BioReader: a Retrieval-Enhanced Text-to-Text Transformer for Biomedical Literature

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, Lorenzo Valgimigli
Department of Computer Science and Engineering, University of Bologna, Italy

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Paper

Motivations

- Encoding all factual domain-specific competencies into opaque weight matrices is inefficient for dynamic and trust-demanding fields like biomedicine
  - ➤ Capturing more world facts = training ever-larger networks
  - ➤ Changing what a PLM knows = retraining on new documents
- Retrieval is a complementary path to architectural scaling
- No semi-parametric models (closed-book + open-book) for biomedicine

Contribution

- BioReader, the first retrieval-enhanced transformer for bio-literature
  - ➤ A text-to-text language model empowered by a differentiable access towards an explicit large-scale text memory grounded on PubMed

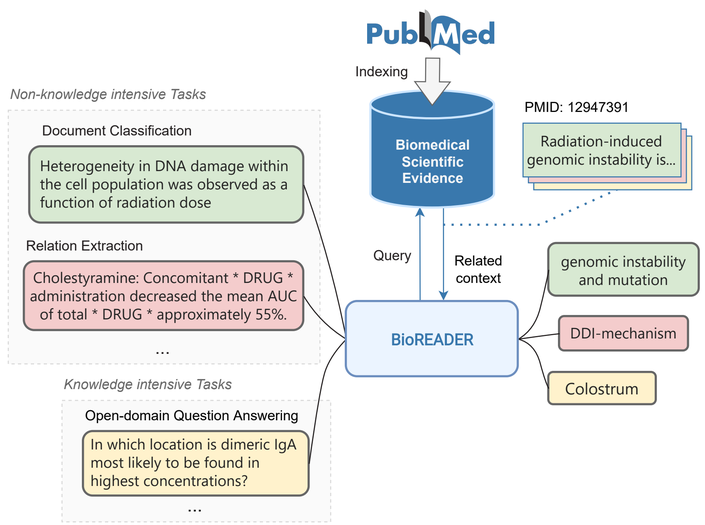

Figure 1: Illustration of BioReader; every biomedical task is cast as translating text spans with the help of external scientific evidence retrieved on the fly

Model	Granularity	Retriever training	Retrieval integration	Unsupervised Retriever	Retrieval Source	Task(s)
kNN-LM	Token	Frozen (Transformer)	Add to probs	✓	Wikipedia, Books	LM
Spalm	Token	Frozen (Transformer)	Gated logits	✓	Wikipedia	OpenQA
Dpr	Prompt	Contrastive proxy	Extractive QA		Wikipedia	OpenQA
Realm	Prompt	End-to-End	Prepend to prompt	✓	Wikipedia	OpenQA
Rag	Prompt	Fine-tuned Dpr	Cross-attention (concatenation)		Wikipedia	OpenQA, QG, FV
FiD	Prompt	Fine-tuned Dpr	Cross-attention		Wikipedia	OpenQA
Emdr^2	Prompt	End-to-end	Cross-attention	✓	Wikipedia	OpenQA
Retro	Chunk	Frozen (Bert)	Chunked cross-attention	✓	Web, Books, News, Wikipedia, GitHub	OpenQA
BioReader (ours)	Chunk	Frozen (Contriever)	Chunked cross-attention	✓	PubMed†	NER, RE, DC, NLI, QA, OpenQA

Table 1: Comparison of BioReader with existing retrieval approaches. LM = language modeling, QG = question generation, FV = fact verification, NER = named entity recognition, RE = relation extraction, DC = document classification, NLI = natural language inference, (Open)QA = (open-domain) question answering. † highlights retrieval sources that are different from training data.

Architecture

- Retro and T5 blocks are interleaved in the decoder stack, with the firsts placed at layers 9 and 12; external knowledge is merged via Cca

$$\textsc{Retro}(H, E) = \textsc{Ffw}(\textsc{Cca}(\textsc{Att}(H), E)) \quad (1)$$

$$\textsc{T5}(H) = \textsc{Ffw}(\textsc{Att}(H)) \quad (2)$$

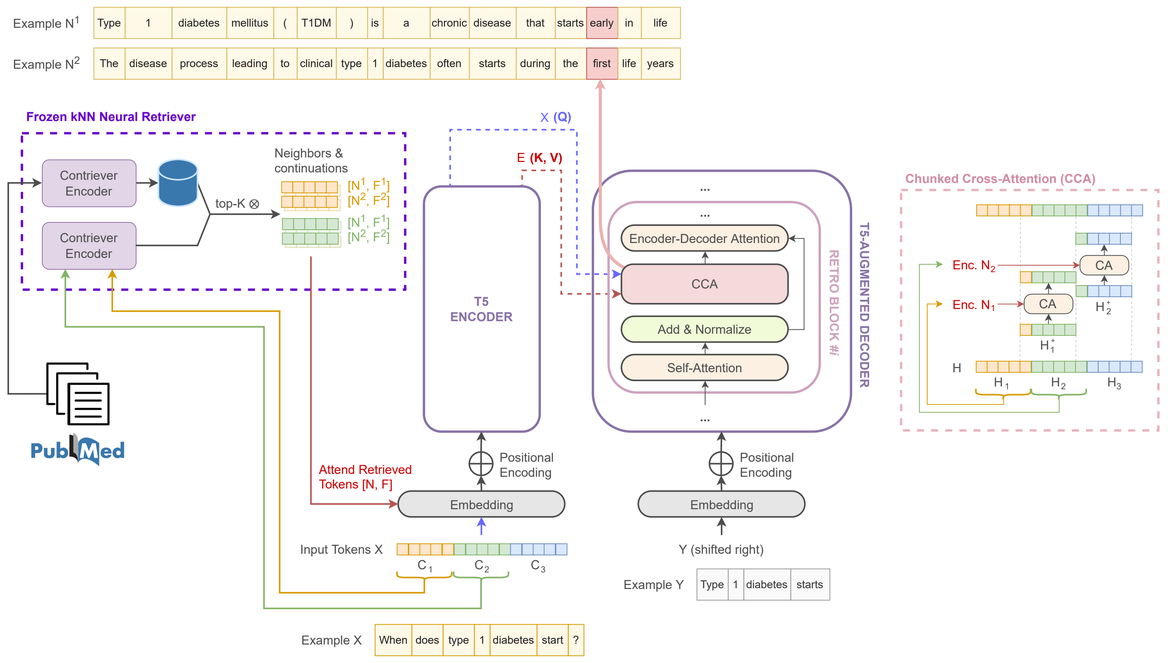

Figure 2: An illustration of the BioReader architecture. *Left:* simplified version where a sequence X of length n=15 is divided into l=3 chunks of size m=5. For each chunk, we retrieve k=2 scientific evidence neighbors of r=10 tokens each (including continuations). The current input prompt X and the fetched tokens are given as input to our encoder-decoder architecture based on T5. The fusion of their learned representations is done in the decoder via chunked-cross attention (Cca). *Right:* autoregressive Cca interaction details.

Method

- **Input Segmentation.** We split the tokenized input X (max-length n) into a sequence of l chunks of size $m = n/l$, w/ $n = 512$ and $m = 16$
- **Evidence Datastore.** $\mathcal{D}$ is an external *key-value store* composed by $\approx$ 200K English abstracts (60M tokens) of 286 randomized controlled trials from 2016 MEDLINE/PubMed Database
  - ➤ Key: $f(N)$, Value: $[N, F]$ where N=neighbor chunk, F=N continuation in the abstract, $f(\cdot)$=Contriever encoder, hidden states mean
- **Nearest Neighbor Retrieval.** For each input chunk C_u, we select its top k most similar documents using $d(C_u, N) = f(C_u) \otimes f(N)$
  - ➤ FAISS for approximate nearest neighbor search (sub-linear memory access)

Training

- Pre-training (only Cca parameters, <5% of total weights)
  - ➤ Cleaned and masked version of the PubMed database (>32M abstracts)
  - ➤ T5-blocks initialization = pretrained weights of SciFive(PubMed)-base
  - ➤ Span-based masked learning
- Fine-tuning (all the layers)
  - ➤ 18 widespread biomedical datasets—mostly from the Biomedical Language Understanding and Reasoning Benchmark (BLURB)
  - ➤ 6 downstream task categories (multi-task learning for NER)
  - ➤ We retrieve 9 neighbors from $\mathcal{D}$ (different values tested for evaluation)
- Objective: maximum likelihood with teacher forcing

$$PX_{u,i} = x_{(u-1)\ m+i} | (x_j)_{j<(u-1)\ m+i}, \qquad PN_u = (\textsc{Ret}(C_{u'}))_{u'<u}, \quad (3)$$

$$L(X|\theta, \mathcal{D}) = \sum_{u=1}^{l} \sum_{i=1}^{m} \ell_\theta(PX_{u,i}, PN_u). \quad (4)$$

Results

- We push the state-of-the-art on 2/7 NER, 1/2 RE, 1/1 DC, and 3/3 QA datasets, staying highly competitive in all other cases
- We beat SciFive-large (3x our size) on 5 different tasks, outperforming models with a comparable number of parameters to ours
- Although not retrained, BioReader adapts correctly to unseen questions on the COVID-19 literature in *zero-shot datastore* settings

Model	#params	In-Domain Vocabulary	NER (F1)							RE (F1)		DC (F1*)	NLI (Acc.)
			NCBI disease	BC5CDR disease	BC5CDR chemical	BC4CHEMD	BC2GM	JNLPBA	Species-800	ChemProt	DDI	HoC	MedNLI
BioBert	110M	✓	**89.71**	87.15	93.47	92.36	84.72	77.49	74.06	76.46	80.88	81.54	—
SciBert	110M	✓	88.25	84.70	92.51	—	83.36	78.51	—	75.00	81.22	81.16	—
BlueBert-base	110M	×	88.04	83.69	91.19	—	81.87	77.71	—	71.46	77.78	80.48	—
ClinicalBert	110M	×	86.32	83.04	90.80	—	81.71	78.07	—	72.04	78.20	80.74	—
PubMedBert-base	110M	✓	87.82	85.62	93.33	—	84.52	79.10	—	77.24	—	—	—
PubMedBert-large	340M	✓	88.25	85.77	93.22	—	84.72	79.44	—	78.77	82.39	82.57	—
PubMedElectra-base	110M	✓	87.68	84.99	93.19	—	83.79	78.60	—	76.54	80.58	81.45	—
PubMedElectra-large	340M	✓	87.93	84.82	92.90	—	83.87	78.77	—	76.80	78.92	82.37	—
BioLinkBert-base	110M	×	88.18	86.10	93.75	—	84.90	79.03	—	77.57	82.72	84.35	—
BioLinkBert-large	340M	×	88.76	86.39	94.04	—	**85.18**	**80.06**	—	79.98	83.35	84.87	—
BioMegatron	345M	✓	87.10	**88.50**	92.90	—	—	—	—	77.00	—	—	—
T5-base	220M	×	88.54	86.83	93.61	89.73	82.29	74.56	74.32	84.82	82.04	85.22	83.90
T5-large	770M	×	88.78	86.31	94.22	89.96	82.36	75.83	74.66	85.41	83.35	85.68	83.80
SciFive-base	220M	×	87.96	87.44	94.35	92.02	83.92	75.60	76.55	**88.83**	83.15	85.89	85.30
SciFive-large	770M	×	89.17	86.98	**94.66**	91.96	83.60	76.08	75.50	87.88	83.67	86.36	**86.36**
BioReader (ours)	229.5M	×	88.90	87.62	94.43	**92.81**	84.77	77.82	**77.44**	88.16	**84.34**	**87.78**	85.76

Table 2: Test results on NER, RE, DC, and NLI after fine-tuning. F1* is F1 on sample average. **Bold** and underline denote the best and second best scores; the gradient of green indicates our improvement compared to the previous state-of-the-art (the deeper, the more).

Model	# params	In-Domain Vocabulary	Automatic Evaluation				Human Evaluation				
			QA			OpenQA	QA			OpenQA	Fluency
			BioAsq 4b	BioAsq 5b	BioAsq 6b	MedQA-USMLE	BioAsq 4b	BioAsq 5b	BioAsq 6b	MedQA-USMLE	(Avg)
BioLinkBert-base	110M	×	—	—	—	40.00	—	—	—	—	—
BioLinkBert-large	340M	×	—	—	—	**44.60**	—	—	—	—	—
SciFive-base	220M	×	60.80	59.53	55.56	34.57	79.98	80.02	70.05	38.03	2.49
SciFive-large	770M	×	62.98	61.67	61.74	35.12	80.23	80.12	71.54	39.78	2.65
BioReader (ours)	229.5M	×	**64.13**	**62.02**	**62.18**	42.96	**82.12**	**81.88**	**73.35**	**48.57**	**2.86**

Table 3: Exact Match accuracy (left) and human-evaluated scientific accuracy (right) on QA and OpenQA.

Question	BioReader w/ $\mathcal{D}$	BioReader w/ $\mathcal{D}'$
medqa: question*: January 2020. A 69-year-old Chinese man comes to the physician with fever, tiredness, cough, dyspnoea, and severe respiratory issues. The clinical picture suggests an infectious disease. What is the most likely diagnosis?	✗ bronchiolitis	✓ COVID-19
medqa: question*: Coronaviruses are viruses that can cause illnesses in humans, including severe respiratory disease and even death. Corona disease-19 virus (COVID-19) spread and caused a pandemic that affected people all over the world. As COVID-19 cases continue to rise globally, which are the most effective options to prevent contamination and infection transmission?	✗ disinfect the respiratory tract	✓ vaccinate against COVID-19

Table 4: Answers generated by BioReader to context-free COVID-19 questions before ($\mathcal{D}$) and after ($\mathcal{D}'$) integrating SARS-CoV-2 evidence into the datastore.

Contacts: {giacomo.frisoni, gianluca.moro}@unibo.it