore lose the opportunity to learn rich dependencies (e.g. hyperlinks, references) and knowledge that span across documents. Yasunaga et al. (2021) consider that document links like hyperlinks and references are ubiquitous. Such document dependencies help learn retrievers because we humans are similarly guided by them to acquire knowledge and even make discoveries.

Given a text corpus, LinkBERT obtains links between documents such as hyperlinks, and creates LM inputs by placing linked documents in the same context, besides the existing option of placing a contiguous segment from the same document or a segment from random documents as in BERT. Then, replacing the Next Sentence Prediction of BERT, LinkBERT uses a Document Relation Predction (DRP) objective, which classifies the relation of the second segment to the first anchored segment (contiguous, random, or linked). DRP encourages learning the relevance and bridging concepts between documents. At building the knowledge graph of the corpus documents, instead of viewing the pretraining corpus as a set of documents X = X(i), we view it as a graph of documents, G = (X,E), where E = (X(i),X(j)) denotes hyperlinks between documents.

For pretraining tasks, LinkBERT first sample an anchor text segment from the corpus (Segment A; $X_A \subseteq X(i)$). For the next segment (Segment B; X_B), LinkBERT either (1) use the contiguous segment from the same document ($X_B \subseteq X(i)$), (2) sample a segment from a random document($X_B \subseteq X(j)$ where $j \neq i$), or (3) sample a segment from one of the documents linked from Segment A ($X_B \subseteq X(j)$ where $(X(i), X(j)) \in E$). We join the two segments via special tokens to form an input instance: $[CLS]\ X_A\ [SEP]\ X_B\ [SEP]$. LinkBERT optimizes the entire loss containing the MLM loss, and DRP loss that compares the prediction based on the $[CLS]$ token. As a result, LinkBERT attains notably large gains for multi-hop reasoning, multi-document understanding, and few-shot question answering, outperforming BERT on MRQA benchmark and GLUE benchmark. This suggests that LinkBERT internalizes significantly more knowledge than existing LMs by pretraining with document link information.

There are many ways of doing retrieval-based augmentation. The "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks " proposed by Lewis et al. (2020) has tried placing related documents in the same LM contexts to improve model inference. Many retrieving methods also exist, including via hyperlinks or reference. For example, in "Learning to retrieve reasoning paths over wikipedia graph for question answering" by Asai et al. (2019). sentences of high lexical similarity are placed into the same context. LinkBERT uses both hyperlinks in Wikipedia and references in biomedicine articles. We choose to focus on designing retrieval techniques that represent the spectrum of document relationships via hyperlinks because hyperlinks are usually included because the content in the linked pages provides useful extension on the topic for readers. According to the paper of LinkBERT, hyperlinks represent both relevant information but salient knowledge — which might not be obvious via lexical similarity — that is helpful at complementing original contexts.

4 Approach

We present four model variants of the LinkBERT model, which modifies the Document Relation Prediction training objective proposed by Yasunaga et al. (2021). We aim to establish variant relations between documents through hyperlinks. Let's consider a Segment P from Document A and a Segment Q from Document B. Now, we define the following relation between these two segments:

1. Direct Linked: Segment P contains hyperlinks that point to Document B.
2. Indicrect Linked: Segment P contains hyperlinks that point to Document C and a segment in Document C contains hyperlinks that point to Document B.
3. Bidirectional Linked: Segment P contains hyperlinks that point to Document B and Segment Q contains hyperlinks that point to Document A.
4. Broad Linked: Document A contains hyperlinks to Document B.

Furthermore, we define a "relevance spectrum" which specifies the degree of relevance of a segment pair from strongest to weakest: {*contiguous, bidirectional linked, direct linked, broad linked, indirect linked, random*}. Based on these new relations, we propose the following variant Document Relation Prediction (DRP) tasks:

1. DDRP: classifies the relations r of Segment P and Segment Q
   ($r \in \{contiguous, random, direct\ linked, indirect\ linked\}$)
   DDRP encourages the model to infer relevant knowledge from the context, besides the capability learned in the DRP objective.

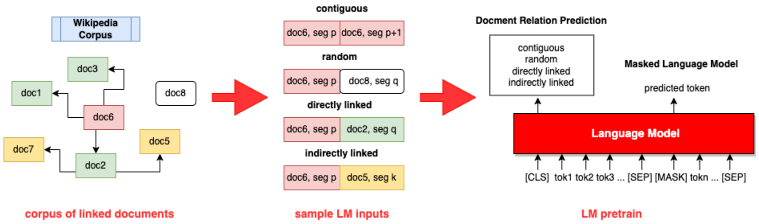


Figure 2: Neural Architecture of DeepLinkBERT

2. BiDRP: classifies the relations r of Segment P and Segment Q
   ($r \in \{contiguous, random, direct\ linked, bidirectional\ linked\}$)
   BiDRP encourages the model to learn the most relevant knowledge from the context.

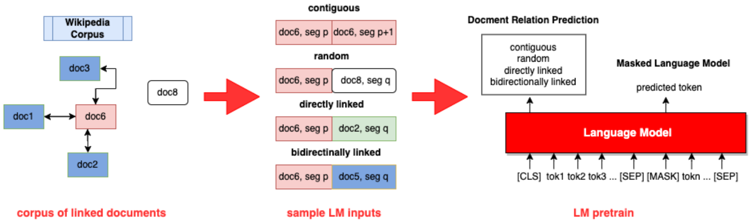


Figure 3: Neural Architecture of BiLinkBERT

3. BDRP: classifies the relations r of Segment P and Segment Q
   ($r \in \{contiguous, random, direct\ linked, broad\ linked\}$)
   BDRP encourages the model to learn more global knowledge related to the context.

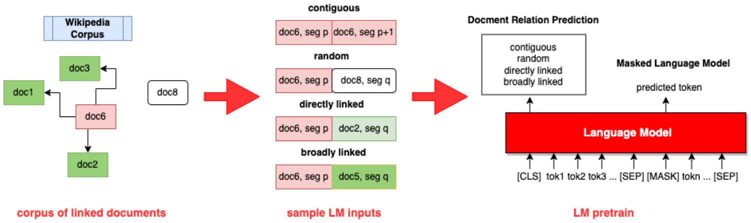


Figure 4: Neural Architecture of BroadLinkBERT

4. MDRP: classifies the relations r of Segment P and Segment Q
   ($r \in \{contiguous, random, direct\ linked, indirect\ linked, bidirectional\ linked, broad\ linked\}$)
   MDRP (Multi-DRP) encourages the model to learn the degree of relevance.

We use two objective functions to train each model: one is masked language modeling and the other is the corresponding document relation prediction. MLM follows the original task defined in (Devlin et al., 2018), which is to predict original tokens from modified input. For DRP, we use our proposed versions of DRP where we use the representation of the [CLS] to predict the relation r. Thus, we optimize the objective function

$$\mathcal{L} = \mathcal{L}_{MLM} + \mathcal{L}_{DRP} = -\sum_i \log p(x_i|h_i) - \log p(r|h_{[CLS]})$$

where x_i's correspond to tokens in the training input ([CLS] seg$_A$ [SEP] seg$_B$ [SEP]), and h_i's correspond to representations. Overall, we developed four variants of the LinkBERT, named as DeepLinkBERT, BiLinkBERT, BroadLinkBERT and MultiLinkBERT.
Regarding baselines, given that we do not have enough computation power to pretrain our upgraded LinkBERT models on the entire Wikipedia dataset as LinkBERT and BERT are pretrained on, we chose not to take the standard results of BERT$_{base}$ or LinkBERT$_{base}$ as baselines, simply because they are trained on way larger datasets with a significantly larger number of epochs. Hence, we pretrained BERT and LinkBERT from scratch using a smaller dataset with fewer epochs. We also control the convergence of all models (two baseline models and four upgraded models) to a similar level of loss to examine the architecture design of our models without bias.

5 Experiments

5.1 Data

Being precise about the exact form of the input and output can be very useful for readers attempting to understand your work, especially if you've defined your own task.
We use the WikipediaDump (Wikimedia Commons, 2022) as our pretraining corpus, which is the

same as the dataset used by BERT (Devlin et al., 2018). We then use the WikiExtractor (Zhang et al., 2019) to retrieve Wikipedia articles and to extract hyperlinks between documents. The whole Wikipedia dump is roughly 20GB and because of the lack of computational resources, we sample a subset of it for our pretraining task, which is approximately 1GB. Next, we create training inputs for each model variant by sampling *contiguous, random, directly linked, indirectly linked, bidirectionally linked,* or *broadly linked* segments as described in the [4] section. The specifications are:

1. BERT$_{tiny}$: Contiguous and Random Segments (50%, 50%)
2. LinkBERT$_{tiny}$: Contiguous, Random and Directly-Linked Segments (33%, 33%, 33%)
3. DeepLinkBERT$_{tiny}$: Contiguous, Random, Directly-Linked and Indirectly-Linked Segments (28%, 28%, 28%, 16%)
4. BiLinkBERT$_{tiny}$: Contiguous, Random, Directly-Linked and Bidirectionally-Linked Segments (30%, 30%, 30%, 10%)
5. BroadLinkBERT$_{tiny}$: Contiguous, Random, Directly-Linked and Broadly-Linked Segments (24%, 24%, 24%, 28%)
6. MultiLinkBERT$_{tiny}$: Contiguous, Random, Directly-Linked, Indirectly-Linked, Bidirectionally-Linked and Broadly-Linked Segments (20%, 20%, 20%, 12%, 4%, 24%)

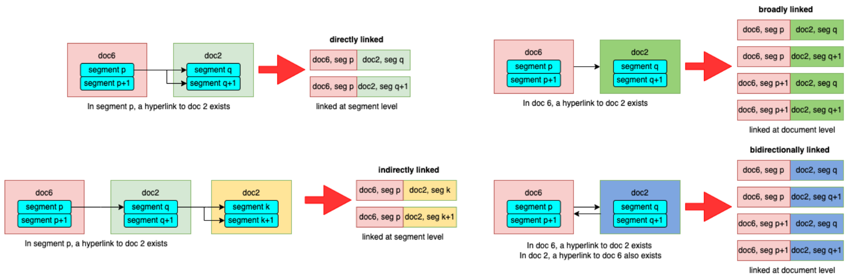


Figure 5: Segment Pair Sampling

Finally, we combine the training inputs in each case to train the corresponding model.

5.2 Evaluation method

We evaluate our model variants mainly on three categories of downstream tasks:
GLUE: The General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019) is a collection of classification tasks that test models on general linguistic knowledge. We evaluate on *CoLA* (Warstadt et al., 2019), *SST-2* (Socher et al., 2013), *MRPC* (Dolan and Brockett, 2005), *STS-B* (Cer et al., 2017), *MNLI* (Williams et al., 2018), *RTE* (Dagan et al., 2005), *WNLI* (Levesque et al., 2011), and calculate the average score.
WikiText: The WikiText dataset is a collection of tokens extracted from verified good articles on Wikipedia (Merity et al., 2016). We evaluate our models on wikipedia-related text-generation tasks and report the score by calculating the average accuracy.
Question Answering (QA): Question answering tasks ask the model to retrieve an answer to a specific question from a given document. We evaluate our models on seven datasets: *WikiQA* (Yang et al., 2015), *SQuAD* (Rajpurkar et al., 2016), *NaturalQA* (Kwiatkowski et al., 2019), *AdversarialQA* (Bartolo et al., 2020), *SocialQA* (Sap et al., 2019), *TweetQA* (Xiong et al., 2019), and *ComQA* (Chaudhri et al., 2021).

5.3 Experimental details

We follow the configurations of BERT$_{tiny}$ (4.4M parameters) defined in Bhargava et al. (2021) to pretrain all models from scratch. We set the learning rate to be 5e-3 with 5000 warmup steps, which linearly decays with weight 0.01 afterwards. The training time varies for each model, as the convergence speeds for each are different and due to the limitation of computational resources, we set training max steps for each model. We train for 10,000 steps with BERT$_{tiny}$, LinkBERT$_{tiny}$, 15,000 steps with DeepLinkBERT$_{tiny}$, BiLinkBERT$_{tiny}$, BroadLinkBERT$_{tiny}$, and 20,000 steps with MultiLinkBERT$_{tiny}$. All the models take roughly 6-8 hours to finish pretraining (without full convergence). We ensure the training losses for all the models to be on the same scale (5.5 $\sim$ 6.0) for fair comparisons.

5.4 Results

Table 1 shows the results on GLUE, where DeepLinkBERT and MultiLinkBERT outperform the rest models. However, the performance of BiLinkBERT is worse than we expected, as it is even far below the BERT baseline model. On average, compared to the result of LinkBERT, the gain is $+0.8\%$ for DeepLinkBERT, -2% for BiLinkBERT, $+0.1\%$ for BroadLinkBERT, and $+0.9\%$ for MultiLinkBERT. These results suggest that LMs can have a better understanding of natural language if they could acquire the ability to infer/reasoning or to identify the relations between sentences/documents.

Table 2 shows the results on WikiText, where DeepLinkBERT and MultiLinkBERT still outperform the rest models. Compared to the result of LinkBERT, the gain is $+0.5\%$ for DeepLinkBERT and MultiLinkBERT, -0.8% for BiLinkBERT, and -0.4% for BroadLinkBERT. The performance of BiLinkBERT and BroadLinkBERT is worse than we expected and we belive the reason is that, unlike Deep and MultiLinkBERT, Bi and BroadLinkBERT are designed for better identifying "direct" relations between documents while being deprived of the ability to reasoning, which is essential in text generation tasks. (Further analysis is in 6 section.)

Table 3 shows the results on different question answering tasks. BroadLinkBERT outperforms the rest in most QA tasks (BiLINKBERT outperforms the rest in some QA tasks). We find it interesting that broad link falls on the middle of the relevance spectrum, which might be an indicator of why it performs well on QA-specific tasks. (Further analysis is in 6 section.)

	cola	sst2	mrpc	stsb	mnli	rte	wnli	avg.
BERT$_{tiny}$	0.335	0.312	0.321	0.301	0.291	0.320	0.380	0.323
LinkBERT$_{tiny}$	0.345	0.323	0.312	0.311	0.291	0.315	0.369	0.324
DeepLinkBERT$_{tiny}$	0.348	0.329	0.324	0.315	0.297	0.333	0.376	0.332
BiLinkBERT$_{tiny}$	0.275	0.291	0.300	0.269	0.279	0.314	0.398	0.304
BroadLinkBERT$_{tiny}$	0.347	0.323	0.313	0.311	0.292	0.318	0.369	0.325
MultiLinkBERT$_{tiny}$	0.352	0.332	0.328	0.304	0.296	0.338	0.383	0.333

Table 1: Results (accuracy) on GLUE benchmark.

	Wikitext score
BERT$_{tiny}$	0.305
LinkBERT$_{tiny}$	0.309
DeepLinkBERT$_{tiny}$	0.314
BiLinkBERT$_{tiny}$	0.301
BroadLinkBERT$_{tiny}$	0.305
MultiLinkBERT$_{tiny}$	0.314

Table 2: Result (average accuracy) on Wikitext dataset

	WikiQA	SQuAD	NaturalQA	AdversarialQA	SocialQA	TweetQA	ComQA
BERT$_{tiny}$	0.253	0.101	0.103	0.100	0.333	0.218	0.106
LinkBERT$_{tiny}$	0.257	0.113	0.116	0.121	0.334	0.198	0.141
DeepLinkBERT$_{tiny}$	0.293	0.116	0.112	0.118	0.328	0.207	0.137
BiLinkBERT$_{tiny}$	0.365	0.112	0.117	0.111	0.282	0.226	0.147
BroadLinkBERT$_{tiny}$	0.376	0.133	0.120	0.126	0.342	0.210	0.155
MultiLinkBERT$_{tiny}$	0.236	0.120	0.108	0.114	0.330	0.191	0.100

Table 3: Results (accuracy) on Question Answering tasks

6 Analysis

6.1 DeepLinkBERT

Improved logical reasoning. From table 2, DeepLinkBERT achieves the highest score on Wikitext evaluation dataset. Wikitext dataset is a text-generation downstream task, which requires models to

have the ability to infer from the given context. The result achieved by DeepLinkBERT suggests that it is good at inference, as it can do multiple hops through hyperlinks to establish relations between documents. To contextualize the idea of inference, we test DeepLinkBERT and LinkBERT on the same text generation prompt and compare the results. The prompt is "The goal of life is <mask>". From the wikipedia page of "goal", there is a hyperlink to "personal life" where "the pursuit of happiness" is mentioned and "happiness" is exactly the answer of LinkBERT which requires 2-hop reasoning. In "personal life" page, there is a hyperlink to "Meaning of Life" page, where "meaning of Life itself no other than Freedom" is mentioned and "freedom" is the answer presented by DeepLinkBERT, which requires 3-hop reasoning. Thus, we can see that DeepLinkBERT makes more logical reasoning as "meaning of life" is more connected to "the goal of life".
Deficient in natural answering. From table 3, DeepLinkBERT does not perform well on question-answering specific task and is even worse than LinkBERT in certain tasks. We further investigate on the SocialQA datset, as DeepLinkBERT shows a significant decline on this specific task (-0.6%). We study in what questions LinkBERT succeeds while DeepLinkBERT fails. One question is "How would you describe S" given the context that "S was a school teacher...". LinkBERT's answer is "As someone that takes teaching..." while DeepLinkBERT's answer is "Like a leader". It is clear that "teaching" is directly linked to "teacher" and "leader" is indirectly linked to "teacher". The poor performance of DeepLinkBERT on question answersing tasks suggests that it fails establishing direct relations sometimes.

6.2 BiLinkBERT

Improved directional reasoning. From table 3, BiLinkBERT achieves the highest or second highest score on WikiQA and TweetQA evaluation dataset, which requires the model to answer questions given texts from Wikipedia and Twitter. The result achieved by BiLinkBERT suggests that it is good at extracting relevant information in Wikipedia texts and tweets to answer the question. However, BiLinkBERT does not perform equally well in other QA tests. We postulate that this disparity exists because the Wikipedia and Twitter corpus contain the most bidirectional links. For example, the Wikipedia pages “Stanford University” and “Silicon Valley” have hyperlinks that connect to each other. Meanwhile, people often quote each other’s tweets and comment on them. This BERT to apply its ability of extracting bidirectionally information to answer the prompt question. In a real example in WikiQA, the question asks “how a water pump works?” LinkBERT may find the following solution from a linked webpage about water pumps in Germany “A large, electrically driven pump (electropump) for waterworks near the Hengsteysee , Germany”. This is not very sensible even if two Wikipedia pages are linked together and both discusses water pumps. However, BiLinkBERT will be able to find a better answer “A pump is a device that moves fluids (liquids or gases), or sometimes slurries , by mechanical action” from the most relevant and linked Wikipedia page that contains a hyperlink connecting back to the anchored document as well.
Deficient in language understanding and generation. From table 1 and table 2, BiLinkBERT does not perform well on GLUE and WikiText tasks and is even worse than LinkBERT and BERT in most case. We think the problem with low GLUE scores comes from the biased distribution of data. Due to the scarcity of “bidirectionally linked” segment pairs in our corpus compared to “linked”, “contiguous”, and “random” segment pairs, “bidirectional” data only counts for 10% after rounding up. This 10% data and the corresponding label in classification then becomes noise in general language understanding and causes the difficulty for BiLinkBERT to evenly learn document/knowledge relationship. For its low performance in WikiText, we believe BiLinkBERT restricts its text generation between two mutually linked documents due to its additional class of “bidirectionally linked” in training tasks. This restriction reduces the diversity and salience of the generated texts, thus causing a low performance in WikiText evaluation.

6.3 BroadLinkBERT

Improved topic reasoning. From table 3, BroadLinkBERT achieves the highest performance on most QA evaluation dataset, which requires the model to answer questions given texts from respective corpus. The result achieved by BroadLinkBERT suggests that it learns the ability to extract relevant information in most corpora to answer the question. We think that BroadLinkBERT’s ability that outperforms all other models in QA exists because of its ability to restrict relevant answers within the linked document pairs and search through the entire page of linked documents. Unlike other segment-focused models, BroadLinkBERT goes through every pairs of segments in two linked documents, creating a larger chance of finding the most desirable answer. Meanwhile, it keeps the searching focus under the same topic (i.e. the same document). For example given the context string about "University of Notre Dame", SQuAD asks the model to answer the question "What is in front of the Notre Dame Main Building?" Our BroadLinkBERT is capable of focusing on the topic "Notre Dame Main Building" and exhaustively search the content under this topic to find the most desirable answer "a copper statue of Christ", while LinkBERT might answer "a golden statue of the Virgin Mary" who "sits on the Notre Dame Main Building" because the information is linked by "Notre Dame Main Building" but with incorrect geographical location.

Deficient in general language generation. From table 2, BroadLinkBERT does not perform well onWikiText task and is even worse than LinkBERT and has the same performance with BERT. We think the reason is that BiLinkBERT restricts its text generation between two linked documents for too long due to its trained habit of exhausting segment pairs within linked documents. Similar to BiLinkBERT, this restriction reduces the diversity and salience of the generated texts, thus causing a low performance in WikiText evaluation.

6.4 MultiLinkBERT

Improved understanding of sentence pair relations Table 1 shows that MultiLinkBERT achieves remarkable gains in almost all GLUE tasks, which suggests that MultiLinkBERT is suitable for general language understanding tasks. The results conform to our expectation, as MultiLinkBERT is the only model that has the access to the whole relevance spectrum, allowing it to have a better understanding of the semantic structure of natural language. Besides, GLUE is a benchmark that targets on sentence-pair language understanding tasks Wang et al. (2019), which helps explain why MultiLinkBERT performs well.

However, MultiLinkBERT's performance on question answering tasks is far below our expectation and we believe it is the consequence of focusing too much on learning general knowledge. For question answering tasks, we are seeking the most "relevant" sentences to be the answers to the questions. MultiLinkBERT's ability to identify a specific relation is weakened because it needs to learn all types of relations, which makes it hard to distinguish between them. (It might also be due to the fact that we did not let the model converge) Furthermore, the imbalanced sizes of the training dataset (of each relation type) might have a negative impact on the performance.

7 Conclusion

We present four variant models of LinkBERT, all of which build on top of the Document Relation Prediction objective proposed by Yasunaga et al. (2021). We define four new concrete relations between documents through hyperlinks and pretrain four variant models of LinkBERT on the Wikipedia corpus from scratch to compare against BERT and LinkBERT. DeepLinkBERT and MultiLinkBERT outperform the rest models in GLUE and WikiText generation downstream tasks, while BiLinkBERT and BroadLinkBERT outperform the rest models in certain question answering tasks. The results suggest that DeepLinkBERT and MultiLinkBERT can effectively learn salient and diverse knowledge, which makes them a perfect choice for global-knowledge related tasks; BiLinkBERT and BroadLinkBERT can be a perfect choice for question-answering specific tasks, as they can retrieve directional and relevant information more efficiently.

However, the primary limitation of our work is that we only tested our models on a subset of the wikipedia corpus and we did not let our models converge, due to the lack of computational powers. All conclusions are based on the performance of the current models and we expect the models to achieve similar results on larger dataset. For future works, we want to evaluate our variant models on larger and more balanced training dataset to test the real performance of each one and we wish to experiment with even more types of DRP: either by adding number of hops (e.g. secondly linked, thirly linked, . . .) or by introducing new relations like the order of segments.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over wikipedia graph for question answering. *CoRR*, abs/1911.10470.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in nli: Ways (not) to go beyond simple heuristics.

Daniel Cer, Mona Diab, Eneko Agirre, and Iñigo Lopez-Gazpio. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Rohit Chaudhri, Xiaodong Song, Saurabh Kumar, Saurabh Tiwary, Abhishek Agarwal, Jayant Krishnamurthy, Jose MF Moura, and Sudipta Kar. 2021. Comqa: A community-sourced dataset for complex factoid question answering with paraphrase clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2760–2772.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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*, pages 4171–4186. Association for Computational Linguistics.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pages 9–16.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments.

Wikimedia Commons. 2022. English Wikipedia Dump File, December 2022. https://dumps.wikimedia.org/20221201/.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Tweetqa: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities.