LinkBERT: Pretraining Language Models with Document Links

Michihiro Yasunaga, Jure Leskovec*, Percy Liang*
Stanford University

Code / Model: https://github.com/michiyasunaga/LinkBERT

Problem

Existing language models (LM) like BERT model a single document at a time, but documents have dependencies (e.g. hyperlinks, citation links), and knowledge can span across multiple documents. How can we incorporate such multihop/meta knowledge into LMs?

Our idea: LinkBERT

We train a knowledgeable LM leveraging document links, e.g. hyperlinks, citation links.

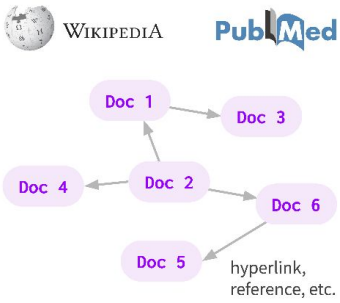
0) ***"Document graph" view of text corpus***
1) ***Link-aware LM input creation:*** sample and place linked documents together in the same context.
2) ***Link-aware LM pretraining:*** train LM with two joint self-supervised tasks: masked LM and document relaton prediction (DRP).

Experiments

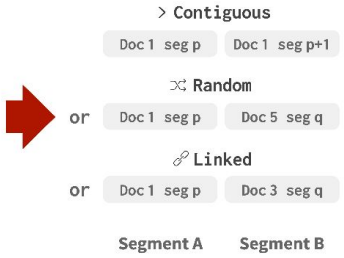
	General domain		Biomedical domain	
Pretraining corpus	Wikipedia + hyperlinks		PubMed + citation links	
Downstream tasks	MRQA	GLUE	BLURB	MMLU
BERT (Baseline)	78.5	80.7	81.0	39.0
LinkBERT (**Ours**)	**81.0**	**81.1**	**83.4**	**50.7**

Benefits

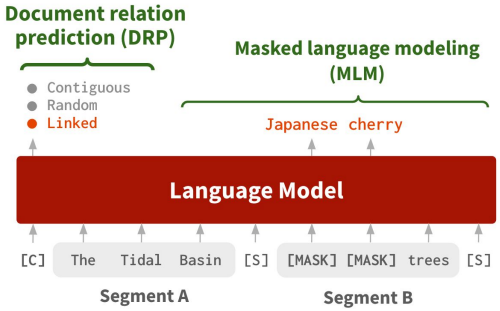
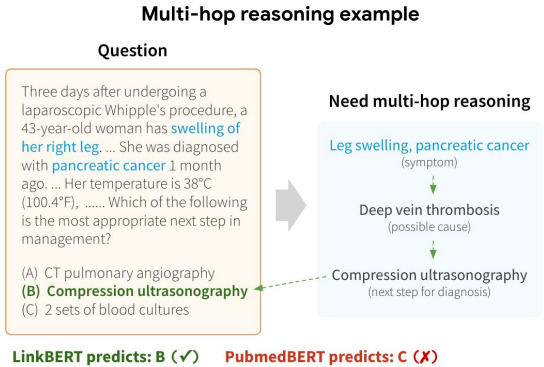
- Improved **multi-hop** reasoning and **cross-doc** understanding
- Effective for **knowledge-intensive** tasks, e.g. few-shot QA



Corpus of linked documents



Create LM inputs



Pretrain the LM